This Week's Finds in Mathematical Physics

Weeks 201 to 250

January 10, 2004 to *April 26, 2007*

by John Baez

Typeset by Tim Hosgood

Contents

Week 201

January 10, 2004

Lately James Dolan and I have been studying number theory. I used to *hate* this subject: it seemed like a massive waste of time. Newspapers, magazines and even lots of math books seem to celebrate the idea of people slaving away for centuries on puzzles whose only virtue is that they're easy to state but hard to solve. For example: are any odd numbers the sum of all their divisors? Are there infinitely many pairs of primes that differ by 2? Is every even number bigger than 2 a sum of two primes? Are there any positive integer solutions to

$$x^n + y^n = z^n$$

for $n > 2$? My response to all these was: WHO CARES?!

Sure, it's noble to seek knowledge for its own sake. But working on a math problem just because it's *hard* is like trying to drill a hole in a concrete wall with your nose, just to prove you can! If you succeed, I'll be impressed — but I'll still wonder why you didn't put all that energy into something more interesting.

Now my attitude has changed, because I'm beginning to see that behind these silly hard problems there lurks an actual *theory*, full of deep ideas and interesting links to other branches of mathematics, including mathematical physics. It just so happens that now and then this theory happens to crack another hard nut.

I'd known for a while that something like this must be true: after all, when Andrew Wiles proved Fermat's Last Theorem, even the newspapers admitted this was just a spinoff of something more important, namely a special case of the Taniyama-Shimura Conjecture. They said this had something to do with elliptic curves and modular forms, which are very nice geometrical things that show up all over in complex analysis and string theory. Unfortunately, the actual statement of this conjecture seemed impenetrable — it didn't resonate with things I understood.

In fact, the Taniyama-Shimura Conjecture is part of a big *network* of problems that are more interesting but harder to explain than the flashy ones I listed above: problems like the Extended Riemann Hypothesis, the Weil Conjecture (now solved), the Birch-Swinnerton-Dyer Conjecture, and bigger projects like the Langlands Program and developing the theory of "motives". And these problems rest on top of a solid foundation of beautiful stuff that's already known, like Galois theory and class field theory, and stuff about modular forms and L-functions.

As I'm gradually beginning to understand little bits of these things, I'm getting really excited about number theory... so I'm dying to *explain* some of it! But where to start? I have to start with something basic that underlies all the fancy stuff. Hmm, I think I'll start with Galois theory.

As you may have heard, Galois invented group theory in the process of showing you can't solve the quintic equation

$$ax^5 + bx^4 + cx^3 + dx^2 + ex + f = 0$$

by radicals. In other words, he showed you can't solve this equation by means of some souped-up version of the quadratic formula that just involves taking the coefficients a, b, c, d, e, f and adding, subtracting, multiplying, dividing and taking nth roots.

The basic idea is something like this. In general, a quintic equation has 5 solutions — and there's no "best one", so your formula has got to be a formula for all five. And there's a puzzle: how do you give one formula for five things?

Well, think about the quadratic formula! It has that "plus or minus" in it, which comes from taking a square root. So, it's really a formula for *both* solutions of the quadratic equation. If there were a formula for the quintic that worked like this, we'd have to get all 5 solutions from different choices of nth roots in this formula.

Galois showed this can't happen. And the way he did it used *symmetry!* Roughly speaking, he showed that the general quintic equation is completely symmetrical under permuting all 5 solutions, and that this symmetry group — the group of permutations of 5 things — can't be built up from the symmetry groups that arise when you take nth roots.

The moral is this: you can't solve a problem if the answer has some symmetry, and your method of solution doesn't let you write down a correct answer that has this symmetry!

An old example of this principle is the medieval puzzle called "Buridan's Ass". Placed equidistant between two equally good piles of hay, this donkey starves to death because it can't make up its mind which alternative is best. The problem has a symmetry, but the donkey can't go to *both* bales of hay, so the only symmetrical thing he can do is stand there.

Buridan's ass would also get stuck if you asked it for *the* solution to the quadratic equation. Galois proof of the unsolvability of the quintic by radicals is just a more sophisticated variation on this theme. (Of course, you *can* solve the quintic if you strengthen your methods.)

A closely related idea is "Curie's principle", named after Marie's husband Pierre. This says that if your problem has a symmetry and it is a unique solution, the solution must be symmetrical.

For example, if some physical system has rotation symmetry and it has a unique equilibrium state, this state must be rotationally invariant.

Now, in the case of a ferromagnet below its "Curie temperature", the equilibrium state is *not* rotationally invariant: the little magnetized electrons line up in some specific direction! But this doesn't contradict Curie's principle, since there's not a unique equilibrium state — there are lots, since the electrons can line up in any direction.

Physicists use the term "spontaneous symmetry breaking" when any *one* solution of a symmetric problem is not symmetrical, but the whole set of them is. This is precisely what happens with the quintic, or even the quadratic equation.

While these general ideas about symmetry apply to problems of all sorts, their application to number theory kicks in when we apply them to *fields*. A "field" is a gadget where you can add, subtract, multiply and divide by anything nonzero, and a bunch of familiar laws of arithmetic hold, which I won't bore you with here. The three most famous fields are the rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, and the complex numbers $\mathbb{C}$. However, there are lots of other interesting fields.

Number theorists are especially fond of algebraic number fields. An "algebraic number" is a solution to a polynomial equation whose coefficients are rational numbers. You get an "algebraic number field" by taking the field of rational numbers, throwing in finitely many algebraic numbers, and then adding, subtracting, multiplying and dividing them to get more numbers until you've got a field.

For example, we could take the rationals, throw in the square root of 2, and get a field consisting of all numbers of the form

$$a + b\sqrt{2}$$

where a and b are rational. Notice: if we add, multiply, subtract or divide two numbers like this, we get another number of this form. So this is really a field — and it's called $\mathbb{Q}(\sqrt{2})$, since we use round pare$n$theses to denote the result of taking a field and "extending" it by throwing in some extra numbers.

More generally, we could throw in the square root of any integer n, and get an algebraic number field called $\mathbb{Q}(\sqrt{n})$, consisting of all numbers

$$a + b\sqrt{n}$$

where a and b are rational. If $\sqrt{n}$ is rational then this field is just $\mathbb{Q}$, which is boring. Otherwise, we call it a "quadratic number field".

Even more generally, we could take the rationals and throw in a solution of any quadratic equation with rational coefficients. But it's easy to see that this doesn't give anything beyond fields like $\mathbb{Q}(\sqrt{n})$. And that's the real reason we call these the "quadratic number fields".

There are also "cubic number fields", and "quartic number fields", and "quintic number fields", and so on. And others, too, where we throw in solutions to a whole bunch of polynomial equations!

Now, it turns out you can answer lots of classic but rather goofy-sounding number theory puzzles like "which integers are a sum of two squares?" by converting them into questions about algebraic number fields. And the good part is, the resulting questions are connected to all sorts of other topics in math — they're not just glorified mental gymnastics! So, from a modern viewpoint, a bunch of classic number theory puzzles are secretly just tricks to get certain kinds of people interested in algebraic number fields.

But right now I *don't* want to explain how we can use algebraic number fields to solve classic but goofy-sounding number theory puzzles. In fact, I want to downplay the whole puzzle aspect of number theory.

Instead, I hope you're reeling with horror at thought of this vast complicated wilderness of fields containing $\mathbb{Q}$ but contained in $\mathbb{C}$. First there's a huge infinite thicket of algebraic number fields... and then, there's an ever scarier jungle of fields that contain transcendental numbers like π and e! I won't even talk about *that* jungle, it's so dark and scary. Physicists usually zip straight past this whole wilderness and work with $\mathbb{C}$.

But in fact, if you stop and carefully examine all the algebraic number fields and how they sit inside each other, you'll find some incredibly beautiful patterns. And these patterns are turning out to be related to Feynman diagrams, topological quantum field theory, and so on...

However, before we can talk about all that, we need to understand the basic tool for analyzing how one field fits inside another: Galois theory!

A function from a field to itself that preserves addition, subtraction, multiplication and division is called an "automorphism". It's just a *symmetry* of the field. But now, suppose we have a field K which contains some smaller field k. Then we define the "Galois group of K over k" to be the group of all automorphisms of K that act as the identity on k. We call this group

$$\mathrm{Gal}(K/k)$$

for short.

The classic example, familiar to all physicists, is the Galois group of the complex numbers, $\mathbb{C}$, over the real numbers, $\mathbb{R}$. This group has two elements: the identity transformation, which leaves everything alone, and complex conjugation, which switches i and $-i$. Since the only group with 2 elements is $\mathbb{Z}/2$, we have

$$\mathrm{Gal}(\mathbb{C}/\mathbb{R}) = \mathbb{Z}/2$$

Where does complex conjugation come from? It comes from the fact that we get $\mathbb{C}$ from $\mathbb{R}$ by throwing in a solution of the quadratic equation

$$x^2 = -1$$

We say $\mathbb{C}$ is a "quadratic extension of $\mathbb{R}$". But as soon as we throw in one solution of this equation, we inevitably throw in another, namely its negative — and there's no way to tell which is which. And complex conjugation is the symmetry that switches them!

Note: we know that i and $-i$ are different, but we can't tell which is which! This sounds a bit odd at first. It's a bit hard to explain precisely in ordinary language, which is part of why Galois had to invent group theory. But it's fun to try to explain it in plain English... so let me try. The complex numbers have two solutions to

$$x^2 = -1$$

By convention, one of them is called "i", and the other is called "$-i$". Having made this convention, there's never any problem telling them apart. But we could reverse our convention and nothing would go wrong. For example, if the ghost of Galois wafted into your office one moonless night and wrote "$-i$" in all your math and physics books wherever there had been "i", everything in these books would still be true!

Here's another way to think about it. Suppose we meet some extraterrestrials and find that they too have developed the complex numbers by taking the real numbers and adjoining a square root of -1, only they call it "@". Then there would be no way for us to tell if their "@" was our "i" or our "$-i$". All we can do is choose an arbitrary convention as to which is which.

Of course, if they put their "@" in the lower halfplane when drawing the complex plane, we might feel like calling it "$-i$"... but here we are secretly making use of a convention for matching their complex plane with ours, and the *other* convention would work equally well! If they drew their real line *vertically* in the complex plane, it would be more obvious that we need a convention to match their complex plane with ours, and that there are two conventions for doing this, both perfectly self-consistent.

If you've studied enough physics, this extraterrestrial scenario should remind you of those thought experiments where you're trying to explain to some alien civilization the difference between left and right... by means of radio, say, where you're *not* allowed to refer to specific objects you both know — so it's cheating to say "imagine you're on Earth looking at the Big Dipper and the handle is pointing down; then Arcturus is to the right."

If the laws of physics didn't distinguish between left and right, you couldn't explain the difference between left and right without "cheating" like this, so the laws of physics would have a symmetry group with two elements: the identity and the transformation that switches left and right. As it turns out, the laws of physics *do* distinguish between left and right — see "Week 73" for more on that. But that's another story. My point here is that the Galois group of $\mathbb{C}$ over $\mathbb{R}$ is a similar sort of thing, but built into the very fabric of mathematics! And that's why complex conjugation is so important.

I could tell you a nice long story about how complex conjugation is related to "charge conjugation" (switching matter and antimatter) and also "time reversal" (switching past and future). But I won't!

Here's another example of a Galois group that physicists should like. Let $\mathbb{C}(z)$ be the field of rational functions in one complex variable z — in other words, functions like

$$f(z) = \frac{P(z)}{Q(z)}$$

where P and Q are polynomials in z with complex coefficients. You can add, subtract, multiply and divide rational functions and get other rational functions, so they form a field. And they contain $\mathbb{C}$ as a subfield, because we can think of any complex number as a *constant* function. So, we can ask about the Galois group of $\mathbb{C}(z)$ over $\mathbb{C}$. What's it like?

It's the Lorentz group!

To see this, it's best to think of rational functions as functions not on the complex plane but on the "Riemann sphere" — the complex plane together with one extra point, the "point at infinity". The only conformal transformations of the Riemann sphere are "fractional linear transformations":

$$T(z) = \frac{az + b}{cz + d}$$

So, the only symmetries of the field of rational functions that act as the identity on constant functions are those coming from fractional transformations, like this:

$$f \mapsto fT \qquad \text{where } fT(z) = f(T(z))$$

If you don't follow my reasoning here, don't worry — the details aren't hard to fill in, but they'd be distracting here.

The last step is to check that the group of fractional linear transformations is the same as the Lorentz group. You can do this algebraically, but you can also do it geometrically by thinking of the Riemann sphere as the "heavenly sphere": that imaginary sphere the stars look like they're sitting on. The key step is to check this remarkable fact: if you shoot past the earth near the speed of light, the constellations will look distorted by a Lorentz transformation — but if you draw lines connecting the stars, all the *angles* between these lines will remain the same; only their *lengths* will get messed up!

Moreover, it's obvious that if you rotate your head, both angles and lengths on the heavenly sphere are preserved. So, any rotation or Lorentz boost gives an angle-preserving transformation of the heavenly sphere — that is, a conformal transformation! And this must be a fractional linear transformation.

Summarizing, the Galois group of $\mathbb{C}(z)$ over $\mathbb{C}$ is the Lorentz group, or more precisely, its connected component, $\mathrm{SO}_0(3,1)$:

$$\mathrm{Gal}(\mathbb{C}(z)/\mathbb{C}) = \mathrm{SO}_0(3,1).$$

We've talked about the Galois group of $\mathbb{C}(z)$ over $\mathbb{C}$ and the Galois group of $\mathbb{C}$ over $\mathbb{R}$. What about the Galois group of $\mathbb{C}(z)$ over $\mathbb{R}$? Unsurprisingly, this is the group of transformations of the Riemann sphere generated by fractional linear transformations *and* complex conjugation. And physically, this corresponds to taking the connected component of the Lorentz group and throwing in *time reversal!* So you see, complex conjugation is related to time reversal. But I promised not to go into that. . . .

I've been talking about Galois groups that physicists should like, but you're probably wondering where the number theory went! Well, it's all part of the same big story. In number theory we're especially interested in Galois groups like

$$\mathrm{Gal}(K/k)$$

where K is some algebraic number field and k is some subfield of K. For starters, consider this example:

$$\mathrm{Gal}(\mathbb{Q}(\sqrt{n})/\mathbb{Q})$$

where $\sqrt{n}$ is irrational. I've already hinted at what this group is! $\mathbb{Q}(\sqrt{n})$ has $\sqrt{n}$ in it, so it also has $-\sqrt{n}$ in it, and there's an automorphism that switches these two while leaving all the rational numbers alone, namely

$$a + b\sqrt{n} \mapsto a - b\sqrt{n} \qquad (a, b \text{ in } \mathbb{Q})$$

So, we have:

$$\mathrm{Gal}(\mathbb{Q}(\sqrt{n}))/\mathbb{Q}) = \mathbb{Z}/2$$

just like the Galois group of $\mathbb{C}$ over $\mathbb{R}$.

To get some bigger Galois groups, let's take $\mathbb{Q}$ and throw in a "primitive nth root of unity". Hmm, I may need to explain what that means. There are n different nth roots of 1 — but unlike the two square roots of -1, these are not all created equal! Only some are "primitive".

For example, among the 4th roots of unity we have 1 and -1, which are actually square roots of unity, and i and $-i$, which aren't. A "primitive nth root of unity" is an nth root of 1 that's not an kth root for any $k < n$. If you take all the powers of any primitive nth root of unity, you get *all* the nth roots of unity. So, if we take some primitive nth root of unity, call it

$$1^{\frac{1}{n}}$$

for lack of a better name, and extend the rationals by this number, we get a field

$$\mathbb{Q}(1^{\frac{1}{n}})$$

which contains all the nth roots of unity. Since the nth roots of unity are evenly distributed around the unit circle, this sort of field is called a "cyclotomic field", for the Greek word for "circle cutting". In fact, one can apply Galois theory to this field to figure out which regular n-gons one can construct with a ruler and compass!

But what's the Galois group

$$\mathrm{Gal}(\mathbb{Q}(1^{\frac{1}{n}})/\mathbb{Q})$$

like? Any symmetry in this group must map $1^{\frac{1}{n}}$ to some root of unity, say $1^{\frac{m}{n}}$ — and once you know which one, you completely know the symmetry. But actually, this symmetry must map $1^{\frac{1}{n}}$ to some *primitive* root of unity, so m has to be relatively prime to n. Apart from that, though, anything goes — so the size of

$$\mathrm{Gal}(\mathbb{Q}(1^{\frac{1}{n}})/\mathbb{Q})$$

is just the number of guys m less than n that are relatively prime to n. And if you think about it, these numbers relatively prime to n are just the same as elements of $\mathbb{Z}/n$ that have multiplicative inverses! So if you think some more, you'll see that

$$\mathrm{Gal}(\mathbb{Q}(1^{\frac{1}{n}})/\mathbb{Q}) = (\mathbb{Z}/n)^{\times}$$

where $(\mathbb{Z}/n)^{\times}$ is the "multiplicative group" of $\mathbb{Z}/n$ — that is, the elements of $\mathbb{Z}/n$ that have multiplicative inverses, made into a group via multiplication!

This group can be big, but it's still abelian. Can we get some nonabelian Galois groups from algebraic number fields?

Sure! Let's say you take some polynomial equation with rational coefficients, take *all* its solutions, throw them into the rationals - and keep adding, subtracting, multiplying and dividing until you get some field K. This K is called the "splitting field" of your polynomial.

But here's the interesting thing: if you pick your polynomial equation at random, the chances are really good that it has n different solutions if the polynomial is of degree n, and that *any* permutation of these solutions comes from a unique symmetry of the field K. In other words: barring some coincidence, all roots are created equal! So in general we have

$$\mathrm{Gal}(K/\mathbb{Q}) = S_n$$

where S_n is the group of all permutations of n things.

Sometimes of course the Galois group will be smaller, since our polynomial could have repeated roots or, more subtly, algebraic relations between roots — as in the cyclotomic case we just looked at.

But, we can already start to see how to prove the unsolvability of the general quintic! Pick some random 5th-degree polynomial, let K be its splitting field, and note

$$\mathrm{Gal}(K/\mathbb{Q}) = S_5$$

Then, show that if we build up an algebraic number field by starting with $\mathbb{Q}$ and repeatedly throwing in nth roots of numbers we've already got, we just can't get S_5 as its Galois group over the rationals! We've already seen this in the case where we throw in a square root of n, or an nth root of 1. The general case is a bit more work. But instead of giving the details, I'll just mention a good textbook on Galois theory for beginners:

1) Ian Stewart, *Galois Theory*, 3rd edition, Chapman and Hall, New York, 2004.

Ian Stewart is famous as a popularizer of mathematics, and it shows here — he has nice discussions of the history of the famous problems solved by Galois theory, and a nice demystification of the Galois' famous duel. But, this is a real math textbook — so you can really learn Galois theory from it! Make sure to get the 3rd edition, since it has more examples than the earlier ones.

Having given Ian Stewart the dirty work of explaining Galois theory in the usual way, let me say some things that few people admit in a first course on the subject.

So far, we've looked at examples of a field k contained in some bigger field K, and worked out the group $\mathrm{Gal}(K/k)$ consisting of all automorphisms of K that fix everything in k.

But here's the big secret: this has NOTHING TO DO WITH FIELDS! It works for ANY sort of mathematical gadget! If you've got a little gadget k sitting in a big gadget K, you get a "Galois group" $\mathrm{Gal}(K/k)$ consisting of symmetries of the big gadget that fix everything in the little one.

But now here's the cool part, which is also very general. Any subgroup of $\mathrm{Gal}(K/k)$ gives a gadget containing k and contained in K: namely, the gadget consisting of all the elements of K that are fixed by everything in this subgroup.

And conversely, any gadget containing k and contained in K gives a subgroup of $\mathrm{Gal}(K/k)$: namely, the group consisting of all the symmetries of K that fix every element of this gadget.

This was Galois' biggest idea: we call this a GALOIS CORRESPONDENCE. It lets us use *group theory* to classify gadgets contained in one and containing another. He applied it to fields, but it turns out to be useful much more generally.

Now, it would be great if the Galois corresondence were always a perfect 1-1 correspondence between subgroups of $\mathrm{Gal}(K/k)$ and gadgets containing k and contained in K. But, it ain't true. It ain't even true when we're talking about fields!

However, that needn't stop us. For example, we can restrict ourselves to cases when it *is* true. And this is where the Fundamental Theorem of Galois Theory comes in! It's easiest to state this theorem when k and K are algebraic number fields, so that's what I'll do. In this case, there's a 1-1 correspondence between subgroups of $\mathrm{Gal}(K/k)$ and extensions of k contained in K if:

i) K is a "finite" extension of k. In other words, K is a finite-dimensional vector space over k.

ii) K is a "normal" extension of k. In other words, if a polynomial with coefficients in k can't be factored at all in k, but has one root in K, then all its roots are in K.

For general fields we also need another condition, namely that K be a "separable" extension of k. But this is automatic for algebraic number fields, so let's not worry about it.

At this point, if we had time, we could work out a bunch of Galois groups and see a bunch of patterns. Using these, we could see why you can't solve the general quintic using radicals, why you can't trisect the angle or double the cube using ruler-and-compass constructions, and why you can draw a regular pentagon using ruler and compass, but not a regular heptagon. Basically, to prove something is impossible, you just show that some number can't possibly lie in some particular algebraic number field, because it's the root of a polynomial whose splitting field has a Galois group that's "fancier" than the Galois group of that algebraic number field.

For example, ruler-and-compass constructions produce distances that lie in "iterated quadratic extensions" of the rationals — meaning that you just keep throwing in square roots of stuff you've got. Doubling the cube requires getting your hands on the cube root of 2. But the Galois group of the splitting field of

$$x^3 = 2$$

has size divisible by 3, while an iterated quadratic extension has a Galois group whose size is a power of 2. Using the Galois correspondence, we see there's no way to stuff the former field into the latter.

But you can read about this in any good book on Galois theory, so I'd rather dive right into that thicket I was hinting at earlier: the field of ALL algebraic numbers! The roots of any polynomial with coefficients in this field again lie in this field, so we say this field is "algebraically closed". And since it's the smallest algebraically closed field containing $\mathbb{Q}$, it's called the "algebraic closure of $\mathbb{Q}$", or $\overline{\mathbb{Q}}$ for short — that is, $\mathbb{Q}$ with a bar over it. (I can't quite draw it here.)

This field $\overline{\mathbb{Q}}$ is huge. In particular, it's an infinite-dimensional vector space over $\mathbb{Q}$. So, condition i) in the Fundamental Theorem of Galois Theory doesn't hold. But that's no disaster: when this happens, we just need to put a topology on the group $\mathrm{Gal}(K/k)$ and set up the Galois correspondence using *closed* subgroups of $\mathrm{Gal}(K/k)$. Using this trick, every algebraic number field corresponds to some closed subgroup of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$.

So, for people studying algebraic number fields,

$$\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$$

is like the holy grail. It's the symmetry group of the algebraic numbers, and the key to how all algebraic number fields sit inside each other! But alas, this group is devilishly complicated. In fact, it has literally driven men mad. One of my grad students knows someone who had a breakdown and went to the mental hospital while trying to understand this group!

(There may have been other reasons for his breakdown, too, but as readers of E. T. Bell's book "Men in Mathematics" know, the facts should never get in the way of a good anecdote.)

If $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ were just an infinitely tangled thicket, it wouldn't be so tantalizing. But there are things we can understand about it! To describe these, I'll have to turn up the math level a notch. . .

First of all, an extension K of a field k is called "abelian" if $\mathrm{Gal}(K/k)$ is an abelian group. Abelian extensions of algebraic number fields can be understood using something called class field theory. In particular, the Kronecker-Weber theorem says that every finite abelian extension of $\mathbb{Q}$ is contained in a cyclotomic field. So, they all sit inside a field called $\mathbb{Q}^{\mathrm{cyc}}$, which is gotten by taking the rationals and throwing in *all* nth roots of unity for *all* n. Since

$$\mathrm{Gal}(\mathbb{Q}(1^{\frac{1}{n}})/\mathbb{Q}) = (\mathbb{Z}/n)^\times$$

we know from Galois theory that $\mathrm{Gal}(\mathbb{Q}^{\mathrm{cyc}}/\mathbb{Q})$ must be a big group containing all the groups $(\mathbb{Z}/n)^\times$ as closed subgroups. It's easy to see that $(\mathbb{Z}/n)^\times$ is a quotient group of $(\mathbb{Z}/m)^\times$ if m is divisible by n; this lets us take the "inverse limit" of all the groups $(\mathbb{Z}/m)^\times$ — and that's $\mathrm{Gal}(\mathbb{Q}^{\mathrm{cyc}}/\mathbb{Q})$. This inverse limit is also the multiplicative group of the ring $\widehat{\mathbb{Z}}$, the inverse limit of all the rings $\mathbb{Z}/n$. $\widehat{\mathbb{Z}}$ is also called the "profinite completion of the integers", and I urge you to play around with it if you never have! It's a cute gadget.

In short:

$$\mathrm{Gal}(\mathbb{Q}^{\mathrm{cyc}}/\mathbb{Q}) = \widehat{\mathbb{Z}}^\times$$

and if we stay inside $\mathbb{Q}^{\mathrm{cyc}}$, we're in a zone where the pattern of algebraic number fields can be understood. This stuff was worked out by people like Weber, Kronecker, Hilbert and Takagi, with the final keystone, the Artin reciprocity theorem, laid in place by Emil Artin in 1927. In a certain sense $\mathbb{Q}^{\mathrm{cyc}}$ is to $\overline{\mathbb{Q}}$ as homology theory is to homotopy theory: it's all about *abelian* Galois groups, so it's manageable.

People now use $\mathbb{Q}^{\mathrm{cyc}}$ as a kind of base camp for further expeditions into the depths of $\overline{\mathbb{Q}}$. In particular, since

$$\mathbb{Q} \subset \mathbb{Q}^{\mathrm{cyc}} \subset \overline{\mathbb{Q}}$$

we get an exact sequence of Galois groups:

$$1 \to \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}^{\mathrm{cyc}}) \to \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{Gal}(\mathbb{Q}^{\mathrm{cyc}}/\mathbb{Q}) \to 1$$

So, to understand $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ we need to understand $\mathrm{Gal}(\mathbb{Q}^{\mathrm{cyc}}/\mathbb{Q})$, $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}^{\mathrm{cyc}})$ and how they fit together! The last two steps are not so easy. Shafarevich has conjectured that $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}^{\mathrm{cyc}})$ is the profinite completion of a free group, say $\widehat{F}$. This would give

$$1 \to \widehat{F} \to \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathbb{Z}^\times \to 1$$

but I have no idea how much evidence there is for Shafarevich's conjecture, or how much people know or guess about this exact sequence.

More recently, Deligne has turned attention to a certain "motivic" version of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, which is a proalgebraic group scheme. This sort of group has a *Lie algebra*, which makes it more tractable. And there are a bunch of fascinating conjectures about this Lie algebra is related to the Riemann zeta function at odd numbers, Connes and Kreimer's work on Feynman diagrams, Drinfeld's work on the Grothendieck-Teichmueller group, and more!

I really want to understand this stuff better — right now, it's a complete muddle in my mind. When I do, I will report back to you. For now, though, let me give you some references.

For two very nice but very different introductions to algebraic number fields, try these:

2) H. P. F. Swinnerton-Dyer, *A Brief Guide to Algebraic Number Theory*, Cambridge U. Press, Cambridge 2001.

3) Juergen Neukirch, *Algebraic Number Theory*, trans. Norbert Schappacher, Springer, Berlin, 1986.

Both assume you know some Galois theory or at least can fake it. Neukirch's book is good for the all-important analogy between Galois groups and fundamental groups, which I haven't even touched upon here! Swinnerton-Dyer's book has the virtue of brevity, so you can see the forest for the trees. Both have a friendly, slightly chatty style that I like.

For Shafarevich's conjecture, try this:

4) K. Iwasawa, "On solvable extensions of algebraic number fields", *Ann. Math.* **58** (1953) 548–572.

For Deligne's motivic analogue, try this:

5) Pierre Deligne, "Le groupe fondamental de la droite projective moins trois points", in *Galois Groups over* $\mathbb{Q}$, MSRI Publications **16** (1989), 79–313.

This stuff has a lot of relationships to 3d topological quantum field theory, braided monoidal categories, and the like... and it all goes back to the Grothendieck-Teichmueller group. To learn about this group try this book, and especially this article in it:

6) Leila Schneps, "The Grothendieck-Teichmuller group: a survey", in *The Grothendieck Theory of Dessins D'Enfants*, London Math. Society Notes **200**, Cambridge U. Press, Cambridge 1994, pp. 183–204.

To hear and watch some online lectures on this material, try:

7) Leila Schneps, "The Grothendieck-Teichmuller group and fundamental groups of moduli spaces", MSRI lecture available at `http://www.msri.org/publications/ln/msri/1999/vonneumann/schneps/1/`

   "Grothendieck-Teichmuller group and Hopf algebras", MSRI lecture available at `http://www.msri.org/publications/ln/msri/1999/vonneumann/schneps/2/`

For a quick romp through many mindblowing ideas which touches on this material near the end:

8) Pierre Cartier, "A mad day's work: from Grothendieck to Connes and Kontsevich — the evolution of concepts of space and symmetry", *Bulletin of the AMS* **38** (2001), 389–408. Also available at http://www.ams.org/

For even more mindblowing ideas along these lines:

9) Jack Morava, "The motivic Thom isomorphism", talk at the Newton Institute, December 2002, also available at `math.AT/0306151`.

Addendum: I received the following email from Avinoam Mann, which corrects some mistakes I made:

> *Dear John,*
>
> *It's very nice that you've come to appreciate the beauties of number theory, and I enjoyed reading your description of Galois theory, but I hope that you would not mind if I ask you not to help spread some common misunderstandings about it. First, it was not Galois who proved the impossibility of solving the quintic by radicals. This was attempted first by Ruffini, I think in 1799, and the proof by Abel, about ten years befors Galois, was the one that the mathematical community accepted. While I often teach Galois theory (e.g. next semester), I never studied Ruffini's and Abel's work in detail. What Galois did was to give a criterion checking for an arbitrary equation whether it is soluble by radicals or not.*
>
> *Another point: there is no need for Galois theory to prove that duplication of the cube and trisection of an angle cannot be done by ruler and compass. Since*

> *ruler and compass constructions are equivalent to solving a series of quadratics, they can lead only to fields $\mathbb{F}$ of dimension 2^n, for some n, over the rationals. But the two problems that I mentioned lead to extensions of dimension 3. All this is very elementary. Similar considerations lead to necessary conditions for the constructibility of regular polygons, but proving these conditions sufficient does require more theory (unless, I guess, you provide directly the relevant system of quadratics; I think that is what Gauss did — his proof also preceded Galois). Squaring the circle is, of course, a different matter. Here we need the transcendence of π.*
>
> *Best wishes from wet Jerusalem,*
>
> *Avinoam Mann*

It's true that Abel and Ruffini beat Galois when it came to the quintic; the details of this history are covered pretty well by Ian Stewart's book, I think. And, it's quite true that one doesn't need of Galois theory to solve a bunch of these problems: for example, to show one can't duplicate the cube, we just need to see that $\mathbb{Q}(2^{1/3})$ has dimension 3 as a vector space over $\mathbb{Q}$, while quadratic extensions have dimension 2^n. My use of the Galois correspondence to express this in terms of the size of certain Galois groups was overkill! The real point of Galois theory is that it provides a unified framework for tackling a wide range of problems.

> Paris, 1 June — A deplorable duel yesterday has deprived the exact sciences of a young man who gave the highest expectations, but whose celebrated precosity was lately overshadowed by his political activities. The young Evariste Galois... was fighting with one of his old friends, a young man like himself, like himself a member of the Society of Friends of the People, and who was known to have figured equally in a political trial. It is said that love was the cause of the combat. The pistol was the chosen weapon of the adversaries, but because of their old friendship they could not bear to look at one another and left their decision to blind fate.
>
> *— Le Precursor, June 4, 1832*

Week 202

February 21, 2004

This week I'll deviate from my plan of discussing number theory, and instead say a bit about something else that's been on my mind lately: structure types. But, you'll see my fascination with Galois theory lurking beneath the surface.

Andre Joyal invented structure types in 1981 — he called them "espces de structure", and lots of people call them "species". Basically, a structure type is just any sort of structure we can put on finite sets: an ordering, a coloring, a partition, or whatever. In combinatorics we count such structures using "generating functions". A generating function is a power series where the coefficient of x^n keeps track of how many structures of the given kind we can put on an n-element set. By playing around with these functions, we can often figure out the coefficients and get explicit formulas — or at least asymptotic formulas — that count the structures in question.

The reason this works is that operations on generating functions come from operations on structure types. For example, in "Week 190", I described how addition, multiplication and composition of generating functions correspond to different ways to get new structure types from old.

Joyal's great contribution was to give structure types a rigorous definition, and use this to show that many calculations involving generating functions can be done directly with structure types. It turns out that just as generating functions form a *set* equipped with various operations, structure types form a *category* with a bunch of completely analogous operations. This means that instead of merely proving *equations* between generating functions, we can construct *isomorphisms* between their underlying structure types — which imply such equations, but are worth much more. It's like the difference between knowing two things are equal and knowing a specific reason WHY they're equal!

Of course, this business of replacing equations by isomorphisms is called "categorification". In this lingo, structure types are categorified power series, just as finite sets are categorified natural numbers.

A while back, James Dolan and I noticed that since you can use power series to describe states of the quantum harmonic oscillator, you can think of structure types as states of a categorified version of this physical system! This gives new insights into the combinatorial underpinnings of quantum physics.

For example, the discrete spectrum of the harmonic oscillator Hamiltonian can be traced back to the discreteness of finite sets! The commutation relations between annihilation and creation operators boil down to a very simple fact: there's one more way to put a ball in a box and then take one out, than to take one out and then put one in. Even better, the whole theory of Feynman diagrams gets a simple combinatorial interpretation. But for this, one really needs to go beyond structure types and work with a generalization called "stuff types".

I've been thinking about this business for a while now, so last fall I decided to start giving a year-long course on categorification and quantization. The idea is to explain bunches of quantum theory, quantum field theory and combinatorics all from this new point of view. It's fun! Derek Wise has been scanning in his notes, and a bunch of people have been putting their homework online. So, you can follow along if you want:

1) John Baez and Derek Wise, "Quantization and Categorification".
Fall 2003 notes: `http://math.ucr.edu/home/baez/qg-fall2003`
Winter 2004 notes: `http://math.ucr.edu/home/baez/qg-winter2004/`
Spring 2004 notes: `http://math.ucr.edu/home/baez/qg-spring2004/`

I'd like to give you a little taste of this subject now. But, instead of explaining it in detail, I'll just give some examples of how structure types yield some far-out generalizations of the concept of "cardinality". This stuff is a continuation of some themes developed in "Week 144", "Week 147", "Week 185", "Week 190", so I'll start with a review.

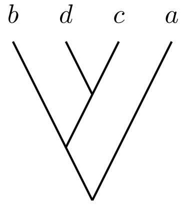
Suppose F is a structure type. Let F_n be the *set* of ways we can put this structure on a n-element set, and let $|F_n|$ be the *number* of ways to do it. In combinatorics, people take all these numbers $|F_n|$ and pack them into a single power series. It's called the generating function of F, and it's defined like this:

$$|F|(x) = \sum \frac{|F_n|}{n!} x^n$$

It may not converge, so in general it's just a "formal" power series — but for interesting structure types it often converges to an interesting function.

What's good about generating functions is that simple operations on them correspond to simple operations on structure types. We can use this to count structures on finite sets. Let me remind you how it works for binary trees!

There's a structure type T where a T-structure on a set is a way of making it into the leaves of a binary tree drawn in the plane. For example, here's one T-structure on the set $\{a, b, c, d\}$:



Thanks to the choice of different orderings, the number of T-structures on an n-element set is $n!$ times the number of binary trees with n leaves. Annoyingly, the latter number is traditionally called the $(n-1)$st Catalan number, C_{n-1}. So, we have:

$$|T|(x) = \sum_{n=1} C_{n-1} x^n.$$

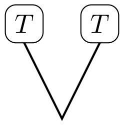
There's a nice recursive definition of T:

> *"To put a T-structure on a set, either note that it has one element, in which case there's just one T-structure on it, or chop it into two subsets and put a T-structure on each one."*

In other words, any binary tree is either a degenerate tree, with just one leaf:

$$X$$

or a pair of binary trees stuck together at the root:



We can write this symbolically as

$$T \cong X + T^2$$

Here's why: X is a structure type called "being the one-element set", $+$ means "exclusive or", and squaring a structure type means you chop your set in two parts and put that structure on each part. (I explained these rules more carefully in "Week 190".)

Note that we only have an *isomorphism* between structure types here, not an equation. But if we take the generating function of both sides we get an actual equation, and the notation is set up to make this really easy:

$$|T| = x + |T|^2$$

In "Week 144" I showed how you can solve this using the quadratic equation:

$$|T| = \frac{1 - \sqrt{1-4x}}{2}$$

and then do a Taylor expansion to get

$$|T| = x + x^2 + 2x^3 + 5x^4 + 14x^5 + 42x^6 + \ldots$$

Lo and behold! The coefficient of x^n is the number of binary trees with n leaves!

There's also another approach where we work directly with the structure types themselves, instead of taking generating functions. This is harder because we can't subtract structure types, or divide them by 2, or take square roots of them — at least, not without stretching the rules of this game. All we can do is use the isomorphism

$$T \cong X + T^2$$

and the basic rules of category theory. It's not as efficient, but it's illuminating. It's also incredibly simple: we just keep sticking in "$X + T^2$" wherever we see "T" on the right-hand side, over and over again. Like this:

$$\begin{aligned} T &\cong X + T^2 \\ &\cong X + (X + T^2)^2 \\ &\cong X + (X + (X + T^2)^2)^2 \end{aligned}$$

and so on. You might not think we're getting anywhere, but if you stop at the nth stage and expand out what we've got, you'll get the first n terms of the Taylor expansion we had before! At least, you will if you count "stages" and "terms" correctly.

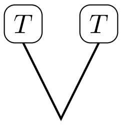
I won't actually do this, because it's better if you do it yourself. When you do, you'll see it captures the recursive process of building a binary tree from lots of smaller binary trees. Each time you see a "T" and replace it with an "$X + T^2$", you're really taking a little binary tree:

$$\boxed{T}$$

and replacing it with either a degenerate tree with just a single leaf:

$$X$$

or a pair of binary trees:



So, each term in the final result actually corresponds to a specific tree! This is a good example of categorification: when we calculate the coefficient of x^n this way, we're not just getting the *number* of binary planar trees with n leaves — we're getting an actual explicit description of the *set* of such trees.

Now, what happens if we take the generating function $|T|(x)$ and evaluate it at $x = 1$? On the one hand, we get a divergent series:

$$|T|(1) = 1 + 1 + 2 + 5 + 14 + 42 + \dots$$

This is the sum of all Catalan numbers — or in other words, the number of binary planar trees. On the other hand, we can use the formula

$$|T| = \frac{1 - \sqrt{1 - 4x}}{2}$$

to get

$$|T| = \frac{1 - \sqrt{-3}}{2}$$

It may seem insane to conclude

$$1 + 1 + 2 + 5 + 14 + 42 + \dots = \frac{1 - \sqrt{-3}}{2}$$

but Lawvere noticed that there's a kind of strange sense to it.

The trick is to work not with generating function $|T|$ but with the structure type T itself. Since $|T|(1)$ is equal to the *number* of planar binary trees, $T(1)$ should be naturally isomorphic to the *set* of planar binary trees. And it is — it's obvious, once you think about what it really means.

The number of binary planar trees is not very interesting, but the set of them is. In particular, if we take the isomorphism

$$T \cong X + T^2$$

and set $X = 1$, we get an isomorphism

$$T(1) \cong 1 + T(1)^2$$

which says

> *"a planar binary tree is either the tree with one leaf or a pair of planar binary trees."*

Starting from this, we can derive lots of other isomorphisms involving the set $T(1)$, which turn out to be categorified versions of equations satisfied by the number

$$|T|(1) = \frac{1-\sqrt{-3}}{2}$$

For example, this number is a sixth root of unity. While there's no one-to-one correspondence between 6-tuples of trees and the 1 element set, which would categorify the formula

$$|T|(1)^6 = 1$$

there *is* a very nice one-to-correspondence between 7-tuples of trees and trees, which categorifies the formula

$$|T|(1)^7 = |T|(1)$$

Of course the set of binary trees is countably infinite, and so is the set of 7-tuples of binary trees, so they can be placed in one-to-one correspondence — but that's boring. When I say "very nice", I mean something more interesting: starting with the isomorphism

$$T \cong X + T^2$$

we get a one-to-one correspondence

$$T(1) \cong 1 + T(1)^2$$

which says that any binary planar tree is either degenerate or a pair of binary planar trees... and using this we can *construct* a one-to-one correspondence

$$T(1)^7 \cong T(1)$$

The construction is remarkably complicated. Even if you do it as efficiently as possible, I think it takes 18 steps, like this:

$$\begin{aligned} T(1)^7 &\cong T(1)^6 + T(1)^8 \\ &\cong \dots \\ &\ \ \vdots \\ &\cong 1 + T(1) + T(1)^2 + T(1)^4 \\ &\cong 1 + T(1) + T(1)^3 \\ &\cong 1 + T(1)^2 \\ &\cong T(1). \end{aligned}$$

I'll let you fill in the missing steps — it's actually quite fun if you like puzzles.

If you get stuck, you can find the answer online in a couple of different places:

2) Andreas Blass, "Seven trees in one", *Jour. Pure Appl. Alg.* **103** (1995), 1–21. Also available at `http://www.math.lsa.umich.edu/~ablass/cat.html`

3) Marcelo Fiore, "Isomorphisms of generic recursive polynomial types", to appear in *31st Symposium on Principles of Programming Languages (POPL04)*. Also available at `http://www.cl.cam.ac.uk/~mpf23/papers/Types/recisos.ps.gz`

Or, take a peek at the "Addenda" down below.

Robbie Gates, Marcelo Fiore and Tom Leinster have also proved some very general theorems about this sort of thing. Gates focused on "distributive categories" (categories with with products and coproducts, the former distributing over the latter), while the work of Fiore and Leinster applies to more general "rig categories":

4) Robbie Gates, "On the generic solution to $P(X) = X$ in distributive categories", *Jour. Pure Appl. Alg.* **125** (1998), 191–212.

5) Marcelo Fiore and Tom Leinster, "Objects of categories as complex numbers", available as math.CT/0212377.

A rig category is basically the most general sort of category in which we can "add" and "multiply" as we do in a ring — but without negatives, hence the missing letter "n". It turns out that whenever we have an object Z in a rig category and it's equipped with an isomorphism

$$Z = P(Z)$$

where P is a polynomial with natural number coefficients, we can associate to it a "cardinality" $|Z|$, namely any complex solution of the equation

$$|Z| = P(|Z|)$$

Which solution should we use? Well, for simplicity let's consider the case where P has degree at least 2 and the relevant Galois group acts transitively on the solutions of this equation, so "all roots are created equal". Then we can pick *any* solution as the cardinality $|Z|$. Any polynomial equation with natural number coefficients satisfied by one solution will be satisfied by *all* solutions, so it won't matter which one we choose.

Now suppose the cardinality $|Z|$ satisfies such an equation:

$$Q(|Z|) = R(|Z|)$$

where neither Q nor R is constant. Then the results of Fiore and Leinster say we can construct an isomorphism

$$Q(Z) = R(Z)\ !$$

In other words, a bunch of equations satisfied by the object's cardinality automatically come from isomorphisms involving the object itself.

This explains why the set $T(1)$ of binary trees acts like it has cardinality

$$|T|(1) = \frac{1-\sqrt{-3}}{2}$$

or equally well,

$$|T|(1) = \frac{1+\sqrt{-3}}{2}$$

(Since the relevant Galois group interchanges these two numbers, we can use either one.) More generally, the set $T(n)$ consisting of binary trees with n-colored leaves acts a lot like the number $|T|(n)$.

This has gotten me interested in trying to find a nice example of a "Golden Object": an object G in some rig category that's equipped with an isomorphism

$$G^2 = G + 1$$

The Golden Object doesn't fit into Fiore and Leinster's formalism, since this isomorphism is not of the form $G = P(G)$ where P has natural number coefficients. But, it still seems that such an object deserves to have a "cardinality" equal to the golden number:

$$|G| = \frac{1+\sqrt{5}}{2} = 1.61803398874989484820458683436 5\ldots$$

James Propp came up with an interesting idea related to the Golden Object: consider what happens when we evaluate the generating function for binary trees at -1. On the one hand we get an alternating sum of Catalan numbers:

$$|T|(-1) = -1 + 1 - 2 + 5 - 14 + 42 + \ldots$$

On the other hand, we can use the formula

$$|T| = \frac{1-\sqrt{1-4x}}{2}$$

to get

$$|T|(-1) = \frac{1-\sqrt{5}}{2}$$

which is -1 divided by the golden number. Of course, it's possible we should use the other sign of the square root, and get

$$|T|(-1) = 1 + \sqrt{5}2$$

which is just the golden number! Galois theory says these two roots are created equal. Either way, we get a bizarre and fascinating formula:

$$-1 + 1 - 2 + 5 - 14 + 42 + \ldots = \frac{1 \pm \sqrt{5}}{2}$$

Can we fit this into some clear and rigorous framework, or is it just nuts? We'd like some generalization of cardinality for which "the set of binary trees with -1-colored leaves" has cardinality equal to the golden number.

James Propp suggested one avenue. A while back, Steve Schanuel made an incredibly provocative observation: if we treat "Euler measure" as a generalization of cardinality, it makes sense to treat the real line as a "space of cardinality -1":

6) Stephen H. Schanuel, "What is the length of a potato?: an introduction to geometric measure theory", in *Categories in Continuum Physics*, Springer Lecture Notes in Mathematics **1174**, Springer, Berlin, 1986, pp. 118–126.

7) Stephen H. Schanuel, "Negative sets have Euler characteristic and dimension", Lecture Notes in Mathematics **1488**, Springer Verlag, Berlin, 1991, pp. 379–385.

James Propp has developed this idea in a couple of fascinating papers:

8) James Propp, "Euler measure as generalized cardinality", available as arXiv:math/0203289.

9) James Propp, "Exponentiation and Euler measure", available as arXiv:math/0204009.

Using this idea, it seems reasonable to consider the space of binary trees with leaves labelled by real numbers as a rigorous version of "the set of binary trees with -1-colored leaves". So, we just need to figure out what generalization of Euler characteristic gives this space an Euler characteristic equal to the golden number. It would be great if we could make this space into a Golden Object in some rig category, but that may be asking too much.

Whew! There's obviously a lot of work left to be done here. Here's something easier: a riddle. What's this sequence?

un, dos, tres, quatre, cinc, sis, set, vuit, nou, deu,...

The answer is at the end of this article.

Now I'd like to mention some important papers on n-categories. You may think I'd lost interest in this topic, because I've been talking about other things. But it's not true!

Most importantly, Tom Leinster has come out with a big book on n-categories and operads:

10) Tom Leinster, *Higher Operads, Higher Categories*, Cambridge U. Press, Cambridge, 2003. Also available as arXiv:math.CT/0305049.

As you'll note, he managed to talk the press into letting him keep his book freely available online! We should all do this. Nobody will ever make much cash writing esoteric scientific tomes — it takes so long, you could earn more per hour digging ditches. The only *financial* benefit of writing such a book is that people will read it, think you're smart, and want to hire you, promote you, or invite you to give talks in cool places. So, maximize your chances of having people read your books by keeping them free online! People will still buy the paper version if it's any good....

And indeed, Leinster's book has many virtues besides being free. He gracefully leads the reader from the very basics of category theory straight to the current battle front of weak n-categories, emphasizing throughout how operads automatically take care of the otherwise mind-numbing thicket of "coherence laws" that inevitably infest the subject. He doesn't take well-established notions like "monoidal category" and "bicategory" for granted — instead, he dives in, takes their definitions apart, and compares alternatives to see what makes these concepts tick. It's this sort of careful thinking that we desperately need if we're ever going to reach the dream of a clear and powerful theory of higher-dimensional algebra. He does a similar careful analysis of "operads" and "multicategories" before presenting a generalized theory of operads that's powerful enough to support various different approaches to weak n-categories. And then he describes and compares some of these different approaches!

In short: if you want to learn more about operads and n-categories, this is *the* book to read.

Leinster doesn't say too much about what n-categories are good for, except for a nice clear introduction entitled "Motivation for Topologists", where he sketches their relevance to homology theory, homotopy theory, and cobordism theory. But this is understandable, since a thorough treatment of their applications would vastly expand an already hefty 380-page book, and diffuse its focus. It would also steal sales from *my* forthcoming book on higher-dimensional algebra — which would be really bad, since I plan to retire on the fortune I'll make from this.

Secondly, Michael Batanin has worked out a beautiful extension of his ideas on n-categories which sheds new light on their applications to homotopy theory:

11) Michael A. Batanin, "The Eckmann-Hilton argument, higher operads and E_n spaces", available as arXiv:math.CT/0207281.

Michael A. Batanin, "The combinatorics of iterated loop spaces", available as arXiv:math.CT/0301221.

Getting a manageable combinatorial understanding of the space of loops in the spaces of loops in the space of loops... in some space has always been part of the dream of higher-dimensional algebra. These "k-fold loop spaces" or have been important in homotopy theory since the 1970s — see the end of "Week 199" for a little bit about them. People know that k-fold loop spaces have k different products that commute up to homotopy in a certain way that can be summarized by saying they are algebras of the E_k operad, also called the "little k-cubes operad". However, their wealth of structure is still a bit mind-boggling. James Dolan and I made some conjectures about their relation to k-tuply monoidal categories in our paper "Categorification" (see "Week 121"), and now Batanin is making this more precise using his approach to n-categories — which is one of the ones described in Leinster's book.

There's also been a lot of work applying higher-dimensional algebra to topological quantum field theory — that's what got me interested in n-categories in the first place, but a lot has happened since then. For a highly readable introduction to the subject, with tons of great pictures, try:

12) Joachim Kock, *Frobenius Algebras and 2D Topological Quantum Field Theories*, Cambridge U. Press, Cambridge, 2003.

This is mainly about 2d TQFTs, where the concept of "Frobenius algebra" reigns supreme, and everything is very easy to visualize.

When we go up to 3-dimensional spacetime life gets harder, but also more interesting. This book isn't so easy, but it's packed with beautiful math and wonderfully drawn pictures:

13) Thomas Kerler and Volodymyr L. Lyubashenko, *Non-Semisimple Topological Quantum Field Theories for 3-Manifolds with Corners*, Lecture Notes in Mathematics **1765**, Springer, Berlin, 2001.

The idea is that if we can extend the definition of a quantum field theory to spacetimes that have not just boundaries but *corners*, we can try to build up the theory for arbitrary spacetimes from its behavior on simple building blocks — since it's easier to chop manifolds up into a few basic shapes if we let those shapes have corners. However, it takes higher-dimensional algebra to describe all the ways we can stick together manifolds with corners! Here Kerler and Lyubashenko make 3-dimensional manifolds going between 2-manifolds with boundary into a "double category"… and make a bunch of famous 3d TQFTs into "double functors".

Closely related is this paper by Kerler:

14) Thomas Kerler, "Towards an algebraic characterization of 3-dimensional cobordisms", *Contemp. Math.* **318** (2003) 141–173. Also available as `arXiv:math/0106253`.

It relates the category whose objects are 2-manifolds with a circle as boundary, and whose morphisms are 3-manifolds with corners going between these, to a braided monoidal category "freely generated by a Hopf algebra object". (I'm leaving out some fine print here, but probably putting in more than most people want!) It comes close to showing these categories are the same, but suggests that they're not quite — so the perfect connection between topology and higher categories remains elusive in this important example.

Answer to the riddle: these are the Catalan numbers — i.e., the natural numbers as written in Catalan. This riddle was taken from the second volume of Stanley's book on enumerative combinatorics (see "Week 144").

Addenda: Long after this issue was written, we had a discussion on the n-Category Caf about the "seven trees in one" problem. Let B be the set of binary planar trees — the set I was calling $T(1)$ above. Starting from the isomorphism

$$B \cong B^2 + 1$$

we want to construct an isomorphism

$$B \cong B^7$$

Here is the proof in Marcelo Fiore's paper:

$$\begin{array}{llllllllll}
\underline{B} \cong & 1+\underline{B^2} & \cong & 1+B+\underline{B^3} & \cong & \underline{1}+B+\underline{B^2}+B^4 & \cong & B+B+\underline{B^4} & \cong & B+\underline{B}+\underline{B^3}+B^5 \\
\cong & B+B^2+\underline{B^5} & \cong & B+\underline{B^2}+\underline{B^4}+B^6 & \cong & \underline{B}+\underline{B^3}+B^6 & \cong & B^2+\underline{B^6} & \cong & B^2+\underline{B^5}+B^7 \\
\cong & \underline{B^2}+\underline{B^4}+B^6+B^7 & \cong & B^3+\underline{B^6}+B^7 & \cong & \underline{B^3}+\underline{B^5}+B^7+B^7 & \cong & B^4+\underline{B^7}+B^7 & \cong & \underline{B^4}+\underline{B^6}+B^7+B^8 \\
\cong & \underline{B^5}+\underline{B^7}+B^8 & \cong & \underline{B^6}+\underline{B^8} & \cong & B^7 & & & &
\end{array}$$

At each step he either replaces B^n by $B^{n-1} + B^{n+1}$, or the reverse. The underlined portion shows where this will be done. Over at the *n*-Caf, Stuart Presnell made a beautiful picture of this proof:

$$\begin{tikzcd}
& B \arrow[dl] \arrow[dr] \\
1 \arrow[d] & & B^2 \arrow[dl] \arrow[dr] \\
1 \arrow[d] & B \arrow[d] & & B^3 \arrow[dl] \arrow[dr] \\
1 \arrow[dr] & B \arrow[d] & B^2 \arrow[dl] & & B^4 \arrow[d] \\
& 2B \arrow[d] & & & B^4 \arrow[dl] \arrow[dr] \\
& 2B \arrow[d] \arrow[dr] & & B^3 \arrow[dl] & & B^5 \arrow[d] \\
& B \arrow[d] & B^2 \arrow[d] & & & B^5 \arrow[dl] \arrow[dr] \\
& B \arrow[d] & B^2 \arrow[dr] & & B^4 \arrow[dl] & & B^6 \arrow[d] \\
& B \arrow[dr] & & B^3 \arrow[dl] & & & B^6 \arrow[d] \\
& & B^2 \arrow[d] & & & & B^6 \arrow[dl] \arrow[dr] \\
& & B^2 \arrow[d] & & & B^5 \arrow[dl] \arrow[dr] & & B^7 \arrow[d] \\
& & B^2 \arrow[dr] & & B^4 \arrow[dl] & & B^6 \arrow[d] & B^7 \arrow[d] \\
& & & B^3 \arrow[d] & & & B^6 \arrow[dl] \arrow[dr] & B^7 \arrow[d] \\
& & & B^3 \arrow[dr] & & B^5 \arrow[dl] & & 2B^7 \arrow[d] \\
& & & & B^4 \arrow[d] & & & 2B^7 \arrow[dl] \arrow[d] \arrow[dr] \\
& & & & B^4 \arrow[dr] & & B^6 \arrow[dl] & B^7 \arrow[d] & B^8 \arrow[d] \\
& & & & & B^5 \arrow[dr] & & B^7 \arrow[dl] & B^8 \arrow[d] \\
& & & & & & B^6 \arrow[dr] & & B^8 \arrow[dl] \\
& & & & & & & B^7
\end{tikzcd}$$

He also made a picture of another proof, which is on page 29 of Pierre Ageron's book

Logiques, Ensembles, Catgories: Le Point de Vue Constructif:

$$\begin{tikzcd}
& B \arrow[dl] \arrow[dr] \\
1 \arrow[d] & & B^2 \arrow[dl] \arrow[dr] \\
1 \arrow[d] & B \arrow[d] & & B^3 \arrow[dl] \arrow[dr] \\
1 \arrow[d] & B \arrow[d] & B^2 \arrow[d] & & B^4 \arrow[dl] \arrow[dr] \\
1 \arrow[d] & B \arrow[d] & B^2 \arrow[d] & B^3 \arrow[d] & & B^5 \arrow[dl] \arrow[dr] \\
1 \arrow[d] & B \arrow[d] & B^2 \arrow[d] & B^3 \arrow[d] & B^4 \arrow[d] & & B^6 \arrow[dl] \arrow[dr] \\
1 \arrow[d] & B \arrow[d] & B^2 \arrow[d] & B^3 \arrow[d] & B^4 \arrow[d] & B^5 \arrow[d] & & B^7 \arrow[dl] \arrow[dr] \\
1 \arrow[d] & B \arrow[d] & B^2 \arrow[d] & B^3 \arrow[d] & B^4 \arrow[d] & B^5 \arrow[d] & B^6 \arrow[dl] \arrow[dr] & & B^8 \arrow[d] \\
1 \arrow[d] & B \arrow[d] & B^2 \arrow[d] & B^3 \arrow[d] & B^4 \arrow[d] & 2B^5 \arrow[dl] \arrow[d] \arrow[dr] & & B^7 \arrow[d] & B^8 \arrow[d] \\
1 \arrow[d] & B \arrow[d] & B^2 \arrow[dr] & B^3 \arrow[d] & 2B^4 \arrow[dl] \arrow[d] & B^5 \arrow[d] & B^6 \arrow[d] & B^7 \arrow[d] & B^8 \arrow[d] \\
1 \arrow[d] & B \arrow[dr] & & 2B^3 \arrow[dl] \arrow[d] & B^4 \arrow[d] & B^5 \arrow[d] & B^6 \arrow[d] & B^7 \arrow[d] & B^8 \arrow[d] \\
1 \arrow[dr] & & B^2 \arrow[dl] & B^3 \arrow[d] & B^4 \arrow[d] & B^5 \arrow[d] & B^6 \arrow[d] & B^7 \arrow[d] & B^8 \arrow[d] \\
& B \arrow[dr] & & B^3 \arrow[dl] & B^4 \arrow[d] & B^5 \arrow[d] & B^6 \arrow[d] & B^7 \arrow[d] & B^8 \arrow[d] \\
& & B^2 \arrow[dr] & & B^4 \arrow[dl] & B^5 \arrow[d] & B^6 \arrow[d] & B^7 \arrow[d] & B^8 \arrow[d] \\
& & & B^3 \arrow[dr] & & B^5 \arrow[dl] & B^6 \arrow[d] & B^7 \arrow[d] & B^8 \arrow[d] \\
& & & & B^4 \arrow[dr] & & B^6 \arrow[dl] & B^7 \arrow[d] & B^8 \arrow[d] \\
& & & & & B^5 \arrow[dr] & & B^7 \arrow[dl] & B^8 \arrow[d] \\
& & & & & & B^6 \arrow[dr] & & B^8 \arrow[dl] \\
& & & & & & & B^7
\end{tikzcd}$$

You can watch a *movie* of a proof here:

15) Dan Piponi," Arboreal isomorphisms from nuclear pennies", *A Neighborhood of*

Infinity, September 30, 2007. Available at `http://blog.sigfpe.com/2007/09/arboreal-isomorphisms-from-nuclear.html`.

It was in the ensuing discussion on this blog that George Bell came up with his more efficient proof. For a bit more discussion, see:

16) John Baez, 'Searching for a video proof of "seven trees in one"', n-Category Caf, July 16, 2009. Available at `http://golem.ph.utexas.edu/category/2009/07/searching_for_a_video_proof_of.html`.

Now, on to some older addenda!

My pal Squark pointed out that if we try to compute the generating function for binary trees by making an initial guess for $|T|(x)$, say t, and repeatedly improving this guess via

$$t \mapsto x + t^2$$

the guess will converge to the right answer if t is small — but the process will fail miserably, with t approaching ∞, if and only if the complex number x lies outside the Mandelbrot set!

After an earlier version of this Week appeared on the category theory mailing list, Steve Schanuel posted some corrections. I've tried to correct the text above as much as possible without making it too technical — for example, by citing the important work of Robbie Gates, and distinguishing more clearly between his work on distributive categories and the paper by Fiore and Leinster, which applies to rig categories. I tend to talk about 3 different sorts of ring-like categories in This Week's Finds:

- Rig categories. A **rig category** is one equipped with a symmetric monoidal structure called $+$ and a monoidal structure called $\otimes$, with all the usual rig axioms holding up to natural isomorphism, and these isomorphisms satisfying a set of coherence laws worked out by Laplaza and Kelly:

  17) M. Laplaza, "Coherence for distributivity", Lecture Notes in Mathematics **281**, Springer Verlag, Berlin, 1972, pp. 29–72.

  18) G. Kelly, "Coherence theorems for lax algebras and distributive laws", Lecture Notes in Mathematics **420**, Springer Verlag, Berlin, 1974, pp. 281–375.

  (These authors spoke of "ring categories", but the term "rig category" is more appropriate since, as in a rig, there need be no additive inverses.)

- 2-Rigs. A **2-rig** is a symmetric monoidal cocomplete category where the monoidal structure, which we call $\otimes$, distributes over the colimits, which we think of as a generalized form of addition. For more on rigs, 2-rigs and structure types see week191. In a 2-rig, distributivity is just a property of the monoidal structure, rather than a structure, as it is in a rig category. However, by choosing a particular coproduct for each pair of objects, and a particular initial object, we can promote any 2-rig to a rig category. To get an example of a rig category that's not a 2-rig, just take any rig and think of it as a discrete category (a category with only identity morphisms). Another example would be the category of finite-dimensional vector spaces, since this only has finite colimits. (Of course, we could make up some sort of "finitary 2-rig" that only had finite colimits, but the profusion of terminology is already annoying.)

- Distributive categories. A **distributive category** is a category with finite products and coproducts, the products distributing over the coproducts. Here again, distributivity is just a property. But, by choosing specified products and coproducts for every pair of objects, and choosing terminal and initial objects, we can promote any distributive category into a rig category. A good example of a 2-rig that is not a distributive category is the category Vect, with direct sum and tensor product as $+$ and $\otimes$. Another example is the discrete category on a rig.

By not distinguishing these, the original version of "Week 202" made it sound as if Fiore and Leinster had simply redone Gates' work on distributive categories. I hope this is a bit clearer now. Schanuel's remarks are still worth reading for their description of what Gates actually did:

> *Dear colleagues,*
>
> *For those who read the most recent long discursion of John Baez, a few of the errors in the section on distributive categories merit correction:*
>
> *(1) J. B. suggests that Blass published what Lawvere had already worked out. In fact, Lawvere (partly to counteract some incorrect uses of infinite series in analyses of 'data types' in computer science) had worked out the algebra of the rig presented by one generator $ $and one relation $X = 1 + X^2$, roughly by the method in (3) below, and conjectured that this rig could be realized as the isomorphism classes in a distributive (even extensive) category, which conjecture Blass then proved (and a bit more) in "Seven Trees. . . ".*
>
> *(2) The generalization of Blass's theorem to one generator ond one polynomial relation of the 'fixed-point' form $X = p(X)$, where p is a polynomial with natural number coefficients and nonzero constant term is not, as J. B. seems to suggest, due to Fiore and Leinster; it was part of the prize-winning doctoral thesis of Robbie Gates, who (using a calculus of fractions) described explicitly the free distributive category on one object X together with an isomorphism from $p(X)$ to X, proving that this category is extensive and that its rig of isomorphism classes satisfies no further relations, i.e. is the rig R presented by one generator and the one relation above.*
>
> *(3) If p is as in (2) and of degree at least 2, the algebra of the rig R is made by J. B. to seem mysterious. It is more easily understood in the way the $X = 2^X + 1$ case was treated in my "Negative Sets. . . " paper; just show that the Euler and dimension homomorphisms, tensoring with $\mathbb{Z}$ and with 2 (the rig true/false) respectively, are jointly injective. In this case the dimension rig has only three elements, which explains why the Euler characteristic captures almost, but not quite, everything.*
>
> *Greetings to all,*
> *Steve Schanuel*

A traveller who refuses to pass over a bridge until he personally tests the soundness of every part of it is not likely to go far; something must be risked, even in mathematics.

— *Horace Lamb*

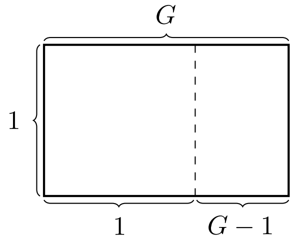
Week 203

February 28, 2004

Last week I posed this puzzle: to find a “Golden Object”.

A couple days ago I got a wonderful solution from Robin Houston, a computer science grad student at the University of Manchester. So, I want to say a bit more about the golden number, then describe his solution, and then describe how he found it.

Supposedly the Greeks thought the most beautiful rectangle was one such that when you chop a square off one end, you’re left with a rectangle of the same shape. If your original rectangle was 1 unit across and G units long, after you chop a 1-by-1 square off the end you’re left with a rectangle that’s $G-1$ units across and 1 unit long:



So, to make the proportions of the little rectangle the same as those of the big one, you want

“1 is to G as G-1 is to 1”

or in other words:

$$\frac{1}{G} = G - 1$$

or after a little algebra,

$$G^2 = G + 1$$

so that

$$G = \frac{1+\sqrt{5}}{2} = 1.618033988749894848204586834365\ldots$$

while

$$\frac{1}{G} = 1.618033988749894848204586834365\ldots$$

and

$$G^2 = 2.618033988749894848204586834365\ldots$$

(At this point I usually tell my undergraduates that the pattern continues like this, with $G^3 = 3.618\ldots$ and so on — just to see if they’ll believe anything I say.)

These days, the number G is called the Golden Number, the Golden Ratio, or the Golden Section. It’s often denoted by the Greek letter Φ, after the Greek sculptor Phidias. Phidias helped design the Parthenon - and supposedly packed it full of golden rectangles, to make it as beautiful as possible.

The golden number is a great favorite among amateur mathematicians, because it has a flashy sort of charm. You can find it all over the place if you look hard enough — and if you look too hard, you'll find it even in places where it's not. It's the ratio of the diagonal to the side of a regular pentagon! If you like the number 5, you'll be glad to know that

$$G = \sqrt{\frac{5+\sqrt{5}}{5-\sqrt{5}}}$$

If you don't, maybe you'd prefer this:

$$G = \exp\left(\operatorname{arcsinh}\left(\frac{1}{2}\right)\right)$$

My favorite formulas for the golden number are

$$G = \sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\dots}}}}}}$$

and the continued fraction:

$$G = 1 + \cfrac{1}{1+\frac{1}{1+\frac{1}{1+\frac{1}{1+}}}}$$

These follow from the equations $G^2 = G + 1$ and $G = 1 + 1/G$, respectively. If you chop off the continued fraction for G at any point, you'll see that G is also the limit of the ratios of successive Fibonacci numbers. See "Week 190" for a very different proof of this fact.

However, don't be fooled! The charm of the golden number tends to attract kooks and the gullible — hence the term "fool's gold". You have to be careful about anything you read about this number. In particular, if you think ancient Greeks ran around in togas philosophizing about the "golden ratio" and calling it "Φ", you're wrong. This number was named Φ after Phidias only in 1914, in a book called *The Curves of Life* by the artist Theodore Cook. And, it was Cook who first started calling $1.618\ldots$ the golden ratio. Before him, $\ldots 0.618$ was called the golden ratio! Cook dubbed this number "φ", the lower-case baby brother of Φ.

In fact, the whole "golden" terminology can only be traced back to 1826, when it showed up in a footnote to a book by one Martin Ohm, brother of Georg Ohm, the guy with the law about resistors. Before then, a lot of people called $1/G$ the "Divine Proportion". And the guy who started *that* was Luca Pacioli, a pal of Leonardo da Vinci who translated Euclid's *Elements*. In 1509, Pacioli published a 3-volume text entitled *Divina Proportione*, advertising the virtues of this number. Some people think da Vinci used the divine proportion in the composition of his paintings. If so, perhaps he got the idea from Pacioli.

Maybe Pacioli is to blame for the modern fascination with the golden ratio; it seems hard to trace it back to Greece. These days you can buy books and magazines about "Elliot Wave Theory", a method for making money on the stock market using patterns related to the golden number. Or, if you're more spiritually inclined, you can go to workshops on “Sacred Geometry” featuring talks about the healing powers of the golden ratio. But Greek texts seem remarkably quiet about this number.

The first recorded hint of it is Proposition 11 in Book II of Euclid’s *Elements*. It also shows up elsewhere in Euclid, especially Proposition 30 of Book VI, where the task is “to cut a given finite straight line in extreme and mean ratio”, meaning a ratio $A : B$ such that

$$A : B :: (A + B) : A$$

$$\text{i.e. ”}A\text{ is to }B\text{ as }A + B\text{ is to }A\text{”}$$

This is later used in Proposition 17 of Book XIII to construct the pentagonal face of a regular dodecahedron.

Of course, Euclid wasn’t the first to do all these things; he just wrote them up in a nice textbook. By now it’s impossible to tell who discovered the golden ratio and just what the Greeks thought about it. For a sane and detailed look at the history of the golden ratio, try this:

1) J. J. O’Connor and E. F. Robertson, “The Golden Ratio”, `http://www-gap.dcs.st-and.ac.uk/~history/HistTopics/Golden_ratio.html`

While I’m at it, I should point out that Theodore Cook’s book introducing the notation “Φ” is still in print:

2) Theodore A. Cook, *The Curves of Life: Being an Account of Spiral Formations and Their Application to Growth in Nature, to Science, and to Art: with Special Reference to the Manuscripts of Leonardo da Vinci*, Dover Publications, New York, 1979.

If you want to see what Euclid said about the golden ratio, you can also pick up a cheap copy of the Elements from Dover — but it’s probably quicker to go online. There are a number of good places to find Euclid’s Elements online these days.

Topologists know David Joyce as the inventor of the “quandle” — an algebraic structure that captures most of the information in a knot. Now he’s writing a high-tech edition of Euclid, complete with Java applets:

3) David E. Joyce’s edition of Euclid’s Elements, `http://aleph0.clarku.edu/~djoyce/java/elements/toc.html`

Joyce is carrying on a noble tradition: back in 1847, Oliver Byrne did a wonderful edition of Euclid complete with lots of beautiful color pictures and even some pop-up models. You can see this online at the Digital Mathematics Archive:

4) Oliver Byrne’s edition of Euclid’s Elements, online at the Digital Mathematics Archive, `http://www.sunsite.ubc.ca/DigitalMathArchive/`

The most famous scholarly English translation of Euclid was done by Sir Thomas Heath in 1908. You can find it together with an edition in Greek and a nearly infinite supply of other classical texts at the Perseus Digital Library:

5) Thomas L. Heath’s edition of Euclid’s Elements, online at The Perseus Digital Library, `http://www.perseus.tufts.edu/`

But I'm digressing! My main point was that while $G = (1+\sqrt{5})/2$ is a neat number, it's a lot easier to find nuts raving about it on the net than to find truly interesting mathematics associated with it — or even interesting references to it in Greek mathematics! The cynic might conclude that the charm of this number is purely superficial. However, that would be premature.

First of all, there's a certain sense in which G is "the most irrational number". To get the best rational approximations to a number you use its continued fraction expansion. For G, this converges as slowly as possible, since it's made of all 1's:

$$G = 1 + \cfrac{1}{1 + \frac{1}{1+\frac{1}{1+\frac{1}{1+}}}}$$

We can make this more precise. For any number x there's a constant $c(x)$ that says how hard it is to approximate x by rational numbers, given by

$$\liminf_{q\to\infty} \left| x - \frac{p}{q} \right| = \frac{c(x)}{q^2}$$

where q ranges over integers, and p is an integer chosen to minimize $|x - p/q|$. This constant is as big as possible when x is the golden ratio!

It'd be ironic if the famously "rational" Greeks, who according to legend even drowned the guy who proved $\sqrt{2}$ was irrational, chose the most irrational number as the proportions of their most beautiful rectangle! But, it wouldn't be a coincidence. Their obsession with ratios and proportions led them to ponder the situation where $A : B :: (A+B) : A$, and this proportion instantly implies that A and B are incommensurable, since if you assume A and B are integers and try to find their greatest common divisor using Euclid's algorithm, you get stuck in an infinite loop. Euclid even mentions this idea in Proposition 2 of Book X:

> *If, when the less of two unequal magnitudes is continually subtracted in turn from the greater that which is left never measures the one before it, then the two magnitudes are incommensurable.*

He doesn't explicitly come out and apply it to what we now call the golden ratio — but how could he not have made the connection? For more info on the Greek use of continued fractions and the Euclidean algorithm, check out the chapter on "antihyphairetic ratio theory" in this book:

6) D. H. Fowler, *The Mathematics of Plato's Academy: A New Reconstruction*, Oxford U. Press, Oxford, 1987.

Anyway, it's actually important in physics that the golden number is so poorly approximated by rationals. This fact shows up in the Kolmogorov- Arnold-Moser theorem, or "KAM theorem", which deals with small perturbations of completely integrable Hamiltonian systems. Crudely speaking, these are classical mechanics problems that have as many conserved quantities as possible. These are the ones that tend to show up in textbooks, like the harmonic oscillator and the gravitational 2-body problem. The reason is that you can solve such problems if you can do a bunch of integrals — hence the term "completely integrable".

The cool thing about a completely integrable system is that time evolution carries states of the system along paths that wrap around tori. Suppose it takes n numbers to describe the position of your system. Then it also takes n numbers to describe its momentum, so the space of states is $2n$-dimensional. But if the system has n different conserved quantities — that's basically the maximum allowed — the space of states will be foliated by n-dimensional tori. Any state that starts on one of these tori will stay on it forever! It will march round and round, tracing out a kind of spiral path that may or may not ever get back to where it started.

Things are pretty simple when $n = 1$, since a 1-dimensional torus is a circle, so the state *has* to loop around to where it started. For example, when you have a pendulum swinging back and forth, its position and momentum trace out a loop as time passes.

When n is bigger, things get trickier. For example, when you have n pendulums swinging back and forth, their motion is periodic if the ratios of their frequencies are rational numbers.

This is how it works for any completely integrable system. For any torus, there's an n-tuple of numbers describing the frequency with which paths on this torus wind around in each of the n directions. If the ratios of these frequencies are all rational, paths on this torus trace out periodic orbits. Otherwise, they don't!

KAM theory says what happens when you perturb such a system a little. It won't usually be completely integrable anymore. Interestingly, the tori with rational frequency ratios tend to fall apart due to resonance effects. Instead of periodic orbits, we get chaotic motions instead. But the "irrational" tori are more stable. And, the "more irrational" the frequency ratios for a torus are, the bigger a perturbation it takes to disrupt it! Thus, the most stable tori tend to have frequency ratios involving the golden number. As we increase the perturbation, the last torus to die is called a "golden torus".

You can actually *watch* tori breaking into chaotic dust if you check out the applet illustrating the "standard map" on this website:

7) Takashi Kanamaru and J. Michael T. Thompson, "Introduction to Chaos and Nonlinear Dynamics", standard map applet, `http://brain.cc.kogakuin.ac.jp/~kanamaru/Chaos/e/Standard/`

The "standard map" is a certain dynamical system that's good for illustrating this effect. You won't actually see 2d tori, just 1d cross-sections of them — but it's pretty fun. For more details, try this:

8) M. Tabor, *Chaos and Integrability in Nonlinear Dynamics: An Introduction*, Wiley, New York, 1989.

In short, the golden number is the best frequency ratio for avoiding resonance!

Some audiophiles even say this means the best shaped room for listening to music is one with proportions $1 : G : G^2$. I leave it to you to find the flaw in this claim. For more dubious claims, check out the ad for expensive speaker cables at the end of this article.

KAM theory is definitely cool, but we shouldn't rest content with this when skeptics ask if the golden number is all it's cracked up to be. I figure it's part of our job as mathematicians to keep on discovering mind-blowing facts about the golden number. A small part, but part: we shouldn't give up the field to amateurs!

Penrose has done his share. His "Penrose tiles" take crucial advantage of the self-similarity embodied by the golden number to create nonperiodic tilings of the plane. This helped spawn a nice little industry, the study of "quasicrystals" with 5-fold symmetry. Here's a good introduction for mathematicians:

9) Andre Katz, "A short introduction to quasicrystallography", in *From Number Theory to Physics*, eds. M. Waldschmit et al, Springer, Berlin, 1992, pp. 496–537.

By the way, this same book has some nice stuff on the role of the golden number in KAM theory and the theory of iterated maps from the circle to itself:

10) Predrag Cvitanovic, "Circle maps: irrationally winding", in *From Number Theory to Physics*, eds. M. Waldschmit et al, Springer, Berlin, 1992, pp. 631–658.

11) Jean-Christophe Yoccoz, "Introduction to small divisors problems", in *From Number Theory to Physics*, eds. M. Waldschmit et al, Springer, Berlin, 1992, pp. 659–679.

Conway and Sloane are also pulling their weight. Starting from the relation between the golden ratio, the isosahedron, and the 4-dimensional big brother of the icosahedron (the "600-cell"), they've described how to construct the coolest lattices in 8 and 24 dimensions using "icosians" — which are certain quaternions built using the golden ratio. I discussed this circle of ideas in "Week 20", "Week 59" and "Week 155".

But if you want some really scary formulas involving the golden ratio, Ramanujan is the one to go to. Check these out:

$$\frac{1}{1+\frac{\exp(-2\pi)}{1+\frac{\exp(-4\pi)}{1+\frac{\exp(-6\pi)}{1+\dots}}}} = \exp\left(\frac{2\pi}{5}\right)\left(\sqrt{G\sqrt{5}}-5\right)$$

and

$$1+\frac{\exp(-2\pi\sqrt{5})}{1+\frac{\exp(-4\pi\sqrt{5})}{1+\frac{\exp(-6\pi\sqrt{5})}{1+\dots}}} = \exp\left(\frac{2\pi}{5}\right)\left(\frac{\sqrt{5}}{1+(5^{\frac{3}{4}}(G-1)^{\frac{5}{2}}-1)^{\frac{1}{5}}}-G\right)$$

These are special cases of a monstrosity called the Rogers-Ramanujan continued fraction, which is a kind of "q-deformation" of the continued fraction for the golden ratio. For details, start here:

12) Eric W. Weisstein, "Rogers-Ramanujan Continued Fraction", `http://mathworld.wolfram.com/Rogers-RamanujanContinuedFraction.html`

It's these two formulas, and one other like it, that led Hardy to write the famous lines:

> *A single look at them is enough to show that they could only be written down by a mathematician of the highest class. They must be true because, if they were not true, no one would have had the imagination to invent them.*

For more by Hardy on these continued fractions, see section 1 and section 6.17 of his book:

13) G. H. Hardy, *Ramanujan: Twelve Lectures on Subjects Suggested by His Life and Work*, Chelsea Publishing Co., New York, 1959.

The golden number also shows up in the theory of quantum groups. I talked about this in "Week 22" so I won't explain it again here. But, I can't resist mentioning that Freedman, Larsen and Wang have subsequently shown that a certain topological quantum field theory called Chern-Simons theory, built using the quantum group $\mathrm{SU}_q(2)$, can serve as a universal quantum computer when the parameter q is a fifth root of unity. And, this is exactly the case where the spin-$1/2$ representation of $\mathrm{SU}_q(2)$ has quantum dimension equal to the golden number!

14) Michael Freedman, Michael Larsen, Zhenghan Wang, "A modular functor which is universal for quantum computation", available at quant-ph/0001108.

But don't get the wrong idea: it's not that some magic feature of the golden number is required to build a universal quantum computer! It's just that the 5 seems to be the *smallest* number n such that $\mathrm{SU}_q(2)$ Chern-Simons theory is computationally universal when q is an nth root of 1.

That's pretty much everything I know about the golden number. So now, what about this "Golden Object" puzzle?

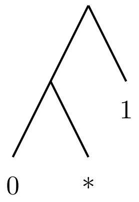
Basically, the problem was to find an object that acts like the golden number. The golden number has $G^2 = G + 1$, so we want to find a object G equipped with a nice isomorphism between G^2 and $G + 1$.

If G is just a set, this means we want a nice one-to-one correspondence between pairs of elements of G, and elements of G together with one other thing. It doesn't matter what that other thing is, so let's call it "@".

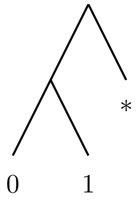
(You may be wondering about the word "nice". The point is, the problem is too easy if we don't demand that the solution be nice in some way — some way that I don't feel like making precise.)

Here's Robin Houston's answer:

Define a "bit" to be either 0 or 1. Define a "golden tree" to be a (planar) binary tree with leaves labelled by 0, 1, or $*$, where every node has at most one bit-child. For example:



is a golden tree, but



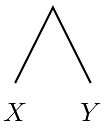
is not. Let G be the set of golden trees. We define an isomorphism

$$f\colon G^2 \to G + \{@\}$$

as follows. First we define $f(X,Y)$ when both X and Y are golden trees with just one node, this node being labelled by a bit. We can identify such a tree with a bit, and doing this we set

$$\begin{aligned} f(0,0) &= 0 \\ f(0,1) &= 1 \\ f(1,0) &= * \\ f(1,1) &= @ \end{aligned}$$

In the remaining case, where the golden trees X and Y are not just bits, we set

$f(X,Y) \;=$ 

There are different ways to show this function f is a one-to-one correspondence, but the best way is to see how Houston came up with this answer! He didn't just pull it out of a hat; he tackled the problem systematically, and that's why his solution counts as "nice".

It's easy to find a set S equipped with an isomorphism

$$S = P(S)$$

where P is some polynomial with natural number coefficients. You just use the fixed-point principle described in "Week 108". Namely, you start with the empty set, keep hitting it with P forever, and take a kind of limit. This is how I built the set of binary trees last week, as a solution of $T = T^2 + 1$.

The problem is that the isomorphism we seek now:

$$G^2 = G + 1 \tag{1}$$

is not of this form. So, what Houston does is to make a substitution:

$$G = H + 2$$

Given this, we'd get (1) if we had

$$H^2 + 4H + 4 = H + 3 \tag{2}$$

and we'd get (2) if we had

$$H^2 + 4H + 1 = H \tag{3}$$

which is of the desired form.

We can rewrite (3) as

$$H = 1 + H^2 + 2H + H2$$

and in English this says "an element of H is either a *, or a pair consisting of two guys that are either bits or elements of H — but not both bits". So, a guy in H is a golden tree! But, if it has just one node, that node can only be labelled by a $*$, not a 0 or 1. This means there are precisely 2 golden trees not in H. So, $G = H + 2$ is the set of all golden trees, and our calculation above gives an isomorphism $G^2 = G + 1$.

Voila!

Note that to derive (3) from (1) we need to subtract, which in general is not allowed in this game. Here we are subtracting constants, and Houston says that's allowed by the "Garsia-Milne involution theorem". I don't know this theorem, so I'll make a note to myself to learn it. But luckily, we don't really need it here: we only need to derive (1) from (3), and that involves addition, so it's fine.

Part of what makes Houston's solution "nice" is that it suggests a general method for turning polynomial equations into recursive definitions of the form $S = P(S)$. Another nice thing is that his trick delivers a structure type $G(X)$ that reduces to G when $X = 1$. To get this, first use the fixed-point method to construct a structure type $H(X)$ with an isomorphism

$$H(X) = (H(X) + X)^2 + 2H(X)$$

Then, define

$$G(X) = H(X) + X + 1$$

and note that this gives

$$G(X)^2 = G(X) + X$$

which reduces to Gˆ2 = G + 1 when X = 1.

As if this weren't enough, Houston also gave another solution to the puzzle. He showed that James Propp's proposed Golden Object, described last week, really is a Golden Object! Maybe Propp already knew this, but I sure didn't.

The idea of the proof is pretty general. Suppose we're in some category that's a "2-rig" in the sense of "Week 191". And, suppose we've got an object X equipped with an isomorphism

$$X = 1 + 2X \qquad (4)$$

so that X acts like "-1". For example, following Schanuel and Propp, we can take the category of "σ-polytopes" and let X be the open interval: then isomorphism (4) says

$$(0, 1) = (0, 1/2) + \{1/2\} + (1/2, 1)$$

Or, following Houston, we can take the category of sets and let X be the set of finite bit-strings. Then (4) says "a finite bit-string is either the empty bit-string, or a bit followed by a finite bit-string". The relation between these two examples is puzzling to me — if anyone understands it, let me know! But anyway, either one works.

Now let G be the object of "binary trees with X-labelled leaves":

$$G = X + X^2 + 2X^3 + 5X^4 + 14X^5 + 42X^6 + \ldots$$

where the coefficients are Catalan numbers. Let's show that G is a Golden Object. To do this, we'll use (4) and this isomorphism:

$$G = G^2 + X \qquad (5)$$

which says "a binary tree with X-labelled leaves is a pair of such trees, or a degenerate tree with just one X-labelled node". The formula for G involving Catalan numbers is really just the fixed-point solution to this!

Here is Houston's fiendishly clever argument. Suppose Z is any type equipped with an isomorphism

$$Z = Z' + X$$

for some Z'. Then

$$\begin{aligned} Z + X + 1 &= Z' + 2X + 1 \\ &= X' + X \\ &= Z \end{aligned}$$

This applies to $Z = G^2$, since

$$G^2 = (X + G^2)^2 = (2X + 1 + G^2)^2$$

has a X term in it when you multiply it out, so it's of the form $Z' + X$. Therefore we have an isomorphism

$$G^2 = G^2 + X + 1$$

But we also have an isomorphism $G + 1 = G^2 + X + 1$ by (5). Composing these, we get our isomorphism

$$G^2 = G + 1.$$

Golden! I'll stop here.

Addendum: The computer scientist Sebastiano Vigna pointed out this paper:

15) Paolo Boldi, Massimo Santini, and Sebastiano Vigna, "Measuring with jugs, or: what if mathematicians were asked to defuse bombs?", *Theoret. Comput. Sci.* **2** (2002). Also available at `http://vigna.dsi.unimi.it/papers.php`

which shows that if you want to approximately measure an arbitrary amount of water using only two jugs, it's best if they have capacity 1 and G. This paper cites a a charming result by Swierczkowski which picks up where a famous theorem due to Dedekind leaves off. Dedekind showed that if x is any irrational number, the numbers $nx \mod 1$ are uniformly distributed in the interval $[0, 1]$. But if $x = 1/G$, these numbers have an especially nice property: each new point in the sequence $(nx \mod 1)$ lands in one of the *longest* intervals not containing a previous point! And, it chops this interval in a golden way.

Stephen Schanuel said some things about "Week 203" on the category theory mailing list, so I'll include his post here along with various replies, concluding with my own.

> *From: Stephen Schanuel Subject: categories: mystification and categorification Date: Thu, 4 Mar 2004 00:44:46 -0500*
>
> *I was unable to understand John Baez' golden object problem, nor his description of the solutions. He refuses to tell us what 'nice' means, but let me at least propose that to be 'tolerable' a solution must be an object in a category, and John doesn't tell us what category is involved in either of the solutions; at*

least I couldn't find a specification of the objects, nor the maps, so I found the descriptions 'intolerable', in the technical sense defined above. He is very generous, allowing one to use a category with both plus and times as extra monoidal structures. (Does anyone know an example of interest in which the plus is not coproduct?) This freedom is unnecessary; a little algebra plus Robbie Gates' theorem provides a solution G to $G^2 = G + 1$ which satisfies no additional equations, in an extensive category (with coproduct as plus, cartesian product as times).

Briefly, here it is. A primitive fifth root of unity z is a root of the polynomial $1 + X + X^2 + X^3 + X^4$, hence satisfies $1 + z + z^2 + z^3 + z^4 + z = z$, which is of the 'fixed point' form $p(z) = z$ with p in $\mathbb{N}[X]$ and $p(0)$ not 0. Gates' theorem then says that the free distributive category containing an object Z and an isomorphism from $p(Z)$ to Z is extensive, and its Burnside rig B (of isomorphism classes of objects) is, as one would hope, $\mathbb{N}[X]/(p(X) = X)$; that is, Z satisfies no unexpected equations. Since the degree of p is greater than 1, an easy general theorem tells us (from the joint injectivity of the Euler and dimension homomorphisms) that two polynomials agree at the object Z if and only if either they are the same polynomial or both are non-constant and they agree at the number z. Now the 'algebra': the golden number is $1 + z + z^4$. So $G = 1 + Z + Z^4$ satisfies $G^2 = G + 1$, as desired. It satisfies no unexpected equations, because the relation $X^2 = X + 1$ reduces any polynomial in $\mathbb{N}[X]$ to a linear polynomial, and these reduced forms have distinct Euler characteristics, i.e. differ at z. Thus the homomorphism from $\mathbb{N}[X]/(X^2 = X+1)$ to B (sending X to G) is injective, and that's all I wanted.

Since in the category of sets, any nasty old infinite set satisfies the golden equation and many others, I have taken the liberty of interpreting 'nice' to mean at least 'satisfying no unexpected equations'. One could ask for more; the construction above has produced a distributive, but not extensive, category whose Burnside rig is $\mathbb{N}[X]/(X^2 = X + 1)$, the full subcategory with objects polynomials in G. (If it were extensive, it would be closed under taking summands, but every object in the larger category is a summand of G.) I don't know whether there is an extensive category with $\mathbb{N}[X]/(X^2 = X + 1)$ as its full Burnside rig; perhaps one or both of the examples John mentioned would do, if I knew what they were.

While I'm airing my confusions, can anyone tell me what 'categorification' means? I don't know any such process; the simplest exanple, 'categorifying' natural numbers to get finite sets, seems to me rather 'remembering the finite sets and maps which gave rise to natural numbers by the abstraction of passing to isomorphism classes'.

Finally, a note to John: While you're trying to give your audience some feeling for the virtues of n-categories, couldn't you give them a little help with $n = 1$, by being a little more precise about objects and maps?

Greetings to all, and thanks for your patience while I got this stuff off my chest,

Steve Schanuel

From: David Yetter Subject: categories: Re: mystification and categorification Date: Fri, 5 Mar 2004 10:55:26 -0600

Categorification is a bit like quantization: it isn't a construction so much as a desideratum for a relationship between one thing and another (in the case of categorification an $(n+1)$-categorical structure and an n-categorical structure; in the case of quantization a quantum mechanical system and a classical mechanical system).

Categorification wants to find a higher-dimensional categorical structure corresponding to a lower-dimensional one by weakening equations to natural isomorphisms and imposing new, sensible, coherence conditions. In general, for the original purpose for which it was proposed–constructions of TQFT's and models of quantum gravity–one wants the highest categorical level to have a linear structure (hence Baez wanting tensor product and a sum it distributes over, rather than cartesian product and coproduct). Specific lower-dimensional categories with structure are 'categorified' by finding a higher-dimensional category with the new structure which 'lies over' the lower dimensional one in the way an additive monoidal category lies over its Grothendieck rig.

For instance any (k-linear) monoidal category with monoid of isomorphism classes M is a categorification of M, and more generally (k-linear) monoidal categories are a categorification of monoids.

A simple example shows why it is not a construction: commutative monoids (as rather special categories with one object) admit two different categorifications: symmetric monoidal categories and braided monoidal categories (each regarded as a kind of bicategory with one object). There is a good argument for regarding braided monoidal categories as the 'correct' categorification: the Eckmann-Hilton theorem ('a group in GROUPS *is an abelian group' or, really as the proof shows, 'a monoid in* MONOIDS *is a commutative monoid') 'categorifies' to: A monoidal category in* MONCAT *is a braided monoidal category.*

From: Vaughan Pratt Subject: categories: Re: mystification and categorification Date: Fri, 05 Mar 2004 22:49:56 -0800

> *While I'm airing my confusions, can anyone tell me what 'categorification' means? I don't know any such process; the simplest exanple, 'categorifying' natural numbers to get finite sets, seems to me rather 'remembering the finite sets and maps which gave rise to natural numbers by the abstraction of passing to isomorphism classes'.*

A fair question. I attended John's Coimbra lectures on this stuff in 1999 but a lot of it leaked out afterwards. If I had to guess I'd say he was categorifying the free monoid on one generator to make it a monoidal category, but then how did the monoid end up as coproduct and the generator as the final object? One suspects some free association there — John, how do you make that connection?

With regard to categorification in general, sets seem to play a central role in at least one development of category theory. The homobjects of "ordinary" categories are homsets. (In that sense I guess "ordinary" must entail "locally small.") 2-categories are what you get if instead you let them be homcats, suitably elaborated.

Going in the other direction, if you take homsets to be vacuous, not in the sense that they are empty but rather that they are all the same, then you get sets. One more step in that direction makes everything look the same, which may have something to do with the Maharishi Yogi hiring category theorists for the math dept. of his university in Fairfield, Iowa. (When I spoke last with the MY's "Minister of World Health," an MD who like Ross Street was a classmate of mine but eight years earlier starting in 1957, the entire conversation seemed to be largely a skirting of the minefield of the sameness of everything, which may subconsciously have been behind my obscure reply to Peter Freyd's posting a while back about unique existence going back to Descartes, where I tried to one-up him by claiming it went much further back.)

Categorification isn't the only way to get to 2-categories, which can be understood instead in terms of the interchange law as a two-dimensional associativity principle. However John has got a lot of mileage out of the categorification approach, which one can't begrudge in an era where mileage and minutes are as integral to a balanced life as one's checkbook. (Q: How many minutes in a month? A: Depends on your plan.)

> *Since in the category of sets, any nasty old infinite set satisfies the golden equation and many others, I have taken the liberty of interpreting 'nice' to mean at least 'satisfying no unexpected equations'.*

Quite right. I would add to this "and satisfying the expected equations." The "nasty sets" of which Steve speaks fail to satisy such expected equations as $2^{2^X} \sim X$. (The power set of a set is a Boolean algebra, for heaven's sake. Why on earth forget that structure prior to taking the second exponentiation? Set theorists seem to think that they can simply forget structure without paying for it, but in the real world it costs $kT/2$ joules per element of X to forget that structure. If set theorists aren't willing to pay real-world prices in their modeling, why should we taxpayers pay them real-world salaries? Large cardinals are a figment of their overactive imaginations, and the solution to consistency concerns is not to go there.)

Vaughan Pratt

From: Tom Leinster Subject: Re: categories: mystification and categorification Date: 07 Mar 2004 20:50:39 +0000

Steve Schanuel wrote: > a category with both plus and times as extra monoidal structures. > (Does anyone know an example of interest in which the plus is not > coproduct?)

Here are two examples that I've come across previously of rig categories in which the plus is not coproduct:

(i) the category of finite sets and bijections, with $+$ and $\times$ inherited from the category of sets;

(ii) discrete rig categories, which are of course the same thing as rigs.

> *This freedom is unnecessary; a little algebra plus Robbie Gates' theorem provides a solution G to $G^2 = G+1$ which satisfies no additional equations, in an extensive category (with coproduct as plus, cartesian product as times).*

If you do allow yourself the freedom to use any rig category then an even simpler solution exists, also satisfying no additional equations: just take the rig freely generated by an element G satisfying $G^2 = G + 1$ and regard it as a discrete rig category.

> *Since in the category of sets, any nasty old infinite set satisfies the golden equation and many others, I have taken the liberty of interpreting 'nice' to mean at least 'satisfying no unexpected equations'.*

I'd interpret "nice" differently. (Apart from anything else, the trivial example in my previous paragraph would otherwise make the golden object problem uninteresting.) "Nice" as I understand it is not a precise term — at least, I don't know how to make it precise. Maybe I can explain my interpretation by analogy with the equation $T = 1 + T^2$. A nice solution T would be the set of finite, binary, planar trees together with the usual isomorphism $T \xrightarrow{\sim} 1 + T^2$; a nasty solution would be a random infinite set T with a random isomorphism to $1 + T^2$. (Both these solutions are in the rig category Set *with its standard $+$ and $\times$.) I regard the first solution as nice because I can see some combinatorial content to it (and maybe, at the back of my mind, because it has a constructive feel), and the second as nasty because I can't. I'm not certain what I think of the solution given by the set of not-necessarily-finite binary planar trees (nice?), or by the set of binary planar trees of cardinality at most $\aleph_5$ (probably nasty).*

Maybe the finding of a "nice" solution is similar in spirit to the finding of a "concrete interpretation" of a combinatorial identity. As an extremely simple example, consider the identity saying that each entry in Pascal's triangle is the sum of the two above it,

$$\binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r}.$$

This is a doddle to prove, but all the same you'd be missing something if you didn't know the standard "concrete interpretation": choosing r objects out of $n+1$ objects amounts to EITHER choosing the first one and then choosing $r-1$ of the remaining n OR . . . . Even if the challenge of finding a "nice solution" or "concrete interpretation" isn't made precise, I think there is a shared sense *of what would count as an answer, and finding an answer is in general not straightforward.*

Best wishes, Tom

From: John Baez Subject: golden objects Date: Sun, 7 Mar 2004 12:50:29 -0800 (PST)

Dear Categorists —

Sorry to take a while to respond. People at UCR have been unable to receive posts on the category theory mailing list, due to problems with our internet connection.

I'd asked for some nice examples of an object G in a rig category equipped with an isomorphism from G^2 to $G + 1$. Steve Schanuel replied:

> *I was unable to understand John Baez' golden object problem, nor his description of the solutions. He refuses to tell us what 'nice' means, [. . .]*

The problem was deliberately open-ended, but you seem to have understood it perfectly, since you've given a nice solution, including a precise specification of what you consider "nice".

Let me repeat the two solutions given by Robin Houston:

1) *The first solution works in any rig category having an object H equipped with an isomorphism to $H^2 + 4H + 1$. The solution is to take*

$$G = H + 2.$$

I described how Houston uses the isomorphism $H \to H^2 + 4H + 1$ to construct an isomorphism $G^2 \to G + 1$. What's nice about this is that it reduces a problem that's not obviously of fixed-point form to one that is.

2) *Houston's second solution works in any monoidal cocomplete category, tensor product distributing over colimits, that contains an object X equipped with an isomorphism to $2X + 1$. The solution is to let G be the object of "binary planar rooted trees with X-labelled leaves", i.e.*

$$G = X + X^2 + 2X^3 + 5X^4 + 14X^5 + 42X^6 + \dots$$

where the coefficients are Catalan numbers. He uses the obvious isomorphism $G \to G^2 + X$ to construct an isomorphism $G^2 \to G + 1$. What's nice about this is that it shows Propp's originally proposed golden object really is one: just take the category of σ-polytopes with its cartesian product, and let X be the open interval! And, it makes precise the sense in which the alternating sum of Catalan numbers equals the golden ratio.

Steve writes:

> *I don't know whether there is an extensive category with $\mathbb{N}[X]/(X^2 = X+1)$ as its full Burnside rig; perhaps one or both of the examples John mentioned would do, if I knew what they were.*

I think example 1) does the job if we take the free distributive category on an object H equipped with an isomorphism to $H^2 + 4H + 1$. Right?

Steve also writes:

> *He is very generous, allowing one to use a category with both plus and times as extra monoidal structures. (Does anyone know an example of interest in which the plus is not coproduct?) This freedom is unnecessary [. . .]*

It's unnecessary, but handy: I think there's also an golden object in the rig category of reps of quantum $\mathrm{SU}(2)$ at a suitable value of q. Here the tensor product is not cartesian.

In the lingo of quantum group theory, this object has "quantum dimension" equal to the golden number. It's interesting how such nonintegral but algebraic "dimensions" show up naturally in quantum group theory, just as nonintegral but algebraic "cardinalities" show up in the theory of distributive categories.

I don't know any golden objects in rig categories where the plus is not coproduct, and I agree that such rig categories arise less often than those where times is not product. But, if you use the obvious way of making the groupoid of finite sets into a rig category, $+$ isn't coproduct, nor is $\times$ product.

> *While I'm airing my confusions, can anyone tell me what 'categorification' means? I don't know any such process; the simplest exanple, 'categorifying' natural numbers to get finite sets, seems to me rather 'remembering the finite sets and maps which gave rise to natural numbers by the abstraction of passing to isomorphism classes'.*

You're right: categorification is not a systematic process! I explained this idea back in "Week 121", and also in my paper "Categorification", at `http://www.arXiv.org/abs/math.QA/9802029`*. Here's what I said:*

> *If one studies categorification one soon discovers an amazing fact: many deep-sounding results in mathematics are just categorifications of facts we learned in high school! There is a good reason for this. All along, we have been unwittingly decategorifying mathematics by pretending that categories are just sets. We 'decategorify' a category by forgetting about the morphisms and pretending that isomorphic objects are equal. We are left with a mere set: the set of isomorphism classes of objects.*
>
> *To understand this, the following parable may be useful. Long ago, when shepherds wanted to see if two herds of sheep were isomorphic, they would look for an explicit isomorphism. In other words, they would line up both herds and try to match each sheep in one herd*

> *with a sheep in the other. But one day, along came a shepherd who invented decategorification. She realized one could take each herd and 'count' it, setting up an isomorphism between it and some set of numbers, which were nonsense words like 'one, two, three, . . . ' specially designed for this purpose. By comparing the resulting numbers, she could show that two herds were isomorphic without explicitly establishing an isomorphism! In short, by decategorifying the category of finite sets, the set of natural numbers was invented.*
>
> *According to this parable, decategorification started out as a stroke of mathematical genius. Only later did it become a matter of dumb habit, which we are now struggling to overcome by means of categorification. While the historical reality is far more complicated, categorification really has led to tremendous progress in mathematics during the 20th century. For example, Noether revolutionized algebraic topology by emphasizing the importance of homology groups. Previous work had focused on Betti numbers, which are just the dimensions of the rational homology groups. As with taking the cardinality of a set, taking the dimension of a vector space is a process of decategorification, since two vector spaces are isomorphic if and only if they have the same dimension. Noether noted that if we work with homology groups rather than Betti numbers, we can solve more problems, because we obtain invariants not only of spaces, but also of maps.*
>
> *In modern language, the nth rational homology is a functor defined on the category of topological spaces, while the nth Betti number is a mere function, defined on the set of isomorphism classes of topological spaces. Of course, this way of stating Noether's insight is anachronistic, since it came before category theory. Indeed, it was in Eilenberg and Mac Lane's subsequent work on homology that category theory was born!*
>
> *Decategorification is a straightforward process which typically destroys information about the situation at hand. Categorification, being an attempt to recover this lost information, is inevitably fraught with difficulties.*
>
> *Finally, a note to John: While you're trying to give your audience some feeling for the virtues of n-categories, couldn't you give them a little help with $n = 1$, by being a little more precise about objects and maps?*

I hope it's clearer now.

Best, jb

As a high-end cable manufacturer, Cardas Audio strives to address every detail of cable and conductor construction, no matter how small. An elegant solution deals with quality, not quantity. Cable geometry problems are resolved in the cable's design, not after the fact with filters. George Cardas received U.S. Patent Number 4,628,151 for creating Golden Section Stranding Audio Cable. It is truly unique. George introduced the concept of Golden Section Stranding to high-end audio, but the Golden Ratio, $1.6180339887\ldots : 1$, is as old as nature itself. The Golden Ratio is the mathematical proportion nature uses to shape leaves and sea shells, insects and people, hurricanes and galaxies, and the heart of musical scales and chords. "Discovered" by the Greeks, but used by the Egyptians in the Great Pyramid centuries before, man has employed the Golden Ratio to create his most beautiful and naturally pleasing works of art and architecture

— *Cardas Audio speaker cable advertisement*

Week 204

March 24, 2004

The star we know as GRB030329 was named after the day the news of its death reached Earth. About 2,650 million years ago, this star exploded. For thirty seconds it put out more power in the form of γ rays than everything else in the visible universe combined. These γ rays reached us on March 3rd, 2003, and they were detected by a satellite called HETE-II: the High-Energy Transient Explorer.

The detection of GRB03029 set off a frenzy of activity among astronomers all over the world. As the closest γ-ray burster to be seen by well-prepared earthlings, GRB030329 taught us a lot. We're not completely sure what it was — but we have a pretty good guess, and it makes a nice story, so I'll recount it as if it were a fact.

As far as we can tell, GRB030329 was a Wolf-Rayet star before it exploded. Wolf-Rayet stars are very rare: only 200 have been seen in our galaxy. They're huge and very bright — up to a million times as bright as the Sun. They're surrounded by enormous bluish-purple nebulae like the one in this picture:

1) NGC 2359, the nebula around the Wolf-Rayet star HD56925, picture at `http://cfa-www.harvard.edu/cfa/hotimage/n2359.html`

But what makes them really special is that their spectral lines show *little or no hydrogen*. Since most of the universe is made of hydrogen, a star without hydrogen is like a dry fish. How can this be?

Well, the life of a star is largely determined by its mass. Small stars last a long time and fade away inconspicuously, while big stars live fast and die with a bang. Wolf-Rayet stars are among the biggest, about 60 times as heavy as the Sun. Like the sun, they begin life as cloud of gas that collapses and heats up until the hydrogen in its core "catches fire" and starts fusing into helium, like a gigantic H-bomb held together by its own gravity. The core is surrounded by a envelope of cooler gas that transmits energy to the surface by convection and radiation, but doesn't actually do any fusion itself.

This stage of a star's life is called the "core burning phase". But after a while helium builds up and sinks to the center, forming an inert helium core, with all the fusion going on in the layer of hydrogen right next to the core. This is called the "shell burning phase".

What next? Well, for the Sun, as its hydrogen gradually runs out it'll become a "red giant", expanding to engulf the Earth... while meanwhile its helium core shrinks to a ball twice the size of the Earth and about 100 times the density of water, turning from ordinary plasma into something called a "degenerate electron gas", where the Pauli exclusion principle limits further compression. As the core shrinks it'll heat up, and when it reaches a temperature of 100 million kelvin the helium will catch fire and start fusing — mainly into carbon. Models predict that this happens in a runaway reaction called the "helium flash", which puts out about 100 billion times the power of the present-day sun for a few hours — zounds! — but gradually settles down into a more stable phase of helium burning that lasts for tens of millions of years. During this phase, the Sun will not be a red giant anymore, but instead a hotter "yellow giant".

The Sun will never get hot enough to burn elements heavier than helium, so eventually it'll develop an inert core of carbon and other junk, surrounded by a helium burning shell, surrounded by a hydrogen burning shell. Then the outer layers will peel off and expand to form a huge nebula, leaving the core as a tiny "white dwarf"... which will cool, after eons, to a "black dwarf". Here's a nice chart of the whole story:

2) Sloan Digital Sky Survey, Evolutionary track of a sun-like star, `http://skyserver.sdss.org/dr1/en/astro/stars/images/starevol.jpg`

Bigger stars do more exciting things. In particular, stars heavier than about 5 solar masses undergo a "carbon flash" when the carbon-rich core reaches 600 million kelvin and starts fusing into heavier elements. Heavier stars then go on to an oxygen-burning phase. Even heavier ones go on to a silicon-burning phase.

But when silicon fuses, it forms highly stable nuclei like iron that don't want to fuse any further. So, silicon burning is the end of the line. And it doesn't last long! For example, a star 25 times the mass of the Sun is expected to spend about 5 to 10 million years burning hydrogen, 0.5 to 1 million years burning helium, 500 to 1000 years burning carbon, 6 to 12 months burning oxygen... but just a day or so burning silicon!

Then what? Well, the details depend on the star's mass. But when a star of at least 8 solar masses runs out of fuel, its core is made mainly of iron, and heavier than our Sun. When it cools, it reaches a point where all of a sudden it collapses — in about a tenth of a second. When it crashes in on itself, it gets so hot that the iron nuclei disintegrate and the whole mess explodes in a "type II supernova". The star's outer layers get thrown off at high speeds, while the core itself gets crushed into a neutron star... or, for truly heavy stars, a black hole!

Type II supernovae are among the most violent events in the cosmos. They can easily reach a temperature of about 50 billion kelvin and emit 10^{46} joules of energy, which is what our galaxy puts out in 10 years! 99% of this energy is in the form of neutrinos, emitted when protons in the iron core absorb electrons and turn into neutrons. But, the remaining 1% in the form of electromagnetic radiation is still enough to fry anything in the vicinity. The supernova in the Crab Nebula was about 6,300 light years away, but when its light reached us in 1054 AD, Chinese astronomers could see it in the daytime for 23 days!

You may think I've forgotten about GRB030329 and Wolf-Rayet stars, but I haven't. This big digression was just to set the stage. I've sketched what stars of up to 25 solar masses will do, but remember, Wolf-Rayets are a lot bigger: they begin life at about 60 solar masses. And astronomy resembles opera in this way: the bigger the star, the more noise they make in their final scene. So, the stuff about supernovae was just to whet your appetite.

So, let's sit back and watch the thrilling life story of GRB030329... assuming that it began its days as most Wolf-Rayets do.

As a child, it burnt hydrogen in a huge core of about 50 solar masses. After a while helium "ashes" built up in this core, so it moved on to burning hydrogen in a shell. But this process put out so much energy that the envelope started getting blown away in a powerful stellar wind!

Since the helium wasn't burning, the core contracted until the temperature hit 40 million kelvin and the helium caught fire. It started burning into carbon-12, but some hydrogen got into the core and made carbon-13 and nitrogen-14, and later — when the helium was almost all burnt — oxygen-16.

All the while the stellar wind was increasing, and eventually almost all the hydrogen was blown away, leaving only a bluish-white core full of helium, carbon, nitrogen and a little oxygen. Now you see how a star gets rid of its hydrogen! At this point GRB030329 was a classic Wolf-Rayet star: almost no hydrogen in the star itself, lots of stellar wind, and surrounded by a big nebula of gas and dust that had been blown off.

When all its helium was burnt, our hero's days were numbered. In an ever-accelerating frenzy, it spent its last centuries burning carbon, then oxygen, then silicon. Meanwhile its stellar wind kept picking up speed, up to 5 or 10 thousand kilometers per second, blowing away more and more gas and dust. By the time all the silicon had burnt to iron, the core had shrunk down to about 10 solar masses.

And when the fuel ran out, the core cooled down and collapsed.

The core was so big, and its collapse so drastic, that it didn't "bounce back" and explode outwards, as in a supernova. Instead, gravity triumphed! A black hole formed, sucking down a hefty amount of the core in less than a tenth of a second.

As several solar masses of iron rapidly spiralled down the throat of this growing black hole, it formed a pancake-like "accretion disk", which emitted powerful jets of radiation and matter in both directions along its axis of rotation. In a few seconds, these jets passed through the outer shell of the star and, together with a blast of newly created radioactive nickel-56, shattered it completely. Our star became a "hypernova"!

Meanwhile, the jets plowed into the material surrounding the star and created highly directional beams of γ rays shooting in opposite directions... one of which just happened to be pointed directly at the Earth.

2,650 million years later, the γ rays reached us, and were detected on March 23, 2003 by HETE-II. Hundreds of such bursts are detected each year, but this one was closer than most, and a whole system had recently been devised for quickly turning the attention of the world's telescopes to the spot where a γ-ray burster had been seen — in this case, within the constellation Leo.

So, within 90 min, a 40-inch telescope at the Siding Spring Observatory in Australia was looking at this spot. So was a telescope in Japan. They saw the optical afterglow of the γ-ray burster and watched how its brightness changed with time. And within 24 hours, a spectrograph on a telescope in Chile made detailed readings of the spectrum, measuring the redshift ($z = 0.1685$) and thus the distance of the burst, and seeing signs of radioactive nickel — about $1/3$ of a solar mass of the stuff, according to one estimate! Later, more telescopes probed the event in different ways.

The details of what was seen gave a lot of credence to the "hypernova" or "collapsar" model of γ-ray bursters, championed by Stan Woosley of U. C. Santa Cruz, among others. But much remains mysterious about γ-ray bursters. Nobody knows exactly how the energy from the jet gets turned into γ rays! And, the hypernova model only fits "long" γ-ray bursters, where the burst lasts about 2 seconds or more. There are also "short" ones, which may work some other way.

So, the hows and whys of γ ray bursts remain one of the most fascinating mysteries in physics. And since we can't actually peek inside a star, a lot of the attempts to study these things involve complicated mathematical models... very technical stuff, when you actually try to read it. So, I really *am* talking about mathematical physics. Honest!

Here are some ways to learn more, starting with the fun easy stuff.

The online version of the Messier Catalog — a famous old catalog of galaxies and nebulae — is a really fun way to learn some astronomy:

3) The Messier Catalog, `http://www.maa.agleia.de/Messier/`

It's packed with interesting stuff. For example, here's a great page about nebulae like the one the sun will form after it becomes a yellow giant:

4) The Messier Catalog, "Planetary nebulae", `http://www.maa.agleia.de/Messier/planetar.html`

(They're misleadingly called "planetary nebulae", though they don't have anything to do with planets.) And here's a nice page about the Crab Nebula, which is now a pulsar — a rapidly spinning neutron star left over from a supernova:

5) The Messier Catalog, "The crab nebula (M1)", `http://www.maa.agleia.de/Messier/E/m001.html`

This website is good for Wolf-Rayet stars and other things:

6) Chris Clowes' Astronomy Page, `http://www.peripatus.gen.nz/Astronomy/`

I learned a lot about GRB030329 from this page:

7) European Southern Observatory (ESO), "Cosmological γ-ray bursts and hypernovae conclusively linked", June 18, 2003, `http://www.eso.org/outreach/press-rel/pr-2003/pr-16-03.html`

For another key moment in the history of γ-ray bursters, try this:

8) Burst and Transient Source Experiment (BATSE), "GOTCHA! — The big one that didn't get away", January 27, 1999, `http://www.batse.com/jan27.html`

For more on γ-ray bursters, try these:

9) NASA, γ ray bursts, `http://imagine.gsfc.nasa.gov/docs/introduction/bursts.html`

10) Edo Berger, Gamma-ray burst FAQ, `http://www.astro.caltech.edu/~ejb/faq.html`

If you get more serious, there are lots of conference proceedings to read, like this:

11) M. Livio, N. Panagia and K. Sahu, editors, *Supernovae and Gamma-Ray Bursts: The Greatest Explosions since the Big Bang*, Cambridge U. Press, 2001.

There must be a bunch of conference proceedings written after the March 2003 burster, but maybe they haven't been published yet, since I haven't been able to find them!

If you're looking for a more general background in astrophysics, this hefty tome is supposed to be a good intro, though I haven't tried it yet:

12) Bradley W. Carroll and Dale A. Ostlie, *Introduction to Modern Astrophysics*, Addison Wesley, 1996.

Personally I've found these helpful in writing the above stuff, though they're full of equations, so I find myself yearning for some purple prose here and there:

13) R. J. Tayler, *The Stars: Their Structure and Evolution*, 2nd edition, Cambridge U. Press, Cambridge, 1994.

14) R. Kippenhahn and A. Weigert, *Stellar Structure and Evolution*, Springer Verlag, Berlin, 1991.

By the way, I don't know much about astrophysics, so I'd love to hear from any experts out there who'd like to correct or add detail to my description of Wolf-Rayet stars, the "hypernova" scenario, or γ-ray bursters in general. I've been fond of Wolf-Rayet stars ever since I wrote a few little articles on weird kinds of stars:

15) John Baez, "Stuff about Stars", `http://math.ucr.edu/home/baez/stars.html`

back before anyone suspected they were related to γ-ray bursters! My interest in them was rekindled while revising the physics FAQ on open questions in physics:

16) John Baez, "Open Questions in Physics", `http://math.ucr.edu/home/baez/open.questions.html`

It hadn't been rewritten since 1997, and it was interesting to see how outdated it had become, particularly in the area of cosmology and astrophysics! Here's the current list of problems:

Condensed Matter and Nonlinear Dynamics

1) *What causes sonoluminescence?*
2) *What causes high temperature superconductivity?*
3) *How can turbulence be understood and its effects calculated?*
4) *The Navier-Stokes equations are the basic equations describing fluid flow. Does these equations have solutions that last for all time, given arbitrary sufficiently nice initial data?*

Quantum Mechanics

1) *How should we think about quantum mechanics?*
2) *Can we build a working quantum computer big enough to do things ordinary computers can't easily do?*

Cosmology and Astrophysics

1) *What happened at or before the Big Bang?*
2) *Are there really three dimensions of space and one of time? If so, why? Or is spacetime higher-dimensional, or perhaps not really a manifold at all when examined on a short enough distance scale?*

3) *Why is there an arrow of time; that is, why is the future so much different from the past?*
4) *Is the Universe infinite in spatial extent? More generally: what is the topology of space?*
5) *Will the future of the Universe go on forever or not?*
6) *Is the universe really full of "dark energy"? If so, what causes it?*
7) *Why does it seem like the gravitational mass of galaxies exceeds the mass of all the stuff we can see, even taking into account our best bets about invisible stuff like brown dwarfs, "Jupiters", and so on?*
8) *The Horizon Problem: why is the Universe almost, but not quite, homogeneous on the very largest distance scales?*
9) *Why are the galaxies distributed in clumps and filaments?*
10) *When were the first stars formed, and what were they like?*
11) *What are γ ray bursters?*
12) *What is the origin and nature of ultra-high-energy cosmic rays?*
13) *Do gravitational waves really exist? If so, can we detect them? If so, what will they teach us about the universe? Will they mainly come from expected sources, or will they surprise us?*
14) *Do black holes really exist? (It sure seems like it.) Do they really radiate energy and evaporate the way Hawking predicts? If so, what happens when, after a finite amount of time, they radiate completely away? What's left? Do black holes really violate all conservation laws except conservation of energy, momentum, angular momentum and electric charge? What happens to the information contained in an object that falls into a black hole?*
15) *Is the Cosmic Censorship Hypothesis true? Roughly, for generic collapsing isolated gravitational systems are the singularities that might develop guaranteed to be hidden beyond a smooth event horizon?*

Particle Physics

1) *Why are the laws of physics not symmetrical between left and right, future and past, and between matter and antimatter?*
2) *Why is there more matter than antimatter, at least around here?*
3) *Are there really just three generations of leptons and quarks? If so, why?*
4) *Why does each generation of particles have precisely this structure: two leptons and two quarks?*
5) *Do the quarks or leptons have any substructure, or are they truly elementary particles?*
6) *Is there really a Higgs boson, as predicted by the Standard Model of particle physics? If so, what is its mass?*

7) *What is the correct theory of neutrinos? Why are they almost but not quite massless? Do all three known neutrinos — electron, muon, and τ — all have a mass?*

8) *Is quantum chromodynamics (QCD) a precise description of the behavior of quarks and gluons? Can we prove that quarks are gluons are confined at low temperatures using QCD? Is it possible to calculate masses of hadrons (such as the proton, neutron, pion, etc.) correctly from the Standard Model, with the help of QCD? Does QCD predict that quarks and gluons become deconfined and form plasma at high temperature? If so, what is the nature of the deconfinement phase transition?*

9) *Is there a mathematically rigorous formulation of a relativistic quantum field theory describing interacting (not free) fields in four spacetime dimensions? For example, is the Standard Model mathematically consistent? How about Quantum Electrodynamics?*

10) *Is the proton really stable, or does it eventually decay?*

11) *Why do the particles have the precise masses they do?*

12) *Why are the strengths of the fundamental forces (electromagnetism, weak and strong forces, and gravity) what they are?*

13) *Are there important aspects of the Universe that can only be understood using the Anthropic Principle? Or is this principle unnecessary, or perhaps inherently unscientific?*

14) *Do the forces really become unified at sufficiently high energy?*

15) *Does some version of string theory or M-theory give specific predictions about the behavior of elementary particles? If so, what are these predictions? Can we test these predictions in the near future? And: are they correct?*

The Big Question

1) *How can we merge quantum theory and general relativity to create a quantum theory of gravity? How can we test this theory?*

A bunch of these questions could turn out to be a bit silly — a good answer might require changing the question. But that's always how it goes for really interesting puzzles. I should also warn you that the statements above are deliberately a bit naive-sounding: as the example of γ ray bursters shows, we actually do know a lot about all these questions — we're just not sure about the answers! So, see the webpage itself for a bit more information on these questions, and the links for even more...

Hmm. I was going to say something about number theory, but I'm out of time!

> Of course: abstraction, irrelevance, purity, formalism make for good mathematics.... But sadly, they make for bad mathematics education. Each one of these concepts — abstract, irrelevance, purity, formalism - pushes mathematics further away from a growing human being, a being whose psyche is

in the phase of it development that no soft-brained psychologist but a great mathematician, Alfred North Whitehead, calls the Romantic Phase.

— *Apostolos Doxiadis*

Week 205

April 11, 2004

This week I'd like to say more about number theory, but first — here's the most fun book on astronomy I've ever seen:

1) James B. Kaler, *The Hundred Greatest Stars*, Copernicus Books (Springer Verlag), New York, 2002.

It's just what the title says: a compilation of the author's 100 favorite stars, each with a picture and a one-page description of what makes that star interesting. They're incredibly diverse, from the mammoth Eta Carinae to tiny brown dwarf Gliese 229B. You'll see soft gamma repeaters, yellow hypergiants, pulsars, Mira-type variables, barium stars, symbiotic stars, and more. There's also an introduction that explains the concepts needed to enjoy all these different kinds of stars, like the Hertzsprung-Russell diagram and a bit of nuclear physics. With the help of this, the whole book forms a wonderful taxonomy of stellar astrophysics. I suggest keeping it by your bed and reading one star a night — though you may wind up staying up late and devouring the whole book.

On to number theory. . . .

There's a widespread impression that number theory is about *numbers*, but I'd like to correct this, or at least supplement it. A large part of number theory — and by the far the coolest part, in my opinion — is about a strange sort of *geometry*. I don't understand it very well, but that won't prevent me from taking a crack at trying to explain it. . . .

The basic idea is to push the analogy between integers and polynomials as far as it will go. They're similar because you can add, subtract and multiply them, and these operations satisfy the usual rules we all learned in high school:

- $x + y = y + x$
- $(x + y) + z = x + (y + z)$
- $x + 0 = x$
- $x + (-x) = 0$
- $xy = yx$
- $(xy)z = x(yz)$
- $x1 = x$
- $x(y + z) = xy + xz$

Anything satisfying these rules is called a "commutative ring". There are also a lot of deeper similarities between integers and polynomials, which I'll talk about later. But, there's a big difference! Polynomials are functions on the line, whereas the integers aren't functions on some space — at least, not in any instantly obvious way.

The fact that polynomials are functions on a space is what lets us graph them. This lets us think about them using *geometry* — and also think about geometry using *them*. This was the idea behind Descartes' "analytic geometry", and it was immensely fruitful.

So, it would be cool if we could also think about the integers using geometry. And it turns out we can, but only if we stretch our concept of geometry far enough!

If we do this, we'll see some cool things. First of all, we'll see that algebra is just like geometry, only backwards.

What do I mean by this? Well, whenever you have a map $T\colon X \to Y$ going from the space X to the space Y, you can use it to take functions on Y and turn them into functions on X. Since this goes backwards, it's called "pulling back along T". Here's how it goes: if f is a function on Y, we get a function $T^*(f)$ on X given by:

$$T^*(f)(x) = f(T(x))$$

Moreover, functions on a space form a commutative ring, since you can add and multiply them pointwise, and pulling back is a "homomorphism", meaning that it preserves all the structure of a commutative ring:

- $T^*(f+g) = T^*(f) + T^*(g)$
- $T^*(0) = 0$
- $T^*(fg) = T^*(f)T^*(g)$
- $T^*(1) = 1$

Conversely, any sufficiently nice homomorphism from functions on Y to functions on X will come from some map $T\colon X \to Y$ this way! Here I'm summarizing a whole bunch of different theorems, each of which goes along with its own precise definition of "space", "map", and "nice".

Some of these theorems are technical, but the basic idea is simple: we can translate back and forth between the study of commutative rings (algebra) and the study of spaces (geometry) and by thinking of commutative rings as consisting of functions on spaces. We get a little dictionary for translating between geometry and algebra, like this:

Geometry	Algebra
spaces	commutative rings
maps	homomorphisms

But be careful: this translation turns maps into homomorphisms going backwards: it's "contravariant". This is really important in two ways. First, suppose we have a point x in a space X. This gives a map

$$i\colon \{x\} \to X$$

This, in turn, gives a homomorphism i^* sending functions on X to functions on $\{x\}$. Functions on a one-point space are like numbers, so i^* acts like "evaluation at x". Moreover, i^* will tend to be onto: that's the backwards analogue of the fact that i is one-to-one!

Second, suppose we have a map from a space E onto the space X:

$$p\colon E \to X.$$

If you know some topology, think of E as a "covering space" of X. Then we get a homomorphism p^* from functions on X to functions on E. Moreover p^* will tend to be one-to-one: that's the backwards version of the fact that p was onto!

We can use these examples to figure out the analogue of a "point" or a "covering space" in the world of commutative rings! And the resulting ideas turn out to be crucial to modern number theory.

In "Week 199" I explained the analogue of a "point" for commutative rings: it's a "prime ideal". So, now I want to explain the analogue of a "covering space". This will expand our dictionary so that it relates Galois groups to fundamental groups of topological spaces. . . and so on.

But, we won't get too far if we don't remember why a "prime ideal" is like a "point"! So, I guess I'd better review some of "Week 199" before charging ahead into the beautiful wilderness.

What's special about the ring of functions on a space consisting of just one point? Take real- or complex-valued functions, for example. How do these differ from the functions on a space with lots of points?

The answer is pretty simple: on a space with just one point, a function that vanishes anywhere vanishes everywhere! So, the only function that fails to have a multiplicative inverse is 0. For bigger spaces, this isn't true.

A commutative ring where only 0 fails to have a multiplicative inverse is called a "field". So, the algebraic analogue of a one-point space is a field.

This means that the algebraic analogue of a map from a one-point space into some other space:

$$i\colon \{x\} \to X$$

should be a homomorphism from a commutative ring R to a field k:

$$f\colon R \to k.$$

Our translation dictionary now looks like this:

Geometry	Algebra
spaces	commutative rings
maps	homomorphisms
one-point spaces	fields
maps from one-point spaces	homomorphisms to fields

It's worth noting some subtleties here. In the geometry we learned in high school, once we see one point, we've seen 'em all: all one-point spaces are isomorphic. But not all fields are isomorphic! So, if we're trying to think of algebra as geometry, it's a funny sort of geometry where points come in different flavors!

Moreover, there are homomorphisms between different fields. These act like "flavor changing" maps — maps from a point of one flavor to a point of some other flavor.

If we have a homomorphism $f\colon R \to k$ and a homomorphism from k to some other field k', we can compose them to get a homomorphism $f'\colon R \to k'$. So, we're doing some funny sort of geometry where if we have a point mapped into our space, we can convert it into a point of some other flavor, using a "flavor changing" map.

Now let's take this strange sort of geometry really seriously, and figure out how to actually turn a commutative ring into a space! First I'll describe what people usually do. Eventually I'll describe what perhaps they really should do — but maybe you can guess before I even tell you.

People usually cook up a space called the "spectrum" of the commutative ring R, or $\mathrm{Spec}(R)$ for short. What are the points of $\mathrm{Spec}(R)$? They're not just all possible homomorphisms from R to all possible fields. Instead, we count two such homomorphisms as the same point of $\mathrm{Spec}(R)$ if they're related by a "flavor changing process". In other words, $f'\colon R \to k'$ gives the same point as $f\colon R \to k$ if you can get f' by composing f with a homomorphism from k to k'.

This is a bit ungainly, but luckily there's a quick and easy way to tell when $f\colon R \to k$ and $f'\colon R \to k'$ are related by such a flavor changing process, or a sequence of such processes. You just see if they have the same kernel! The "kernel" of $f\colon R \to k$ is the subset of R consisting of elements r with

$$f(r) = 0$$

The kernel of a homomorphism to a field is a "prime ideal", and two homomorphisms are related by a sequence of flavor changing processes iff they have the same kernel. Furthermore, every prime ideal is the kernel of a homomorphism to some field. So, we can save time by defining $\mathrm{Spec}(R)$ to be the set of prime ideals in R.

For completeness I should remind you what a prime ideal is! An "ideal" in a ring R is a set closed under addition and closed under multiplication by anything in R. It's "prime" if it's not all of R, and whenever the product of two elements of R lies in the ideal, at least one of them lies in the ideal.

So, we have something like this:

Geometry	Algebra
spaces	commutative rings
maps	homomorphisms
one-point spaces	fields
maps from one-point spaces	homomorphisms to fields
points of a space	prime ideals of a commutative ring

Now let's use these ideas to study "branched covering spaces" and their analogues in algebra. This week I'll talk about two examples. The first is very geometrical, and it should be familiar to anyone who has studied a little complex analysis. The second is more algebraic, and it's important in number theory. But, the cool part is that they fit into the same formalism!

If you don't know what a branched covering space is, don't worry: we'll start with the very simplest example. We'll look at this map from the complex plane to itself:

$$\begin{aligned} p\colon \mathbb{C} &\to \mathbb{C} \\ z &\mapsto z^2 \end{aligned}$$

Except for zero, every complex number has two square roots, so this map is two-to-one and onto away from the origin. In fact, away from the origin you can visualize this thing locally as two sheets of paper sitting above one. But these two sheets have a global complication: if you start on the top sheet and hike once around the origin, you wind up on the bottom sheet — and vice versa! In topology we call this sort of thing a "double cover". When we include the point $z = 0$ things get even more complicated, since the two sheets meet there. So we have something trickier: a "branched cover". In general, a branched cover is like a covering space except that the different "sheets" can merge together at certain points, called "branch points".

Now let's think about this algebraically. To keep from getting confused, let's write

$$z^2 = w$$

so that p is a map from the "z-plane" down to the "w-plane", sending each point z to the point $z^2 = w$. The ring of polynomial functions on the z-plane is called $\mathbb{C}[z]$; the ring of polynomial functions on the w-plane is called $\mathbb{C}[w]$. We can pull functions from the w-plane back up to the z-plane:

$$p^* : \mathbb{C}[w] \to \mathbb{C}[z]$$

and p^* works in the obvious way, taking any function $f(w)$ to the function $f(z^2)$.

Just as p is onto, p^* is one-to-one! So, we can think of $\mathbb{C}[w]$ as sitting inside $\mathbb{C}[z]$, consisting of those polynomials in z that only depend on z^2: the even functions. We say $\mathbb{C}[w]$ is a "subring" of $\mathbb{C}[z]$, or equivalently, that $\mathbb{C}[z]$ is an "extension" of $\mathbb{C}[w]$.

In this example we can get the bigger ring from the smaller one by throwing in solutions of some polynomial equations, so we call it an "algebraic extension". We've already seen some algebraic extensions, namely algebraic number fields, where take the field of rational numbers and throw in some solutions of polynomial equations. Algebraic extensions can be complicated, but this one is really simple: we just start with $\mathbb{C}[w]$ and throw in the solution of *one* polynomial equation, namely

$$z^2 = w$$

It turns out that quite generally, algebraic extensions of commutative rings act a lot like branched covering spaces. I probably don't have the technical details perfectly straight, but let's add this to our translation dictionary, because it's an important idea:

Geometry	Algebra
spaces	commutative rings
maps	homomorphisms
one-point spaces	fields
maps from one-point spaces	homomorphisms to fields
points of a space	prime ideals of a commutative ring
branched covering spaces	algebraic extensions of commutative rings

Now let's have some fun: let's see how our algebraic concept of "point", namely "prime ideal", interacts with our branched double cover of the complex plane. There's something straightforward going on, but also something more subtle and interesting.

The straightforward thing is that any point up on the z-plane maps to one down on the w-plane. We don't need fancy algebra to see this! But, it's worth doing algebraically. According to the fancy algebraic definition, a "point" in the spectrum of the commutative ring $\mathbb{C}[z]$ is a prime ideal. But as you might hope, these are the same as good old-fashioned points in the complex plane!

It works like this: given any point x in $\mathbb{C}$, we get a homomorphism from $\mathbb{C}[z]$ to $\mathbb{C}$ called "evaluation at x", which sends any polynomial f to the number $f(x)$. The kernel of this is the prime ideal consisting of all polynomials that vanish at x. These are just the polynomials containing a factor of $z - x$, so we call this ideal

$$\langle z - x\rangle$$

So, we get some prime ideals in $\mathbb{C}[z]$ from points of $\mathbb{C}$ this way. But in fact there's a theorem that *every* prime ideal in $\mathbb{C}[z]$ is of this form! So, we get a one-to-one correspondence

$$\mathrm{Spec}(\mathbb{C}[z]) = \mathbb{C}$$

Similarly,

$$\mathrm{Spec}(\mathbb{C}[w]) = \mathbb{C}$$

Now let's think about our branched cover

$$p\colon \mathbb{C} \to \mathbb{C}$$

in different ways. It starts out life as a map from the z-plane down to the w-plane. We can use this to pull back functions on the w-plane up to the z-plane:

$$p^*\colon \mathbb{C}[w] \to \mathbb{C}[z]$$

But then, by general abstract baloney, the inverse image under p^* of any prime ideal in $\mathbb{C}[z]$ is a prime ideal back in $\mathbb{C}[w]$. This gives a map from $\mathrm{Spec}(\mathbb{C}[z])$ to $\mathrm{Spec}(\mathbb{C}[w])$. But this is just a map from the z-plane to the w-plane! And it's the same map p we started with. If you don't see why, it's a good exercise to check this.

So: we translated from geometry to algebra and back to geometry, and we got right back where we started. Note that each time we translated, our description of the map p got turned around backwards.

But there's a subtler and more interesting thing we can do with our branched cover. We can take a point down on the w-plane and look at the points up on the z-plane that map down to it!

Usually there will be two, but for the origin there's just one. This much is clear from thinking geometrically. But if we think algebraically, we'll see something funny going on at the origin. We can already see it geometrically: the origin is where the two sheets of our branched cover meet, so we call it a "branch point". But the algebraic viewpoint sheds an interesting new light on this.

What we'll do is take a prime ideal in $\mathbb{C}[w]$ and push it forwards via

$$p^*\colon \mathbb{C}[w] \to \mathbb{C}[z]$$

The resulting subset won't be an ideal, but it will "generate" an ideal, meaning we can take the smallest ideal containing it. This ideal won't be prime, but we can "factor" it into prime ideals: there's a fairly obvious way to multiply ideals, and we happen to be working with rings where there's a unique way to factor any ideal into prime ideals.

Let's try it. First pick a number x that's not zero. It gives a prime ideal in $\mathbb{C}[w]$, namely

$$\langle w - x\rangle$$

Next push this ideal forwards via p^* and let it generate an ideal in $\mathbb{C}[z]$, namely

$$\langle z^2 - x\rangle$$

This is not prime, but we can factor it, which in this case simply amounts to factoring the polynomial that generates it:

$$\langle z^2 - x\rangle = \langle (z - \sqrt{x})(z + \sqrt{x})\rangle = \langle z - \sqrt{x}\rangle\langle z + \sqrt{x}\rangle$$

We get a product of two prime ideals, corresponding to two points in the z-plane, namely $+\sqrt{x}$ and $-\sqrt{x}$. These are the two points that map down to x.

In this sort of situation, we say the prime ideal $\langle w - x\rangle$ "splits" into the prime ideals $\langle z - \sqrt{x}\rangle$ and $\langle z + \sqrt{x}\rangle$ when we go from $\mathbb{C}[w]$ to the extension $\mathbb{C}[z]$. This is just an overeducated way of saying the number x has two different square roots.

But suppose $x = 0$. This doesn't have two square roots! Everything works the same except we get

$$\langle z^2\rangle = \langle z\rangle\langle z\rangle$$

We say the prime ideal $\langle w\rangle$ "ramifies" when we go from $\mathbb{C}[w]$ to the extension $\mathbb{C}[z]$. We still get a product of prime ideals; they just happen to be the same. This is a way of making sense of the funny notion that the number 0 has two square roots... which just happen to be the same! Lots of mathematicians and physicists talk about "repeated roots" when an equation has "two solutions that just happen to be equal". This is just a way of making that precise.

But all this algebraic machinery must seem like overkill if this is the first time you've seen it. It pays off when we get to more algebraic examples. So, let me sketch the simplest one.

Let $\mathbb{Z}$ be the ring of integers, and let $\mathbb{Z}[i]$ be the ring of Gaussian integers, namely numbers of the form $a + bi$ where a and b are integers. $\mathbb{Z}[i]$ is an algebraic extension of $\mathbb{Z}$, since we can get it by throwing in a solution z of the polynomial equation

$$z^2 = -1$$

This equation is quadratic, just like it was in the example we just did! Now we're throwing in a square root of -1 instead of a square root of some function on the complex plane. But if we take the analogy between geometry and algebra seriously, this extension should still give some sort of "branched double cover"

$$p\colon \mathrm{Spec}(\mathbb{Z}[i]) \to \mathrm{Spec}(\mathbb{Z})$$

What's this like?

It's actually really interesting, but I'll just *sketch* how it works.

The points of $\mathrm{Spec}(\mathbb{Z})$ are prime ideals in $\mathbb{Z}$. In "Week 199" we saw that except for the prime ideal $\langle 0\rangle$, these are generated by prime numbers.

Similarly, except for $\langle 0\rangle$, the prime ideals in $\mathbb{Z}[i]$ are generated by "Gaussian primes": Gaussian numbers that have no factors except themselves and the "units" 1, -1, i and $-i$. (A "unit" in a ring is an element with a multiplicative inverse; we don't count units as primes.)

The map p sends each Gaussian prime to a prime, and it's fun to work out how this goes... but it's even more fun to work backwards! Let's take primes in the integers and see what happens when we let them generate ideals in the Gaussian integers! This is like taking points in the base space of a branched cover and seeing what's sitting up above them.

For example, the prime 5 "splits". It has two prime factors in the Gaussian integers:

$$5 = (2+i)(2-i)$$

so in $\mathbb{Z}[i]$ the ideal it generates is a product of two prime ideals:

$$\langle 5\rangle = \langle 2+i\rangle\langle 2-i\rangle$$

This means that two different points in $\mathrm{Spec}(\mathbb{Z}[i])$ map down to the point $\langle 5\rangle$ in $\mathrm{Spec}(Z)$, namely $\langle 2+i\rangle$ and $\langle 2-i\rangle$. So we indeed have something like a double cover!

On the other hand, the prime 2 "ramifies". It has two prime factors in the Gaussian integers:

$$2 = (1+i)(1-i)$$

but these two Gaussian primes generate the same prime ideal:

$$\langle 1+i\rangle = \langle 1-i\rangle$$

since if we multiply $1+i$ by the unit $-i$ we get $1-i$. So, in the Gaussian integers we have

$$\langle 2\rangle = \langle 1+i\rangle\langle 1+i\rangle$$

A repeated factor! This is just what happened to the branch point in our previous example: it had "two points sitting over it, which happen to be the same".

So far, everything seems to be working nicely. But, besides splitting and ramification, there's a third thing that happens here, which didn't happen in our example involving the complex plane. In fact, this third option never happens when we're doing algebraic geometry over the complex numbers!

Here's how it works. Consider the prime 3. This is still prime in the Gaussian integers! It doesn't split, and it doesn't ramify. If we factorize the ideal generated by 3 in $\mathbb{Z}[i]$ we just get

$$\langle 3\rangle = \langle 3\rangle$$

It doesn't do anything — it just sits there! So, we say this prime is "inert".

This may seem boring, but it's actually mysterious — and downright MADDENING if we take the analogy between geometry and algebra seriously. It's weird enough to have a "branched" cover where sheets merge at certain points, but at least in that case we can *see* they've merged: a prime ideal in our subring generates an ideal in the extension that's not prime, but is a product of several prime factors, some of which happen to be the same. But when a prime ideal in our subring generates a *prime* ideal in the extension, it's as if our "cover" has just *one* sheet over this point in the base space! And if this happens for a quadratic extension — as it just did — something seems to have gone horribly wrong with the nice idea that "quadratic extensions are like branched double covers".

Luckily, this puzzle has a nice resolution. We shouldn't have decategorified! When we started discussing "points" for a commutative ring, we saw they form a category in a nice way: there are points of different "flavors", with "flavor-changing operations" going between them. Then we freaked out and turned this category into a set by decreeing that two point are the same whenever there's a morphism between them. If we hadn't done this, we'd have seen more clearly how "inert" primes fit into a nice pattern along with "split" and "ramified" ones.

I'll probably talk about this more sometime, and also look more carefully at what happens to all the different primes when we go to the Gaussian integers — to show you that we are, indeed, doing number theory!

But for now, I just want to make a few comments about this idea of points of different "flavors".

In fact Grothendieck proposed an even more general idea of this sort in his second approach to "schemes", which is simpler but much less widely discussed than his first approach. Basically, he said that given a commutative ring R, we should not only consider points that are homomorphisms from R to any *field*, but also to any *commutative ring*. For each commutative ring k we get a set consisting of all "k-points" of R, namely homomorphisms

$$f\colon R \to k$$

And, for each homomorphism $g\colon k \to k'$ we get a "flavor changing operation" that sends k-points to k'-points. So, we get a functor from CommRing to Set! He called such a functor a "scheme". We can get schemes from commutative rings as just described — these are called "affine schemes" — but there are also others, for example those coming from projective varieties.

Anyway, here are some places to read more about number theory... mostly with an emphasis on the geometric viewpoint and the issue of "splitting, ramification and inertia".

For a really quick and friendly no-nonsense introduction, try this:

2) Harold M. Stark, "Galois Theory, Algebraic Number Theory, and Zeta Function", in *From Number Theory to Physics*, eds. M. Waldschmit et al, Springer, Berlin, 1992, pp. 313–393.

To dig a lot deeper, try this book by Neukirch:

3) Juergen Neukirch, *Algebraic Number Theory*, trans. Norbert Schappacher, Springer, Berlin, 1986.

I already mentioned it, but it's worth mentioning again, because it's pretty elementary, and very clear on the analogy between "function fields" (fields of functions on Riemann surfaces) and "number fields" (algebraic number fields).

This book by Borevich and Shafarevich doesn't make the analogy to geometry explicit:

4) Z. I. Borevich and I. R. Shafarevich, *Number Theory*, trans. Newcomb Greenleaf, Academic Press, New York, 1966.

However, it has a nice concept of a "theory of divisors" for a commutative ring — and if you know a bit about divisors from algebraic geometry, you'll see that this is *secretly* very geometrical! They show how to classify algebraic extensions of commutative rings using a theory of divisors, and show how to get a theory of divisors using "valuations". This manages to accomplish a lot of what other texts do using "adeles", without actually mentioning adeles. I find this instructive.

This book goes much further in the geometric direction, but still without introducing schemes:

5) Dino Lorenzini, *An Invitation to Arithmetic Geometry*, American Mathematical Society, Providence, Rhode Island, 1996.

It's really great — very pedagogical! It develops number fields and function fields in parallel. You'll need to be pretty comfy with commutative algebra to work all the way through it, though.

If you want to learn about schemes — not the kind I just talked about, just the usual sort, which still includes cool "spaces" like $\mathrm{Spec}(Z)$ — try these:

6) V. I. Danilov, V. V. Shokurov, and I. Shafarevich, *Algebraic Curves, Algebraic Manifolds and Schemes*, Springer, Berlin, 1998.

7) David Eisenbud and Joe Harris, *The Geometry of Schemes*, Springer, Berlin, 2000.

Schemes have a reputation for being scary, but both these books try hard to make them less so, including lots of actual *pictures* of things like $\mathrm{Spec}(\mathbb{Z}[i])$ sitting over $\mathrm{Spec}(Z)$.

To wrap things up, I just want to mention two papers on subjects I'm fond of. . . .

In "Week 172" I discussed Tarski's "high school algebra problem". This asks whether every identity involving 1, $+$, $\times$, and exponentials that holds in the positive natural numbers follows from the eleven we learned in high school:

- $x + y = y + x$
- $(x + y) + z = x + (y + z)$
- $xy = yx$
- $(xy)z = x(yz)$
- $1x = x$
- $x^1 = x$
- $1^x = 1$
- $x(y + z) = xy + xz$
- $x^{y+z} = x^y x^z$
- $(xy)^z = x^z y^z$
- $x^{yz} = (x^y)^z$

The rules of this game allow only purely equational reasoning — not stuff like mathematical induction. The reason is that this is secretly a problem about "universal algebra" or "algebraic theories", as explained in "Week 200".

It turns out the answer is *no!* In fact there are infinitely many more independent identities! Here is the first one, due to Wilkies:

$$\begin{aligned}&[(x+1)^x+(x^2+x+1)^x]^y[(x^3+1)^y+(x^4+x^2+1)^y]^x\\ =&[(x+1)^y+(x^2+x+1)^y]^x[(x^3+1)^x+(x^4+x^2+1)^x]^y\end{aligned}$$

I just found a paper, apparently written after "Week 172", which gives a very detailed account of this problem:

8) Stanley Burris and Karen Yeats, "The saga of the high school identities", available at `http://web.archive.org/web/20070212200835/http://www.math.uwaterloo.ca/~snburris/htdocs/MYWORKS/PREPRINTS/saga.ps`

It includes some new results, like the smallest known algebraic gadget satisfying all the high school identities but not Wilkies' identity — but also more interesting things that are a bit harder to describe.

Also, here's a cool paper relating some of Ramanujan's work to string theory:

9) Antun Milas, 'Ramanujan's "Lost Notebook" and the Virasoro Algebra', available as `math.QA/0309201`.

A *lot* of Ramanujan's weird identities turn out to be related to concepts from string theory, suggesting that he was born about a century too soon to be fully appreciated... but this paper tackles an identity of his that nobody had managed to explain using string theory before.

Addendum: Here's something a friend of mine wrote, and an expanded version of my reply.

> *By the way, I very much liked your explanation of points and prime ideals. Up until now I haven't seen a satisfactory explanation of why points correspond to* prime *rather than* maximal *ideals, and although I haven't completely digested what you wrote, it looks like it might do the job...*

Both here and in my discussion of spectra in "Week 199", I've been avoiding saying the things people usually say. People usually note that a maximal ideal is the same as the kernel of a homomorphism ONTO a field, while a prime ideal is the same as the kernel of a homomorphism ONTO an integral domain. (Recall that an integral domain is a commutative ring where $xy = 0$ implies that x or y is zero.) If we define the "points" of a commutative ring R to be its maximal or prime ideals, we can therefore think of these as the kernels of homomorphisms from R onto fields or integral domains.

However, defining points in terms of homomorphisms ONTO a given sort of commutative ring is rather irksome, because it doesn't tell us how points transform under homomorphisms of commutative rings, nor how they transform under the "flavor-changing operations" I was describing. The problem is that the composite of a homomorphism with an onto homomorphism needn't be onto!

So, what really matters is that a prime ideal is the same as the kernel of a homomorphism TO a field. To see how this follows from the usual story, note that any integral domain is contained in a field called its "field of fractions" — just as $\mathbb{Z}$ is contained in $\mathbb{Q}$. Any homomorphism ONTO the integral domain thus becomes a homomorphism TO this field, with the same kernel. Conversely, any homorphism TO a field becomes a homomorphism ONTO its image, with the same kernel — and this image is always an integral domain.

Much later, in 2013, someone wrote:

> *Dear Professor Baez,*
>
> *In TWF 205 you discuss the puzzle of inert primes and indicate that this puzzle has a nice resolution in terms of decategorification. You also indicated then that you may talk further about this at a later date. By any chance did you return to this topic and if so might you be able to point me to the appropriate TWF (or other reference)?*
>
> *My apologies for asking about a nine year old remark.*

I replied:

Sorry, I never wrote more about that. I sort of forget what I was talking about, because it comes from a time long ago when I was discussing number theory with James Dolan, and he understood this stuff much better than me. But I think I can remember a bit if I try...

> *[...] you discuss the puzzle of inert primes and indicate that this puzzle has a nice resolution in terms of decategorification.*

I remember that the resolution of the puzzle is to *refrain* from decategorifying.

In Grothendieck's approach, the "k-points" of a ring R of algebraic numbers, for various choices of field k, are homomorphisms

$$R \to k$$

These are the objects of a category, in the manner I described, with "flavor-changing operations" coming from homomorphisms $k \to k'$ as the morphisms.

If we restrict our attention to the case I was talking about, where $R = \mathbb{Z}[i]$ is the Gaussian integers and k is a finite field, the "inert" primes correspond to situations where a k-point down in $\mathbb{Z}$ lifts to two different but isomorphic k-points in $\mathbb{Z}[i]$. When we decategorify, as we do in the usual approach, it looks like there's just one point in $\mathbb{Z}[i]$ sitting over this point in $\mathbb{Z}$.

Let me warm up by considering the prime 5, which is not inert. The ideal in $\mathbb{Z}$ generated by 5 splits into two ideals in $\mathbb{Z}[i]$, $\langle 2-i\rangle$ and $\langle 2+i\rangle$. Correspondingly, there are two k-points

$$\mathbb{Z}[i] \to \mathbb{F}_5$$

one having $\langle 2-i\rangle$ as its kernel (it sends i to 2), and having $\langle 2+i\rangle$ as its kernel (it sends i to -2). These k-points are not isomorphic.

On the other hand, 3 is inert. The ideal in $\mathbb{Z}$ generated by 3 does not split in $\mathbb{Z}[i]$. Now we can find a field $k = \mathbb{F}_9$ such that $\mathbb{Z}[i]$ has two different k-points with the ideal $\langle 3\rangle$ as kernel. One k-point

$$\mathbb{Z}[i] \to \mathbb{F}_9$$

sends i to one of the square roots of -1 in $\mathbb{F}_9$, the other sends it to the other. (We have to go up to $\mathbb{F}_9$ since there's no square root of -1 in $\mathbb{F}_3$.) However, these two k-points are isomorphic, since there's an automorphism of $\mathbb{F}_9$ interchanging its two square roots of -1.

Moral: When we work with prime ideals as points instead of homomorphisms whose kernels are these prime ideals, it looks like there's just one point in $\mathrm{Spec}(\mathbb{Z}[i])$ sitting over the prime ideal $\langle 3\rangle$. But when we define points to be homomorphisms, we see $\mathbb{Z}[i]$ has two isomorphic points $\mathbb{Z}[i] \to \mathbb{F}_9$ sitting over the unique point $\mathbb{Z} \to \mathbb{F}_9$ whose kernel is that prime ideal.

> One discovery opens another, and then another. Everything in this country is nested like Russian dolls. Even a solid artifact in front of me drew back into other levels. Schemes within schemes.
>
> — *Craig Childs, Soul of Nowhere*

Week 206

May 10, 2004

I just got back from Marseille, where Carlo Rovelli, Laurent Freidel and Phillipe Roche held the first really big conference on loop quantum gravity and spin foams since the 2nd Warsaw workshop run by Jerzy Lewandowski back in 1997:

1) *Non Perturbative Quantum Gravity: Loops and Spin Foams*, 3–7 May 2004, CIRM, Luminy, Marseille, France, `http://w3.lpm.univ-montp2.fr/~philippe/quantumgravitywebsite/`

It was good to see old friends and talk about quantum gravity near the "Calanques" — the rugged limestone cliffs lining the Mediterranean coastline. It was good to meet lots of young people who have recently entered this difficult field: about 100 people attended, considerably more than at any previous meeting. But most of all, it was good to see some progress on the tough problem of understanding dynamics in nonperturbative quantum gravity.

Can we get the 4-dimensional spacetime we know and love, whose geometry is described by general relativity, to emerge from some theory that takes quantum physics into account? And can we do it *nonperturbatively?*

In other words, can we do quantum physics without choosing some fixed spacetime geometry from the start, a "background" on which small perturbations move like tiny quantum ripples on a calm pre-established lake? A background geometry is convenient: it lets us keep track of times and distances. It's like having a fixed stage on which the actors — gravitons, strings, branes, or whatever — cavort and dance. But, the main lesson of general relativity is that spacetime is *not* a fixed stage: it's a lively, dynamical entity! There's no good way to separate the ripples from the lake. This distinction is no more than a convenient approximation — and a dangerous one at that.

So, we should learn to make do without a background when studying quantum gravity. But it's tough! There are knotty conceptual issues like the "problem of time": how do we describe time evolution without using a fixed background to measure the passage of time? There are also practical problems: in most attempts to describe spacetime from the ground up in a quantum way, all hell breaks loose!

We can easily get spacetimes that crumple up into a tiny blob... or spacetimes that form endlessly branching fractal "polymers" of Hausdorff dimension 2... but it seems hard to get reasonably smooth spacetimes of dimension 4. It's even hard to get spacetimes of dimension 10 or 11... or *anything* remotely interesting!

It almost seems as if we need a solid background as a bed frame to keep the mattress of spacetime from rolling up, getting all lumpy, or otherwise misbehaving. Unfortunately, even *with* a background there are serious problems: we can use perturbation theory to write the answers to physics questions as power series, but these series diverge and nobody knows how to resum them.

String theorists are pragmatic in a certain sense: they don't mind using a background, and they don't mind doing what physicists always do: approximating a divergent series by the sum of the first couple of terms. But this attitude doesn't solve everything, because right now in string theory there is an enormous "landscape" of different backgrounds, with no firm principle for choosing one. Some estimates guess there are over 10^{100}. Leonard Susskind guesses there are 10^{500}, and argues that we'll need the anthropic principle to choose the one describing our world:

2) Leonard Susskind, "The Landscape", article and interview on John Brockman's "EDGE" website, `http://www.edge.org/3rd_culture/susskind03/susskind_index.html`

This position is highly controversial, but my point here shouldn't be: developing a background-free theory of quantum gravity is tough, but working *with* a background has its own difficulties. And let's face it: we haven't spent nearly as much time thinking about background-free or nonperturbative physics as we've spent on background-dependent or perturbative physics. So, it's quite possible that our failures with the former are just a matter of inexperience.

Given all this, I'm delighted to see some real progress on getting 4d spacetime to emerge from nonperturbative quantum gravity:

3) Jan Ambjorn, Jerzy Jurkiewicz and Renate Loll, "Emergence of a 4d world from causal quantum gravity", available as `hep-th/0404156`.

This trio of researchers have revitalized an approach called "dynamical triangulations" where we calculate path integrals in quantum gravity by summing over different ways of building spacetime out of little 4-simplices. They showed that if we restrict this sum to spacetimes with a well-behaved concept of causality, we get good results. This is a bit startling, because after decades of work, most researchers had despaired of getting general relativity to emerge at large distances starting from the dynamical triangulations approach. But, these people hadn't noticed a certain flaw in the approach... a flaw which Loll and collaborators noticed and fixed!

If you don't know what a path integral is, don't worry: it's pretty simple. Basically, in quantum physics we can calculate the expected value of any physical quantity by doing an average over all possible histories of the system in question, with each history weighted by a complex number called its "amplitude". For a particle, a history is just a path in space; to average over all histories is to integrate over all paths — hence the term "path integral". But in quantum gravity, a history is nothing other than a SPACETIME.

Mathematically, a "spacetime" is something like a 4-dimensional manifold equipped with a Lorentzian metric. But it's hard to integrate over all of these — there are just too darn many. So, sometimes people instead treat spacetime as made of little discrete building blocks, turning the path integral into a sum. You can either take this seriously or treat it as a kind of approximation. Luckily, the calculations work the same either way!

If you're looking to build spacetime out of some sort of discrete building block, a handy candidate is the "4-simplex": the 4-dimensional analogue of a tetrahedron. This shape is rigid once you fix the lengths of its 10 edges, which correspond to the 10 components of the metric tensor in general relativity.

There are lots of approaches to the path integrals in quantum gravity that start by chopping spacetime into 4-simplices. The weird special thing about dynamical triangulations is that here we usually assume every 4-simplex in spacetime has the same shape.

The different spacetimes arise solely from different ways of sticking the 4-simplices together.

Why such a drastic simplifying assumption? To make calculations quick and easy! The goal is get models where you can simulate quantum geometry on your laptop — or at least a supercomputer. The hope is that simplifying assumptions about physics at the Planck scale will wash out and not make much difference on large length scales.

Computations using the so-called "renormalization group flow" suggest that this hope is true *if* the path integral is dominated by spacetimes that look, when viewed from afar, almost like 4d manifolds with smooth metrics. Given this, it seems we're bound to get general relativity at large distance scales — perhaps with a nonzero cosmological constant, and perhaps including various forms of matter.

Unfortunately, in all previous dynamical triangulation models, the path integral was *not* dominated by spacetimes that look like nice 4d manifolds from afar! Depending on the details, one either got a "crumpled phase" dominated by spacetimes where almost all the 4-simplices touch each other, or a "branched polymer phase" dominated by spacetimes where the 4-simplices form treelike structures. There's a transition between these two phases, but unfortunately it seems to be a 1st-order phase transition — not the sort we can get anything useful out of. For a nice review of these calculations, see:

4) Renate Loll, "Discrete approaches to quantum gravity in four dimensions", available as gr-qc/9805049 or as a website at *Living Reviews in Relativity*, `http://www.livingreviews.org/Articles/Volume1/1998-13loll/`

Luckily, all these calculations shared a common flaw!

Computer calculations of path integrals become a lot easier if instead of assigning a complex "amplitude" to each history, we assign it a positive real number: a "relative probability". The basic reason is that unlike positive real numbers, complex numbers can cancel out when you sum them!

When we have relative probabilities, it's the *highly probable* histories that contribute most to the expected value of any physical quantity. We can use something called the "Metropolis algorithm" to spot these highly probable histories and spend most of our time focusing on them.

This doesn't work when we have complex amplitudes, since even a history with a big amplitude can be canceled out by a nearby history with the opposite big amplitude! Indeed, this happens all the time. So, instead of histories with big amplitudes, it's the *bunches of histories that happen not to completely cancel out* that really matter. Nobody knows an efficient general-purpose algorithm to deal with this!

For this reason, physicists often use a trick called "Wick rotation" that converts amplitudes to relative probabilities. To do this trick, we just replace time by imaginary time! In other words, wherever we see the variable "t" for time in any formula, we replace it by "it". Magically, this often does the job: our amplitudes turn into relative probabilities! We then go ahead and calculate stuff. Then we take this stuff and go back and replace "it" everywhere by "t" to get our final answers.

While the deep inner meaning of this trick is mysterious, it can be justified in a wide variety of contexts using the "Osterwalder-Schrader theorem". Here's a pretty general version of this theorem, suitable for quantum gravity:

5) Abhay Ashtekar, Donald Marolf, Jose Mourao and Thomas Thiemann, "Constructing Hamiltonian quantum theories from path integrals in a diffeomorphism invariant context", *Class. Quant. Grav.* **17** (2000) 4919–4940. Also available as quant-ph/9904094.

People use Wick rotation in all work on dynamical triangulations. Unfortunately, this is *not* a context where you can justify this trick by appealing to the Osterwalder-Schrader theorem. The problem is that there's no good notion of a time coordinate "t" on your typical spacetime built by sticking together a bunch of 4-simplices!

The new work by Ambjorn, Jurkiewiecz and Loll deals with this by restricting to spacetimes that *do* have a time coordinate. More precisely, they fix a 3-dimensional manifold and consider all possible triangulations of this manifold by regular tetrahedra. These are the allowed "slices" of spacetime — they represent different possible geometries of space at a given time. They then consider spacetimes having slices of this form joined together by 4-simplices in a few simple ways.

The slicing gives a preferred time parameter "t". On the one hand this goes against our desire in general relativity to avoid a preferred time coordinate — but on the other hand, it allows Wick rotation. So, they can use the Metropolis algorithm to compute things to their hearts' content and then replace "it" by "t" at the end.

When they do this, they get convincing good evidence that the spacetimes which dominate the path integral look approximately like nice smooth 4-dimensional manifolds at large distances! Take a look at their graphs and pictures — a picture is worth a thousand words.

Naturally, what *I'd* like to do is use their work to develop some spin foam models with better physical behavior than the ones we have so far. If you look at my talk you can see some of the problems we've encountered:

6) John Baez, "Spin foam models", talk at *Non Perturbative Quantum Gravity: Loops and Spin Foams*, May 4, 2004, transparencies available at `http://math.ucr.edu/home/baez/spin_foam_models/`

Now that Loll and her collaborators have gotten something that works, we can try to fiddle around and make it more elegant while making sure it still works. In particular, I'm hoping we can get well-behaved models that don't introduce a preferred time coordinate as long as they rule out "topology change" — that is, slicings where the topology of space changes. After all, the Osterwalder-Schrader theorem doesn't require a *preferred* time coordinate, just *any* time coordinate together with good behavior under change of time coordinate. For this we mainly need to rule out topology change. Moreover, Loll and her collaborators have argued in 2d toy models that topology change is one thing that makes models go bad: the path integral can get dominated by spacetimes where "baby universes" keep branching off the main one:

7) Jan Ambjorn, Jerzy Jurkiewicz and Renate Loll, "Non-perturbative Lorentzian quantum gravity, causality and topology change", *Nucl. Phys.* **B536** (1998) 407–434. Also available as hep-th/9805108.

   Renate Loll and W. Westra, "Space-time foam in 2d and the sum over topologies", *Acta Phys. Polon.* **B34** (2003) 4997–5008. Also available as `hep-th/0309012]`(https://arxiv.org/abs/hep-th/0309012).

By the way, it's also reading about their 3d model:

8) Jan Ambjorn, Jerzy Jurkiewicz and Renate Loll, "Non-perturbative 3d Lorentzian quantum gravity", *Phys. Rev.* **D64** (2001) 044011. Also available as hep-th/0011276.

and for a general review, try this:

9) Renate Loll, "A discrete history of the Lorentzian path integral", Lecture Notes in Physics **631**, Springer, Berlin, 2003, pp. 137–171. Also available as hep-th/0212340.

All this is great, but don't get me wrong — there were a lot of *other* cool talks at the conference besides Loll's. I'll just mention a few.

Laurent Freidel spoke on his work on spin foam models. Especially exciting is how David Louapre and he have managed to "sum over topologies" in 3d Riemannian quantum gravity with vanishing cosmological constant — otherwise known as the Ponzano-Regge model He has to subtract out a counterterm that would otherwise lead to a bubble divergence, but then he gets a beautiful theory where the sum over spin foams is Borel summable:

10) Laurent Freidel and David Louapre, "Non-perturbative summation over 3D discrete topologies", *Phys. Rev.* **D68** (2003) 104004. Also available as hep-th/0211026.

Their work on gauge-fixing and the inclusion of spinning point particles in the Ponzano-Regge model is also very impressive, especially given how long this model has been studied. It shows we have lots left to learn!

11) Laurent Freidel and David Louapre, "Ponzano-Regge model revisited I: Gauge fixing, observables and interacting spinning particles", available as hep-th/0401076.

The title suggests we're in for more treats to come.

Kirill Krasnov gave a talk entitled simple "$\ln(3)$" — it was all about the appearance of this constant in the work of Hod, Dreyer, Motl and Neitzke on black hole entropy and the ringing of black holes. I've discussed all this at length in "Week 198", but Krasnov has given an elegant new proof of Hod's conjecture using Riemann surface theory. One can even think of this as a "stringy" explanation of the quasinormal modes of black holes — but much remains mysterious here:

12) Kirill Krasnov, "Black hole thermodynamics and Riemann surfaces", *Class. Quant. Grav.* **20** (2003) 2235–2250. Also available as gr-qc/0302073.

Kirill Krasnov and Sergey N. Solodukhin, "Effective stringy description of Schwarzschild black holes", available as hep-th/0403046.

While I'm at it, I can't resist mentioning Krasnov's work on including point particles in 3d Lorentzian quantum gravity with negative cosmological constant, since it has close connections with that of Freidel and Louapre, though the context is a bit different:

13) Kirill Krasnov, "$\Lambda < 0$ quantum gravity in 2+1 dimensions I: quantum states and stringy S-matrix", *Class. Quant. Grav.* **19** (2002) 3977–3998. Also available as hep-th/0112164.

    Kirill Krasnov, "$\Lambda < 0$ quantum gravity in 2+1 dimensions II: black hole creation by point particles", *Class. Quant. Grav.* **19** (2002) 3999–4028. Also available as hep-th/0202117.

If I could duplicate myself, I'd have one copy write a book on 3d quantum gravity that would synthesize all these wonderful results in a nice big picture. It's not realistic physics; it's just a toy model. But the math is *so* nice, and so enlightening for real-world physics in some ways, that it's hard to resist pondering it! TQFTs, Riemann surfaces, hyperbolic geometry, spinning point particles colliding and creating black holes — a wonderful stew! Alas, I don't have time to savor it.

There were a lot of other interesting talks — but I don't have time to go through and describe all of them, either. So, I'll wrap up with something very different!

Lee Smolin told me some neat stuff about MOND — that's "Modified Newtonian Dynamics", which is Mordehai Milgrom's way of trying to explain the strange behavior of galaxies without invoking dark matter. The basic problem with galaxies is that the outer parts rotate faster than they should given how much mass we actually see.

If you have a planet in a circular orbit about the Sun, Newton's laws say its acceleration is proportional to $1/r^2$, where r is its distance to the Sun. Similarly, if almost all the mass in a galaxy were concentrated right at the center, a star orbiting in a circle at distance r from the center would have acceleration proportional to $1/r^2$. Of course, not all the mass is right at the center! So, the acceleration should drop off more slowly than $1/r^2$ as you go further out. And it does. But, the observed acceleration drops off a lot more slowly than the acceleration people calculate from the mass they see. It's not a small effect: it's a *huge* effect!

One solution is to say there's a lot of mass we don't see: "dark matter" of some sort. If you take this route, which most astronomers do, you're forced to say that *most* of the mass of galaxies is in the form of dark matter.

Milgrom's solution is to say that Newton's laws are messed up.

Of course this is a drastic, dangerous step: the last guy who tried this was named Einstein, and we all know what happened to him. Milgrom's theory isn't even based on deep reasoning and beautiful math like Einstein's! Instead, it's just a blatant attempt to fit the experimental data. And it's not even elegant. In fact, it's downright ugly.

Here's what it says: the usual Newtonian formula for the acceleration due to gravity is correct as long as the acceleration is bigger than

$$a = 2 \times 10^{-10} \text{ m/sec}^2$$

But, for accelerations less than this, you take the geometric mean of the acceleration Newton would predict and this constant a.

In other words, there's a certain value of acceleration such that above this value, the Newtonian law of gravity works as usual, while below this value the law suddenly changes.

Any physicist worth his salt who hears this modification of Newton's law should be overcome with a feeling of revulsion! There just *aren't* laws of physics that split a situation in two cases and say "if this is bigger than that, then do X, but if it's smaller, then do Y." Not in fundamental physics, anyway! Sure, water is solid below 0 centigrade and fluid above this, but that's not a fundamental law — it presumably follows from other stuff. Not that anyone has derived the melting point of ice from first principles, mind you. But we think we could if we were better at big messy calculations.

Furthermore, you can't easily invent a Lagrangian for gravity that makes it fall off more *slowly* than $1/r^2$. It's easy to get it to fall off *faster* — just give the graviton a mass, for example! But not more slowly. It turns out you can do it — Bekenstein and Milgrom have a way — but it's incredibly ugly.

So, MOND should instantly make any decent physicist cringe. Esthetics alone would be enough to rule it out, except for one slight problem: it seems to fit the data! In some cases it matches the observed rotation of galaxies in an appallingly accurate way, fitting every wiggle in the graph of stellar rotation velocity as a function of distance from the center.

So, even if MOND is wrong, there may need to be some reason why it *acts* like it's right! Apparently even some proponents of dark matter agree with this.

But: take everything I'm saying here with a grain of salt. I'm no expert on this stuff, so if you know any astrophysics you should read the literature and make up your own mind.

Here are two reviews that Smolin especially recommended:

14) Robert H. Sanders and Stacy S. McGaugh, "Modified Newtonian Dynamics as an Alternative to Dark Matter", available as astro-ph/0204521.

15) Anthony Aguirre, "Alternatives to dark matter (?)", available as astro-ph/0310572.

Here's McGaugh's website with links to many papers on MOND, including Milgrom's original papers:

16) "The MOND pages", `http://www.astro.umd.edu/~ssm/mond/litsub.html`

McGaugh is a strong proponent of MOND — though he didn't start out that way — so the selection may be biased. Does anyone know an intelligent detailed critique of MOND? If so, I want to see it! We can't throw out Newton's law of gravity (or more precisely, general relativity, which has Newtonian gravity as a limiting case for low densities and low velocities) unless we have *very* good reasons! So we have to think about things carefully, and weigh the evidence on both sides.

If I could duplicate myself, I'd have one copy try to get to the bottom of this dark matter / MOND puzzle. But I can't...

... so if you're an expert who knows a lot about this, let me know what you think — or better yet, post an article about this to `sci.physics.research`!

By the way, you can see lots of photos of the Marseille conference here:

17) John Baez, Marseille, `http://math.ucr.edu/home/baez/marseille/`

Almost everyone working on loop quantum gravity and spin foams can be seen here!

Addendum: A few people took me up on my request.

Steve Carlip wrote:

> *John Baez wrote:*
>
> > *So, even if MOND is wrong, there may need to be some reason why it acts like it's right! Apparently even some proponents of dark matter agree with this.*
>
> *Try this:*
>
> *M. Kaplinghat and M. S. Turner, "How Cold Dark Matter Theory Explains Milgrom's Law," astro-ph/0107284,* Astrophys. J. ***569*** *(2002) L19. Note that this analysis also explains why the "critical acceleration" in MOND does* not *apply at cluster scales. There is some debate over these results, but the paper is certainly worth reading.*
>
> *Steve Carlip*

Rein Halbersma wrote:

> *Dear John Baez,*
>
> *Your writings in Week 206 brought back some vivid memories from the good old days in graduate school with all-night philosophical discussions! In your Finds in Week 206 you discuss the MOND-framework of Milgrom and asked for detailed critique of it. A few years ago the authors Scott, White, Cohn and Pierpaoli (astro-ph/0104435) published precisely such an account. Hopefully it is of use to you.*
>
> *As an aside, my connection with the whole MOND story is this: I have a PhD in high-energy physics from Groningen University (my advisor was Eric Bergshoeff, one of the inventors of the supermembrane). While I was working as a graduate student in string theory & conformal supergravity, a roommate of mine, Roland Eppinga, was an undergraduate student for Robert Sanders, who is an astronomy professor in Groningen. My friend was assigned a project involving cosmological simulations within the MOND-framework. Needless to say, we had many discussions on MOND in which my esthetical views of general relativity were put to the test by the need to fit a damn rotational curve.*
>
> *My personal view is that MOND is indeed too ugly to be true. Or as Einstein would have said, if Nature is not described by relativity, then God designed it badly!*
>
> *Best wishes,*
>
> *Rein Halbersma*

Christine Dantas wrote:

> *Hello all,*
>
> *Concerning MOND x GR, the recent paper by Bekenstein seems to be a relevant contribution to this issue (see below).*
>
> *Regards,*
> *Christine Dantas*

> *INPE/Brazil*
>
> `astro-ph/0403694`
> *Relativistic gravitation theory for the MOND paradigm*
> *Jacob D. Bekenstein*
>
> *The modified newtonian dynamics (MOND) paradigm of Milgrom can boast of a number of successful predictions regarding galactic dynamics; these are made without the assumption that dark matter plays a significant role. MOND requires gravitation to depart from Newtonian theory in the extragalactic regime where dynamical accelerations are small. So far relativistic gravitation theories proposed to underpin MOND have either clashed with the post-Newtonian tests of general relativity, or failed to provide significant gravitational lensing, or violated hallowed principles by exhibiting superluminal scalar waves or an a priori vector field. We develop a relativistic MOND inspired theory which resolves these problems. In it gravitation is mediated by metric, a scalar field and a 4-vector field, all three dynamical. For a simple choice of its free function, the theory has a Newtonian limit for nonrelativistic dynamics with significant acceleration, but a MOND limit when accelerations are small. We calculate the β and γ PPN coefficients showing them to agree with solar system measurements. The gravitational light deflection by nonrelativistic systems is governed by the same potential responsible for dynamics of particles. Consequently, the new theory predicts gravitational lensing by extragalactic structures that cannot be distinguished from that predicted within the dark matter paradigm by general relativity. Cosmological models based on the theory are quite similar to those based on general relativity; they predict slow evolution of the scalar field. For a range of initial conditions, this last result makes it easy to rule out superluminal propagation of metric, scalar and vector waves.*

Ethan Vishniac wrote:

> *Hi,*
>
> *I don't have a reference for a skeptical review of MOND. As you might expect, this is considered a fringe hypothesis by most. However, there is an interesting paper you should see:* `astro-ph/0312273`.
>
> *Briefly, they examine a galaxy cluster with a strong sub-cluster, which has just passed through the main cluster for the first time (probably). Most of the baryonic mass is in the hot gas (by a factor of ten) so the initial pass has stripped the gas out of the sub-cluster. In fact, in the X-rays the subcluster is not evident. If stars and gas are all there is then there is no significant mass concentration associated with the sub-cluster. However, the sub-cluster is quite easy to see in the gravitational lensing map. Evidently, the mass of the sub-cluster has not been significantly reduced by losing all of the gas. (That is, the mass to light ratio for the sub-cluster is what one would expect for an isolated system.) This looks like a simple demonstration that most of the mass in galaxy clusters is non-luminous and dissipationless. There have also been attempts to disprove*

> *MOND by comparing time delays in strong lensing systems with MOND based models. Unfortunately, the real problem here is that there is no clear set of predictions for MOND.*
>
> *Ethan Vishniac*

He also wrote:

> *BTW, one way to address MOND on its own terms is to try to follow galactic rotation curves out to very very great distances. If the dark matter model is correct, they will eventually turn over and fall as* $r^{-1/2}$. *This is hard, perhaps impossible, using gas. There is some work using the velocity dispersion of satellite galaxies around otherwise isolated bright galaxies (Prada et al.,* ApJ ***598**, 260–,2003). (The Sloan Digital Sky Survey makes it possible to get good statistics for very weak signals.) They claim to have detected a drop in the velocity dispersion by a factor of 2 between 20 and 350 kpc. This is roughly in line with expectations from cosmological simulations, and stands in contradiction to what one would expect from MOND.*
>
> *Finally, one can try to measure the shape of galaxy halos using weak lensing. The line of reasoning is a bit indirect, but the point is that an elliptical or disklike distribution of mass at small radii gives rise to nearly spherical equipotential surfaces at large radii. On the other hand, dark matter halos are generally triaxial, and will appear elliptical on the sky. Hoekstra et al. (*Astrophysical Journal*, **606**, 67–77, 2004) have done this and claim a strong elliptical signal in the weak lensing data.*

Finally, Renate Loll corrected an oversimplification in my account of her model:

> *[. . .] I never claimed the geometries we find are nice* and smooth*, I think they almost certainly will be fairly wild individually, even at relatively large scales. Like the particle paths in the quantum mechanical path integral, the individual histories should not be taken too literally, the physics'll all be in suitable expectation values of course.*

When I am working on a problem, I never think about beauty. I think only about how to solve the problem. But when I have finished, if the solution is not beautiful, I know it is wrong.

— *Buckminister Fuller*

Week 207

July 25, 2004

I'm spending the summer in Cambridge, but last week I was in Dublin attending "GR17", which is short for the 17th International Conference on General Relativity and Gravitation:

1) GR17 homepage, `http://www.dcu.ie/~nolanb/gr17.htm`

This is where Stephen Hawking decided to announce his solution of the black hole information loss problem. Hawking is a media superstar right up there with Einstein and Michael Jackson, so when reporters heard about this, the ensuing hoopla overshadowed everything else in the conference.

As soon I arrived, one of the organizers complained to me that they'd had to spend 4000 pounds on a public relations firm to control the reporters and other riff-raff who would try to attend Hawking's talk. Indeed, there seemed to be more than the usual number of crackpots floating about, though I admit I haven't been to this particular series of conferences before — perhaps general relativity attracts such people? The public lecture by Penrose on the last day of the conference may have helped lure them in. He spoke on "Fashion, Faith and Fantasy in Theoretical Physics", and people by the door sold copies of his brand new thousand-page blockbuster:

2) Roger Penrose, The Road To Reality: A Complete Guide to the Physical Universe, Jonathan Cape, 2004.

(You may enjoy guessing which popular theories he classified under the three categories of fashion, faith and fantasy.) After his talk, *all* the questions were actually harangues from people propounding idiosyncratic theories of their own, and the question period was drawn to an abrupt halt in the middle of one woman's rant about fractal cosmology. But I bumped into the saddest example when I was having a chat with some colleagues at a local pub. A fellow with long curly grey locks and round horn-rimmed glasses sat down beside me. I'd seen him around the conference, so I said hello. He asked me if I'd like to hear about his theory; at this point my internal alarm bells started ringing. I told him I was busy, but said I'd take a look at his manuscript later.

It turned out to describe an idea I'd never even dreamt of before: a heliocentric cosmology in which the planets move along circular orbits with epicycles a la Ptolemy! And his evidence comes from a neolithic Irish tomb called Newgrange. This tomb may have been aligned to let in the sun on the winter solstice, but some people doubt this, because it seems the alignment would have been slightly off back in 3200 BC when Newgrange was built. However, it's slightly off only if you work out the precession of the equinox using standard astronomy. If you use his theory, it lines up perfectly! Pretty cute. The only problem is that his paper contains no evidence for this claim. Instead, it's only a short note sketching the idea, followed by lengthy attachments containing his correspondence with the Dublin police. In these, he complained that people were trying to block his patent on a refrigerator that produces no waste heat. They were constantly flying airplanes over his house, and playing pranks like boiling water in his teakettle when he was away, trying to drive him insane.

Anyway, on Wednesday the 21st the whole situation built to a head when Hawking gave his talk in the grand concert hall of the Royal Dublin Society. As we had been warned, the PR firm checked our badges at the door. Reporters with press badges were also allowed in, so the aisles were soon lined with cameras and recording equipment. I got there half an hour early to get a good seat, and while I was waiting, Jenny Hogan from the New Scientist asked if she could interview me for my reaction afterwards. In short, a thoroughly atypical physics talk!

But you shouldn't imagine the mood as one of breathless anticipation. At least for the physicists present, a better description would be something like "skeptical curiosity". None of them seemed to believe that Hawking could suddenly shed new light on a problem that has been attacked from many angles for several decades. One reason is that Hawking's best work was done almost 30 years ago. A string theorist I know said that thanks to work relating anti-deSitter space and conformal field theory — the so-called "AdS-CFT" hypothesis — string theorists had become convinced that no information is lost by black holes. Thus, Hawking had been feeling strong pressure to fall in line and renounce his previous position, namely that information *is* lost. A talk announcing this would come as no big surprise.

After a while Kip Thorne, John Preskill, Petros Florides and Hawking's grad student Christophe Galfard came on stage. Then, amid a burst of flashbulbs, Hawking's wheelchair gradually made its way down the aisle and up a ramp, attended by a nurse — possibly his wife, I don't know. He had been recently sick with pneumonia.

Once Hawking was on stage, the conference organizer Petros Florides made an introduction, joking that while physicists believe no information can travel faster than light, this seems to have been contradicted by the speed with which the announcement of Hawking's talk spread around the globe. Then he recalled the famous bet that Preskill made with Hawking and Thorne. In case you don't know, John Preskill is a leader in quantum computation at Caltech. Kip Thorne is an expert on relativity, also at Caltech, one of the authors of the famous textbook "Gravitation", and now playing a key role in the LIGO project to detect gravitational waves.

The bet went like this:

> *Whereas Stephen Hawking and Kip Thorne firmly believe that information swallowed by a black hole is forever hidden from the outside universe, and can never be revealed even as the black hole evaporates and completely disappears,*
>
> *And whereas John Preskill firmly believes that a mechanism for the information to be released by the evaporating black hole must and will be found in the correct theory of quantum gravity,*
>
> *Therefore Preskill offers, and Hawking/Thorne accept, a wager that:*
>
> > *When an initial pure quantum state undergoes gravitational collapse to form a black hole, the final state at the end of black hole evaporation will always be a pure quantum state.*
>
> *The loser(s) will reward the winner(s) with an encyclopedia of the winner's choice, from which information can be recovered at will.*
>
> *Stephen W. Hawking, Kip S. Thorne, John P. Preskill*
> *Pasadena, California, 6 February 1997*

It's signed by Thorne and Preskill, with a thumbprint of Hawking's.

After a bit of joking around and an explanation of how the question session would work, Hawking began his talk. Since it's fairly short and not too easy to summarize, I think I'll just quote the whole transcript which I believe Sean Carroll got from the New York Times science reporter Dennis Overbye. I've made a few small corrections.

There were also some slides, but you're not missing a lot by not seeing them. The talk was not easy to understand, so unless quantum gravity is your speciality you may feel like just skimming it to get the flavor, and then reading my attempt at a summary.

The talk began with Hawking's trademark introduction, uttered as usual in his computer-generated voice:

> *Can you hear me?*
>
> *I want to report that I think I have solved a major problem in theoretical physics, that has been around since I discovered that black holes radiate thermally, thirty years ago. The question is, is information lost in black hole evaporation? If it is, the evolution is not unitary, and pure quantum states, decay into mixed states.*
>
> *I'm grateful to my graduate student Christophe Galfard for help in preparing this talk.*
>
> *The black hole information paradox started in 1967, when Werner Israel showed that the Schwarzschild metric, was the only static vacuum black hole solution. This was then generalized to the no hair theorem: the only stationary rotating black hole solutions of the Einstein-Maxwell equations are the Kerr-Newman metrics. The no hair theorem implied that all information about the collapsing body was lost from the outside region apart from three conserved quantities: the mass, the angular momentum, and the electric charge.*
>
> *This loss of information wasn't a problem in the classical theory. A classical black hole would last for ever, and the information could be thought of as preserved inside it, but just not very accessible. However, the situation changed when I discovered that quantum effects would cause a black hole to radiate at a steady rate. At least in the approximation I was using, the radiation from the black hole would be completely thermal, and would carry no information. So what would happen to all that information locked inside a black hole, that evaporated away, and disappeared completely? It seemed the only way the information could come out would be if the radiation was not exactly thermal, but had subtle correlations. No one has found a mechanism to produce correlations, but most physicists believe one must exist. If information were lost in black holes, pure quantum states would decay into mixed states, and quantum gravity wouldn't be unitary.*
>
> *I first raised the question of information loss in '75, and the argument continued for years, without any resolution either way. Finally, it was claimed that the issue was settled in favour of conservation of information, by AdS/CFT. AdS/CFT is a conjectured duality between supergravity in anti-deSitter space and a conformal field theory on the boundary of anti-deSitter space at infinity. Since the conformal field theory is manifestly unitary, the argument is that supergravity must be information preserving. Any information that falls in a black hole in*

anti-deSitter space, must come out again. But it still wasn't clear how information could get out of a black hole. It is this question I will address.

Black hole formation and evaporation can be thought of as a scattering process. One sends in particles and radiation from infinity, and measures what comes back out to infinity. All measurements are made at infinity, where fields are weak, and one never probes the strong field region in the middle. So one can't be sure a black hole forms, no matter how certain it might be in classical theory. I shall show that this possibility allows information to be preserved and to be returned to infinity.

I adopt the Euclidean approach, the only sane way to do quantum gravity non-perturbatively. [He grinned at this point.] In this, the time evolution of an initial state is given by a path integral over all positive definite metrics that go between two surfaces that are a distance T apart at infinity. One then Wick rotates the time interval, T, to the Lorentzian.

The path integral is taken over metrics of all possible topologies that fit in between the surfaces. There is the trivial topology: the initial surface cross the time interval. Then there are the nontrivial topologies: all the other possible topologies. The trivial topology can be foliated by a family of surfaces of constant time. The path integral over all metrics with trivial topology, can be treated canonically by time slicing. In other words, the time evolution (including gravity) will be generated by a Hamiltonian. This will give a unitary mapping from the initial surface to the final.

The nontrivial topologies cannot be foliated by a family of surfaces of constant time. There will be a fixed point in any time evolution vector field on a nontrivial topology. A fixed point in the Euclidean regime corresponds to a horizon in the Lorentzian. A small change in the state on the initial surface would propagate as a linear wave on the background of each metric in the path integral. If the background contained a horizon, the wave would fall through it, and would decay exponentially at late time outside the horizon. For example, correlation functions decay exponentially in black hole metrics. This means the path integral over all topologically nontrivial metrics will be independent of the state on the initial surface. It will not add to the amplitude to go from initial state to final that comes from the path integral over all topologically trivial metrics. So the mapping from initial to final states, given by the path integral over all metrics, will be unitary.

One might question the use in this argument, of the concept of a quantum state for the gravitational field on an initial or final spacelike surface. This would be a functional of the geometries of spacelike surfaces, which is not something that can be measured in weak fields near infinity. One can measure the weak gravitational fields on a timelike tube around the system, but the caps at top and bottom, go through the interior of the system, where the fields may be strong.

One way of getting rid of the difficulties of caps would be to join the final surface back to the initial surface, and integrate over all spatial geometries of the join. If this was an identification under a Lorentzian time interval, T, at infinity, it would introduce closed timelike curves. But if the interval at infinity is the Eu-

clidean distance, beta, the path integral gives the partition function for gravity at temperature $1/\beta$.

The partition function of a system is the trace over all states, weighted with $e^{-\beta H}$. *One can then integrate* β *along a contour parallel to the imaginary axis with the factor* $e^{-\beta E}$. *This projects out the states with energy* E_0. *In a gravitational collapse and evaporation, one is interested in states of definite energy, rather than states of definite temperature.*

There is an infrared problem with this idea for asymptotically flat space. The Euclidean path integral with period β *is the partition function for space at temperature* $1/\beta$. *The partition function is infinite because the volume of space is infinite. This infrared problem can be solved by a small negative cosmological constant. It will not affect the evaporation of a small black hole, but it will change infinity to anti-deSitter space, and make the thermal partition function finite.*

The boundary at infinity is then a torus, S^1 *cross* S^2. *The trivial topology, periodically identified anti-deSitter space, fills in the torus, but so also do nontrivial topologies, the best known of which is Schwarzschild anti-deSitter. Providing that the temperature is small compared to the Hawking-Page temperature, the path integral over all topologically trivial metrics represents self-gravitating radiation in asymptotically anti-deSitter space. The path integral over all metrics of Schwarzschild AdS topology represents a black hole and thermal radiation in asymptotically anti-deSitter.*

The boundary at infinity has topology S^1 *cross* S^2. *The simplest topology that fits inside that boundary is the trivial topology,* S^1 *cross* D^3, *the three-disk. The next simplest topology, and the first nontrivial topology, is* S^2 *cross* D^2. *This is the topology of the Schwarzschild anti-deSitter metric. There are other possible topologies that fit inside the boundary, but these two are the important cases: topologically trivial metrics and the black hole. The black hole is eternal. It cannot become topologically trivial at late times.*

In view of this, one can understand why information is preserved in topologically trivial metrics, but exponentially decays in topologically non trivial metrics. A final state of empty space without a black hole would be topologically trivial, and be foliated by surfaces of constant time. These would form a 3*-cycle modulo the boundary at infinity. Any global symmetry would lead to conserved global charges on that* 3*-cycle. These would prevent correlation functions from decaying exponentially in topologically trivial metrics. Indeed, one can regard the unitary Hamiltonian evolution of a topologically trivial metric as the conservation of information through a* 3*-cycle.*

On the other hand, a nontrivial topology, like a black hole, will not have a final 3*-cycle. It will not therefore have any conserved quantity that will prevent correlation functions from exponentially decaying. One is thus led to the remarkable result that late time amplitudes of the path integral over a topologically non trivial metric, are independent of the initial state. This was noticed by Maldacena in the case of asymptotically anti-deSitter in 3d, and interpreted as implying that information is lost in the BTZ black hole metric. Maldacena was* *able to show that topologically trivial metrics have correlation functions that do not decay, and have amplitudes of the right order to be compatible with a unitary evolution. Maldacena did not realize, however that it follows from a canonical treatment that the evolution of a topologically trivial metric, will be unitary.*

So in the end, everyone was right, in a way. Information is lost in topologically nontrivial metrics, like the eternal black hole. On the other hand, information is preserved in topologically trivial metrics. The confusion and paradox arose because people thought classically, in terms of a single topology for spacetime. It was either $\mathbb{R}^4$, or a black hole. But the Feynman sum over histories allows it to be both at once. One can not tell which topology contributed the observation, any more than one can tell which slit the electron went through, in the two slits experiment. All that observation at infinity can determine is that there is a unitary mapping from initial states to final, and that information is not lost.

My work with Hartle showed the radiation could be thought of as tunnelling out from inside the black hole. It was therefore not unreasonable to suppose that it could carry information out of the black hole. This explains how a black hole can form, and then give out the information about what is inside it, while remaining topologically trivial. There is no baby universe branching off, as I once thought. The information remains firmly in our universe. I'm sorry to disappoint science fiction fans, but if information is preserved, there is no possibility of using black holes to travel to other universes. If you jump into a black hole, your mass-energy will be returned to our universe, but in a mangled form, which contains the information about what you were like, but in an unrecognisable state.

There is a problem describing what happens, because strictly speaking the only observables in quantum gravity are the values of the field at infinity. One cannot define the field at some point in the middle, because there is quantum uncertainty in where the measurement is done. However, in cases in which there are a large number, N, of light matter fields, coupled to gravity, one can neglect the gravitational fluctuations, because they are only one among N quantum loops. One can then do the path integral over all matter fields, in a given metric, to obtain the effective action, which will be a functional of the metric.

One can add the classical Einstein-Hilbert action of the metric to this quantum effective action of the matter fields. If one integrated this combined action over all metrics, one would obtain the full quantum theory. However, the semiclassical approximation is to represent the integral over metrics by its saddle point. This will obey the Einstein equations, where the source is the expectation value of the energy momentum tensor, of the matter fields in their vacuum state.

The only way to calculate the effective action of the matter fields, used to be perturbation theory. This is not likely to work in the case of gravitational collapse. However, fortunately we now have a non-perturbative method in AdS/CFT. The Maldacena conjecture says that the effective action of a CFT on a background metric is equal to the supergravity effective action of anti-deSitter space with that background metric at infinity. In the large N limit, the supergravity effective action is just the classical action. Thus the calculation of the quantum

> *effective action of the matter fields, is equivalent to solving the classical Einstein equations.*
>
> *The action of an anti-deSitter-like space with a boundary at infinity would be infinite, so one has to regularize. One introduces subtractions that depend only on the metric of the boundary. The first counter-term is proportional to the volume of the boundary. The second counter-term is proportional to the Einstein-Hilbert action of the boundary. There is a third counter-term, but it is not covariantly defined. One now adds the Einstein-Hilbert action of the boundary and looks for a saddle point of the total action. This will involve solving the coupled four- and five-dimensional Einstein equations. It will probably have to be done numerically.*
>
> *In this talk, I have argued that quantum gravity is unitary, and information is preserved in black hole formation and evaporation. I assume the evolution is given by a Euclidean path integral over metrics of all topologies. The integral over topologically trivial metrics can be done by dividing the time interval into thin slices and using a linear interpolation to the metric in each slice. The integral over each slice will be unitary, and so the whole path integral will be unitary.*
>
> *On the other hand, the path integral over topologically nontrivial metrics, will lose information, and will be asymptotically independent of its initial conditions. Thus the total path integral will be unitary, and quantum mechanics is safe.*
>
> *It is great to solve a problem that has been troubling me for nearly thirty years, even though the answer is less exciting than the alternative I suggested. This result is not all negative however, because it indicates that a black hole evaporates, while remaining topologically trivial. However, the large N solution is likely to be a black hole that shrinks to zero. This is what I suggested in 1975.*
>
> *In 1997, Kip Thorne and I bet John Preskill that information was lost in black holes. The loser or losers of the bet are to provide the winner or winners with an encyclopaedia of their own choice, from which information can be recovered with ease. I'm now ready to concede the bet, but Kip Thorne isn't convinced just yet. I will give John Preskill the encyclopaedia he has requested. John is all-American, so naturally he wants an encyclopaedia of baseball. I had great difficulty in finding one over here, so I offered him an encyclopaedia of cricket, as an alternative, but John wouldn't be persuaded of the superiority of cricket. Fortunately, my assistant, Andrew Dunn, persuaded the publishers Sportclassic Books to fly a copy of "Total Baseball: The Ultimate Baseball Encyclopedia" to Dublin. I will give John the encyclopaedia now. If Kip agrees to concede the bet later, he can pay me back.*

At this point the encyclopedia was brought on stage and given to John Preskill, who waved it over his head in a parody of athletic triumph. The order of events is a bit fuzzy in my mind, but sometime around then he said "I always hoped that when Stephen conceded, there would be a witness — this really exceeds my expectations."

After this, Kip Thorne ran a question and answer period, saying that he would alternate between questions from conference participants, which Hawking's grad student would answer, and questions from the press, which Hawking would answer — after

Thorne checked Hawking's facial expressions to see whether he felt they were worth answering.

First, a correspondent from the BBC asked Stephen Hawking what the significance of this result was for "life, the universe and everything". (Here I'm using John Preskill's humorous paraphrase.) Hawking agreed to answer this, and while he began laboriously composing a reply using the computer system on his wheelchair, his grad student Christophe Galfard fielded three questions from experts: Bill Unruh, Gary Horowitz and Robb Mann. I didn't find the replies terribly illuminating, except that when asked if information would be lost if we kept feeding the black hole matter to keep it from evaporating away, Galfard said "yes". Everyone afterwards commented on what a tough job it would be for a student to field questions in front of about 800 physicists and the international press.

At this point Kip Thorne checked to see if Hawking was done composing his reply. He was not. To fill time, Thorne explained why he hadn't yet conceded the bet, saying "This looks to me, on the face of it, to be a lovely argument. But I haven't seen all the details." He took this opportunity to tell the reporters a bit about how science was done: we don't just listen to Hawking and take his word for everything, we have to go off and check things ourselves. He told a nice story about how when Hawking first showed that black holes radiate, everyone with their own approach to quantum field theory on curved spacetime needed to redo this calculation their own way to be convinced — with Yakov Zeldovich, who'd gotten the game started by showing that energy could be extracted from *rotating* black holes in the form of radiation, being one of the very last to agree! Preskill chimed in, saying "I'll be honest — I didn't understand the talk", and that he too would need to see more details.

After a bit more of this sort of thing, Hawking was ready to answer the BBC reporter's question. His answer was surprisingly short, and it went something like this (I can't find an exact quote): "This result shows that everything in the universe is governed by the laws of physics." A suitably grandiose answer for a grandiose question! One can imagine better explanations of unitarity, but not quicker ones.

At this point Kip Thorne solicited more questions from the press but said they should confine themselves to yes-or-no questions, so Hawking could answer them more efficiently. Jenny Hogan got the first question, asking what Hawking would do now that he'd solved this problem. Kip Thorne pointed out that this was not a yes-or-no question. Hogan replied that it shouldn't take long to reply; Thorne was doubtful, but in the midst of the ensuing conversation Hawking shot off an unexpectedly rapid response: "I don't know." Everyone laughed, and at this point the public question period was called to a close, though reporters were allowed to stay and pester Hawking some more.

At the time Hawking's talk seemed very cryptic to me, but in the process of editing the above transcript it's become a lot clearer, so I'll try to give a quick explanation.

I should start by saying that the jargon used in this talk, while doubtless obscure to most people, is actually quite standard and not very difficult to anyone who has spent some time studying the Euclidean path integral approach to quantum gravity. The problem is not the jargon so much as the lack of detail, which requires some imagination to fill in. When I first heard the talk, this lack of detail had me completely stumped. But now it makes a little more sense. . . .

He's studying the process of creating a black hole and letting it evaporate away. He's imagining studying this in the usual style of particle physics, as a "scattering experiment", where we throw in a bunch of particles and see what comes out. Here we throw in a bunch of particles, let them form a black hole, let the black hole evaporate away, and examine the particles (typically photons for the most part) that shoot out.

The rules of the game in a "scattering experiment" are that we can only talk about what's going on "at infinity", meaning very far from where the black hole forms — or more precisely, where it may or may not form!

The advantage of this is that physics at infinity can be described without the full machinery of quantum gravity: we don't have to worry about quantum fluctuations of the geometry of spacetime messing up our ability to say where things are. The disadvantage is that we can't actually say for sure whether or not a black hole formed. At least this *seems* like a "disadvantage" at first — but a better term for it might be a "subtlety", since it's crucial for resolving the puzzle:

> *Black hole formation and evaporation can be thought of as a scattering process. One sends in particles and radiation from infinity, and measures what comes back out to infinity. All measurements are made at infinity, where fields are weak, and one never probes the strong field region in the middle. So one can't be sure a black hole forms, no matter how certain it might be in classical theory. I shall show that this possibility allows information to be preserved and to be returned to infinity.*

Now, the way Hawking likes to calculate things in this sort of problem is using a "Euclidean path integral". This is a rather controversial approach — hence his grin when he said it's the "only sane way" to do these calculations — but let's not worry about that. Suffice it to say that we replace the time variable "t" in all our calculations by "it", do a bunch of calculations, and then replace "it" by "t" again at the end. This trick is called "Wick rotation". In the middle of this process, we hope all our formulas involving the geometry of 4d *spacetime* have magically become formulas involving the geometry of 4d *space*. The answers to physical questions are then expressed as integrals over all geometries of 4d space that satisfy some conditions depending on the problem we're studying. This integral over geometries also includes a sum over topologies.

That's what Hawking means by this:

> *I adopt the Euclidean approach, the only sane way to do quantum gravity non-perturbatively. In this, the time evolution of an initial state is given by a path integral over all positive definite metrics that go between two surfaces that are a distance T apart at infinity. One then Wick rotates the time interval, T, to the Lorentzian. The path integral is taken over metrics of all possible topologies that fit in between the surfaces.*

Unfortunately, nobody knows how to define these integrals. However, physicists like Hawking are usually content to compute them in a "semiclassical approximation". This means integrating not over all geometries, but only those that are close to some solution of the classical equations of general relativity. We can then do a clever approximation to get a closed-form answer.

(Nota bene: here I'm talking about the equations of general relativity on 4d *space*, not 4d spacetime. That's because we're in the middle of this Wick rotation trick.)

Actually, I'm oversimplifying a bit. We don't get "the answer" to our physics question this way: we get one answer for each solution of the equations of general relativity that we deem relevant to the problem at hand. To finish the job, we should add up all these partial answers to get the total answer. But in practice this last step is always too hard: there are too many topologies, and too many classical solutions, to keep track of them all.

So what do we do? We just add up a few of the answers, cross our fingers, and hope for the best! If this procedure offends you, go do something easy like math.

In the problem at hand here, Hawking focuses on two classical solutions, or more precisely two classes of them. One describes a spacetime with no black hole, the other describes a spacetime with a black hole which lasts forever. Each one gives a contribution to the semiclassical approximation of the integral over all geometries. To get answers to physical questions, he needs to sum over *both*. In principle he should sum over infinitely many others, too, but nobody knows how, so he's probably hoping the crux of the problem can be understood by considering just these two.

He says that if you just do the integral over geometries near the classical solution where there's no black hole, you'll find - unsurprisingly — that no information is lost as time passes.

He also says that if you do the integral over geometries near the classical solution where there is a black hole, you'll find - surprisingly — that the answer is *zero* for a lot of questions you can measure the answers to far from the black hole. In physics jargon, this is because a bunch of "correlation functions decay exponentially".

So, when you add up both answers to see if information is lost in the real problem, where you can't be sure if there's a black hole or not, you get the same answer as if there were no black hole!

> *So in the end, everyone was right, in a way. Information is lost in topologically nontrivial metrics, like the eternal black hole. On the other hand, information is preserved in topologically trivial metrics. The confusion and paradox arose because people thought classically, in terms of a single topology for spacetime. It was either $\mathbb{R}^4$, or a black hole. But the Feynman sum over histories allows it to be both at once. One can not tell which topology contributed the observation, any more than one can tell which slit the electron went through, in the two slits experiment. All that observation at infinity can determine is that there is a unitary mapping from initial states to final, and that information is not lost.*

The mysterious part is why the geometries near the classical solution where there's a black hole don't contribute at all to information loss, even though they do contribute to other important things, like the Hawking radiation. Here I'd need to see an actual calculation. Hawking gives a nice hand-wavy topological argument, but that's not enough for me.

Since this issue is long enough already and I want to get it out soon, I won't talk about other things that happened at this conference — nor will I talk about the conference on n-categories earlier this summer! I just want to say a few elementary things about the topology lurking in Hawking's talk... since some mathematicians may enjoy it.

As he points out, the answers to a bunch of questions diverge unless we put our black hole in a box of finite size. A convenient way to do this is to introduce a small negative cosmological constant, which changes our default picture of spacetime from Minkowski spacetime, which is topologically $\mathbb{R}^4$, to anti-deSitter spacetime, which is topologically $\mathbb{R} \times D^3$ after we add the "boundary at infinity". Here $\mathbb{R}$ is time and the 3-disk D^3 is space. This is a Lorentzian manifold with boundary, but when we do Wick rotation we get a Riemannian manifold with boundary having the same topology.

However, when we are doing Euclidean path integrals at nonzero temperature, we should replace the time line $\mathbb{R}$ here by a circle whose radius is the reciprocal of that temperature. (Take my word for it!) So now our Riemannian manifold with boundary is $S^1 \times D^3$. This is what Hawking uses to handle the geometries without a black hole. The boundary of this manifold is $S^1 \times S^2$. But there's another obvious manifold with this boundary, namely $D^2 \times S^2$. And this corresponds to the geometries with a black hole! This is cute because we see it all the time in surgery theory. In fact I commented on Hawking's use of this idea a long time ago, in "Week 67".

In his talk, Hawking points out that $S^1 \times D^3$ has a nontrivial 3-cycle in it if we use relative cohomology and work relative to the boundary $S^1 \times S^2$. But, $D^2 \times S^2$ does not. When spacetime is n-dimensional, conservation laws usually come from integrating closed $(n-1)$-forms over cycles that correspond to "space", so we get interesting conservation laws when there are nontrivial $(n-1)$-cycles. Here Hawking is using this to argue for conservation of information when there's no black hole — namely for $S^1 \times D^3$ — but not when there is, namely for $D^2 \times S^2$.

All this is fine and dandy; the hard part is to see why the case when there *is* a black hole doesn't screw things up! This is where his allusions to "exponentially decaying correlation functions come in" — and this is where I'd like to see more details. I guess a good place to start is Maldacena's papers on the black hole in 3d spacetime — the so-called Banados-Teitelboim-Zanelli or "BTZ" black hole. This is a baby version of the problem, one dimension down from the real thing, where everything should get much simpler. For a bunch about the BTZ black hole, try:

3) Maximo Banados, Marc Henneaux, Claudio Teitelboim, and Jorge Zanelli, "Geometry of the 2+1 black hole", *Phys. Rev.* **D48** (1993) 1506–1525, also available as gr-qc/9302012.

The relevant paper by Maldacena seems to be:

4) Juan Maldacena, "Eternal Black Holes in AdS", *JHEP* **0304** (2003) 021, also available as hep-th/0106122.

You can also see a talk he gave on this at the Institute for Theoretical Physics at U. C. Santa Barbara:

5) Juan Maldacena, "Eternal Black Holes in AdS", `http://online.itp.ucsb.edu/online/mtheory_c01/maldacena/`.

By the way, here are some photos of the conference. . .

6) John Baez, Dublin, `http://math.ucr.edu/home/baez/dublin/`

. . . and also photos of the plaque on the bridge where Hamilton carved his defining relations for the quaternions!

Addendum: My friend Ted Bunn filled a gap in my understanding of the history of astronomy. I had written:

> *It turned out to describe an idea I'd never even dreamt of before: a heliocentric cosmology in which the planets move along circular orbits with epicycles a la Ptolemy!*

to which he replied:

> *There is nothing new under (or orbiting) the Sun. This idea is originally due to Copernicus. Thomas Kuhn's book "The Copernican Revolution" has a nice discussion.*

In retrospect it's obvious that *someone* had to try this idea before Kepler came up with elliptical heliocentric orbits. In fact, Kepler tried ellipses only because the epicycle theory didn't work well for Mars.

> I'm not that good at math, but I do know that the universe is formed with mathematical principles whether I understand them or not, and I was going to let that guide me.
>
> *— Bob Dylan, Chronicles (vol. 1)*

Week 208

November 6, 2004

Last weekend I went to a conference at the Perimeter Institute:

1) Workshop on Quantum Gravity in the Americas, `http://www.perimeterinstitute.ca/activities/scientific/PI-WORK-2/`

It was great to see the new building. I'd visited this institute before in its temporary location, which was a funky old hotel building complete with pool tables and a bar. The new building is very different: four stories of intensely modern architecture overlooking a lake, consisting mainly of an enormous atrium lined with walkways and glass-walled offices. There's also a big lecture theater, a couple of smaller seminar rooms, a library, a restaurant whose walls are all blackboards, a reflecting pool, and lots of little places to sit and talk, complete with espresso machines.

In short, a theoretical physicist's idea of heaven!

But perhaps the design of heaven shouldn't be left to theoretical physicists. Some aspects of the setup don't seem very comfortable. Like most modern architecture, the place is short on coziness — there's too much glass, metal and concrete for my taste. You also find yourself spending a lot of time climbing up and down uncomfortably narrow staircases.

The last, at least, is no accident: they made the stairs skinny on purpose, so you have to say hello to anyone you meet going the other way. It'll be interesting to see how many collaborative papers come out of this.

Abhay Ashtekar was supposed to give the first talk, but he got lost walking to the new building, so suddenly I had to give the first talk. Yikes! Jet-lagged and not fully awake, I sketched the problem of dynamics in quantum gravity: the problem of describing motion in a world where the geometry of spacetime is quantum-mechanical and interacts with matter. I gave a generally downbeat assessment of the progress so far in all known approaches:

2) John Baez, "The problem of dynamics in quantum gravity", `http://math.ucr.edu/home/baez/dynamics/`

Even though the last few Weeks have been on quantum gravity conferences, I've been mainly working on n-categories lately, because I've been sort of fed up with quantum gravity. I did, however, sketch some avenues for progress — and later in this workshop I saw some work that really cheered me up!

For example, I've always been fascinated by John Wheeler's old dream of "matter without matter". In its original version, the idea was to describe elementary particles as the ends of wormholes: if there's an electric field going in one end and out the other, the ends will look like particles of equal and opposite charge! So, the formation of a wormhole could mimic the creation of a particle-antiparticle pair. But there were big problems with this idea: for example, getting the wormhole ends to act like spin-$1/2$ particles.

More recently this idea was reincarnated in the spin network formalism by Lee Smolin, with spin network edges replacing wormholes:

3) Lee Smolin, "Fermions and topology", available as gr-qc/9404010.

A spin network is a gadget with vertices and edges, where the edges represent "field lines" — lines of the electric field or the analogous thing for other forces, including gravity. If a spin network edge goes between vertices that would otherwise be far apart, it acts a bit like a wormhole. It will be hidden from observers in the rest of spacetime, and its ends will look like particles of equal and opposite charge. Even better, it seems easy to get spin-$1/2$ particles this way: they don't call them "spin networks" for nothing!

A variant on this idea is to have spin networks with "loose ends": edges that just fizzle out. This is more ad hoc, but easier to study in some ways. A decade ago, Kirill Krasnov and I showed how to describe the kinematics of charged spin-$1/2$ particles this way:

4) John Baez and Kirill Krasnov, "Quantization of diffeomorphism-invariant theories with fermions", hep-th/9703112.

However, the hard problem in quantum gravity is always dynamics.

Does the dynamics of spin networks with loose ends actually mimic that of particles? Recently Krasnov and other people have begun tackling this question in a toy model, 3-dimensional Lorentzian gravity:

5) Kirill Krasnov, "$\Lambda < 0$ Quantum Gravity in 2+1 Dimensions I: Quantum States and Stringy S-Matrix", *Class. Quant. Grav.* **19** (2002) 3977–3998, also available as hep-th/0112164.

Kirill Krasnov, "$\Lambda < 0$ Quantum Gravity in 2+1 Dimensions II: Black Hole Creation by Point Particles", *Class. Quant. Grav.* **19** (2002) 3999–4028, also available as hep-th/0202117.

He saw that in this theory you can indeed think of particles as spin network ends — though you don't need to emphasize that viewpoint, since there are also other nice ways to think about what's going on, using hyperbolic geometry and complex analysis. It all fits together in a beautiful picture. In principle you can even calculate the amplitude for particles to form black holes when they collide!

In this conference, Laurent Freidel explained how this idea works in 3-dimensional Riemannian gravity — a less physical but mathematically more tractable spin foam model. Some but not all of his work can be found here:

6) Laurent Freidel and David Louapre, "Ponzano-Regge model revisited I: Gauge fixing, observables and interacting spinning particles", available as hep-th/0401076.

Laurent Freidel and David Louapre, "Ponzano-Regge model revisited II: Equivalence with Chern-Simons", available as gr-qc/0410141.

Freidel showed that if you take this theory and allow spin networks with loose ends, they'll act like particles. The spin of these particles is automatically quantized. More surprisingly, so is their mass — and there's an upper bound on the mass! That's because when we quantize this theory, its gauge group automatically gets replaced by a "quantum group". Physically, this means that spacetime becomes quantum-mechanical in such a way that it no longer makes sense to talk about times shorter than about the Planck time. Since the energy of a particle is proportional to the rate at which its wavefunction oscillates, this puts an upper bound on the energy of a particle. And since $E = mc^2$, this means there's an upper bound on the mass a particle can have.

Mathematically, part of the point is that we can describe 3d Riemannian gravity as a gauge theory where the gauge group is the double cover of the 3d Euclidean group — the analogue of the Poincare group in this context. But when we quantize the theory, this gets replaced by a quantum group: the "quantum double" of $\mathrm{SU}(2)$. As with the 3d Euclidean group, unitary representations of this quantum group are classified by mass and spin. . . but now both mass and spin are discrete, and both are bounded above.

Anyway, what's great about Freidel and Louapre's work is that it gives a simplified but mathematically rigorous testbed in which loose ends of spin networks act like particles. We can also think about spin networks with "hidden edges" in this setup. So, we should be able to do calculations and see if a spin network with a hidden edge acts like a spin network with a pair of loose ends — and thus, a particle-antiparticle pair.

Unfortunately, all this work on gravity in 3d spacetime doesn't easily generalize to 4d spacetime. The reason is that gravitational waves are only possible when spacetime has dimension 4 or more. . . so 3d gravity theories have no local degrees of freedom until we include particles: all the fun comes from global topology, like wormholes. That's why 3d theories are easy to calculate with — we can use ideas from topological quantum field theory. The danger, though, is that these calculations are misleading it comes to real-world physics. Indeed, that's precisely the sort of thing I was worrying about in my talk.

So, it really cheered me up when a young guy named Artem Starodubtsev spoke about a promising new spin foam model of quantum gravity in 4 dimensions! He's working on it now with Laurent Freidel. He has a couple of papers out that *hint* at the main ideas, but you'll have to wait to see what they're up to now:

7) Artem Starodubtsev, "Topological excitations around the vacuum of quantum gravity I: The symmetries of the vacuum", available as hep-th/0306135.

Artem Starodubtsev and Lee Smolin, "General relativity with a topological phase: an action principle", available as hep-th/0311163.

The basic idea is to treat 4d general relativity with positive cosmological constant as a perturbation of a topological quantum field theory. The topological theory has a single state, which corresponds to a quantum version of "deSitter space": an exponentially expanding universe similar to the one we see today, but with no matter. To calculate in full-fledged gravity, we then use perturbation theory, getting answers as power series in a coupling constant. But the cool part is that unlike ordinary perturbative quantum gravity this perturbation theory is manifestly diffeomorphism invariant term by term. And each term is a sum over spin foams!

Even better, the coupling constant in this theory is the cosmological constant in Planck units! That's an incredibly small dimensionless number: about 10^{-123}. Physicists like perturbation theory when the coupling constant is small, since then the first few terms tend to give reasonably accurate answers — even if the whole series diverges. For example, quantum electrodynamics gives high-precision answers because the fine structure constant is about $1/137$, which is pretty small. But 10^{-123} is *really* small.

I'd seen Starodubtsev talk about this in Marseille (see "Week 206") but now he and Freidel have done calculations recovering Newton's law of gravity in an appropriate approximation from this theory. That may not seem like a big deal, but it's actually very cool to see Newton's law reemerge from a manifestly diffeomorphism-invariant theory of quantum gravity: no model had ever managed this feat before.

For those of you hungering for technical details, I'll just say that the topological theory in question is BF theory with the symmetry group of deSitter spacetime, namely $\mathrm{SO}(4,1)$, as the gauge group. General relativity can be regarded as a perturbation of this BF theory by borrowing some ideas from the "MacDowell-Mansouri" formulation of general relativity. If you haven't heard of that, well, neither had I. It's a sort of old idea:

8) S. W. MacDowell and F. Mansouri, "Unified geometric theory of gravity and supergravity", *Phys. Rev. Lett.* **38** (1977), 739–742.

... but here we aren't using anything anything about supergravity, just the fact that ordinary general relativity can be treated as a theory with gauge group $\mathrm{SO}(4,1)$ and a Lagrangian that breaks this symmetry down to the Lorentz group $\mathrm{SO}(3,1)$. The paper by Smolin and Starodubtsev listed above describes the details, but in the case of going from $\mathrm{SO}(5)$ down to $\mathrm{SO}(4)$.

When we quantize BF theory in 4 dimensions we get a spin foam model called the Crane-Yetter model, where the spin foams are defined using the representation theory of a quantum group: a "q-deformed version" of the original gauge group. So, the real engine behind Freidel and Starodubtsev's calculations are spin foams involving the q-deformed version of $\mathrm{SO}(4,1)$, called $\mathrm{SO}_q(4,1)$. This is technically tricky because $\mathrm{SO}(4,1)$ is noncompact, and noncompact quantum groups are just beginning to be understood. So, there's probably still tons of mathematical work left to be done. But, the upshot is that Freidel and Starodubtsev calculate stuff as power series in the cosmological constant where each term is computed using $\mathrm{SO}_q(4,1)$ spin foams. It's sort of like a souped-up Feynman diagram expansion, but with spin foams replacing Feynman diagrams.

Now that I've thrown around enough buzzwords to scare off the kids, I can tell you about Lee Smolin's talk, which was definitely X-rated: for adults only, people who can listen to speculations with just the right mixture of disbelief and open-mindedness. It was the last talk in the conference. And it was about the possible physical effects of spin networks with "hidden edges"!

First of all, he reminded us how these can mimic particles, and even some of the usual particle interactions. But then he went ahead and suggested that hidden edges can cause nonlocal effects in physics, like the force of gravity decaying more slowly than $1/r^2$ — just as it does in MOND, the wacky but strangely accurate explanation of galactic rotation curves that uses a modification of Newtonian gravity instead of positing dark matter! (See "Week 206" for more on MOND.) It's hard to make up sensible theories of forces that decay more slowly than $1/r^2$, but nonlocal interactions would be one way to do it... and hidden spin network edges might cause those.

There are a million things that could go wrong with this idea, but I like it, because it suggests a way quantum gravity might try to explain one of the big mysteries of physics — dark matter. And until we get our theories to make contact with experiment, it'll be very hard for us to tell if they're on the right track.

Anyway, Smolin hasn't come out with a paper on this stuff yet, so we'll have to wait for more details.

By the way:

In what I've written this week, I've had to seriously downplay the cool math involved, to give (I hope) some inkling of the cool physics. Krasnov work on $2+1$-dimensional Lorentzian gravity with positive cosmological constant uses the fact that the phase space of this theory is closely related to "Teichmueller space" — the space of complex structures mod diffeomorphisms that are connected to the identity. I talked about this space in "Week 28", but I forgot to say that we can think of it as a space of flat $\mathrm{SO}(2,1)$ connections mod gauge transformations. Here $\mathrm{SO}(2,1)$ is just the Lorentz group in 3 dimensions. So, if we quantize 2+1 Lorentzian gravity with positive cosmological constant, we get a theory where states are described by $\mathrm{SO}_q(2,1)$ spin networks… but this is also a theory of "quantum Teichmueller space". Again this is tricky because $\mathrm{SO}(2,1)$ is noncompact, but people have made a lot of progress lately, thanks in part to a line of work started by Kashaev:

9) R. M. Kashaev, "Quantization of Teichmueller spaces and the quantum dilogarithm", available as q-alg/9705021.

10) L. Chekhov and V. V. Fock, "Quantum Teichmueller space", *Theor. Math. Phys.* **120** (1999) 1245–1259, also available as math.QA/9908165.

You can get a sense of who's working on this stuff and what they're doing by looking at the references for this recent conference on 3d quantum gravity in Edinburgh, which unfortunately took place when I was in Hong Kong:

11) Workshop on physics and geometry of 3-dimensional quantum gravity, http://www.ma.hw.ac.uk/~bernd/references.html

I should also add that people don't usually don't talk about the 3d Lorentz group $\mathrm{SO}(2,1)$ here; they talk about its double cover $\mathrm{SL}(2,\mathbb{R})$.

Anyway, I'll quit here. The next conference on loops and spin foams will probably happen in Berlin at the Albert Einstein Institute in 2005, which happens to be the hundredth birthday of special relativity. I hope we can make a lot of progress before then and make Al proud.

> The best way to have a good idea is to have a lot of ideas.
>
> — *Linus Pauling*

(Not necessarily true, but worth keeping in mind.)

Week 209

November 21, 2004

Time flies! This June, Peter May and I organized a workshop on n-categories at the Institute for Mathematics and its Applications:

1) n-Categories: Foundations and Applications, `http://www.ima.umn.edu/categories/`

I've been meaning to write about it ever since, but I keep putting it off because it would be so much work. The meeting lasted almost two weeks. It was an intense, exhausting affair packed with talks, conversations, and "Russian-style seminars" where the audience interrupted the speakers with lots of questions. I took about 50 pages of notes. How am I supposed to describe all that?!

Oh well... I'll just dive in. I'll quickly list all the official talks in this conference. I won't describe the many interesting "impromptu talks", some of which you can see on the above webpage. Nor will I explain what n-categories are, or what they're good for! If you want to learn what they're good for, you should go back to "Week 73" and read "The Tale of n-Categories". And if you want to know what they *are*, try this brand-new book:

2) Eugenia Cheng and Aaron Lauda, *Higher-Dimensional Categories: an Illustrated Guide Book*, available free online at: `http://www.dpmms.cam.ac.uk/~elgc2/guidebook/`

Eugenia and Aaron wrote it specially for the workshop! It's packed with pictures and it's lots of fun.

I'm just going to list the talks. . . .

Throwing etiquette to the winds, I kicked off the conference myself with two talks explaining some reasons why n-categories are interesting and what they should be like:

3) John Baez, "Why n-Categories?" and "What n-categories should be like". Notes available at `http://www.ima.umn.edu/categories/#mon`

If you're a long-time reader of This Week's Finds you'll know what I said: n-categories give a new world of math in which equations are always replaced by isomorphisms, and this world is incredibly rich in structure. The n-categories called "n-groupoids" magically know everything there is to know about homotopy theory, while those called "n-categories with duals" know everything there is to know about the topology of manifolds. There are, unfortunately, some details that still need to be worked out!

After my talks there was a reception. Later, over dinner, Tom Leinster gave a "Russian style seminar" outlining the different approaches to n-categories:

4) Tom Leinster, "Survey and Taxonomy". Talk based on chapter 10 of his book *Higher Operads, Higher Categories*, Cambridge U. Press, Cambridge, 2004, also available free online at `math.CT/0305049`.

You'll notice these young n-category people are smart: they force their publishers to keep their books available for free online! All scientists should do this, since the only people who make serious money from scientific monographs are the publishers. What scientists get from writing technical books is not money but attention. As George Franck said, "Attention is a mode of payment... reputation is the asset into which the attention received from colleagues crystallizes."

The next morning began with a triple-header talk on "weak categories":

5) Andre Joyal, Peter May and Timothy Porter, "Weak categories". Notes available at `http://www.ima.umn.edu/categories/#tues`

Here a "weak category" means a category where the usual laws hold only up to homotopy, where the homotopies satisfy laws of their own up to homotopy, ad infinitum. If you know what weak ∞-categories are, you can define a weak category to be one of these where all the j-morphisms are equivalences for $j > 1$. But, the nice thing is that there are ways to define weak categories without the full machinery of ∞-categories! People have come up with different approaches: "categories enriched over simplicial sets", "Segal categories", "A_∞ categories" and also Joyal's "quasicategories". The talk was a nice introduction to all these approaches.

Then Michael Batanin explained his definition of ∞-categories. This was a blackboard talk, so there are no notes on the web, but you can try his original paper:

6) Michael Batanin, "Monoidal globular categories as natural environment for the theory of weak n-categories", *Adv. Math.* **136** (1998), 39–103, also available at `http://www.ics.mq.edu.au/~mbatanin/papers.html`

and when you get stuck, try the books by Cheng-Lauda and Leinster.

Over dinner, Eugenia Cheng and Tom Leinster explained the concepts of "operad" and "multicategory" which play such an important role in so much work on n-categories. Again there are no notes, so try their books.

I forget when it happened, but sometime around the second or third day of the conference people decided it was too much of a nuisance listening to math lectures while eating dinner — mainly because there wasn't enough room in the dining hall to take notes, and the blackboards weren't big enough. So at that point, we switched to having lectures *after* dinner. As I said, this workshop was not for wimps!

The morning of the third day began with a no-holds-barred minicourse on model categories by Peter May:

7) Peter May, "Model categories". Notes available at `http://www.ima.umn.edu/categories/#wed`

Model categories are a wonderful framework for relating different approaches to homotopy theory, and a bunch of people hope they can also be used to relate different approaches to n-categories.

Then Clemens Berger explained Andre Joyal's approach to weak n-categories:

8) Clemens Berger, "Cellular definitions". Notes available at `http://www.ima.umn.edu/categories/#wed`

Then, either during or after dinner, Eugenia Cheng explained various "opetopic" approaches to weak n-categories. Again, the best way to learn about these is to read the book she wrote with Lauda, or else the book by Leinster.

On the morning of the fourth day, Andre Joyal explained his work on quasicategories — an approach to weak categories in which they are simplicial sets satisfying a restricted version of the Kan condition. They've been around a long time, but Joyal is redoing all of category theory in this context! He's been writing a book about this, which deserves to be called "Quasicategories for the Working Mathematician". Since Joyal is a perfectionist, this will take forever to finish. However, we're hoping to extract a preliminary version from him for the proceedings of this conference. For now, you can read a bit about quasicategories in Tim Porter's notes mentioned in item 5) above.

Then Tom Leinster and Nick Gurski spoke about Ross Street's definition to weak ∞-categories, where they are simplicial sets satisfying an even more subtly restricted version of the Kan condition.

9) Nick Gurski and Tom Leinster, "Simplicial definition". Notes available at `http://www.ima.umn.edu/categories/#thur`

Street's definition is tough to understand at first, but it should eventually include Joyal's quasicategories as a special case, which is nice. For Street's own discussion, see:

10) Ross Street, "Weak ω-categories", in *Diagrammatic Morphisms and Applications*, eds. David Radford, Fernando Souza, and David Yetter, Contemp. Math. **318**, AMS, Providence, Rhode Island, 2003, pp. 207–213. Also available as `www.maths.mq.edu.au/\~street/Womcats.pdf`

It relies on some work by Dominic Verity which has finally been written up after many years of unpublished limbo:

11) Dominic Verity, "Complicial sets", available as `math.CT/0410412`

After dinner we took a turn towards applications, and Larry Breen explained his work on n-stacks and n-gerbes. An n-stack is like a sheaf that has an $(n-1)$-category of sections, while an n-gerbe has an $(n-1)$-groupoid of sections. Such things show up a lot in algebraic geometry, and more recently in mathematical physics inspired by string theory. Alas, the audience was rather tired this evening, so Larry only got to 1-stacks and 1-gerbes! But he gave an impromptu talk later where he reached $n = 2$, and the notes for both talks are available in combined form here:

12) Larry Breen, "n-Stacks and n-gerbes: homotopy theory". Notes available at `http://www.ima.umn.edu/categories/#thur`

You've heard about David Corfield's quest for a philosophy of real mathematics in "Week 198". He's one of the few philosophers who understands enough math to realize how cool n-categories are — which may explain why he's having trouble getting a job. On the morning of the fourth day, he gave a talk on the impact n-categories could have in philosophy:

13) David Corfield, "n-Category theory as a catalyst for change in philosophy". Notes available at `http://www.ima.umn.edu/categories/#fri`

Later that day, Bertrand Toen explained Segal categories, which are another popular approach to weak categories:

14) Bertrand Toen, "Segal categories". Notes by Joachim Kock available at `http://www.ima.umn.edu/categories/#fri`

After dinner, he spoke about n-stacks and n-gerbes:

15) Bertrand Toen, "n-Stacks and n-gerbes: algebraic geometry". Notes by Joachim Kock available at `http://www.ima.umn.edu/categories/#fri`

Everyone slept all weekend long. Then on Monday of the second week, the homotopy theorist Zbigniew Fiedorowicz spoke about his work on a kind of n-fold monoidal category that has an n-fold loop space as its nerve. He has some good papers on the web about this, too:

16) Zbigniew Fiedorowicz, "n-Fold categories". Notes available at `http://www.ima.umn.edu/categories/#mon2`

    C. Balteanu, Z. Fiedorowicz, R. Schwaenzl and R. Vogt, "Iterated monoidal categories", available at `math.AT/9808082`

    Z. Fiedorowicz, "Constructions of E_n operads", available at `math.AT/9808089`.

Stefan Forcey continued this theme by discussing enrichment over n-fold monoidal categories. He also has a number of papers about this on the arXiv, of which I'll just mention one:

17) Stefan Forcey, "Higher enrichment: n-fold operads and enriched n-categories, delooping and weakening". Notes available at `http://www.ima.umn.edu/categories/\#mon2`

    Stefan Forcey, "Enrichment over iterated monoidal categories", *Algebraic and Geometric Topology* **4** (2004), 95–119, available online at `http://www.maths.warwick.ac.uk/agt/AGTVol4/agt-4-7.abs.html`. Also available as `math.CT/0403152`.

After dinner we discussed how to relate different definitions of weak n-category.

On Tuesday of the second week, the logician Michael Makkai presented his astounding project of redoing logic in a way that completely eliminates the concept of "equality". This *forces* you to do all of mathematics using weak ∞-categories. I thought this stuff was great, in part because I finally understood it, and in part because it leads naturally to the "opetopic" definition of n-categories that James Dolan and I introduced. The idea of eliminating equality was very much on our mind in inventing this definition, but we didn't create a system of logic that systematizes this idea.

There are no notes for Makkai's talk online, but you can get a lot of good stuff from his website, including:

18) Michael Makkai, "On comparing definitions of weak n-category", available at `http://www.math.mcgill.ca/makkai/`

and this more technical paper which works out the details of his vision:

19) Michael Makkai, "The multitopic ω-category of all multitopic ω-categories", available at http://www.math.mcgill.ca/makkai/

After Makkai's talk, Mark Weber spoke on n-categorical generalizations of the concept of "monad", which is a nice way of describing mathematical gadgets. There are no notes for this talk, but his work on higher operads is at least morally related:

20) Mark Weber, "Operads within monoidal pseudo algebras", available as math.CT/0410230.

Again, after dinner we talked about how to relate different definitions of weak n-category.

On Wednesday of the second week, Michael Batanin spoke about his recent work relating n-categories to n-fold loop spaces. Again no notes, but you can read these papers:

21) Michael Batanin, "The Eckmann-Hilton argument, higher operads and E_n-spaces", available at http://www.ics.mq.edu.au/~mbatanin/papers.html

    Michael Batanin, "The combinatorics of iterated loop spaces", available at http://www.ics.mq.edu.au/~mbatanin/papers.html

Then Joachim Kock laid the ground for a discussion of n-categories and topological quantum field theories, or "TQFTs", by explaining the definition of a TQFT and the classification of 2d TQFTs:

22) Joachim Kock, "Topological quantum field theory primer". Notes available at http://www.ima.umn.edu/categories/#wed2

In the evening, Marco Mackaay and I said more about the relation between TQFTs and n-categories:

23) Marco Mackaay, "Topological quantum field theories". Notes available at http://www.ima.umn.edu/categories/#wed2

24) John Baez, "Space and state, spacetime and process". Notes available at http://www.ima.umn.edu/categories/#wed2

On Thursday, Ross Street started the day in a pleasantly different way — he gave a historical account of work on categories and n-categories in Australia! Australia is home to much of the best work on these subjects, so if you can understand his history you'll wind up understanding these subjects pretty well:

25) Ross Street, "An Australian conspectus of higher category theory". Notes available at http://www.ima.umn.edu/categories/#thur2

As a younger exponent of the Australian tradition, it was then nicely appropriate for Steve Lack to speak about ways of building a model category of 2-categories:

26) Steve Lack, "Higher model categories". Notes available at http://www.ima.umn.edu/categories/#thur2

In the afternoon we had a blast of computer science. First John Power gave a hilarious talk phrased in terms of how one should convince computer theorists to embrace categories, then 2-categories, and then maybe higher categories:

27) John Power, "Why tricategories?". Notes available at `http://www.ima.umn.edu/categories/#thur2`

I spoke about Power's paper with this title back in "Week 53"; now you can get it online!

Then Philippe Gaucher, Lisbeth Fajstrup and Eric Goubault spoke about higher-dimensional automata and directed homotopy theory:

28) Philippe Gaucher, "Towards a homotopy theory of higher dimensional automata". Notes available at `http://www.ima.umn.edu/categories/#thur2`

Lisbeth Fajstrup, "More on directed topology and concurrency", Notes available at `http://www.ima.umn.edu/categories/#thur2`

Eric Goubault, "Directed homotopy theory and higher-dimensional automata". Notes available at `http://www.ima.umn.edu/categories/#thur2`

On Friday, Martin Hyland and Tony Elmendorf gave a double-header talk on higher-dimensional linear algebra and how some concepts in this subject can be simplified using symmetric multicategories. There are, alas, no notes for this talk. You just had to be there.

Finally, my student Alissa Crans gave a talk on higher-dimensional linear algebra, with an emphasis on categorified Lie algebras:

29) Alissa Crans, "Higher linear algebra". Notes available at `http://www.ima.umn.edu/categories/#fri2`

Hers was the last talk in the workshop! I would like to say more about it, but I'm exhausted… and her talk fits naturally into a discussion of "higher gauge theory", which deserves a Week of its own.

By the way, you can see pictures of this workshop here:

30) John Baez, IMA, `http://math.ucr.edu/home/baez/IMA/`

If you want to see what these crazy n-category people look like, you can see most of them here.

Hmm. If you wanted me to actually *explain* something this week, I'm afraid you'll be rather disappointed — so far everything has just been pointers to other material.

Luckily, while I was at this workshop I wrote a little explanation of some material on Picard groups and Brauer groups. There's a Spanish school of higher-dimensional algebra, centered in Granada, and this spring Aurora del Rio Cabeza came from Granada to visit UCR. She and James Dolan spent a lot of time talking about categorical groups (also known as "2-groups") and cohomology theory. I was, alas, too busy to keep up with their conversations, but I learned a little from listening in… and here's my writeup!

Higher categories show up quite naturally in the study of commutative rings and associative algebras over commutative rings. I'd heard of things called "Brauer groups" and "Picard groups" of rings, and something called "Morita equivalence", but I only understood how these fit together when I learned they were part of a marvelous thing: a weak 3-groupoid!

Here's how it goes. You don't need to know much about higher categories for this to make some sense. . . at least, I hope not.

Starting with a commutative ring R, we can form a weak 2-category $\mathsf{Alg}(R)$ where:

- an object A is an associative algebra over R,
- a 1-morphism $M\colon A \to B$ is an (A, B)-bimodule,
- a 2-morphism $f\colon M \to N$ is a homomorphism between (A, B)-bimodules.

This has all the structure you need to get a 2-category. In particular, we can "compose" an (A, B)-bimodule and a (B, C)-bimodule by tensoring them over B, getting an (A, C)-bimodule. But since tensor products are only associative up to isomorphism, we only get a *weak* 2-category, not a strict one.

This weak 2-category has a tensor product, since we can tensor two associative algebras over R and get another one. All the stuff listed above gets along with this process! When an n-category has a well-behaved tensor product we call it "monoidal", so $\mathsf{Alg}(R)$ is a weak monoidal 2-category. But using a standard trick we can reinterpret this as a weak 3-category with one object, as follows:

- there's only one object, R
- a 1-morphism $A\colon R \to R$ is an associative algebra over R
- a 2-morphism $M\colon A \to B$ is an (A, B)-bimodule
- a 3-morphism $f\colon M \to N$ is a homomorphism between (A, B)-bimodules.

Note how all the morphisms have shifted up a notch. What used to be called objects, the associative algebras over R, are now called 1-morphisms. We "compose" them by tensoring them over R.

Next, recall a bit of n-category theory from "Week 35". In an n-category we define a j-morphism to be an "equivalence" iff it's invertible. . . up to equivalence! This definition may sound circular, but really just recursive. To start it off we just need to add that an n-morphism is an equivalence iff it's invertible.

What does equivalence amount to in the 3-category $\mathsf{Alg}(R)$? It's easiest to figure this out from the top down:

- A 3-morphism $f\colon M \to N$ is an equivalence iff it's invertible, so it's an isomorphism between (A, B)-bimodules.
- A 2-morphism $M\colon A \to B$ is an equivalence iff it's invertible up to isomorphism, meaning there exists $N\colon B \to A$ such that:

  — $M \otimes_B N$ is isomorphic to A as an (A, A)-bimodule, — $N \otimes_A M$ is isomorphic to B as a (B, B)-bimodule.

  In this situation people say M is a "Morita equivalence" from A to B.

- A 1-morphism $A\colon R \to R$ is an equivalence iff it's invertible up to Morita equivalence, meaning there exists a 1-morphism $B\colon x \to x$ such that:

  — $A \otimes_R B$ is Morita equivalent to R as an associative algebra over R, — $B \otimes_R A$ is Morita equivalent to R as an associative algebra over R.

  In this situation people say A is an "Azumaya algebra".

Here's a nice example of how Morita equivalence works. Over any commutative ring R there's an algebra $R[n]$ consisting of $n \times n$ matrices with entries in R. $R[n]$ isn't usually isomorphic to $R[m]$, but they're always Morita equivalent! To see this, suppose

- $M\colon R[n] \to R[m]$ is the space of $n \times m$ matrices with entries in R,
- $N\colon R[m] \to R[n]$ is the space of $m \times n$ matrices with entries in R.

These become bimodules in an obvious way via matrix multiplication, and a little calculation shows that they're inverses up to isomorphism!

So, all the algebras $R[n]$ are Morita equivalent. In particular this means that they're all Morita equivalent to R, so they are Azumaya algebras of a rather trivial sort.

If we take R to be the real numbers there is also a more interesting Azumaya algebra over R, namely the quaternions $\mathbb{H}$. This follows from the fact that

$$\mathbb{H} \otimes_{\mathbb{R}} \mathbb{H} = \mathbb{R}[4]$$

This says $\mathbb{H} \otimes_{\mathbb{R}} \mathbb{H}$ is Morita equivalent to $\mathbb{R}$ as an associative algebra over $\mathbb{R}$, which implies (by the definition above) that $\mathbb{H}$ is an Azumaya algebra.

Morita equivalence is really important in the theory of C^*-algebras, Clifford algebras, and things like that. Someday I want to explain how it's connected to Bott periodicity! Oh, there's so much I want to explain. . . .

But right now I want to take our 3-category $\mathsf{Alg}(R)$, massage it a bit, and turn it into a topological space! Then I'll look at the homotopy groups of this space and see what they have to say about our ring R.

To do this, we need a bit more n-category theory. A weak n-category where all the 1-morphisms, 2-morphisms and so on are equivalences is called a "n-groupoid". For example, given any weak n-category, we can form a weak n-groupoid called its "core" by throwing out all the morphisms that aren't equivalences.

So, let's take the core of $\mathsf{Alg}(R)$ and get a weak 3-groupoid. Here's what it's like:

- There's one object, R.
- The 1-morphisms $A\colon x \to x$ are Azumaya algebras over R.
- The 2-morphisms $M\colon A \to B$ are Morita equivalences.
- The 3-morphisms $f\colon M \to N$ are bimodule isomorphisms.

Since as a groupoid with one object is a group, this weak 3-groupoid with one object deserves to be called a "3-group".

Next, given a weak n-groupoid with one object, it's very nice to compute its "homotopy groups". These are easy to define in general, but I'll just do it for the core of $\mathsf{Alg}(R)$ and let you guess the general pattern. First, notice that:

- The identity 1-morphism $1_R\colon R \to R$ is just R, regarded as an associative algebra over itself in the obvious way.
- The identity 2-morphism $1_{1_R}\colon 1_R \to 1_R$ is just R, regarded as an (R, R)-bimodule in the obvious way.
- The identity 3-morphism $1_{1_{1_R}}\colon 1_{1_R} \to 1_{1_R}$ is just the identity function on R, regarded as an isomorphism of (R, R)-bimodules.

At this point we let out a cackle of n-categorical glee. Then, we define the homotopy groups of the core of $\mathsf{Alg}(R)$ as follows:

- The 1st homotopy group consists of equivalence classes of 1-morphisms from R to itself.
- The 2nd homotopy group consists of equivalence classes of 2-morphisms from 1_R to itself
- The 3rd homotopy group consists of equivalence classes of 3-morphisms from 1_{1_R} to itself.

Here we say two morphisms in an n-category are "equivalent" if there is an equivalence from one to the other (or if they're equal, in the case of n-morphisms).

I hope the pattern in this definition of homotopy groups is obvious. In fact, n-groupoids are secretly "the same" — in a subtle sense I'd rather not explain — as spaces whose homotopy groups vanish above dimension n. Using this, the homotopy groups as defined above turn out to be same as the homotopy groups of a certain space associated with the ring R! So, we're doing something very funny: we're using algebraic topology to study algebra.

But, we don't need to know this to figure out what these homotopy groups are like. Unraveling the definitions a bit, one sees they amount to this:

- The 1st homotopy group consists of Morita equivalence classes of Azumaya algebras over R. This is also called the BRAUER GROUP of R.
- The 2nd homotopy group consists of isomorphism classes of Morita equivalences from R to R. This is also called the PICARD GROUP of R.
- The 3rd homotopy group consists of invertible elements of R. This is also called the UNIT GROUP of R.

People had been quite happily studying these groups for a long time without knowing they were the homotopy groups of the core of a weak 3-category associated to the commutative ring R! But, the relationships between these groups are easier to explain if you use the n-categorical picture. It's a great example of how n-categories unify mathematics.

For example, everything we've done is functorial. So, if you have a homomorphism between commutative rings, say

$$f\colon R \to S$$

then you get a weak 3-functor

$$\mathsf{Alg}(f)\colon \mathsf{Alg}(R) \to \mathsf{Alg}(S)$$

This gives a weak 3-functor from the core of $\mathsf{Alg}(R)$ to the core of $\mathsf{Alg}(S)$, and thus a map between spaces. . . which in turn gives a long exact sequence of homotopy groups! So, we get interesting maps going from the unit, Picard and groups of R to those of S — and these fit into an interesting long exact sequence.

For more, try the following papers. The first paper is actually about a generalization of Azumaya algebras called "Azumaya categories", but it starts with a nice quick review of Azumaya algebras and Brauer groups:

31) Francis Borceux and Enrico Vitale, "Azumaya categories", available at `http://www.math.ucl.ac.be/AGEL/Azumaya_categories.pdf`

Category theorists will enjoy the generalization: since algebras are just one-object categories enriched over Vect, the concept of Azumaya algebra really *wants* to generalize to that of an Azumaya category. I'm sure most of the Brauer-Picard-Morita stuff generalizes too, but I haven't checked that out yet.

This second paper makes the connection between Picard and Brauer groups explicit using categorical groups:

32) Enrico Vitale, "A Picard-Brauer exact sequence of categorical groups", *Journal of Pure and Applied Algebra* **175** (2002) 383–408. Also available as `http://www.math.ucl.ac.be/membres/vitale/cat-gruppi2.pdf`

Addendum: It turns out that the Picard-Brauer 3-group has a long and illustrious history. Ross Street explained this to me; I've taken the liberty of numbering the references in his email.

> *Dear John*
>
> *It is great that you jumped in and started writing that report on the Minneapolis meeting. "A journey of a thousand miles . . .".*
>
> *[Carrying on the IMA Russian spirit, I just got back from Christchurch NZ where I gave 11 hours (in 2 days) of lectures on topos theory to a very patient group of physicists, philosophers, mathematicians, and even one economist.]*
>
> *It is also great that you promoted the work of the Granada School. That subject is particularly close to my heart. So here goes another personal history. Probably back at Tulane U in 1969-70, Jack Duskin (who was a great source of inspiration to me and, I believe, to the Granada School) would have pointed me to the papers*
>
> 33) *Grothendieck, Alexander. "Le groupe de Brauer. III. Exemples et complements." (French) 1968.* Dix Exposes sur la Cohomologie des Schemas *pp. 88–188 North-Holland, Amsterdam; Masson, Paris*

34) *Grothendieck, Alexander. "Le groupe de Brauer. II. Theorie cohomologique." (French) 1968.* Dix Exposes sur la Cohomologie des Schemas *pp. 67–87 North-Holland, Amsterdam; Masson, Paris*

35) *Grothendieck, Alexander. "Le groupe de Brauer. I. Algbres d'Azumaya et interpretations diverses." (French) 1968.* Dix Exposes sur la Cohomologie des Schemas *pp. 46–66 North-Holland, Amsterdam; Masson, Paris*

pushing the Brauer group concept of ring theorists (e.g. Azumaya) into the scheme view of algebraic geometry. I later read papers by category theorists, like

36) *Lindner, Harald, "Morita equivalences of enriched categories". Conferences du Colloque sur l'Algebre des Categories (Amiens, 1973), III.* Cahiers Topologie Geom. Differentielle ***15** (1974), no. 4, 377–397, 449–450.*

37) *Fisher-Palmquist, J.; Palmquist, P. H. "Morita contexts of enriched categories".* Proc. Amer. Math. Soc. ***50** (1975), 55–60.*

which seemed to be the beginning of a simpler understanding. Somehow (?) I obtained an original bound reprint of

38) *Froehlich, A.; Wall, C. T. C. "Graded monoidal categories".* Compositio Math. ***28** (1974), 229–285.*

which I have just looked at and realised I should read again (since Turaev and Mueger have been using G-graded categories to understand the G-equivariant version of Turaev's 3-manifold invariant work). It was forerunner to

39) *Froehlich, A.; Wall, C. T. C. "Equivariant Brauer groups". Quadratic forms and their applications (Dublin, 1999), 57–71,* Contemp. Math. ***272**, Amer. Math. Soc., Providence, RI, 2000.*

On my sabbatical at Wesleyan University (Middletown CT) in 1976-77, I joined in the algebraists' workshop on SLNM 181 on separable algebras over commutative rings which was trying to do some of Grothendieck's stuff without the cohomology and alg geom. Joyal taught me a bit about Brauer too, motivating to some extent the work I did on stacks.

Anyway, out of all this, other stuff I've forgotten, and the experience in module theory for enriched categories, it became clear that Morita contexts were a bit silly and adjunctions of (bi)modules were probably better and less ad hoc. The beginning point should be a particular monoidal bicategory $\mathsf{Alg}(R\text{-}\mathsf{Mod})$ *based on a commutative ring R: objects are R-algebras, morphisms are bimodules, 2-cells are module morphisms. The group of units, Picard group and Brauer group all sat happily in there as homotopy groups of the monoidal bicategory.*

> *I'd heard of things called "Brauer groups" and "Picard groups" of rings, and something called "Morita equivalence", but I only understood how these fit together when I learned they were part of a marvelous thing: a weak 3-groupoid!*

After beginning the work with Joyal on braided monoidal categories and learning of his work with Tierney on homotopy 3*-types, I spoke at the homotopy meeting in Bangor in 1986(?) on this monoidal bicategory* $\mathsf{Alg}(R\text{-}\mathsf{Mod})$ *as a fundamental example. (It is discussed much later in the last part of*

40) *R. Gordon, A.J. Power and R. Street, "Coherence for tricategories",* Memoirs of the American Math. Society ***117*** *(1995) Number 558.)*

At the 1987 Meeting in Louvain-La-Neuve, Duskin (who loves simplicial sets) found a simplicial set whose only non-trivial homotopy groups were the three in question:

41) *Duskin, John W. "The Azumaya complex of a commutative ring".* Categorical algebra and its applications (Louvain-La-Neuve, 1987)*, 107–117, Lecture Notes in Math.* ***1348****, Springer, Berlin, 1988.*

I pointed out to Jack that this was the nerve of $\mathsf{Alg}(R\text{-}\mathsf{Mod})$ *and he included a remark about that in the published version. Also see*

42) *Duskin, J. "An outline of a theory of higher-dimensional descent". Actes du Colloque en l'Honneur du Soixantieme Anniversaire de Rene Lavendhomme (Louvain-la-Neuve, 1989).* Bull. Soc. Math. Belg. Ser. A ***41*** *(1989), no. 2, 249–277.*

The Brauer group section of

43) *"Categorical and combinatorial aspects of descent theory",* Applied Categorical Structures *(to appear; March 2003 preprint available at `math.CT/0303175`).*

gives some more on this.

The article

44) *K.K. Ulbrich, "Group cohomology for Picard categories",* J. Algebra ***91*** *(1984) 464–498.*

should also be mentioned. It is a great, to use your term, "categorification" of usual cohomology with abelian group coefficients: one step towards the grander goal of coefficients in a general weak n*-category.*

The Spanish School (and the Belgian School) is continuing with nice work in this area. For example there is the recent paper by Carrasco/Martinez-Moreno. Here is the review I wrote yesterday.

> *Carrasco/Martinez-Moreno: Simplicial cohomology with coefficients in symmetric categorical groups*

> *The full cohomology theory of simplicial sets with coefficients in a general weak n-category is a long-term goal. The classical cohomology revolves around the fact that an abelian group A can be regarded as an n-category whose simplicial nerve is the combinatorial Eilenberg-Mac Lane space K(A,n). Following Takeuchi and Ulbrich [*J. Pure Appl. Algebra ***27*** *(1983) 61–73; MR84g:18025] and Ulbrich [*J. Algebra ***91*** *(1984) 464–498; MR86h:18003], the present authors develop cohomology where the coefficient object is a symmetric categorical group A. In this important case too, A can be regarded as a weak n-category whose simplicial nerve is here denoted by $K(A, n)$; it has non-vanishing homotopy groups only in dimensions n and $n + 1$, and represents the cohomology of simplicial sets in the homotopy category. This functor $K(-, n)$ essentially has a left-adjoint left-inverse P_n so that homotopy classes of simplicial maps from X to Y are classified by the cohomology of X with coefficients in $P_n(Y)$.*

Back to marking papers.

Best wishes,
Ross

This last paper is:

44) P. Carrasco and J. Martinez-Moreno, "Simplicial cohomology with coefficients in symmetric categorical groups", *Applied Categorical Structures* **12** (2004), 257–286.

Week 210

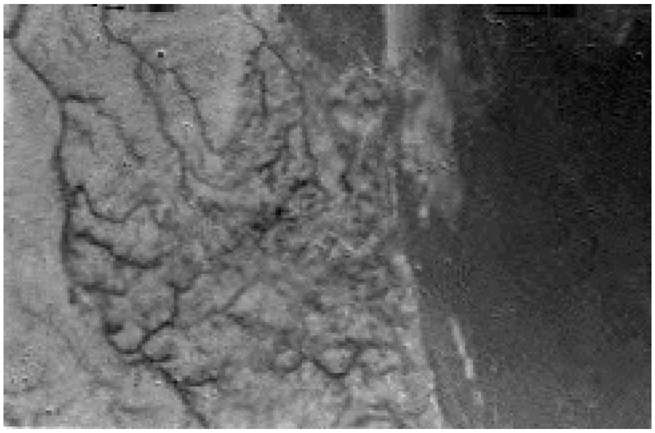
January 25, 2005

As you've probably heard, the Huygens probe has successfully landed on Saturn's moon Titan and is sending back pictures:

1) Huygens Probe Descent, `http://saturn.jpl.nasa.gov/news/events/huygensDescent/index.cfm`

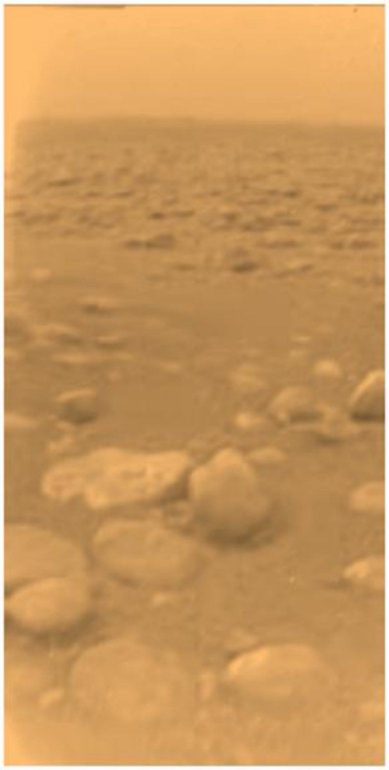
Titan averages a chilly -180 degrees Celsius, and its smoggy orange atmosphere is thicker than the Earth's, mostly nitrogen but 6 percent methane, together with substantial traces of all sorts of other hydrocarbons. The orange color may come from "tholins": polymers made by irradiating a mix of nitrogen and methane. Some other icy moons in the outer solar system are covered with this goop, but Titan is the only moon in the Solar System to have a substantial atmosphere. It even has clouds.

As the Huygens probe parachuted to the surface, it photographed what look like twisty riverbeds flowing into a large lake:



People have long suspected that Titan has lakes of made of methane and/or ethane, but now we may be seeing them. And when Huygens landed, its sensors reported that it broke through a crusty surface and sunk about 20 centimeters into something mushy: probably methane mud!

The first color photo of the surface looks disappointingly like Mars at first sight:



But, the surface is pumpkin-colored due to tholins or something, not rust red. The sky is orange too! The “rocks” could be water ice. And they’ve detected hints of volcanos that spew molten water and ammonia! So, it’s a strange new world.

Back here on Earth, there was a conference in December in honor of Larry Breen’s 60th birthday:

2) Arithmetic, Geometry and Topology: Conference on occasion of Larry Breen’s sixtieth birthday, `http://www-math.univ-paris13.fr/~lb2004/`

It was in Paris. This was my first visit to that city, but luckily I got to stay there an extra week after the conference, so I could focus on the math while it was going on.

But I can’t resist a digression! Paris won my heart, despite my suspicions that it had somehow been hyped all along. First of all, it’s beautiful. Second, it’s nice to be someplace where people take simple foods like bread, cheese, fruits and vegetables really seriously, and don’t settle for the tasteless crud we so often eat in the US.

None of this came as a surprise, of course. What surprised is that I’ve never seen a city with so many bookstores — and good ones, too! They’re clustered thick near the Sorbonne, but the Latin Quarter is dotted with them, and there are even lots along the

Boulevard St-Germain, which is the biggest most fashionable shopping street. I don't think there's any place in the English speaking world with so many bookstores. Not London, not New York... Cambridge Massachusetts used to have lots near Harvard Square, back when I was a grad student, but the high rents have long since squeezed them out, replacing bohemian diversity with clothing shops for boring rich people, like Abercrombie and Fitch. Somehow in Paris fancy clothing and books coexist.

Umm, but what about the conference?

Well, Breen's work is mainly on algebraic geometry a la Grothendieck, with a strong emphasis on category theory. Beautiful stuff, and lately it's it's begun to find applications to string theory — especially his work on gerbes. People at his conference spoke on all sorts of topics, most of which I didn't understand very well — some heavy-duty number theory, for example. I understood a few well enough to really enjoy them, like Mike Hopkins' talk on derived algebraic geometry, Clemens Berger's talk on geometric Reedy categories, and Ieke Moerdijk's talk on the homotopy theory of operads. But I won't try to explain these — I want to explain what a "gerbe" is, so I have my work cut out for me.

One way to get going on the idea of gauge theory is to start with electromagnetism, where the concept of "phase" turns out to play a crucial role. If you move a charged particle through an electromagnetic field, its wavefunction gets multiplied by a unit complex number, or "phase" — and it turns out, rather wonderfully, that all effects of electricity and magnetism on charged particles is due to this!

However, phases are funny. You can't actually measure the phase of a charged particle — at least, there's no such thing as a "phasometer" where you stick in a particle and the dial on the meter points to a unit complex number. Of course a unit complex number is just a fancy name for a point on the circle, and a dial is precisely the right shape for that... but you just can't build this machine.

Instead, you can only measure the *change* in phase of a particle as it goes around a loop. Or, equivalently, the *difference* in phases when a particle takes two different paths from here to there. See, in quantum mechanics you can play tricks like the "double slit experiment", where you coax a particle's wavefunction to smear out and take two routes from here to there... and then when it arrives, it interferes with itself, and if you're smart you can see by these interference effects what the relative phase of the two paths is.

Pretty weird, eh? I'm so used to this that it seems completely normal to me, but I should admit that this way of understanding the electromagnetic field came fairly late. Weyl had a hint of it in 1918 when he invented the term "gauge theory" in his quest to unify electromagnetism and gravity, but he was mixed up in some crucial ways that only got sorted out quite a bit later. For more details, try O'Raiferteagh's book "The Dawning of Gauge Theory", which I discussed in "Week 116".

Anyway, the concept of relative phase, or difference in phase, is nicely captured by the concept of a "torsor". A unit complex number is a point on the unit circle in the complex plane. This circle is a group since we can multiply unit complex numbers and get unit complex numbers back. This group is called $\mathrm{U}(1)$. Like a dial, $\mathrm{U}(1)$ has standard names for all the points on it — and it has one god-given special point, the identity element, namely the number 1.

A "$\mathrm{U}(1)$-torsor" is a lot like $\mathrm{U}(1)$, but subtly different. It's a circle where the points aren't given these standard names... but where you can still tell measure angles, and tell the difference between clockwise and counterclockwise.

You can't get an element of $\mathrm{U}(1)$ from *one* point on a $\mathrm{U}(1)$-torsor. But, if you have *two* points on a $\mathrm{U}(1)$-torsor, you can say how much rotation it takes to get from one to the other, and this give an element of $\mathrm{U}(1)$. In other words, you can describe the "difference in phase" between these two points.

For more on torsors, try this:

3) John Baez, "Torsors made easy", `http://math.ucr.edu/home/baez/torsors.html`

Anyway, the real idea behind electromagnetism is that sitting over each point in spacetime is a $\mathrm{U}(1)$-torsor. If a particle is sitting at some point in spacetime, its phase is not really a number: it's an element of the $\mathrm{U}(1)$-torsor sitting over that point! To get a *number*, we have to carry the particle around a loop! Its phase will change when we do this, so we get *two* points in a $\mathrm{U}(1)$-torsor, and their difference is an element of $\mathrm{U}(1)$.

So while it sounds far-out, the key mathematical structure in electromagnetism is a bunch of $\mathrm{U}(1)$-torsors, one over each point in spacetime. This is called a "principal $\mathrm{U}(1)$-bundle" or sometimes just a "$\mathrm{U}(1)$-bundle" for short.

If we wanted to describe some force other than electromagnetism, we could take this whole setup and replace $\mathrm{U}(1)$ with some other group. In fact, this idea works great: it's the main idea behind gauge theories, which do an excellent job of describing all the forces in nature.

To set up a gauge theory, the first thing you need to do is pick a group G and pick a "principal G-bundle" over spacetime. Spacetime will be some manifold X. A principal G-bundle over X is gadget that assigns a G-torsor to each point of X. A G-torsor is a space where if you pick two points in it, you get an element of G which describes their "difference".

I'm being fairly sloppy here, so don't take these as precise definitions! I give a precise definition of a G-torsor in the above webpage, and any decent book on differential geometry will give you a definition of a principal G-bundle. However, only rather highbrow books define principal G-bundles with the help of G-torsors... which is sad, because it's not that hard, and rather enlightening.

Anyway, in gauge theory the forces of nature are described by "connections" on principal G-bundles. Let's say we have a principal G-bundle P which assigns to each point x of our manifold a G-torsor $P(x)$. Then a "connection" on P is a gadget that says how any path from x to y gives a map from $P(x)$ to $P(y)$. If G is $\mathrm{U}(1)$, for example, this gadget says how the phase of a charged particle changes as we move it along any path from x to y.

Now suppose we have a loop that starts and ends at x. Then our connection gives a map from $P(x)$ to itself. If we start with a point in $P(x)$, and apply this map, we get another point in $P(x)$. Since $P(x)$ is a G-torsor, these two points determine an element of G. This is how we get group elements from loops in gauge theory!

Now let me sketch how gerbes enter the game. First I'll do the case where the group G is abelian, for example $\mathrm{U}(1)$. It's the nonabelian gerbes that really interest me... but the abelian case is a lot easier. The reason is that when G is abelian, the group element we get in the previous paragraph doesn't depend on the choice of a point of $P(x)$.

Gerbes show up when we try to invent a kind of "higher gauge theory" that describes how not just point particles but 1-dimensional objects transform when you move them around. For example, the strings in string theory, or the loops in loop quantum gravity.

This leads to a mind-boggling self-referential twist, which is just the kind of thing I love:

As we've seen, a connection describes how a point particle transforms when you carry it along a path:

$$x \xrightarrow{f} y$$

a path f from the point x to the point y;
we write this as $f\colon x \to y$.

Now we need a gadget that'll describe how a *path* transforms when you carry it along a *path of paths:*

$$\begin{tikzcd}
x \arrow[r, bend left=50, "f"{name=U}] \arrow[r, bend right=50, "g"'{name=D}] & y \arrow[Rightarrow, from=U, to=D, "F"]
\end{tikzcd}$$

a path f from the point x to the point y;
we write this as $f\colon x \to y$.

To do this, we need to boost our level of thinking a notch, working not with "G-torsors" and "principal G-bundles" but instead with "G-2-torsors" and "G-gerbes".

Here's how it goes:

We start by picking an abelian group G and a manifold X.

Then we pick a "G-gerbe" over X, say P.

What's that? It's a thing that assigns to each point x of X a "G-2-torsor", say $P(x)$.

What's that? Well, it's a thing where if you pick two points in it, you get a *G-torsor* describing their difference!

Get it? This is the beginning of a story that goes on forever:

- Two points in a G-torsor determine an element of G;
- two points in a G-2-torsor determine a G-torsor;
- two points in a G-3-torsor determine a G-2-torsor;
- ...

But, you'll probably be relieved to know we won't go beyond G-2-torsors today.

Next, we pick a "connection" on P.

What's that? Well, it's a gadget that for each path from x to y gives us a map from the G-2-torsor $P(x)$ to the G-2-torsor $P(y)$. If we call the path

$$f\colon x \to y$$

then we call this map

$$P(f)\colon P(x) \to P(y)$$

Moroever, this sort of connection also gives a "map between maps" for each path-of-paths! So, from

$$F\colon f \Rightarrow g$$

it gives

$$P(F)\colon P(f) \Rightarrow P(g)$$

I haven't explained enough stuff to say yet what these "maps between maps" are, so let's just see what happens if we have a loop

$$f\colon x \to x$$

and then a loop-of-loops

$$F\colon f \Rightarrow f$$

From the loop $f\colon x \to x$, our connection gives us a map:

$$P(f)\colon P(x) \to P(x)$$

If we start with a point in $P(x)$, and apply this map, we get another point in $P(x)$. Since $P(x)$ is a G-2-torsor, these two points determine a G-torsor. This G-torsor doesn't depend on our initial choice of point, and it completely determines the map $P(f)$. So, we can think of $P(f)$ as *being* this G-torsor, if we like.

From the loop-of-loops $F\colon f \Rightarrow f$, our connection gives us a map:

$$P(F)\colon P(f) \Rightarrow P(f)$$

If we start with a point in $P(f)$, and apply this map, we get another point. Since $P(f)$ is a G-torsor, these two points determine an element of G. This element of G doesn't depend on our initial choice of point, and it completely determines the map $P(F)$. So, we can think of $P(F)$ as *being* this element of G, if we like.

In short, the machinery functions just as you'd hope, giving a group element that describes how a loop of string "changes phase" as you carry it around a loop-of-loops!

So far I've been strenously avoiding the language of categories and 2-categories, but if you're at all familiar with that language, you'll have guessed that it's precisely what we need to make everything I'm saying precise.

It's actually incredibly beautiful... but I'm getting lazy, so I'll explain it very tersely now, in a way that only true lovers of abstraction will enjoy:

If G is a group, it acts on itself by left translation. So, it becomes a left G-set. Any left G-set isomorphic to this one is called a "G-torsor". There's a category G-Tor whose objects are G-torsors and whose morphisms are maps compatible with the action of G. Since all G-torsors are isomorphic, and the automorphism group of any one is just G, this category G-Tor is equivalent to G (regarded as a 1-object category).

If G is abelian, every left G-set becomes a right G-set too. This allows us to define a "tensor product" of G-sets. The tensor product of G-torsors is a G-torsor, so G-Tor becomes a monoidal category. In fact, it's a "2-group": a monoidal category where all the objects and morphisms are invertible.

This allows us to iterate what we've just done:

Since G-Tor is a 2-group, it acts on itself by left translation. So, it becomes a "left G-category". Any left G-category isomorphic to this one is called a "G-2-torsor". There's a 2-category G-2-Tor whose objects are G-2-torsors, whose morphisms are functors compatible the action of G, and whose morphisms are natural transformations compatible with the action of G. Since all G-2-torsors are isomorphic, any the automorphism 2-group of any one is just G-Tor, this 2-category is equivalent to G-Tor (regarded as a 1-object 2-category).

And so on! This infinite hierarchy only works when G is abelian; when G is nonabelian we need a different hierarchy, which uses "bitorsors", where G acts on both left and right, instead of "torsors".

To learn more about this stuff, here are some references. I'll stick to ones I didn't already list in "Week 71" and "Week 151".

First, for physicists, some work on the role of gerbes and 2-gerbes in string theory and M-theory:

4) Paolo Aschieri, Luigi Cantini and Branislav Jurco, "Nonabelian bundle gerbes, their differential geometry and gauge theory", available as hep-th/0312154.

5) Paolo Aschieri and Branislav Jurco, "Gerbes, M5-brane anomalies and E_8 gauge theory", available as hep-th/0409200.

Second, for mathematicians, some classic works by Breen:

6) Lawrence Breen, "Bitorseurs et cohomologie non-abelienne", in *The Grothendieck Festschrift*, eds. P. Cartier et al, Progress in Mathematics vol. **86**, Birkhauser, Boston, 1990, pp. 401–476.

7) Lawrence Breen, "Theorie de Schreier superieure", *Ann. Sci. Ecole Norm. Sup.* **25** (1992), 465–514.

8) Lawrence Breen, "Classification of 2-stacks and 2-gerbes", *Asterisque* **225**, Societe Mathematique de France, 1994.

2-gerbes are what you get if you climb the hierarchy one more step. They should be good for describing the parallel transport of 2-dimensional surfaces, or "2-branes" — and indeed they make an appearance in Aschieri and Jurco's paper for precisely that reason.

Another key reference is Breen's paper with Messing about connections on nonabelian gerbes:

9) Lawrence Breen and William Messing, "The differential geometry of gerbes", available as math.AG/0106083.

and Breen's lecture notes from the IMA workshop on higher categories:

10) Larry Breen, "n-Stacks and n-gerbes: homotopy theory". Notes available at http://www.ima.umn.edu/categories/#thur

I've been working on this stuff myself lately, from a somewhat different viewpoint. So far I've written papers with Aaron Lauda and Alissa Crans about 2-groups and Lie 2-algebras:

11) John Baez and Aaron Lauda, "Higher-dimensional algebra V: 2-groups", *Theory and Applications of Categories* **12** (2004), 423–491. Available online at http://www.tac.mta.ca/tac/volumes/12/14/12-14abs.html or as math.QA/0307200.

12) John Baez and Alissa Crans, "Higher-dimensional algebra VI: Lie 2-algebras", *Theory and Applications of Categories* **12** (2004), 492–528. Available online at http://www.tac.mta.ca/tac/volumes/12/15/12-15abs.html or as math.QA/0307263.

Aaron Lauda was getting a masters degree in physics at UCR when we started our paper on 2-groups. Now he's a grad student in math at the University of Cambridge, working on things related to topological quantum field theory with the category theorist Martin Hyland. Alissa Crans did her PhD in math at UCR, and our paper on Lie 2-algebras contains a lot of stuff from her thesis. Now she has a job at Loyola Marymount University, in Los Angeles.

I've had a huge amount of fun working with both of them! Luckily Alissa lives nearby, and I visit Cambridge most summers. So, we can all keep working on other projects together — and we are.

I also have some gerbe-related projects going on with my grad student Toby Bartels, Danny Stevenson (who is teaching at UCR now) and Urs Schreiber, a fellow moderator of sci.physics.research who will soon be a postdoc at Hamburg with Christoph Schweigert. Urs will be visiting UCR for two weeks in February, and we plan to figure a lot of stuff out. So, I've got gerbes on the brain, and I'll probably be saying more about them in the future, unless I burn up all my expository energy writing papers.

In fact, one of the best places to learn about the differential geometry of abelian gerbes and 2-gerbes is Danny's thesis:

13) Danny Stevenson, *The geometry of bundle gerbes*, Ph.D. thesis, University of Adelaide, 2000. Available as math.DG/0004117.

He's also written lots of other papers on gerbes, which you can find on the arXiv. Physicists may find these the most interesting:

14) Michael K. Murray and Danny Stevenson, "Higgs fields, bundle gerbes and string structures", *Comm. Math. Phys.* **236** (2003), 541–555. Also available as math.DG/0106179.

15) Alan L. Carey, Stuart Johnson, Michael K. Murray, Danny Stevenson and Bai-Ling Wang, "Bundle gerbes for Chern-Simons and Wess-Zumino-Witten models", available as math.DG/0410013.

Toby is doing his thesis on categorified bundles, or "2-bundles", and you can already get a preview here:

16) Toby Bartels, "Categorified gauge theory: 2-bundles", available as math.CT/0410328.

2-bundles are meant to be an alternative to gerbes: although I've done my best to hide it above, a gerbe is really more like a categorified *sheaf* than a bundle. And, just as a bundle has a sheaf of sections, we're hoping that a 2-bundle has a stack of sections, which in certain cases will be a gerbe. That's one of the things we need to figure out, though.

And, while I'm listing the papers of the gerbe gang, I should admit that Urs and I have written a paper about connections on 2-bundles. But, I want to polish this paper a bit before talking about it here.

As for 2-groups, various people have been studying their representations lately, and this should become an important part of higher gauge theory, just as group representations are crucial in gauge theory:

17) Magnus Forrester-Barker, *Representations of crossed modules and cat^1-groups*, Ph.D. thesis, Department of Mathematics, University of Wales, Bangor, 2004. Available at `http://www.maths.bangor.ac.uk/research/ftp/theses/forrester-barker.pdf`

18) John Barrett and Marco Mackaay, "Categorical representations of categorical groups", available as `math.CT/0407463`.

19) Josep Elgueta, "Representation theory of 2-groups on finite dimensional 2-vector spaces", available as `math.CT/0408120`.

20) Louis Crane and David Yetter, "Measurable categories and 2-groups", available as `math.QA/0305176`.

21) David Yetter, "Measurable categories", available as `math.CT/0309185`.

22) Louis Crane and Marnie D. Sheppeard, "2-categorical Poincare representations and state sum applications", available as `math.QA/0306440`.

Hendryk Pfeiffer's papers on higher gauge theory are also very interesting. Since he works on lattice gauge theory and spin foam models, the first two papers here develop higher gauge theory on a discrete spacetime, and then compare it to higher gauge theory on a manifold:

23) Hendryk Pfeiffer, "Higher gauge theory and a non-Abelian generalization of 2-form electrodynamics", *Annals Phys.* **308** (2003), 447–477. Also available as `hep-th/0304074`.

24) Florian Girelli and Hendryk Pfeiffer, "Higher gauge theory — differential versus integral formulation", *Jour. Math. Phys.* **45** (2004), 3949–3971. Also available as `hep-th/0309173`.

25) Hendryk Pfeiffer, "2-groups, trialgebras and their Hopf categories of representations", available as `math.QA/0411468`.

The third one partially fulfills an old dream of Crane and Frenkel — a dream I vaguely hinted at way back in "Week 50". Their dream was to find a concept of "trialgebra" such that a trialgebra has a Hopf category of representations, which in turn can have a monoidal 2-category of representations of its own. This is a kind of aggravated version of a pattern already familiar in algebra, where a Hopf algebra (or bialgebra) has a monoidal category of representations.

Pfeiffer doesn't define general trialgebras, but only "cocommutative trialgebras" and "commutative cotrialgebras". A cocommutative trialgebra is a category in the category of cocommutative Hopf algebras, while a commutative cotrialgebra is a category in the opposite of the category of commutative Hopf algebras. Zounds — say that three times fast!

He shows you can get these two gadgets from 2-groups in analogy to how you get cocommutative or commutative Hopf algebras from groups, by taking the group algebra or the algebra of functions on a group. He also proves a Tannaka- Krein theorem that lets you reconstruct commutative cotrialgebras from their Hopf categories of representations.

Really cool stuff!

By the way, here are some photos of Larry Breen's conference, and of Paris:

26) John Baez, Paris, `http://math.ucr.edu/home/baez/paris/`

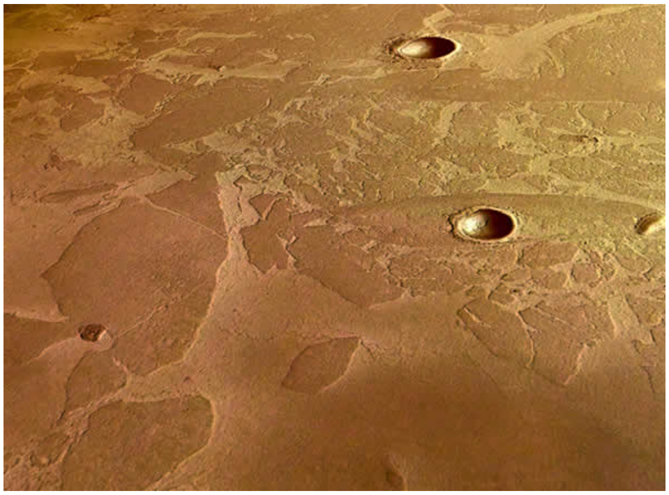
Week 211

March 6, 2005

The last time I wrote an issue of this column, the Huyghens probe was bringing back cool photos of Titan. Now the European "Mars Express" probe is bringing back cool photos of Mars!

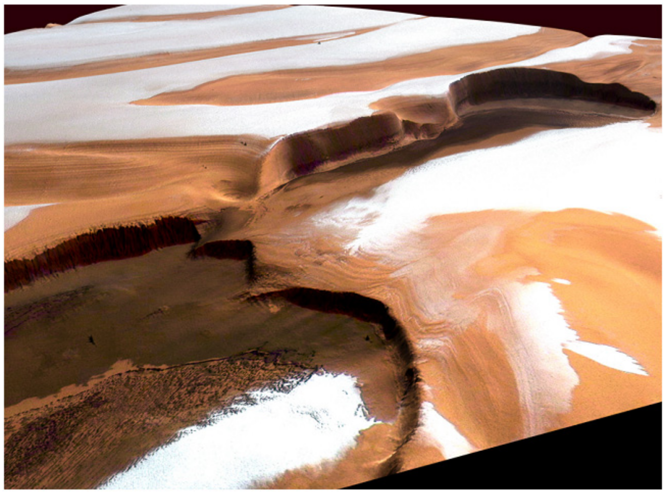
1) Mars Express website, `http://www.esa.int/SPECIALS/Mars_Express/index.html`

There are some tantalizing pictures of what might be a "frozen sea" — water ice covered with dust — near the equator in the Elysium Planitia region:



2) Mars Express sees signs of a "frozen sea", `http://www.esa.int/SPECIALS/Mars_Express/SEMCHPYEM4E_0.html`

Ice has already been found at the Martian poles — it's easily visible there, and Mars Express is getting some amazing closeups of it now - here's a here's a view of some ice on sand at the north pole:



3) Glacial, volcanic and fluvial activity on Mars: latest images, `http://www.esa.int/SPECIALS/Mars_Express/SEMLF6D3M5E_1.html`

What's new is the possibility of large amounts of water in warmer parts of the planet.

Now for some math. It's always great when two subjects you're interested in turn out to be bits of the same big picture. That's why I've been really excited lately about Bott periodicity and the "super-Brauer group".

I wrote about Bott periodicity in "Week 105", and about the Brauer group in "Week 209", but I should remind you about them before putting them together.

Bott periodicity is all about how math and physics in n+8-dimensional space resemble math and physics in n-dimensional space. It's a weird and wonderful pattern that you'd never guess without doing some calculations. It shows up in many guises, which turn out to all be related. The simplest one to verify is the pattern of Clifford algebras.

You're probably used to the complex numbers, where you throw in just *one* square root of -1, called i. And maybe you've heard of the quaternions, where you throw in *two* square roots of -1, called i and j, and demand that they anticommute:

$$ij = -ji$$

This implies that $k = ij$ is another square root of -1. Try it and see!

In the late 1800s, Clifford realized there's no need to stop here. He invented what we now call the "Clifford algebras" by starting with the real numbers and throwing in n square roots of -1, all of which anticommute with each other. The result is closely related to rotations in $n + 1$ dimensions, as I explained in "Week 82".

I'm not sure who first worked out all the Clifford algebras — perhaps it was Cartan — but the interesting fact is that they follow a periodic pattern. If we use C_n to stand for the Clifford algebra generated by n anticommuting square roots of -1, they go like this:

n	C_n
0	$\mathbb{R}$
1	$\mathbb{C}$
2	$\mathbb{H}$
3	$\mathbb{H} \oplus \mathbb{H}$
4	$\mathbb{H}(2)$
5	$\mathbb{C}(4)$
6	$\mathbb{R}(8)$
7	$\mathbb{R}(8) \oplus \mathbb{R}(8)$

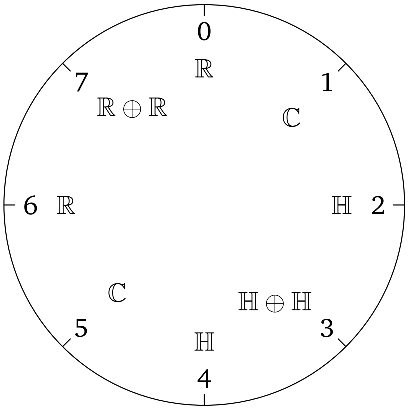
where:

- $\mathbb{R}(n)$ means $n \times n$ real matrices,
- $\mathbb{C}(n)$ means $n \times n$ complex matrices, and
- $\mathbb{H}(n)$ means $n \times n$ quaternionic matrices.

All these become algebras with the usual addition and multiplication of matrices. Finally, if A is an algebra, $A \oplus A$ consists of pairs of guys in A, with pairwise addition and multiplication.

What happens next? Well, from then on things sort of "repeat" with period 8: C_{n+8} consists of 16×16 matrices whose entries lie in C_n!

So, you can remember all the Clifford algebras with the help of this eight-hour clock:



To use this clock, you have to remember to use matrices of the right size to get C_n to have dimension 2^n. So, when I write "$\mathbb{R} \oplus \mathbb{R}$" next to the "7" on the clock, I don't mean C_7 is really $\mathbb{R} \oplus \mathbb{R}$. To get C_7, you have to take $\mathbb{R} \oplus \mathbb{R}$ and beef it up until it becomes an algebra of dimension $2^7 = 128$. You do this by taking $\mathbb{R}(8) \oplus \mathbb{R}(8)$, since this has dimension $8 \times 8 + 8 \times 8 = 128$.

Similarly, to get C_{10}, you note that 10 is 2 modulo 8, so you look at "2" on the clock and see "$\mathbb{H}$" next to it, meaning the quaternions. But to get C_{10}, you have to take $\mathbb{H}$ and beef it up until it becomes an algebra of dimension $2^{10} = 1024$. You do this by taking $\mathbb{H}(16)$, since this has dimension $4 \times 16 \times 16 = 1024$.

This "beefing up" process is actually quite interesting. For any associative algebra A, the algebra $A(n)$ consisting of $n \times n$ matrices with entries in A is a lot like A itself. The reason is that they have equivalent categories of representations!

To see what I mean by this, remember that a "representation" of an algebra is a way for its elements to act as linear transformations of some vector space. For example, $\mathbb{R}(n)$ acts as linear transformations of $\mathbb{R}^n$ by matrix multiplication, so we say $\mathbb{R}(n)$ has a representation on R^n. More generally, for any algebra A, the algebra $A(n)$ has a representation on A^n.

More generally still, if we have any representation of A on a vector space V, we get a representation of $A(n)$ on V^n. It's less obvious, but true, that *every* representation of $A(n)$ comes from a representation of A this way.

In short, just as $n \times n$ matrices with entries in A form an algebra $A(n)$ that's a beefed-up version of A itself, every representation of $A(n)$ is a beefed-up version of some representation of A.

Even better, the same sort of thing is true for maps between representations of $A(n)$. This is what we mean by saying that $A(n)$ and A have equivalent categories of representations. If you just look at the categories of representations of these two algebras as abstract categories, there's no way to tell them apart! We say two algebras are "Morita equivalent" when this happens.

It's fun to study Morita equivalence classes of algebras — say algebras over the real numbers, for example. The tensor product of algebras gives us a way to multiply these classes. If we just consider the invertible classes, we get a *group*. This is called the "Brauer group" of the real numbers.

The Brauer group of the real numbers is just $\mathbb{Z}/2$, consisting of the classes $[\mathbb{R}]$ and $[\mathbb{H}]$. These correspond to the top and bottom of the Clifford clock! Part of the reason is that

$$\mathbb{H} \otimes \mathbb{H} = \mathbb{R}(4)$$

so when we take Morita equivalence classes we get

$$[\mathbb{H}] \times [\mathbb{H}] = [\mathbb{R}]$$

But, you may wonder where the complex numbers went! Alas, the Morita equivalence class $[\mathbb{C}]$ isn't invertible, so it doesn't live in the Brauer group. In fact, we have this little multiplication table for tensor product of algebras:

$\otimes$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{H}$
$\mathbb{R}$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{H}$
$\mathbb{C}$	$\mathbb{C}$	$\mathbb{C} \oplus \mathbb{C}$	$\mathbb{C}(2)$
$\mathbb{H}$	$\mathbb{H}$	$\mathbb{C}(2)$	$\mathbb{R}(4)$

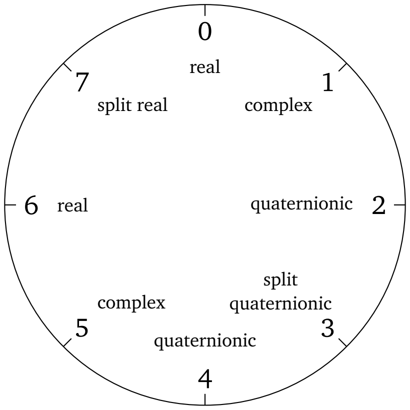
Anyone with an algebraic bone in their body should spend an afternoon figuring out how this works! But I won't explain it now.

Instead, I'll just note that the complex numbers are very aggressive and infectious — tensor anything with a $\mathbb{C}$ in it and you get more $\mathbb{C}$'s. That's because they're a field in their own right — and that's why they don't live in the Brauer group of the real numbers.

They do, however, live in the *super-Brauer* group of the real numbers, which is $\mathbb{Z}/8$

— the Clifford clock itself!

But before I explain that, I want to show you what the categories of representations of the Clifford algebras look like:



Here, the labels at each hour describe the type of vector space, e.g. at 3-o'clock we have split quaternionic vector spaces.

You can read this information off the 8-hour Clifford clock I showed you before, at least if you know some stuff:

- A real vector space is just something like $\mathbb{R}^n$
- A complex vector space is just something like $\mathbb{C}^n$
- A quaternionic vector space is just something like $\mathbb{H}^n$

and a "split" vector space is a vector space that's been written as the direct sum of two subspaces.

Take C_4, for example — the Clifford algebra generated by 4 anticommuting square roots of -1. The Clifford clock tells us this is $\mathbb{H} \oplus \mathbb{H}$. And if you think about it, a representation of this is just a pair of representations of $\mathbb{H}$. So, it's two quaternionic vector spaces — or if you prefer, a "split" quaternionic vector space.

Or take C_7. The Clifford clock says this is $\mathbb{R} \oplus \mathbb{R}$... or at least Morita equivalent to $\mathbb{R} \oplus \mathbb{R}$: it's actually $\mathbb{R}(8) \oplus \mathbb{R}(8)$, but that's just a beefed-up version of $\mathbb{R} \oplus \mathbb{R}$, with an equivalent category of representations. So, the category of representations of C_7 is *equivalent* to the category of split real vector spaces.

And so on. Note that when we loop all the way around the clock, our Clifford algebra becomes 16×16 matrices of what it was before, but this is Morita equivalent to what it was. So, we have a truly period-8 clock of categories!

But here's the really cool part: there are also arrows going clockwise and counterclockwise around this clock! Arrows between categories are called "functors".

Each Clifford algebra is contained in the next one, since they're built by throwing in more and more square roots of -1. So, if we have a representation of C_n, it gives us a representation of C_{n-1}. Ditto for maps between representations. So, we get a functor from the category of representations of C_n to the category of representations of C_{n-1}.

This is called a “forgetful functor”, since it “forgets” that we have representations of C_n and just thinks of them as representations of C_{n-1}.

So, we have forgetful functors cycling around counterclockwise!

Even better, all these forgetful functors have “left adjoints” going back the other way. I talked about left adjoints in “Week 77”, so I won’t say much about them now. I’ll just give an example.

Here’s a forgetful functor:

$$\text{complex vector spaces} \xrightarrow{\text{forget complex structure}} \text{real vector spaces}$$

which is one of the counterclockwise arrows on the Clifford clock. This functor takes a complex vector space and forgets your ability to multiply vectors by i, thus getting a real vector space. When you do this to $\mathbb{C}^n$, you get $\mathbb{R}^{2n}$.

This functor has a left adjoint:

$$\text{complex vector spaces} \xleftarrow{\text{complexify}} \text{real vector spaces}$$

where you take a real vector space and “complexify” it by tensoring it with the complex numbers. When you do this to $\mathbb{R}^n$, you get $\mathbb{C}^n$.

So, we get a beautiful version of the Clifford clock with forgetful functors cycling around counterclockwise and their left adjoints cycling around clockwise! When I realized this, I drew a big picture of it in my math notebook — I always carry around a notebook for precisely this sort of thing. Unfortunately, it’s a bit hard to draw this chart in ASCII, so I won’t include it here.

Instead, I’ll draw something easier. For this, note the following mystical fact: the Clifford clock is symmetrical under reflection around the 3-o’clock/7-o’clock axis. It seems bizarre at first that it’s symmetrical along *this* axis instead of the more obvious 0-o’clock/4-o’clock axis. But there’s a good reason, which I already mentioned: the Clifford algebra C_n is related to rotations in $n+1$ dimensions.

I would be very happy if you had enough patience to listen to a full explanation of this fact, along with everything else I want to say. But I bet you don’t. . . so I’ll hasten on to the really cool stuff.

First of all, using this symmetry we can fold the Clifford clock in half. . . and the forgetful functors on one side perfectly match their left adjoints on the other side!

So, we can save space by drawing this "folded" Clifford clock:

$$\begin{tikzcd}
\text{split real vector spaces} \arrow[d, shift right, "\text{forget splitting}"'] \\
\text{real vector spaces} \arrow[u, shift right, "\text{double}"'] \arrow[d, shift right, "\text{complexify}"'] \\
\text{complex vector spaces} \arrow[u, shift right, "\text{forget complex structure}"'] \arrow[d, shift right, "\text{quaternionify}"'] \\
\text{quaternionic vector spaces} \arrow[u, shift right, "\text{forget quaternionic structure}"'] \arrow[d, shift right, "\text{double}"'] \\
\text{split quaternionic vector spaces} \arrow[u, shift right, "\text{forget splitting}"']
\end{tikzcd}$$

The forgetful functors march downwards on the right, and their left adjoints march back up on the left!

The arrows going between 7 o'clock and 0 o'clock look a bit weird:

$$\begin{tikzcd}
\text{split real vector spaces} \arrow[d, shift right, "\text{forget splitting}"'] \\
\text{real vector spaces} \arrow[u, shift right, "\text{double}"']
\end{tikzcd}$$

Why is "forget splitting" on the left, where the left adjoints belong, when it's obviously an example of a forgetful functor?

One answer is that this is just how it works. Another answer is that it happens when we wrap all the way around the clock — it's like how going from midnight to 1 am counts as going forwards in time even though the number is getting smaller. A third answer is that the whole situation is so symmetrical that the functors I've been calling "left adjoints" are also "right adjoints" of their partners! So, we can change our mind about which one is "forgetful", without getting in trouble.

But enough of that: I really want to explain how this stuff is related to the super-Brauer group, and then tie it all in to the *topology* of Bott periodicity. We'll see how far I get before giving up in exhaustion. . . .

What's a super-Brauer group? It's just like a Brauer group, but where we use superalgebras instead of algebras! A "superalgebra" is just physics jargon for a $\mathbb{Z}/2$-graded algebra — that is, an algebra A that's a direct sum of an "even" or "bosonic" part A_0 and an "odd" or "fermionic" part A_1:

$$A = A_0 \oplus A_1$$

such that multiplying a guy in A_i and a guy in A_j gives a guy in A_{i+j}, where we add the subscripts $\bmod 2$.

The tensor product of superalgebras is defined differently than for algebras. If A and B are ordinary algebras, when we form their tensor product, we decree that everybody in A commutes with everyone in B. For superalgebras we decree that everybody in A "supercommutes" with everyone in B — meaning that

$$ab = ba$$

if either a or b are even (bosonic) while

$$ab = -ba$$

if a and b are both odd (fermionic).

Apart from these modifications, the super-Brauer group works almost like the Brauer group. We start with superalgebras over our favorite field — here let's use the real numbers. We say two superalgebras are "Morita equivalent" if they have equivalent categories of representations. We can multiply these Morita equivalence classes by taking tensor products, and if we just keep the invertible classes we get a group: the super-Brauer group.

As I've hinted already, the super-Brauer group of the real numbers is $\mathbb{Z}/8$ — just the Clifford algebra clock in disguise!

Here's why:

The Clifford algebras all become superalgebras if we decree that all the square roots of -1 that we throw in are "odd" elements. And if we do this, we get something great:

$$C_n \otimes C_m = C_{n+m}$$

The point is that all the square roots of -1 we threw in to get C_n *anticommute* with those we threw in to get C_m.

Taking Morita equivalence classes, this mean

$$[C_n][C_m] = [C_{n+m}]$$

but we already know that

$$[C_{n+8}] = [C_n]$$

so we get the group $\mathbb{Z}/8$. It's not obvious that this is *all* the super-Brauer group, but it actually is — that's the hard part.

Now let's think about what we've got. We've got the super-Brauer group, $\mathbb{Z}/8$, which looks like an 8-hour clock. But before that, we had the categories of representations of Clifford algebras, which formed an 8-hour clock with functors cycling around in both directions.

In fact these are two sides of the same coin — or clock, actually. The super-Brauer group consists of Morita equivalence classes of Clifford algebras, where Morita equivalence means "having equivalent categories of representations". But, our previous clock just shows their categories of representations!

This suggests that the functors cycling around in both directions are secretly an aspect of the super-Brauer group. And indeed they are! The functors going clockwise are just "tensoring with C_1", since you can tensor a representation of C_n with C_1 and get a representation of C_{n+1}. And the functors going counterclockwise are "tensoring with

C_{-1}"... or C_7 if you insist, since C_{-1} doesn't strictly make sense, but 7 equals -1 $\bmod 8$, so it does the same job.

Hmm, I think I'm tired out. I didn't even get to the topology yet! Maybe that'll be good as a separate little story someday. If you can't wait, just read this:

4) John Milnor, *Morse Theory*, Princeton U. Press, Princeton, New Jersey, 1963.

You'll see here that a representation of C_n is just the same as a vector space with n different anticommuting ways to "rotate vector by 90 degrees", and that this is the same as a real inner product space equipped with a map from the n-sphere into its rotation group, with the property that the north pole of the n-sphere gets mapped to the identity, and each great circle through the north pole gives some action of the circle as rotations. Using this, and stuff about Clifford algebras, and some Morse theory, Milnor gives a beautiful proof that

$$\Omega^8(\mathrm{SO}(\infty)) \sim \mathrm{SO}(\infty)$$

or in English: the 8-fold loop space of the infinite-dimensional rotation group is homotopy equivalent to the infinite-dimensional rotation group!

The thing I really like, though, is that Milnor relates the forgetful functors I was talking about to the process of "looping" the rotation group. That's what these maps from spheres into the rotation group are all about... but I want to really explain it all someday!

I learned about the super-Brauer group here:

5) V. S. Varadarajan, *Supersymmetry for Mathematicians: An Introduction*, American Mathematical Society, Providence, Rhode Island, 2004.

though the material here on this topic is actually a summary of some lectures by Deligne in another book I own:

6) P. Deligne, P. Etingof, D.S. Freed, L. Jeffrey, D. Kazhdan, J. Morgan, D.R. Morrison and E. Witten, *Quantum Fields and Strings: A Course For Mathematicians* 2 vols., American Mathematical Society, Providence, 1999. Notes also available at `http://www.math.ias.edu/QFT/`

Varadarajan's book doesn't go as far, but it's much easier to read, so I recommend it as a way to get started on "super" stuff.

Week 212

March 26, 2005

As you may know, theoretical particle physics is highly enamored of "supersymmetry" these days. This is not because there's a shred of experimental evidence for it — there's not — but just because it's such a cool idea from a mathematical point of view. Mathematicians should have gotten this idea and run with it first, but physicists did — and maybe it's turned them into mathematicians.

The unarguable central core of this idea is that everything is made of bosons and fermions. In the Standard Model, most bosons are "force carriers", like photons, which carry the electromagnetic force. Fermions are more like what we'd normally call "matter": leptons and quarks, for example. The one big exception is the Higgs boson, which gives elementary particles their mass and... umm... hasn't been seen yet!

But, at a more fundamental level, the really important thing is that bosons commute:

$$xy = yx$$

while fermions anticommute:

$$xy = -yx$$

Also, in case you're wondering, bosons commute with fermions.

But already, most mathematicians reading this will be confused and unhappy. What does it mean for two particles to commute, much less anticommute? Does an apple commute with a grape? Here in the suburbs of Los Angeles almost everyone commutes, but that's not what we're talking about.

The whole idea of particles commuting or anticommuting only occurred to people after they invented quantum theory, where the state of any system is described by a unit vector in some Hilbert space. In quantum theory, if you have a system in some state x, and you check to see if it's in the state y, your experiment gives you the answer "yes" with probability

$$|\langle x, y\rangle|^2$$

the square of the absolute value of the inner product of x and y.

There! Now you know quantum theory.

Given this setup, when you have a system consisting of two particles, the first in some state x and the second in some state y, it's natural to write the state of the whole system as a kind of product xy. But then you have to figure out what rules you want this product to satisfy!

If you require it to be commutative:

$$xy = yx$$

you're saying that there's no difference between the *first* particle being in state x and the *second* particle being in state y, and the other way around. In other words, the particles don't have little name tags on them saying who they are.

This seems reasonable, and particles satisfying this rule are called "bosons". But, there's another popular option, called "fermions":

$$xy = -yx$$

Here again, the particles don't have name tags, since if we put the whole system in the state xy and check to see if it's in the state yx, we get the same answer as when we check to see if it's in the state xy! See:

$$|\langle xy, yx\rangle|^2 = |\langle xy, -xy\rangle|^2 = |\langle xy, xy\rangle|^2$$

thanks to the absolute value. This means that the states xy and yx are indistinguishable.

Reading what I just said, you'd be forgiven for wondering what's the big difference between fermions and bosons! After all, that absolute value in the formula for probabilities just ignores minus signs.

One difference is the "Pauli exclusion principle". Take a pair of fermions and check to see if they're both in the state x. The probability is always zero, since

$$xx = -xx$$

so $xx = 0$. So, fermions are antisocial: that's why the electrons in an atom form "shells" with different electrons in different states, instead of all hanging out at the lowest energy state.

Bosons, by contrast, are gregarious: when a store clerk uses a laser scanner to ring up your purchases, that beam of red light is a bunch of photons all in the same state! A laser is a quintessentially quantum-theoretic gadget — we live in a marvelous world, where such things are taken for granted.

After getting used to these ideas for a while — Bose and Einstein worked out the idea of bosons in 1924, Pauli came up with his exclusion principle in 1925, and Dirac systematized the whole business in 1926 - physicists eventually started looking for symmetries that relate bosons and fermions. *Supersymmetries!* They're not seen in nature, but physicists were looking to see if they're mathematically possible. They turn out not only to be possible, but fascinating.

Formulating supersymmetries in a slick way requires that we take everything we knew about linear algebra and generalize it by letting all our vector spaces have both an "even" or "bosonic" part and an "odd" or "fermionic" part. Mathematically this just amounts to writing our vector space as a direct sum

$$V = V_0 \oplus V_1$$

where V_0 is the "even part" and V_1 is the "odd part". Such a thing is called a "$\mathbb{Z}/2$-graded vector space", or "super vector space".

So far this is pathetically simple. But then — and this is the really crucial part! — whenever we multiply things, we have to follow this rule:

×	**even**	**odd**
even	even	odd
odd	odd	even

It's a little confusing, since this isn't what happens when you multiply even and odd numbers — it's what happens when you ADD them. But, one quickly adapts.

Also, when we generalize equations involving multiplication, we must remember to stick in an extra minus sign whenever we switch two odd vectors.

So, for example, the usual concept of an algebra gets replaced by that of a "superalgebra". This is a super vector space A equipped with an associative product and unit such that when we multiply even and/or odd vectors, the rules in the above table hold. We say a superalgebra is "supercommutative" if

$$xy = yx$$

when at least one of x, y lives in the even part A_0, while

$$xy = -yx$$

when both x and y live in A_1.

Similarly we can define super Lie algebras, super Lie groups, supermanifolds, and so on. . . .

People have done a lot of work on this stuff: it would take me days to explain it all — even longer if I actually knew something about it.

But right now, I just want to zoom in the direction of super division algebras. These are not the most important aspect of "superalgebra" — but they're pretty cool, and Todd Trimble has been explaining them to me lately. Everything interesting I'm about to say is due to him.

As you know, I'm inordinately fond of the normed division algebras: the real numbers, complex numbers, quaternions and octonions. They're so beautiful, it's a little sad at times that there are only four! Could superalgebra allow for more?

YES! And, they turn out to be related to Bott periodicity.

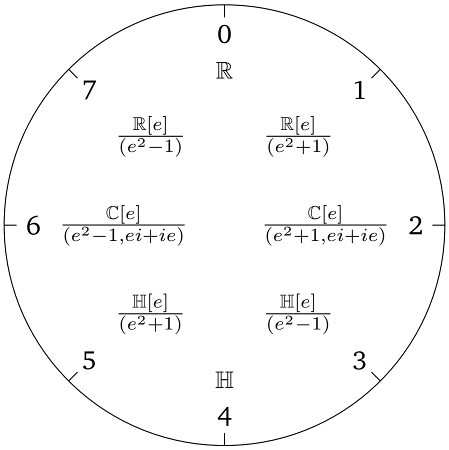
Nobody seems to have pondered *nonassociative* super division algebras yet, but Deligne has a nice article about the associative ones, which I mentioned in "Week 211". I'll give more references later.

So, what's the idea?

I've already told you what a superalgebra is. We say it's a "super division algebra" if every nonzero element that's purely even or purely odd is invertible.

That's pretty easy. What are they like?

Well, I don't completely understand all the options yet, so I'll just list the "central" super division algebras over the real numbers, namely those where the elements that supercommute with everything form a copy of the real numbers. There turn out to be 8, and their beautiful patterns are best displayed in a circular layout:


0
$\mathbb{R}$
7
1
$\frac{\mathbb{R}[e]}{(e^2-1)}$
$\frac{\mathbb{R}[e]}{(e^2+1)}$
6
$\frac{\mathbb{C}[e]}{(e^2-1,ei+ie)}$
$\frac{\mathbb{C}[e]}{(e^2+1,ei+ie)}$
2
$\frac{\mathbb{H}[e]}{(e^2+1)}$
$\frac{\mathbb{H}[e]}{(e^2-1)}$
5
3
$\mathbb{H}$
4

What does this notation mean? Well, as usual $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{H}$ stand for the reals, complex numbers, and quaternions. In all but two cases, we start with one of those algebras and throw in an odd element "e" satisfying the relations listed: e is either a square root of $+1$ or of -1, and in the complex cases it anticommutes with i.

So, for example, super division algebra number 1:

$$\mathbb{R}[e]/(e^2+1)$$

is just the real numbers with an odd element thrown in that satisfies $e^2+1=0$. In other words, it's just the complex numbers made into a superalgebra in such a way that i is *odd*.

The real reason I've arranged these guys in a circle numbered from 0 to 7 is to remind you of the Clifford algebra clock in "Week 210", where I discussed the super Brauer group of the real numbers, and said it was $\mathbb{Z}/8$.

Indeed, the central super division algebras are a complete set of representatives for this super Brauer group! In particular, the Clifford algebra C_n is super Morita equivalent to the nth algebra on this circle:

n	C_n	super-Morita-equivalent algebra
0	$\mathbb{R}$	$\mathbb{R}$
1	$\mathbb{C}$	$\mathbb{R}[e]/(e^2+1)$
2	$\mathbb{H}$	$\mathbb{C}[e]/(e^2+1, ei+ie)$
3	$\mathbb{H}\oplus\mathbb{H}$	$\mathbb{H}[e]/(e^2-1)$
4	$\mathbb{H}(2)$	$\mathbb{H}$
5	$\mathbb{C}(4)$	$\mathbb{H}[e]/(e^2+1)$
6	$\mathbb{R}(8)$	$\mathbb{C}[e]/(e^2-1, ei+ie)$
7	$\mathbb{R}(8)\oplus\mathbb{R}(8)$	$\mathbb{R}[e]/(e^2-1)$

where the notation for Clifford algebras was explained last week.

I think this is cool. I'm not quite sure what to do with it yet, though. How much of what people ordinarily do with division algebras can be done with super division algebras? For example, can we define projective spaces over super division algebras? (See "Week 106" and "Week 145" for why that would be interesting.)

To read more about this, try:

1) Pierre Deligne, "Notes on spinors", in *Quantum Fields and Strings: A Course For Mathematicians*, volume **1**, American Mathematical Society, Providence, 1999. Also available at `http://www.math.ias.edu/QFT/fall/spinors.ps`

A lot of the ideas go back to here:

2) C. T. C. Wall, "Graded Brauer groups", *J. Reine Angew. Math.* **213** (1963/1964), 187–199.

and here's another good reference:

3) Peter Donovan and Max Karoubi, "Graded Brauer groups and K-theory with local coefficients", *Publications Math. IHES* **38** (1970), 5–25. Also available at `http://www.math.jussieu.fr/~karoubi/Donavan.K.pdf`

I should admit that I have a yearning to classify *nonassociative* super division algebras. Has anyone ever tried this? It's already plain to see that we have two 16-dimensional nonassociative super division algebras:

$$\mathbb{O}[e]/(e^2+1)$$

and

$$\mathbb{O}[e]/(e^2-1)$$

where e is an odd element that commutes with all the octonions. (I should have mentioned this before, when talking about $\mathbb{H}[e]$: even though the quaternions are noncommutative, we assume that e commutes with all of them.) Maybe one of these algebras deserves to be called the *superoctonions*. I bet these or something awfully similar are lurking around in string theory.

Hmm. . . next I wanted to write something about the topology of Bott periodicity and how *that* fits into what I've been discussing, but I'm running out of energy. Let me say it briefly, without much detail, just in case I never get around to a decent explanation.

Two super algebras are super Morita equivalent precisely when they have equivalent categories of super representations. So, the super Brauer group really consists of 8 different *categories*: the categories $\mathsf{SuperRep}(C_n)$, where Bott periodicity says

$$\mathsf{SuperRep}(C_{n+8}) \sim \mathsf{SuperRep}(C_n)$$

Moreover these are symmetric monoidal categories, since direct summing lets us "add" objects in these categories in a nice way.

A long time ago, Graeme Segal figured out how to take a symmetric monoidal category and get an infinite loop space from it. I explained this construction in "Week 199", but for a much more detailed and intense treatment with lots of references to earlier work, try:

4) R. W. Thomason, "Symmetric categories model all connective spectra", *Theory and Applications of Categories* **1** (1995), 78–118. Available at `http://www.tac.mta.ca/tac/volumes/1995/n5/1-05abs.html`

If we do this to $\mathsf{SuperRep}(C_n)$, I think we get something like

$$\Omega^n(kO)$$

that is, the n-fold loop space of something called kO, the "connective K-theory spectrum", which I explained in "Week 105". The fact that this repeats with period 8:

$$\Omega^{n+8}(kO) \sim \Omega^n(kO)$$

is the topological version of Bott periodicity — see "Week 105" for more. So, we get the topological version of Bott periodicity from the algebraic version by turning symmetric monoidal categories into infinite loop spaces!

But, the interesting puzzle here is: what process can we do to $\mathsf{SuperRep}(C_n)$ to get $\mathsf{SuperRep}(C_{n+1})$, which is the algebraic version of looping? And I think the answer is: "taking super representations of C_1 in it". You see,

$$C_1 \otimes C_n = C_{n+1}$$

where I'm using the super tensor product of superalgebras, and this means that the category of representations of C_1 in $\mathsf{SuperRep}(C_n)$ is $\mathsf{SuperRep}(C_{n+1})$.

And, if I were trying to really explain this instead of merely scribbling notes about it, I would try to explain why this is because C_1 is the complex numbers, and the unit circle in the complex numbers is related to *loops*.

But, sigh, that will have to wait.

One more thing before I quit for today. . .

I just saw a cool paper by Dror Bar-Natan, Thang Le and Dylan Thurston about the "Duflo isomorphism". This is a cousin of the Poincare-Birkhoff-Witt theorem, which in its best form says that the universal enveloping algebra UL of a Lie algebra L is isomorphic *as a coalgebra* to the symmetric algebra SL. You'll often see worse versions of the PBW theorem in textbooks, and ugly proofs, but James Dolan showed me the nice version and proof a while back.

The kernel of the idea is this: if L is the Lie algebra of a group G, UL consists of left-invariant differential operators on G, and there's a map $UL \to SL$ sending any differential operator to its "symbol". This is an isomorphism of vector spaces and even of coalgebras, but not of algebras.

Anyway, there's something vaguely similar relating the invariant subalgebras of UL and SL. By "invariant" here, I mean that since L acts as derivations of UL and SL, we can look at the subalgebra of either one consisting of guys who are killed by these derivations; such guys are called "invariant". Physicists call invariant elements of UL "Casimirs", after the first physicist to think about this stuff. They commute with everything else in UL. Invariant elements of SL are like classical Casimirs: there's a Poisson bracket on SL, and these are the guys whose Poisson bracket with everyone vanishes.

The Duflo map is an *algebra isomorphism* from the invariant subalgebra of SL to the invariant subalgebra of UL. So, it's like a very nice way to quantize Casimirs, one that gets along with multiplication. It's called the "Duflo map" because it was invented by Harish-Chandra for semisimple Lie algebras and for Kirillov in general. Kirillov conjectured that it was always an isomorphism; what Duflo did is prove it:

5) Michel Duflo, "Operateurs differentiels bi-invariants sur un groupe de Lie", *Ann. Sci. Ecole Norm. Sup.* **10** (1977), 265–288.

Apparently all known proofs are sort of hard! According to Bar-Natan, Le and Thurston:

> *In the book of Dixmier, the proof is given only in the last chapter and it utilizes most of the results developed in the whole book, including many classification results (a situation Godement called "scandalous"). As discussed below, there have been several recent proofs that do not use classification results, but they all use tools from well outside the natural domain of the problem.*

The proof by Bar-Natan, Le and Thurston uses the connection between knot theory and Lie algebras — namely, the theory of Vassiliev invariants. I think there's still something slightly scandalous about this, but it's awfully interesting. Anyway, take a look:

6) Dror Bar-Natan, Thang T. Q. Le and Dylan P. Thurston, "Two applications of elementary knot theory to Lie algebras and Vassiliev invariants", *Geometry and Topology* **7** (2003), 1–31. Available at `http://www.maths.warwick.ac.uk/gt/GTVol7/paper1.abs.html` and also as `math.QG/0204311`.

For more, try Thurston's thesis:

7) Dylan P. Thurston, "Wheeling: a diagrammatic analogue of the Duflo isomorphism", `math.QG/0006083`.

and, just for fun, Deligne's handwritten letter to Bar-Natan:

8) Pierre Deligne, letter to Dror Bar-Natan about the Duflo map, available at `http://www.math.toronto.edu/~drorbn/Deligne/`

Addendum: Todd Trimble has kindly allowed me to append some rough notes in which he outlines proofs of some results above.

> *From: Todd Trimble Subject: notes on super Brauer To: John Baez Cc: James Dolan Date: Sun, 27 Mar 2005 19:30:00 -0500*
>
> *John,*
>
> *These are some notes on some of the super Brauer discussion.*
>
> 1. *Let V be the category of finite-dimensional super vector spaces over $\mathbb{R}$. By super algebra I mean a monoid in this category. There's a bicategory whose objects are super algebras A, whose 1-cells $M\colon A \to B$ are left A- right B-modules in V, and whose 2-cells are homomorphisms between modules. This is a symmetric monoidal bicategory under the usual tensor product on V.*

A and B are super Morita equivalent if they are equivalent objects in this bicategory. Equivalence classes $[A]$ form an abelian monoid whose multiplication is given by the monoidal product. The super Brauer group of $\mathbb{R}$ is the subgroup of invertible elements of this monoid.

If $[B]$ is inverse to $[A]$ in this monoid, then in particular $A \otimes (-)$ can be considered left biadjoint to $B \otimes (-)$. On the other hand, in the bicategory above we always have a biadjunction

$$\frac{A \otimes C \to D}{C \to A^* \otimes D}$$

(essentially because left A-modules are the same as right A^*-modules, where A^* denotes the super algebra opposite to A). Since right biadjoints are unique up to equivalence, we see that if an inverse to $[A]$ exists, it must be $[A^*]$.

This can be sharpened: an inverse to $[A]$ exists iff the unit and counit

$$1 \to A^* \otimes A \qquad A \otimes A^* \to 1$$

are equivalences in the bicategory. Actually, one is an equivalence iff the other is, because both of these canonical 1-cells are given by the same A-bimodule, namely the one given by A acting on both sides of the underlying superspace of A (call it S) by multiplication.

Either is an equivalence if the bimodule structure map

$$A^* \otimes A \to \mathrm{Hom}(S, S),$$

which is a map of super algebras, is an isomorphism.

2. *As an example, let $A = C_1$ be the Clifford algebra generated by the 1-dimensional space $\mathbb{R}$ with the usual quadratic form $Q(x) = |x|^2$, and $\mathbb{Z}/2$-graded in the usual way. Thus, the homogeneous parts of A are 1-dimensional and there is an odd generator i satisfying $i^2 = -1$. The opposite A^* is similar except that there is an odd generator e satisfying $e^2 = 1$. Under the map*

$$A^* \otimes A \to \mathrm{Hom}(S, S),$$

where we write S as a sum of even and odd parts $R + Ri$, this map has a matrix representation

$$e \otimes i \mapsto \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$
$$1 \otimes i \mapsto \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$
$$e \otimes 1 \mapsto \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

which makes it clear that this map is surjective and thus an isomorphism. Hence $[C1]$ is invertible.

One manifestation of Bott periodicity is that $[C1]$ has order 8. We will soon see a very easy proof of this fact. A theorem of C.T.C. Wall is that $[C1]$ in fact generates the super Brauer group; I believe this can be shown by classifying super division algebras, as discussed below.

3. *That $[C1]$ has order 8 is an easy calculation. Let C_r denote the r-fold tensor of C_1. C_2 for instance has two super-commuting odd elements i, j satisfying $i^2 = j^2 = -1$; it follows that $k := ij$ satisfies*
$k^2 = -1$, and we get the usual quaternions, graded so that the even part is the span $\langle 1, k\rangle$ and the odd part is $\langle i, j\rangle$.

C_3 has three super-commuting odd elements i, j, l, all of which are square roots of -1. It follows that $e = ijl$ is an odd central involution (here "central" is taken in the ungraded sense), and also that $i' = jl$, $j' = li$, $k' = ij$ satisfy the Hamiltonian equations

$$(i')^2 = (j')^2 = (k')^2 = i'j'k' = -1,$$

so we have $C_3 = \mathbb{H}[e]/(e2 - 1)$. Note this is the same as

$$\mathbb{H} \otimes (C_1)^*$$

where the $\mathbb{H}$ here is the quaternions viewed as a super algebra concentrated in degree 0 (i.e. is purely bosonic).

Then we see immediately that $C_4 = C_3 \otimes C_1$ is equivalent to purely bosonic $\mathbb{H}$ (since the C_1 cancels $(C_1)^*$ in the super Brauer group).

At this point we are done: we know that conjugation on (purely bosonic) $\mathbb{H}$ gives an isomorphism

$$\mathbb{H}^* \to \mathbb{H}$$

hence $[\mathbb{H}] - 1 = [\mathbb{H}^*] = [\mathbb{H}]$, i.e. $[\mathbb{H}] = [C4]$ has order 2!
Hence $[C1]$ has order 8.

4. *All this generalizes to Clifford algebras: if a real quadratic vector space (V, Q) has signature (r, s), then the super algebra $\mathrm{Cliff}(V, Q)$ is isomorphic to $A_r \otimes (A^*)_s$, where A_r denotes r-fold tensor product of $A = C_1$. By the above calculation we see that $\mathrm{Cliff}(V, Q)$ is equivalent to C_{r-s} where $r - s$ is taken modulo 8.*

For the record, then, here are the hours of the super Clifford clock, where e denotes an odd element, and $\sim$ denotes equivalence:

- $C_0 \sim \mathbb{R}$
- $C_1 \sim \mathbb{R} \oplus \mathbb{R}[e]$, $e2 = -1$
- $C_2 \sim \mathbb{C} \oplus \mathbb{C}[e]$, $e2 = -1, ei = -ie$
- $C_3 \sim \mathbb{H} \oplus \mathbb{H}[e]$, $e2 = 1, ei = ie, ej = je, ek = ke$
- $C_4 \sim \mathbb{H}$

> - $C_5 \sim \mathbb{H} \oplus \mathbb{H}[e]$, $e2 = -1, ei = ie, ej = je, ek = ke$
> - $C_6 \sim \mathbb{C} \oplus \mathbb{C}[e]$, $e2 = 1, ei = -ie$
> - $C_7 \sim \mathbb{R} \oplus \mathbb{R}[e]$, $e2 = 1$
>
> *All the super algebras on the right are in fact super division algebras, i.e. super algebras in which every nonzero homogeneous element is invertible.*
>
> *To prove Wall's result that $[C1]$ generates the super Brauer group, we need a lemma: any element in the super Brauer group is the class of a super division algebra.*
>
> *[To be continued. I had wanted to show that every element in the super Brauer group must be of the form $[A]$ where A is a super division algebra, and then classify super (associative) division algebras, showing on a case by case basis that those not in the super Clifford clock above are seen not to belong to the super Brauer group.]*
>
> *Todd*

Todd finished off the job later... though by this point we had formulated the grander goal of classifying not-necessarily-associative super division algebras!

> *From: Todd Trimble Subject: super division algebras To: John Baez Date: Wed, 27 Apr 2005 22:17:12 EDT*
>
> *John,*
>
> *This is a warm-up to classifying super division algebras over $\mathbb{R}$, where I'll consider just the associative case.*
> *Nothing I say will be deep, but I found it somewhat fun and diverting, and there may be echoes of things to come.*
>
> *I'll take as known that the only associative division algebras over $\mathbb{R}$ are $\mathbb{R}$, $\mathbb{C}$, $\mathbb{H}$ – the even part A of an associative super division algebra must be one of these cases. We can express the associativity of a super algebra (with even part A) by saying that the odd part M is an A-bimodule equipped with a A-bimodule map pairing*
>
> $$\langle -, - \rangle \colon M \otimes_A M \to A$$
>
> *such that:*
>
> $$a\langle b, c\rangle = \langle a, b\rangle c \quad \text{for all } a, b, c \text{ in } M. \qquad (\star\star)$$
>
> *If the super algebra is a super division algebra which is not wholly concentrated in even degree, then multiplication by a nonzero odd element induces an isomorphism*
>
> $$A \to M$$
>
> *and so M is 1-dimensional over A; choose a basis element e for M.*
>
> *The key observation is that for any a in A, there exists a unique a' in A such that*
>
> $$ae = ea'$$

and that the A-bimodule structure forces $(ab)' = a'b'$. Hence we have an automorphism (fixing the real field)

$$(--)'\colon A \to A$$

and we can easily enumerate (up to isomorphism) the possibilities for associative super division algebras over $\mathbb{R}$:

1. *$A = \mathbb{R}$. Here we can adjust e so that either $e^2 := \langle e, e\rangle$ is -1 or 1. The corresponding super division algebras occur at 1 o'clock and 7 o'clock on the super Brauer clock.*
2. *$A = \mathbb{C}$. There are two $\mathbb{R}$-automorphisms $\mathbb{C} \to \mathbb{C}$. In the case where the automorphism is conjugation, condition ($\star\star$) for super associativity gives $\langle e, e\rangle e = e\langle e, e\rangle$ so that $\langle e, e\rangle$ must be real. Again e can be adjusted so that $\langle e, e\rangle$ is -1 or 1. These possibilities occur at 2 o'clock and 6 o'clock on the super Brauer clock.*

For the identity automorphism, we can adjust e so that $\langle e, e\rangle$ is 1. This gives the super algebra $\mathbb{C}[e]/(e^2 - 1)$ (where e commutes with elements in $\mathbb{C}$). This does not occur on the super Brauer clock over $\mathbb{R}$. However, it does generate the super Brauer group over $\mathbb{C}$ (which is of order two).

3. *$A = \mathbb{H}$. Here $\mathbb{R}$-automorphisms $\mathbb{H} \to \mathbb{H}$ are given by $h \mapsto xhx^{-1}$ for x in $\mathbb{H}$. In other words*

$$he = exhx^{-1}$$

whence ex commutes with all elements of $\mathbb{H}$ (i.e. we can assume wlog that the automorphism is the identity). The properties of the pairing guarantee that $h\langle e, e\rangle = \langle e, e\rangle h$ for all h in $\mathbb{H}$, so $\langle e, e\rangle$ is real and again we can adjust e so that $\langle e, e\rangle$ is 1 or -1. These cases occur at 3 o'clock and 5 o'clock on the super Brauer clock.

This appears to be a complete (even if a bit pedestrian) analysis.

Best,

Todd

Week 213

April 3, 2005

Here's a book I've been reading lately:

1) Kenneth S. Brown, *Cohomology of Groups*, Graduate Texts in Mathematics **182**, Springer, 1982.

I should have read this book a long time ago — but I probably wouldn't have enjoyed it as much as I do now. All sorts of things I struggled to learn for years are neatly laid out here. Best of all, he comes right out and admits from the start that the cohomology of groups is really a branch of *topology*, instead of hiding this fact like some people do.

This is something every mathematician should know: you can take any group and turn it into a space, thus "reducing" group theory to topology. In particular, if you have any trick for telling spaces apart, like "cohomology theory", you can apply it to groups as well.

Of course topology is *harder* than group theory in many ways — hence my quotes around "reducing". Indeed, algebraic topology was invented as a trick for reducing topology to group theory! But, the bridge turns out to go both ways, and there's a lot of profitable traffic in both directions.

Ultimately, as James Dolan likes to point out, it's all about the unity of mathematics. Topology is about our concept of *space*, while group theory is about our concept of *symmetry*. . . but the amazing fact is that they turn out to be two aspects of the same big thing! Mathematics is a source of endless surprises, but this is one of the biggest jaw-droppers of all.

The idea goes back at least to Evariste Galois, who noticed that you can classify the ways a little thing can sit in a bigger thing by keeping track of what we now call its "Galois group": the group of all symmetries of the big thing that map the little thing to itself. For example, you can pick out a point or line in the plane by keeping track of which symmetries of the plane map this point or line to itself.

However, the idea of using groups to classify how a little thing sits in a big one was really made explicit in Felix Klein's "Erlangen program", a plan for reducing *geometry* to group theory.

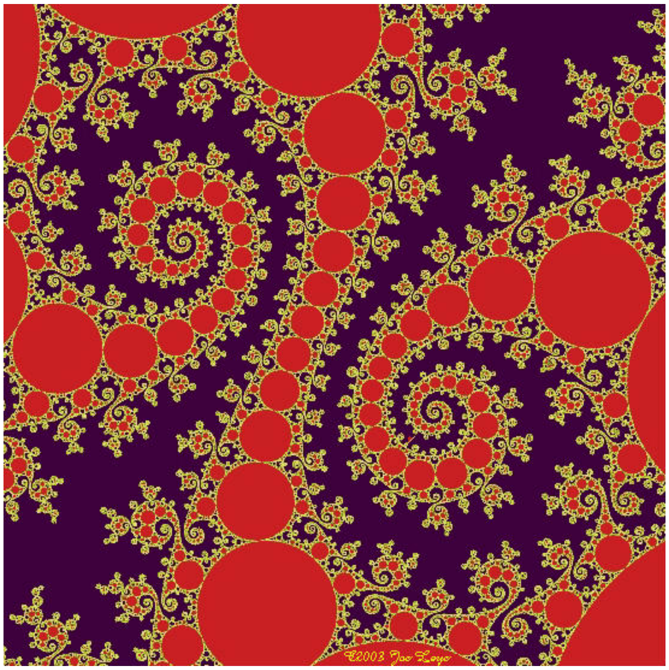
You may know Klein for his famous one-sided bottle:

A mathematician named Klein
Thought the Mbius strip was divine.
Said he: "If you glue
The edges of two
You'll get a weird bottle like mine!"

Or maybe you know that the symmetry group of a rectangle, including reflections, is called the "Klein 4-group":

$\cdot$	1	a	b	c
1	1	a	b	c
a	a	1	c	b
b	b	c	1	a
c	c	b	a	1

He is also known for some other groups called "Kleinian groups", which act as symmetries of fractal patterns like this:



2) Jos Leys, "Kleinian Pages", `http://www.josleys.com/creatures42.htm`

If you like cool pictures, check out this website! I've linked you to the page that most closely connects to Kleinian groups, but there are lots of other more fanciful pictures. And if you get interested in the math lurking behind these fractals, you've *got* to try this book:

3) David Mumford, Caroline Series, and David Wright, *Indra's Pearls: The Vision of Felix Klein*, Cambridge U. Press, Cambridge, 2002.

Mumford is a world-class mathematician, so this book is completely different from the superficial descriptions of fractals one often sees in math popularizations — but it's still readable, and it's packed with beautiful pictures. You can learn a lot about Kleinian groups from this!

The Kleinian groups arose from Klein's studies of complex functions, which he considered his best work. But, he was also a mathematical physicist. Among other things, he wrote a four-volume book on tops with one of the fathers of quantum mechanics, Arnold Sommerfeld:

4) Felix Klein and Arnold Sommerfeld, *ber die Theorie des Kreisels*, 4 vols, 1897–1910. Reprinted by Johnson, New York, 1965. Also available at `http://www.hti.umich.edu/cgi/t/text/text-idx?c=umhistmath;idno=ABV7354.0001.001`

This came after a book he wrote on his own:

5) Felix Klein, *The Mathematical Theory of the Top*, Scribner's, New York, 1887.

This may seem like a lot of books about a kid's toy! But, tops are profoundly related to the rotation group, and the "exactly solvable" tops discovered by Euler, Lagrange/Poisson, and Sofia Kowalevskaya are solvable because of their symmetries — deeply hidden symmetries, in the case of the Kowalevskaya top. So, one can imagine why Klein liked this subject.

Klein also wrote a book on the icosahedron and the quintic equation:

6) Felix Klein, *Lectures on the Icosahedron and the Solution of Equations of the Fifth Degree*, 1888. Reprinted by Dover, New York, 2003. Also available at `http://historical.library.cornell.edu/cgi-bin/cul.math/docviewer?did=03070001&seq=7`

Galois had already noticed that the number field you get by taking the rationals and throwing in the roots of a typical quintic:

$$ax^5 + bx^4 + cx^3 + dx^2 + ex + f = 0$$

has as its symmetry group all the permutations of the 5 roots. Indeed, he saw that the "unsolvability" of this group, in a technical sense, is what makes it impossible to solve the quintic by radicals. It must have been common knowledge that the symmetry group of the icosahedron is the group of all *even* permutations of 5 things. But, Klein took this much further! Alas, I've never really understood what he did. Perhaps if I read these and think hard, I'll understand:

7) Jerry Shurman, *Geometry of the Quintic*, John Wiley and Sons, New York, 1997.

Peter Doyle and Curt McMullen, "Solving the quintic by iteration", *Acta Math.* **163** (1989), 151–180. Available at `http://math.dartmouth.edu/~doyle/docs/icos/icos/icos.html`

Anyway, it should be clear by now that Klein was a lover of symmetry. He was also a bit of a visionary, and his obituary by Grace Chisholm Young shows that this got him in some trouble:

> *One of Weierstrass' pupils, still alive, told me that at Berlin Klein was anathema: it was said that his work was not mathematics at all, but mere talk. This criticism shows a want of appreciation of his rare type of mind. It teemed with ideas and brilliant reflections, but it is true that his work lacks the stern aspects required by mathematical exactitude. It was in personal contact that this was corrected, at least in so far as his students were concerned. His favourite maxim was, "Never be dull".*

In a talk he wrote in 1872 when he was made professor at Erlangen University — a talk he didn't actually give! — Klein outlined what is now called his "Erlangen program". The idea here is that different kinds of geometry correspond to different symmetry groups. Taken to the extreme, this philosophy says that a geometry is just a group! In a given geometry, a "figure" of any kind — like a point or line — can be detected by the subgroup of symmetries that map that figure to itself. So, a figure is just a subgroup!

This program eventually led to a grand theory of groups and geometry based on "flag manifolds", which I tried to sketch in "Week 178", "Week 180", and "Week 181".

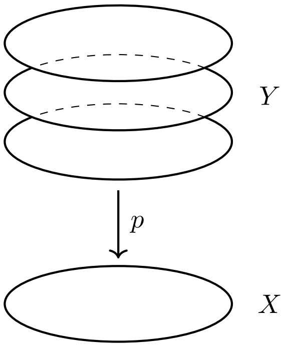
It's important to realize how similar the Erlangen program is to Galois theory. Galois had also used group theory to classify how a little thing can sit in a bigger thing, but in situations where the "things" in question are commutative algebras — for example, the rational numbers with some roots of polynomials thrown in.

Now, commutative algebra is like topology, only backwards. Any space has a commutative algebra consisting of functions on it, and if we're very clever we can think of any commutative algebra as functions on some space — though this was only achieved long after Galois, by Alexander Grothendieck.

What do I mean by "backwards"? Well, suppose you have a "covering space" — a big space sitting over a little one, like a spiral sitting over the circle. In this situation, any function on the little space downstairs defines a function on the big one upstairs. So, the algebra of functions on the little space sits inside the algebra of functions on the big space.

Notice how it's backwards. Classifying how a little commutative algebra can sit in a big one amounts to classifying how a big space can cover a little one! For more details on this analogy, try "Week 198", "Week 201" and especially "Week 205".

I should warn you: the Galois group has a different name when we apply it to the classification of covering spaces — we call it the group of "deck transformations". The idea is pretty simple. Suppose Y is a covering space of X, like this:



We've got a function $p\colon Y \to X$, and sitting over each point of X are the same number of points of Y, living on different "sheets" that look locally just like X. You should imagine the sheets being able to twist around from place to place, like the edges of a Moebius strip.

Anyway, a "deck transformation" is just a way of mapping Y to itself that permutes the different points sitting over each point of X.

The theory of this was worked out by Riemann, Poincare, and others. Poincare showed you could use this idea to turn any connected space X into a group — its "fundamental group". There are different ways to define this, but one is to form the most complicated possible covering space of X that's still connected — its "universal cover". Then, take the group of deck transformations of this! Following Galois' philosophy, all the other connected covering spaces of X correspond to subgroups of this group.

The theory of the fundamental group was just the beginning when it came to groups and topology. One of many later big steps, back in the late 1940s, was due to Sammy Eilenberg and Saunders Mac Lane. They saw how to reverse the "fundamental group" idea and turn any group back into a space!

More precisely: for any group G, there's a space whose fundamental group is G and whose higher homotopy groups vanish. It's sometimes called the "Eilenberg–Mac Lane space" and denoted $K(G, 1)$, but sometimes it's called the "classifying space" and denoted BG. It's pretty easy to build; I described how back in "Week 70".

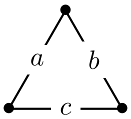
You start with a point:

•

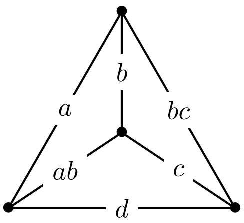
Then you stick on an edge looping from this point to itself for each element a in G. Unrolled, it looks like this:

•— a —

where a is an element of our group. Then, whenever we have $ab = c$ in our group, we stick on a triangle like this:



Then, whenever we have $abc = d$ in our group, we stick on a tetrahedron like this:



And so on, forever! For each list of n group elements, we get an n-dimensional simplex in our Eilenberg-Mac Lane space. The resulting space knows everything about the group we started with. In particular, the fundamental group of this space will be the group we started with!

Using this idea, we can do some fiendish things. For example, for each n we can form a set $C_n(G, A)$ consisting of all functions that eat n-dimensional simplices in the Eilenberg-Mac Lane space of G and spit out elements of some abelian group A. There are maps

$$d\colon C_n(G, A) \to C_{n+1}(G, A)$$

reflecting the fact that each $(n+1)$-simplex has a bunch of n-simplices as its faces. Since the boundary of a boundary is zero,

$$d^2 = 0$$

Guys who live in the kernel of

$$d\colon C_n(G,A)\to C_{n+1}(G,A)$$

are called "n-cocycles", and guys who live in the image of

$$d\colon C_{n-1}(G,A)\to C_n(G,A)$$

are called "n-coboundaries". Since $d^2=0$, every coboundary is a cocycle, but not always vice versa. So, we can form the group of cocycles mod coboundaries. This is called the "nth cohomology group" of G with coefficients in A, and it's denoted

$$H^n(G,A)$$

This sounds unmotivated at first, but the nth cohomology group of a space is really just a clever way of keeping track of n-dimensional holes in that space. So, what we're doing here is cleverly defining a way to study "holes" in a *group!* There are deeper, more conceptual ways of understanding group cohomology, but this is not bad for starters.

For example, let's take the simplest group that's not *utterly* dull — the integers $\bmod 2$, or $\mathbb{Z}/2$. Here we get

$$K(\mathbb{Z}/2,1)=\mathbb{RP}^\infty$$

where $\mathbb{RP}^\infty$ is the space formed by taking an infinite-dimensional sphere and identifying opposite points. This space has holes of arbitrarily high dimension, so the cohomology groups of $\mathbb{Z}/2$ go on being nontrivial for arbitrarily high n. I sketched a "picture proof" here:

8) John Baez, Fall 2004 Quantum Gravity Seminar, week 10, notes by Derek Wise, `http://math.ucr.edu/home/baez/qg-fall2004/`

and I showed that, for example

$$H_n(\mathbb{Z}/2,\mathbb{Z})=\begin{cases}\mathbb{Z} & \text{if } n=0;\\ 0 & \text{if } n \text{ is odd;}\\ \mathbb{Z}/2 & \text{if } n \text{ is even and non-zero.}\end{cases}$$

I also explained how this stuff is related to topological quantum field theory.

Anyway, all this is just the very superficial beginnings of the subject of group cohomology. Read Brown's book to dig deeper!

Personally, what I find most exciting about this book now are the remarks on the "Euler characteristic" of a group. Let me explain this… though now I'll have to pull out the stops and assume you know some group cohomology.

We can try to define the "Euler characteristic" of a group G to be the Euler characteristic of $K(G,1)$. This is the alternating sum of the dimensions of the rational cohomology groups

$$H_n(G,\mathbb{Q})$$

Of course, this alternating sum only converges if the cohomology groups vanish for big enough n. Also, they all need to be finite-dimensional.

Unfortunately, not many groups have well-defined Euler characteristic with this naive definition!

For example, people have studied groups G whose nth cohomology vanishes for $n > d$, regardless of the coefficients. If we take the smallest d for which this holds, such a group G is said to have "cohomological dimension" d. Eilenberg and Ganea showed that for $d \geqslant 3$, a group has cohomological dimension d whenever we can build $K(G, 1)$ as a simplicial complex (or CW complex) with no cells of dimension more than d.

This is a nice geometrical interpretation of the cohomological dimension. But, one can show that groups with torsion never have finite cohomological dimension! We've seen an example already: $\mathbb{Z}/2$, whose Eilenberg-Mac Lane space is infinite-dimensional.

However, it turns out that there's a generalization of the Euler characteristic that makes sense for any group G that has a torsion-free subgroup H whose Euler characteristic is well-defined in the naive way, as long as H has finite index in G. We just define the Euler characteristic of G to be the Euler characteristic of H divided by the index of H in G. The answer doesn't depend on the choice of H!

Take my favorite example, $\mathrm{SL}(2, \mathbb{Z})$. This has torsion, so its cohomological dimension is infinite and its naive Euler characteristic is undefined! Indeed, I wrote a whole issue of This Week's Finds about some elements of orders 4 and 6 sitting inside $\mathrm{SL}(2, \mathbb{Z})$, related to the symmetries of square and hexagonal lattices — see "Week 125".

But, $\mathrm{SL}(2, \mathbb{Z})$ has a torsion-free subgroup of index 12, namely its commutator subgroup — the group you need to quotient by to make $\mathrm{SL}(2, \mathbb{Z})$ be abelian. This subgroup has finite cohomological dimension and its Euler characteristic is -1. I'm not sure why this is true, but Brown says so! This means the Euler characteristic of $\mathrm{SL}(2, \mathbb{Z})$ works out to be $-1/12$.

If you've read my stuff about Euler characteristics in "Week 147", you'll see why this gets me so excited — I can add this stuff to my list of weird ways of calculating the Euler characteristic. Plus, it's related to the magical role of the number "24" in string theory, and also the Riemann zeta function!

Indeed, the Riemann zeta function gives a way to make rigorous Euler's zany observation that

$$1 + 2 + 3 + \ldots = -\frac{1}{12},$$

as I explained here:

9) John Baez, "Euler's Proof that $1 + 2 + 3 + \ldots = -1/12$, Bernoulli Numbers and the Riemann Zeta Function", Winter 2004 Quantum Gravity Seminar, homework for weeks 5,6,7, available at `http://math.ucr.edu/home/baez/qg-winter2004/`

This suggests that there should be a version of the Eilenberg-Mac Lane space for $\mathrm{SL}(2, \mathbb{Z})$ which has 1 cell of dimension 0, 2 cells of dimension 2, 3 cells of dimension 4, and so on. Does anyone know if this is true?

More generally, G. Harder computed the (generalized) Euler characteristic for a large class of arithmetic groups:

10) G. Harder, "A Gauss-Bonnet formula for discrete arithmetically defined groups", *Ann. Sci. Ecole Norm. Sup.* **4** (1971), 409–455.

For example, he looked at the symplectic group defined over the integers, $\mathrm{Sp}(n,\mathbb{Z})$, and showed that its Euler characteristic is this product of values of the Riemann zeta function:

$$\zeta(-1)\zeta(-3)\cdots\zeta(1-2n)$$

In the case $n=1$ we get back $\mathrm{SL}(2,\mathbb{Z})$ and $\zeta(-1)=-1/12$.

In fact, every Chevalley group over the integers has a well-defined Euler characteristic, and Harder was able to compute it in terms of Bernoulli numbers. A Chevalley group is sort of like a simple Lie group, but defined algebraically. $\mathrm{Sp}(n,\mathbb{Z})$ is one example. $\mathrm{SL}(n,\mathbb{Z})$ is another, but it's Euler characteristic turns out to vanish for $n>2$, so it's not too interesting.

Harder worked them all out. For example, he showed the Euler characteristic of the integral form of the exceptional group E_7 is some wacky number like

$$-\frac{691\times 43867}{2^{21}\times 3^9\times 5^2\times 7^3\times 11\times 13\times 19}$$

Serre went even further, computing Euler characteristics of Chevalley groups defined over algebraic number fields. He also noticed that when you write the Euler characteristic of a group as a fraction, the primes in the denominator are precisely the primes p for which the group has p-torsion. He was thus able to conclude, for example, that E_7 defined over the integers has p-torsion for $p=2,3,5,7,11,13,19$.

For more, see:

11) Jean-Pierre Serre, "Cohomologie des groups discretes", *Ann. Math. Studies* **70** (1971), 77–169.

This only takes us up to 1971. I shudder to think what bizarre results along these lines are known by now! Probably they'd seem not bizarre but beautiful if I understood this stuff better: I don't really have a clue how the Riemann zeta function gets into this game, so everything after that seems like black magic to me — bewitching but bewildering.

But, it's clear that the study of groups and symmetry has not lost its ability to turn up surprises.

Addenda: I had often wondered how Klein's name got attached to the pathetic little "4-group" mentioned above, which is just $\mathbb{Z}/2\times\mathbb{Z}/2$. John McKay proffered an explanation:

> *There is a group called the Klein group. It is denoted V_4 = The Vierer-Gruppe (The fours group).*
>
> *Klein worked with the simple group of order 168 and found the "Klein quadric" which has it for symmetry group.*
>
> *The suggestion is that friends decided to call the non-cyclic abelian group of order 4 the "Klein group" = the "little group" as a joke.*
>
> *I have a question you may like to posit to your readers:*
>
> *Is V_4 the abstract group or a permutation group?*

> *There are other points ... I presume you know that your* $-1/12$ *is* $\zeta(-1)$. *There is a paper by Lepowsky on the occurrence of such* $\zeta(-n)$ *involving vertex algebras.*
>
> *I dearly wish I understood cohomology!*
>
> *I am busy tethering moonshine!*
>
> *Best, John*

This group of order 168 has made an appearance here before, in "Week 194": it's $\mathrm{PSL}(3,\mathbb{Z}/2)$ — the group of symmetries of the projective plane over $\mathbb{Z}/2$, or "Fano plane", whose points can also be thought of as imaginary unit octonions. It's also $\mathrm{PSL}(2,\mathbb{Z}/7)$. I've long been mystified by its relation to Klein's quartic, mainly because I've never spent time trying to understand it! — it's just one of those things that's been gnawing at the edges of my consciousness, especially ever since I saw this book come out:

12) Silvio Levy, *The Eightfold Way: the Beauty of Klein's Quartic Curve*, MSRI Research Publications **35**, Cambridge U. Press, Cambridge 1999. Available at `http://www.msri.org/publications/books/Book35/`

It has a translation of Klein's original paper on this subject. Someday I'll break down and study this.

Anyway. . . .

James Dolan mentioned some other folklore saying that the "Kleinian groups" were *also* named after Klein as a joke:

> *by the way, i enjoyed the latest twf a lot (although i don't know why we seem to never get a chance to talk about all this stuff ourselves that much), but i noticed that you (apparently non-ironically) mentioned kleinian groups as a famous thing named after klein, without telling the story that i always hear about how poincare gave kleinian groups the name "kleinian groups" after klein complained to poincare about poincare's use of the terminology "fuchsian groups" for something that fuchs apparently didn't event.*
>
> *i guess that the versions of the story that i'd heard seemed to suggest that klein was complaining because he thought that fuchs hadn't significantly contributed to the study of fuchsian groups, and that poincare may have been naively trying to placate klein and/or not-so-naively twitting him by then giving the name "kleinian groups" to something that klein hadn't significantly contributed to the study of.*
>
> *however i did just look for the story on the web, and the tellings that i found there i guess don't really suggest that klein didn't "significantly contribute to the study of" kleinian groups (or at least not by my standards). it's still not clear though what sort of reaction poincare may have been trying to provoke in klein, and whether he succeeded in provoking it. it's claimed that poincare did come up with the name "kleinian function" later in the same day after klein complained about the name "fuchsian function", and also that klein was subsequently just as vociferous in complaining to poincare about the name "kleinian function" as he was in complaining about the name "fuchsian function". but apparently*

> *klein's complaints were based on* very *exacting concerns about absolute priority, so that the names "fuchsian function" and "kleinian function" can be seen as inappropriate only by the standards of someone with similarly ridiculous concerns about absolute priority, rather than by a reasonable person such as myself.*
>
> *i'd also heard that klein's nervous breakdown was provoked by the stress of trying to keep up with a genius like poincare, but maybe it was actually provoked by poincare's apparently casual attitude towards priority disputes and/or concept-naming.*
>
> *i'd thought of asking you about this issue of whether klein really did have much to do with kleinian groups right after i read the advance copy of twf that you sent me, but i guess that i didn't notice that it was an advance copy. i guess that it doesn't matter though, since apparently there* is *a case to be made that klein had lots to do with developing the theory of kleinian groups; just not by his own apparently ridiculous standards.*

Noam Elkies suggested that the commutator subgroup of $\mathrm{SL}(2,\mathbb{Z})$ has Euler characteristic -1 because it's a a free group on 2 generators, so its classifying space is a figure 8, with Euler characteristic $1-2=-1$ since it has one vertex and two edges.

This sounds right. In particular, I already mentioned how Brown claims the commutator subgroup of $\mathrm{SL}(2,\mathbb{Z})$ is torsion-free. Further, Kevin Buzzard shows below that any torsion-free subgroup of $\mathrm{SL}(2,\mathbb{Z})$ is a free group. So, we just need to check that the commutator subgroup of $\mathrm{SL}(2,\mathbb{Z})$ can be generated by two elements but not by just one.

Laurent Bartholdi just made this job easier; he sent me an email saying these are free generators for the commutator subgroup of $\mathrm{SL}(2,\mathbb{Z})$:

$$\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$$

In fact, Kevin Buzzard's email was packed with wisdom. He wrote:

> *I know one elementary argument which you don't appear to, so I thought I'd fill you in. The argument below is waffly but rather easy really.*
>
> *John Baez wrote:*
>
> > *But,* $\mathrm{SL}(2,\mathbb{Z})$ *has a torsion-free subgroup of index 12, namely its commutator subgroup — the group you need to quotient by to make* $\mathrm{SL}(2,\mathbb{Z})$ *be abelian. This subgroup has finite cohomological dimension and its Euler characteristic is* -1*. I'm not sure why this is true, but Brown says so! This means the Euler characteristic of* $\mathrm{SL}(2,\mathbb{Z})$ *works out to be* $-1/12$*.*
>
> *One doesn't have to use such a "strange" subgroup as the commutator subgroup of* $\mathrm{SL}(2,\mathbb{Z})$*. People who do modular forms, like me, far prefer "congruence subgroups", as these are the ones that show up when you study automorphic forms for* $\mathrm{SL}(2,\mathbb{Z})$*. So here's an easy way to compute the Euler characteristic of* $\mathrm{SL}(2,\mathbb{Z})$*: take your favourite congruence subgroup which has no torsion, work out its Euler characteristic (this is easy, I'll show you how to do it in a second) and then deduce what the Euler characteristic of* $\mathrm{SL}(2,\mathbb{Z})$ *is.*

Here are some examples of congruence subgroups: for any integer $N \geqslant 1$, consider the subgroup $\Gamma_1(N)$ of $\mathrm{SL}(2,\mathbb{Z})$, defined as the matrices

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

in $\mathrm{SL}(2,\mathbb{Z})$ such that $c = 0 \mod N$ and $a = d = 1 \mod N$. It's just the preimage in $\mathrm{SL}(2,\mathbb{Z})$ of the upper triangular unipotent matrices in $\mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z})$ so it's a subgroup of $\mathrm{SL}(2,\mathbb{Z})$. Here's a neat fact that makes life easy:

***Lemma:** if $N \geqslant 5$ then $\Gamma_1(N)$ has no torsion.*

***Proof:** say g in $\mathrm{SL}(2,\mathbb{Z})$ has finite order $d \geqslant 1$. Then its min poly divides $X^d - 1$ so over the complexes it has distinct linear factors so it's diagonalisable with roots of unity z and w on the diagonal. Now $|z| = |w| = 1$ so $|\operatorname{trace}(g)| \leqslant 2$. But it's an integer, so it's $-2, -1, 0, 1, 2$. And for $N \geqslant 5$ the only one of these congruent to $2 \mod N$ is 2. So $z = w = 1$ and so g is the identity.*

Deeper, but also completely standard (and not logically necessary for what follows)—any torsion-free subgroup of $\mathrm{SL}(2,\mathbb{Z})$ is free! This is because $\mathrm{SL}(2,\mathbb{Z})$ acts very naturally on a certain tree in the upper half plane. This is a neat piece of mathematics. $\mathrm{SL}(2,\mathbb{Z})$ acts on the upper half plane $\{z = x + iy : y > 0\}$ via the rule:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

sends z to $(az + b)/(cz + d)$. Now draw dots at the points $i = \sqrt{-1}$ and $\rho = \exp(2\pi i/6)$, the primitive 6th root of unity in the upper half plane, and draw the obvious arc between them (the one that lies on the circle $|z| = 1$), this is our first edge, and now look at the image of what you have under the $\mathrm{SL}(2,\mathbb{Z})$ action. It's a rather pretty tree, with two kinds of vertices—those in the i orbit have valency 2 and stabiliser of order 4, and those in the ρ orbit have valency 3 and stabiliser of order 6. Now a group is free iff it acts freely on a tree, and anything torsion-free in $\mathrm{SL}(2,\mathbb{Z})$ must be acting freely on this tree because the stabiliser of each vertex and edge under the $\mathrm{SL}(2,\mathbb{Z})$ action is finite.

So $\Gamma_1(5)$ is, by this general theorem, free. In fact I don't really need this general nonsense, one can give a hands-on proof of this fact, which I'll do now. We've seen that $\mathrm{SL}(2,\mathbb{Z})$, and hence $\Gamma_1(5)$, acts on the upper half plane. There is no torsion in $\Gamma_1(5)$ so the action is very nice, one checks easily that the action is free in fact by a similar sort of argument to the lemma above, it's the sort of thing you can find in the first few pages of any book on modular forms. So we can quotient out the upper half plane by $\Gamma_1(5)$ and get a quotient Riemann surface. The point is that this computation is very manageable and can be done "in practice". There is a standard argument which shows how to quotient out the upper half plane by $\mathrm{SL}(2,\mathbb{Z})$ — the answer is a Riemann surface isomorphic to the complex plane (although you have to take care at the points where the action isn't free—this is exactly the vertices of the tree above), and the isomorphism can even be given "explicitly" via the j-function coming from the theory of elliptic curves—there is a standard fundamental domain even, the one with corners ρ, ρ^2 and $+i\infty$. I'm sure you'll have come across this sort of thing many times *before. Now* $\mathrm{SL}(2,\mathbb{Z})$ *surjects onto* $\mathrm{SL}(2,\mathbb{Z}/5\mathbb{Z})$ *so the index of* $\Gamma_1(5)$ *in* $\mathrm{SL}(2,\mathbb{Z})$ *is just the index of*

$$\begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix}$$

in $\mathrm{SL}(2,\mathbb{Z}/5Z)$ *and by counting orders this comes out to be 24. Now it's not hard to find explicitly 24 translates of the standard fundamental domain and then glue them together to work out the quotient of the upper half plane by* $\Gamma_1(5)$ *— it turns out that it is isomorphic to the Riemann Sphere minus 4 points.*

In fact there is no need to do this sort of computation — the modular forms people have automated it long ago. The quotient of the upper half plane by $\Gamma_1(N)$ *is a Riemann surface called* $Y_1(N)$ *and I can just ask my computer to compute the genus of its natural compactification (this exists and is called* $X_1(N)$*) and also to compute how many cusps were added to compactify it. So in practice you just have to find a friendly modular forms person and then say "hey, what's the genus of* $X_1(5)$ *and how many cusps does it have?" and then you have a complete description of* $\Gamma_1(5)$ *because it's* π_1 *of the answer.*

OK, the upper half plane modulo $\Gamma_1(5)$ *is the sphere minus 4 points, so* $\Gamma_1(5)$ *is* π_1 *of this, i.e. it's free on three generators. That makes the Euler characteristic of* $\Gamma_1(5)$ *equal to* $1-3=-2$*. And we already checked that the index was 24, so the Euler Characteristic of* $\mathrm{SL}(2,\mathbb{Z})$ *works out to be* $-1/12$*. Grothendieck wouldn't have chosen* $\Gamma_1(5)$*; he would have chosen something called* $\Gamma(2)$*, the subgroup of* $\mathrm{SL}(2,\mathbb{Z})$ *consisting of the matrices which are the identity* $\mod 2$*. There is another classical modular function* λ *inducing an isomorphism of* $Y(2)$*, the quotient of the upper half plane by* $\Gamma(2)$*, with the sphere minus three points— this is what gives the one-line proof of the fact that any analytic function* $\mathbb{C}\to\mathbb{C}$ *that misses two points must be constant, because it then lifts to a function from* $\mathbb{C}$ *to the upper half plane which is the same as the unit disc, so we're done by Liouville. There is a subtlety here though:*

$$\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$

is in $\Gamma(2)$*. So you have to work with* $\mathrm{PSL}(2,\mathbb{Z}) = \mathrm{SL}(2,\mathbb{Z})/\{\pm 1\}$ *instead. Let* $\mathrm{P}\Gamma(2)$ *denote the image of* $\Gamma(2)$ *in* $\mathrm{PSL}(2,\mathbb{Z})$*. Note that* -1 *is kind of a pain in the theory of modular forms sometimes because it acts trivially on everything but isn't the identity. Grothendieck was very interested in the sphere minus three points but it's much older than this that* $\mathrm{P}\Gamma(2)$ *is its fundamental group, so* $\mathrm{P}\Gamma(2)$ *has Euler characteristic* $2-3=-1$ *and index 6 in* $\mathrm{PSL}(2,\mathbb{Z})$*, so* $\mathrm{PSL}(2,\mathbb{Z})$ *has Euler characteristic* $-1/6$*, so* $\mathrm{SL}(2,\mathbb{Z})$ *has Euler characteristic* $-1/12$ *because that's how they work. :-)*

John Baez wrote:

> *This only takes us up to 1971. I shudder to think what bizarre results along these lines are known by now! Probably they'd seem not bizarre but beautiful if I understood this stuff better: I don't really have a clue how the Riemann zeta function gets into this game, so everything after that seems like black magic to me — bewitching but bewildering.*

> *Nowadays almost any analytic function that is involved in number theory, when evaluated at certain "natural" points, gives an answer which has a conjectural interpretation in terms of relations between cohomology theories — this is the subject of many conjectures (Deligne, Beilinson, Bloch-Kato,. . .). It is still absolutely black magic! Actually I'm being unfair, the relation between special values of ζ and Euler characteristics is somehow less profound than this stuff. I wish I knew more about it! It can actually be used to compute certain values of L-functions (things more general than the zeta function but along the same lines). . .*
>
> *Kevin*

I replied:

> *Hi -*
>
> *Thanks VERY much for this email. I was actually wondering why Brown used the commutator subgroup of $\mathrm{SL}(2,\mathbb{Z})$ as a kind of "warmup" for computing the Euler characteristic of $\mathrm{SL}(2,\mathbb{Z})$ instead of one of the congruence subgroups. It seems this subgroup is not any of the beloved congruence subgroups. . . .*
>
> *In fact, I've finally managed to turn up the thing I was looking for. How does this relate to the stuff you're saying? It involves $\Gamma(3)$ rather than the $\Gamma_1(N)$ groups:*
>
> *In "Week 97", I wrote:*
>
>> *Where does the extra 24 come from? I don't know, but Stephan Stolz said it has something to do with the fact that while $\mathrm{PSL}(2,\mathbb{Z})$ doesn't act freely on the upper half-plane — hence these elliptic curves with extra symmetries — the subgroup "$\Gamma(3)$" does. This subgroup consists of integer matrices*
>>
>> $$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$
>>
>> *with determinant 1 such that each entry is congruent to the corresponding entry of*
>>
>> $$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$
>>
>> *modulo 3.*
>>
>> *So, if we form*
>>
>> $$H/\Gamma(3)$$
>>
>> *we get a nice space without any "points of greater symmetry". To get the moduli space of elliptic curves from this, we just need to mod out by the group*
>>
>> $$\mathrm{SL}(2,\mathbb{Z})/\Gamma(3) = \mathrm{SL}(2,\mathbb{Z}/3)$$
>>
>> *But this group has 24 elements!*
>>
>> *In fact, I think this is just another way of explaining the period-24 pattern in the theory of modular forms, but I like it.*

> *Kevin wrote:*
>
> > *It's a rather pretty tree,*
>
> *Yes, there's a picture of it in Brown's book, drawn on top of an old picture by Klein of a triangulation of the hyperbolic plane.*
>
> *What Brown seems to be doing there is showing that this tree is a deformation retract of that triangulation (with its simplicial topology, where the points on the boundary of the hyperbolic plane form a discrete set), and thus proving that the cohomological dimension of* $\mathrm{SL}(2,\mathbb{Z})$ *is just 1.*
>
> *Anyway, this is all great stuff. Do you mind if I attach a copy of your email to the copy of "Week 213" on my website? I think people will find it helpful, especially because of its friendly straight-to-the-point style, which books rarely seem to manage. . . .*
>
> *Best,*
>
> *jb*

Kevin replied:

> *John Baez wrote:*
>
> > *I was actually wondering why Brown used the commutator subgroup of* $\mathrm{SL}(2,\mathbb{Z})$ *as a kind of "warmup" for computing the Euler characteristic of* $\mathrm{SL}(2,\mathbb{Z})$ *instead of one of the congruence subgroups. It seems this subgroup is not any of the beloved congruence subgroups. . . .*
>
> *You're right, I don't think it is. For* $N \geqslant 1$ *define* $\Gamma(N)$ *to be the kernel of the obvious map* $\mathrm{SL}(2,\mathbb{Z}) \to \mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z})$*; a congruence subgroup is any subgroup of* $\mathrm{SL}(2,\mathbb{Z})$ *that contains one of these* $\Gamma(N)$*'s. Clearly such things have finite index in* $\mathrm{SL}(2,\mathbb{Z})$*. But unfortunately there exist subgroups of finite index in* $\mathrm{SL}(2,\mathbb{Z})$ *that are not congruence subgroups. This is a "low-dimensional" phenomenon—the moment you have a bit more freedom, e.g. you're working with* $\mathrm{SL}(3,\mathbb{Z})$ *or indeed* $\mathrm{SL}(n,\mathbb{Z})$ *for any* $n \geqslant 3$*, or even* $\mathrm{SL}(2,\mathbb{Z}[1/p])$ *for some prime* p*, then any subgroup of finite index is a congruence subgroup—these groups satisfy the "congruence subgroup property". But I've never understood the commutator of* $\mathrm{SL}(2,\mathbb{Z})$ *precisely for the reason that it's not a congruence subgroup (this is essentially because the commutator subgroup of* $\mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z})$ *never has index 12 in* $\mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z})$*! The index is always smaller than 12 because* $\mathrm{SL}(2,\mathbb{Z}/p\mathbb{Z})$ *is essentially a simple group.)*
>
> *John Baez wrote:*
>
> > *In fact, I've finally managed to turn up the thing I was looking for. How does this relate to the stuff you're saying? It involves* $\Gamma(3)$ *rather than the* $\Gamma 1(N)$ *groups.*

Anything will do. If you know about $\Gamma(3)$ then great. The same key observation is true — $\Gamma(3)$ contains no elements of finite order, because any finite order element

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

of $\Gamma(3)$ which isn't the identity must have trace in $\{-2,-1,0,1\}$ congruent to $2 \mod 3$, so the trace must be -1, so $d = -1 - a$, so the det is $a(-1-a) \mod 9$, which is never $1 \mod 9$. Now the index of $\Gamma(3)$ in $\mathrm{SL}(2,\mathbb{Z})$ is 24, and the modular curve $X(3)$ has genus 0 (everyone knows this because Wiles needed it to prove Fermat's Last Theorem!) and four cusps (0, 1, $1/2$ and ∞) and hence the Euler Characteristic of $\Gamma(3)$ is $2 - 4 = -2$, so we recover the result that the Euler Characteristic of $\mathrm{SL}(2,\mathbb{Z})$ is $-1/12$ again.

John Baez wrote:

> *Where does the extra 24 come from? I don't know, but Stephan Stolz said it has something to do with the fact that while $\mathrm{PSL}(2,\mathbb{Z})$ doesn't act freely on the upper half-plane — hence these elliptic curves with extra symmetries — the subgroup "$\Gamma(3)$" does.*

One can see that any subgroup of $\mathrm{SL}(2,\mathbb{Z})$ which has finite index and is free, must have index a multiple of 12 (and hence at least 12). Because if it has index d and is free on g generators, when we know $(1-g)/d = -1/12$, so 12 divides the denominator of $(1-g)/d$ in lowest terms. Geometrically what is going on is that perhaps the "correct" quotient of the upper half plane by $\mathrm{SL}(2,\mathbb{Z})$ is not just the complex numbers, it's something that looks a bit like the complex numbers except there is a little bit of extra magic going on at i and ρ, corresponding to the fact that one shouldn't really have attempted to quotient out there, one should just remember that really the quotient is kind of "crumpled up" near there. So for example the fundamental group of the quotient shouldn't be the trivial group — if you take a small loop around i then this should not be regarded as contractible — you have to go around i twice before you can hope to contract the loop. Similarly you have to go around ρ three times. Even worse — if you do this carefully enough then even going around i twice or ρ three times isn't enough to contract the loop — because the resulting loop somehow corresponds to the element -1 in $\mathrm{SL}(2,\mathbb{Z})$, which acts trivially but which isn't the identity! So you have to do everything again before you get to the element 1. Mumford thought hard about how to make all this sort of thing rigorous, and managed in the late 60s to prove that the Picard group of the quotient of the upper half plane by $\mathrm{SL}(2,\mathbb{Z})$ was in fact $\mathbb{Z}/12\mathbb{Z}$.

John Baez wrote:

> *Anyway, this is all great stuff. Do you mind if I attach a copy of your email to the copy of "Week 213" on my website?*

Go ahead!

Kevin

Regarding the fundamental investigations of mathematics, there is no final ending ... no first beginning.

— *Felix Klein*

In point of fact, it has traditionally been the "continuous" aspect of things which has been the central focus of Geometry, while those properties associated with "discreteness", notably computational and combinatorial properties, have been passed over in silence or treated as an afterthought. It was therefore all the more astonishing to me when I made the discovery, about a dozen years ago, of the combinatorial theory of the icosahedron, even though this theory is barely scratched (and probably not even understood) in the classic treatise of Felix Klein on the icosahedron. I see in this another significant indicator of this indifference (of over 2000 years) of geometers vis-a-vis those discrete structures which present themselves naturally in Geometry: observe that the concept of the group (notably of symmetries) appeared only in the last century (introduced by Evariste Galois), in a context that was considered to have nothing to do with Geometry. Even in our own time it is true that there are lots of algebraists who still haven't understood that Galois theory is primarily, in essence, a geometrical vision, which was able to renew our understanding of so-called "arithmetical" phenomena.

— *Alexander Grothendieck*

Week 214

April 20, 2005

What common English slang phrase alludes to the number 168?

I won't tell you — not right away. But, I'll tell you a bunch of other cool stuff about this number, and eventually the answer should jump out at you.

Okay:

Start with a bunch of equilateral triangles. Glue them together so that 3 meet at each corner. You get a regular tetrahedron.

Next, take a bunch of squares. Glue them together so that 3 meet at each corner. You get a cube.

Next, take a bunch of regular pentagons. Glue them together so that three meet at each corner. You get a regular dodecahedron.

This is fun! We're getting a series of Platonic solids.

Next, take a bunch of regular hexagons and glue them together so that three meet at each corner. Now the angles of the hexagons add up to 360 degrees, so we don't get a Platonic solid. Instead, we get a tiling of the plane. It looks like a honeycomb that stretches out forever in all directions.

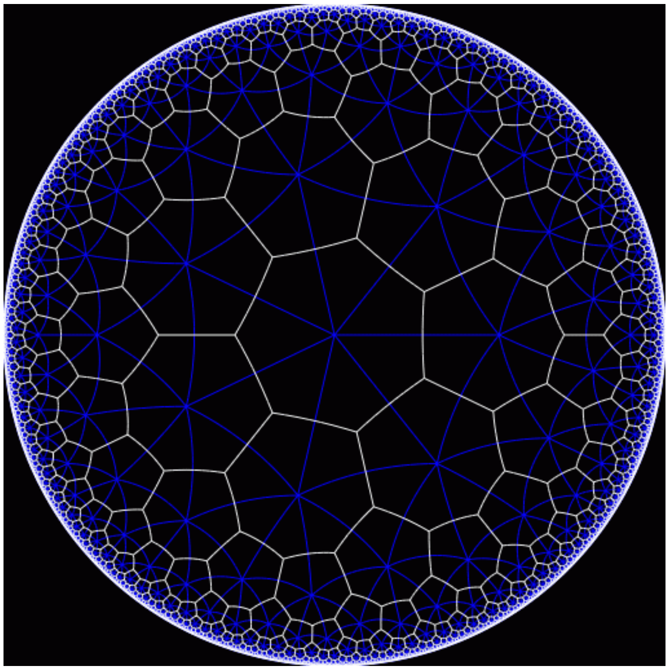
But, if you want something finite in size, you can cut out a portion of this honeycomb and curl it up to get a doughnut, or torus. There are actually lots of ways to do this. You might have fun figuring out what they all are. Can you take just *one* regular hexagon and curl it up to form a torus?

Anyway, these tori deserve to be called "Platonic surfaces", since they are surfaces tiled with regular polygons, with the same number meeting at each vertex.

Next, let's take a bunch of regular *heptagons* and glue them together so that three meet each corner. Now the angles add up to more than 360 degrees, so we get a tiling of the "hyperbolic plane". The hyperbolic plane is like the opposite of a sphere, since it's saddle-shaped at every point instead of bulging out at every point. In fact the sphere and the hyperbolic plane are the two most symmetrical forms of non-Euclidean geometry. The sphere is "positively curved", while the hyperbolic plane is "negatively curved".

You may have trouble visualizing the hyperbolic plane tiled with regular heptagons, but if we distort it, it fits into a disk and looks really pretty! Here it is:

1) Don Hatch, Hyperbolic planar tesselations, `http://www.hadron.org/\~hatch/HyperbolicTesselations/`



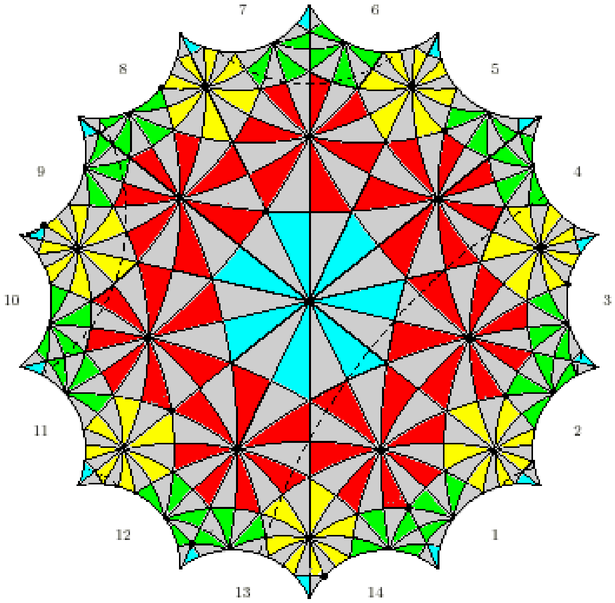
It's called "$\{7,3\}$", since it's made of 7-sided figures with 3 meeting at each corner.

In this picture there's one heptagon at the center, surrounded by rings of heptagons that appear smaller (but aren't really — that's just an effect of the distortion).

Can we cut out a portion of this tiling and curl it up to get a torus? No! But, we can curl up a portion to get a 3-holed torus — like the surface of a doughnut with three holes. But, we can only do this if we use precisely 24 heptagons!

Here's how we do it. Here's a picture of 24 heptagons, taken from an old paper by Klein and Fricke but prettied up a bit:

2) Tony Smith, Klein's quartic surface, `http://www.valdostamuseum.org/hamsmith/cdomain.html#tesselations`



You'll notice they're drawn in a fancy style: each heptagon has been "barycentrically subdivided" into 14 right triangles. But don't worry about that yet; concentrate on the heptagons.

There's a blue heptagon in the middle, 7 red ones touching that, 7 yellow ones touching those, then 7 green ones falling off the edge of the picture, and 2 blue ones broken into bits all around the corners of the picture. That's a total of 24 heptagons.

We wrap this thing up into a 3-holed torus using the numbers on the edges of the picture:

- connect edges 1 and 6
- connect edges 3 and 8
- connect edges 5 and 10
- connect edges 7 and 12
- connect edges 9 and 14
- connect edges 11 and 2
- connect edges 13 and 4

In other words, connect edges $2n+1$ and $2n+6 \mod 14$. To connect them the right way, make sure that triangles of the same color never touch each other.

Here's how to see if you get the idea. Ignore the little triangles; just pay attention to the heptagons! Then:

Start on any edge of any heptagon and march along in either direction.
Then, when you get to the end, turn left.
Then, when you get to the end, turn right.
Then, when you get to the end, turn left.
Then, when you get to the end, turn right.
Then, when you get to the end, turn left.
Then, when you get to the end, turn right.
Then, when you get to the end, turn left.
Then, when you get to the end, turn right.
You should now be back where you started!!!

These are like the driving directions the devil gives people who ask the way out of hell. LRLRLRLR and you're right back where you started.

But the resulting Platonic surface is heavenly. It has lots of symmetries. Each of the 24 heptagons has 7-fold rotational symmetry — and amazingly, all these rotations extend to a symmetry of the Platonic surface!

Now let's talk about those little triangles. Since our surface is made of 24 heptagons, each chopped into 14 right triangles, there are a total of

$$24 \times 14 = 336$$

triangles. And this number is also the number of symmetries of the Klein quartic, including reflections!

This is no coincidence. We can specify a symmetry by saying where it sends our favorite right triangle. Since it can go to any other triangle, there are 336 possibilities. If we exclude reflections, we get half as many symmetries: $24 \times 7 = 168$.

By the way, this trick works for ordinary Platonic solids as well. For example, if we take a dodecahedron and barycentrically subdivide all 12 pentagons, we get $10 \times 12 = 120$ right triangles. If we pick one of these as the "identity element", we can specify any symmetry by saying which triangle this triangle gets sent to. So, the set of triangles becomes a vivid **picture** of the 120-element symmetry group of the dodecahedron. It's called the "Coxeter complex". This idea generalizes in many directions, and is incredibly useful.

Anyway... there is much more to say about this stuff. For example, if we take our hyperbolic plane tiled with heptagons and count them grouped according to how far they are from the central one, we get the sequence

$$7, 7, 14, 21, 35, 56, 91, \ldots$$

These are 7 times the Fibonacci numbers!

To dig a bit deeper, though, it helps to think about complex analysis.

If we think of the hyperbolic plane as the unit disc in the complex plane, this surface becomes a "Riemann surface", meaning that it gets equipped with a complex structure. This was Felix Klein's viewpoint when he discovered all this stuff in about 1878. He showed this surface could be described by an incredibly symmetrical quartic equation in 3 complex variables:

$$u^3v + v^3w + w^3u = 0$$

where we count two solutions as the same if they differ by an overall factor. So, it's called "Klein's quartic curve".

(Why a "curve" and not a surface? Because it takes one *complex* number to say where you are on it. We have 3 unknowns and one equation, but we mod out by an overall factor, so we get something locally parametrized by one complex number... so algebraic geometers call it a curve.)

You can read Klein's original article translated into English. It's available online as part of a whole *book* about his incredible quartic:

3) Silvio Levy, *The Eightfold Way: the Beauty of Klein's Quartic Curve*, MSRI Research Publications **35**, Cambridge U. Press, Cambridge 1999. Also available as `http://www.msri.org/publications/books/Book35/`

This book was put out by the Mathematical Sciences Research Institute in Berkeley, to celebrate the completion of sculpture of Klein's quartic curve made by Helaman Ferguson. I must admit that the sculpture leaves me unmoved. But the curve itself — ah, that's another story!

For example, Klein's quartic curve turns out to have the maximum number of symmetries of any 3-holed Riemann surface.

Let's back up a minute and think about a Riemann surface with no holes: a sphere. There's only one way to make a sphere into a Riemann surface — it's called the Riemann sphere. You can think of it as the complex numbers plus a point at infinity. This has *infinitely* many symmetries. They're called conformal transformations, and they all look like this:

$$z \mapsto \frac{az+b}{cz+d}$$

They form a group called $\mathrm{PSL}(2,\mathbb{C})$, since it's the same as the group of 2×2 complex matrices with determinant 1, mod scalars. It's also the same as the Lorentz group!

There are different ways to make a torus into a Riemann surface, some with more symmetries than others (see "Week 124"). But, there are always translation symmetries in both directions, so the symmetry group is always infinite.

On the other hand, a Riemann surface with 2 or more holes can only have a *finite* group of conformal transformations. In fact, in 1893 Hurwitz proved that a Riemann surface with g holes has at most

$$84(g-1)$$

For $g = 3$, this is 168. So, Klein's quartic surface is as symmetrical as possible! (We don't count reflections here, since they don't preserve the complex structure — they're like complex conjugation.)

Now I should break down and give the best description of Klein's quartic curve as a Riemann surface. Sitting inside $\mathrm{PSL}(2,\mathbb{C})$ is $\mathrm{PSL}(2,\mathbb{Z})$, where we only use integers a, b, c, d in our fractional linear transformation

$$z \mapsto \frac{az+b}{cz+d}$$

This subgroup acts on the upper half-plane H, which is just another way of thinking about the hyperbolic plane.

Sitting inside $\mathrm{PSL}(2,\mathbb{Z})$ is a group $\Gamma(7)$ consisting of guys where the matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is congruent to the identity:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

modulo 7. This is an example of a "congruence subgroup"; these serve to relate complex analysis to number theory in lots of cool ways. In particular, Klein's quartic curve is just

$$H/\Gamma(7)$$

Since $\Gamma(7)$ is a normal subgroup of $\mathrm{PSL}(2,\mathbb{Z})$, the quotient group

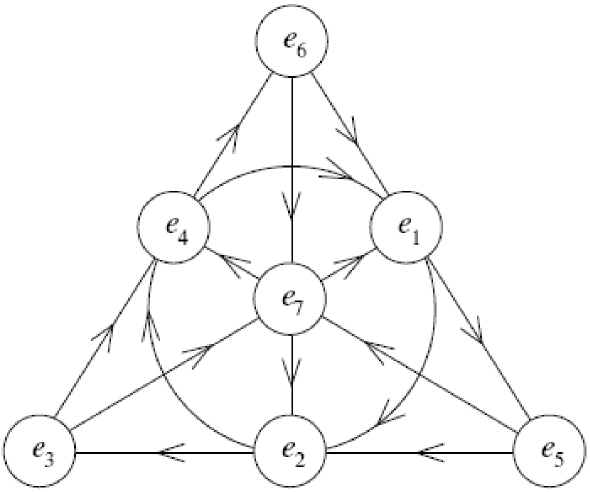
$$\mathrm{PSL}(2,\mathbb{Z})/\Gamma(7) = \mathrm{PSL}(2,\mathbb{Z}/7)$$

acts as symmetries of Klein's quartic curve. And, this group has 168 elements!

In fact, this group is the second smallest nonabelian simple group. The smallest one is the rotational symmetry group of the icosahedron, which has 60 elements. This group is actually $\mathrm{PSL}(2,\mathbb{Z}/5)$, and Klein had run into it in his work on the icosahedron and quintic equations (see "Week 213"). So, it's actually far from sheer luck that he then moved on to $\mathrm{PSL}(2,\mathbb{Z}/7)$ and ran into his wonderful quartic curve.

By the way, this 168-element group is also known as $\mathrm{PSL}(3,\mathbb{Z}/2)$ — the symmetry group of the "Fano plane". This is a name for the projective plane over $\mathbb{Z}/2$. The Fano plane is closely related to the octonions:

3) John Baez, "The Fano plane", `http://math.ucr.edu/home/baez/octonions/node4.html`

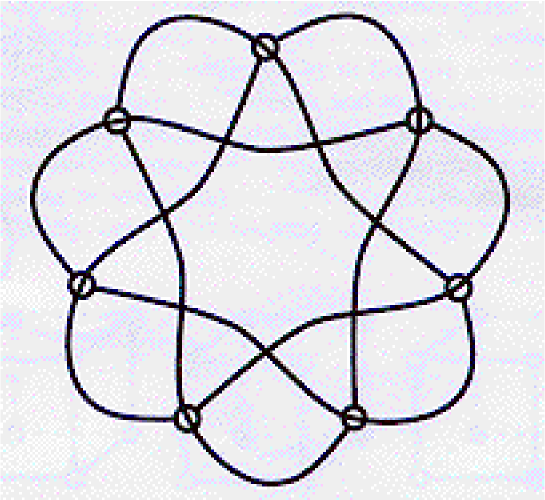


So in fact, our 168-element group acts on the set of octonion multiplication tables:

4) Tony Smith, "Octonion products", `http://www.valdostamuseum.org/hamsmith/480op.html`

5) Geoffrey Dixon, "Octonion multiplication tables", `http://www.7stones.com/Homepage/octotut0.html`

And, as James Dolan just noted today, and Tony Smith seems to have known all along, there's a way to draw the Fano plane that even *looks* like the diagram Klein and Fricke used to build the Klein quartic. Here's a picture drawn by Burkard Polster, author of "The Mathematics of Juggling" and "Geometries on Surfaces":



So, something interesting is going on, and I want to know what it is!

By the way, fans of the quaternions and octonions may like this review of Conway and Smith's book:

6) John Baez, "review of *On Quaternions and Octonions: Their Geometry, Arithmetic and Symmetry*, by John H. Conway and Derek A. Smith", *Bull. Amer. Math. Soc.* **42** (2005), 229–243. Available at `http://www.ams.org/bull/2005-42-02/` and `http://math.ucr.edu/home/baez/octonions/node24.html`

It's packed with cool pictures and weird facts — a more refined version of the material in "Week 193" and "Week 194".

It builds up to a kind of crazy climax in which I describe how when you pack spheres as densely as possible in 8 dimensions, each sphere touches 240 others... and if you look at the 240 neighbors of a given sphere, each one of those neighbors touches 56 other neighbors. Then I explain how this gives rise to a 56-dimensional representation of the exceptional group E_7 — its smallest nontrivial representation! And, how it gives rise to a 57-dimensional manifold on which the exceptional group E_8 acts — the smallest space on which it acts nontrivially!

Bertram Kostant is one of the real gurus of Lie theory. He teaches at MIT, and he has a strong fondness for exceptional Lie groups. When he saw this review of mine, he mentioned a couple of other papers that construct the **57**-dimensional space on which E_8 acts:

7) Ranee Brylinski and Bertram Kostant, "Lagrangian models of minimal representations of E_6, E_7, and E_8", in *Functional Analysis on the Eve of the 21st Century*, vol. **1**, Progress in Math. **131**, Birkhauser, Boston, 1995, pp. 13–53.

   Bertram Kostant, "Minimal coadjoint orbits and symplectic induction", in *The Breadth of Symplectic and Poisson geometry*, 391–422, Progress in Math. **232**, Birkhauser, Boston, 2005. Also available as `http://www.arXiv.org/abs/math.SG/0312252`

I've got to read these sometime.

Having the number 56 on my brain, I can't resist nothing that if you take Klein's quartic curve tiled by heptagons, and you count the vertices, you get

$$24 \times 7/3 = 56$$

since each vertex is shared by 3 heptagons. I'm hoping this is not a coincidence!

Okay, that's all for this week, except for some silly stuff. . . .

First of all, speaking of octonions, Geoff Corbishley just told me that their inventor, John Thomas Graves, is a relative of Robert Graves — the author of "I Claudius".

Second of all, I hope you've figured out the puzzle I gave at the beginning of this Week. The phrase is "24-7", as in "we're working on it 24-7". 24 hours a day, 7 days a week, makes 168 hours per week!

Finally, speaking of numerology, this number 168 is related to why the days of the week have the names they do! I explained why in "Week 175", but I'll remind you:

Astrologers liked to list the planets in order of decreasing orbital period, counting the sun as having a period of one year, and the moon as period of one month:

Planet	Orbital period
Saturn	29 years
Jupiter	12 years
Mars	687 days
Sun	365 days
Venus	224 days
Mercury	88 days
Moon	29.5 days

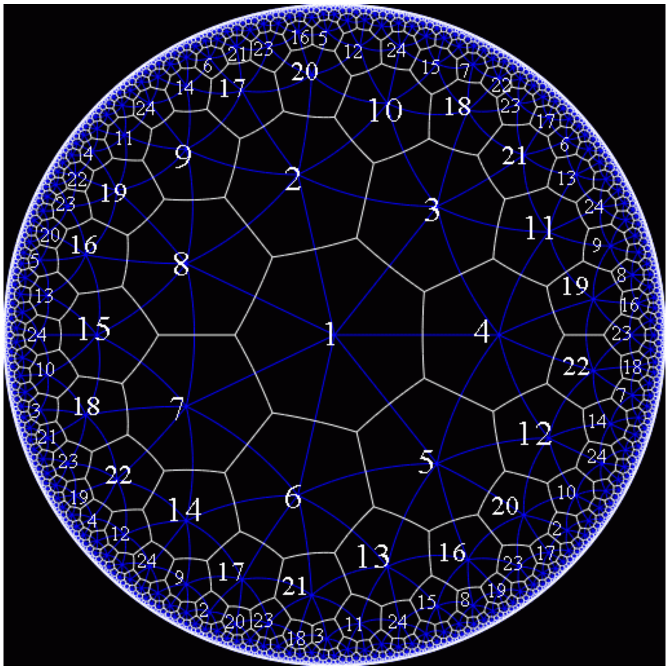
For the purposes of astrology they wanted to assign a planet to each hour of each day of the week. To do this, they assigned Saturn to the first hour of the first day, Jupiter to the second hour of the first day, and so on, cycling through the list of planets over and over, until each of the $24 \times 7 = 168$ hours was assigned a planet. Each day was then named after the first hour in that day. Since 24 mod 7 equals 3, this amounts to taking the above list and cycling around it, reading off every third planet:

Saturn	Saturday
Sun	Sunday
Moon	Monday
Mars	Tuesday
Mercury	Wednesday

Jupiter	Thursday
Venus	Friday

And that's how they got listed in this order! At least, this is what the Roman historian Dion Cassius (AD 150–235) claims. Nobody knows for sure.

Addendum: Mike Stay took Don Hatch's picture and drew numbers from 1 to 24 showing how to identify heptagons in order to get the Klein quartic curve:



Gerard Westendorp had some interesting comments on what I wrote:

If you take Euler's formula

$$V + F - E = 2 - 2 \times holes$$

then you can figure out that for a $(7,3)$ *tiling with* N *heptagons, you have*

$$V = \frac{7N}{3}$$
$$F = N$$
$$E = \frac{7N}{2}$$

so that

$$N = 12 \times (holes - 1)$$

Here's the table of solutions:

holes	*N*
0	*-12*
1	*0*
2	*12*
3	*24*
4	*36*
⋮	⋮

So indeed, there are no solutions for 0 holes (sphere) or 1 hole (torus). But a 2*-holed torus should be possible, as well as the* 24*-faced* 3*-holed one.*

Anyway, see if I can visualise the 3*-holed one.*

If you start with a sphere, i.e. genus 0, and drill a tunnel through it, you will get a genus 1 object. On the outer surface, you see 2 holes, one for each side of the "tunnel". (I use the word "tunnel" for something into a 3D object, and "hole" for something in a 2D surface.)

Next, you can drill a second tunnel, and get a genus 2 object, and you would see 4 holes on the outer surface. But a nice trick is to drill not to the outer surface, but to a secret "cave" in the middle where you meet the first tunnel. Here you stop drilling. To complete the genus-3 object, you drill the third tunnel again not to the outer surface, but to the central cave. Thus, you get and object with genus 3, which has 4 holes on its outer surface, each leading to a central cave.

Confusingly, 4 tunnels to a central cave is topologically the same as 3 separate tunnels! The trick is that tunnels do not have to end on the outer surface, the inner surface is topologically the same.

OK, so we have an object with 4 holes on its outer surface. 4 holes → tetrahedron…

I built a cardboard model of a tetrahedron with a central cave. Truncated tetrahedrons together with tetrahedrons can fill space. So you can stack 4 truncated tetrahedra on top of each other, leaving a hole in the shape of an imaginary 5th one. Then use tetrahedra to fill up some gaps. This was basically the shape I built out of cardboard. Then, I spent rather a long time trying to tile this with heptagons. A clue to a solution was that a triangulation of the surface I made had 120 triangle, and $120 = 24 \times 5$*. What is so good about 5? Well, 5 triangles stuck together have 7 outer sides, sop they are a kind of pseudo heptagons. Anyway, I got a bit frustrated, and did not find a nice tiling.*

As I was trying to figure it out, I found this site:

`http://www.math.uni-siegen.de/wills/klein/`

It has some nice pictures.

> *These are like the driving directions the devil gives people who ask the way out of hell. LRLRLRLR and you're right back where you started.*

> *Btw, this works the same on other polyhedra, e.g. the cube.*
>
> > - *Saturn (Saturday)*
> > - *Sun (Sunday)*
> > - *Moon (Monday)*
> > - *Mars (Tuesday)*
> > - *Mercury (Wednesday)*
> > - *Jupiter (Thursday)*
> > - *Venus (Friday)*
>
> *My French is not so good, but in French some names look more convincing:*
>
> - *tuesday = mardi (Mars?)*
> - *wednesday = mercredi (Mercury?)*
> - *thursday = jeudi (Jove?)*
> - *friday = vendredi (Venus?)*
>
> *Gerard*

I replied:

> *Gerard Westendorp wrote:*
>
> > *John Baez wrote:*
> >
> > > *It's called "$\{7,3\}$", since it's made of 7-sided figures with 3 meeting at each corner.*
> > >
> > > *Can we cut out a portion of this tiling and curl it up to get a torus? No! But, we can curl up a portion to get a 3-holed torus — like the surface of a doughnut with three holes. But, we can only do this if we use precisely 24 heptagons!*
> >
> > *If you take Euler's formula [. . . .]*
> >
> > *So indeed, there are no solutions for 0 holes (sphere), 1 hole(torus). But a 2-holed torus should be possible, as well as the 24-faced 3-holed one.*
>
> *I was going to talk about this, but I figured my article was getting too long.*
>
> *Indeed, Euler's formula also allows the possibility of a* 2-holed *torus tiled with 12 heptagons meeting 3 at each corner.*
>
> *But this does not prove such a tiling is possible. I don't know if it is! Someone should try it.*
>
> *However: even if such a tiling exists, it's not possible for each rotational symmetry of each heptagon to extend to a symmetry of the whole tiled surface. What's marvelous about the 3-holed case is that they all do — at least if you do things correctly. This is what makes the Klein quartic a full-fledged "Platonic surface".*
>
> *If you look here:*

- *Hermann Karcher and Mattias Weber, "The Geometry of Klein's Riemann Surface", in* The Eightfold Way: the Beauty of Klein's Quartic Curve, *ed. Silvio Levy, MSRI Research Publications* ***35****, Cambridge U. Press, Cambridge 1999. Also available as PDF and gzipped Postscript.*

you'll see that Karcher and Weber study Platonic surfaces using Euler's formula.

On pages 13–19 they consider Platonic surfaces with 2 holes. On page 19 they give a clever proof that no tiling of the 2-holed torus by heptagons meeting 3 at each corner can be a Platonic surface. The proof is so clever that I don't understand it.

(Warning: their article starts on page 9.)

> *Anyway, see if I can visualise the 3-holed one.*

I wish I could visualize it myself.

> *As I was trying to figure it out, I found this site:*
>
> `http://www.math.uni-siegen.de/wills/klein/`
>
> *It has some nice pictures.*

These pictures are interesting, but what I'd really like is a nice picture of a 3-holed torus, not weird or crumpled up, which is tiled by 24 heptagons just like the Klein quartic.

The heptagons can't all be regular if the torus is embedded in $\mathbb{R}^3$, since there's no way to embed a compact surface of constant negative curvature in $\mathbb{R}^3$. But, you can get the topology *correct and still have the torus embedded in $\mathbb{R}^3$.*

If anyone draws such a picture, and I think it looks nice, I'd love to put it on This Week's Finds!

If anyone wants instructions on how such a surface should be made, look above, where Mike Stay has kindly drawn numbers from 1–24 on a portion of the hyperbolic plane tiled with heptagons. These numbers indicate how to identify heptagons to get the Klein quartic. For example, all the heptagons labelled "21" are really the same heptagon in the Klein quartic.

> - *Saturn (Saturday)*
> - *Sun (Sunday)*
> - *Moon (Monday)*
> - *Mars (Tuesday)*
> - *Mercury (Wednesday)*
> - *Jupiter (Thursday)*
> - *Venus (Friday)*
>
> *My French is not so good, but in French some names look more convincing:*
>
> - *tuesday = mardi (Mars?)*

- *wednesday = mercredi (Mercury?)*
- *thursday = jeudi (Jove?)*
- *friday = vendredi (Venus?)*

Yes, this because most of the English names for planets come from Latin, and French is more like Latin.

English is more complicated, but I'm so used to it that I forgot people might find the connection to Latin mysterious:

- *"Tuesday" comes from "Tiu" or "Tyr", an ancient Germanic god of war whom the Romans identified with Mars. We see traces of this in the German "Dienstag" as well.*
- *"Wednesday" comes from "Woden" or "Odin", a Germanic god whom the Romans identified with Mercury. Modern German uses "Mittwoch" instead, which means "mid-week".*
- *"Thursday" comes from "Thor", a Germanic thunder god whom the Romans identified with Jupiter. Modern German uses "Donnerstag" instead, with "Donner" meaning "thunder".*
- *"Friday" comes from "Freya" or "Frigga", a Germanic goddess of married love whom the Romans identified with Venus. The German "Freitag" is very similar.*

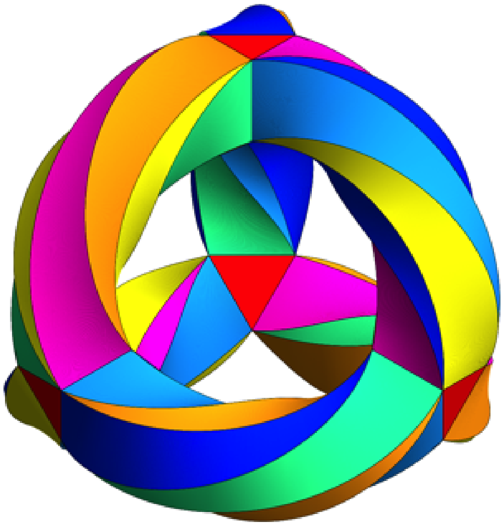
Week 215

May 9, 2005

This week I'd like to report on some cool things people have been explaining to me. The science fiction writer Greg Egan has been helping me understand Klein's quartic curve, and the mathematician Darin Brown has been explaining the analogy between geodesics and prime numbers. The two subjects even overlap slightly!

Last week I talked about Klein's quartic curve. This led Gerard Westendorp and Mike Stay to draw some pictures of it, and their ideas helped Greg Egan create this really nice picture:

1) Greg Egan, "Klein's quartic curve", `http://math.ucr.edu/home/baez/mathematical/KleinDual.png`



It looks sort of tetrahedral at first glance, but if you look carefully you'll see that topologically speaking, it's a 3-holed torus. It's tiled by triangles, with 7 meeting at each vertex. So, it's the Klein quartic curve!

Perhaps I should explain. Last week I talked about a tiling of the hyperbolic plane by regular heptagons with 3 heptagons meeting at each vertex. Dual to this is a tiling of the hyperbolic plane by equilateral triangles with 7 triangles meeting at each vertex. We can take a quotient space of this by a certain symmetry group and get a 3-holed torus tiled by 56 triangles with 7 meeting at each vertex. This is what Egan drew!

With this picture you can almost *see* the 168 symmetries of Klein's quartic curve.

First, you can take any vertex and twist it, causing the 7 triangles that meet at this vertex to cycle around. It's not obvious that this is a symmetry of the whole tiled surface, but it is. This gives a 7-element symmetry group.

Second, the whole thing looks like a tetrahedron, so it inherits the rotational symmetries of a tetrahedron. This gives a more obvious 12-element symmetry group.

$7 \times 12 = 84$, so how do we get a total of 168 symmetries?

Well, there's also a 2-fold symmetry that corresponds to turning the tetrahedron inside out! And Egan made a wonderful *movie* of this. If a picture is worth a thousand words, this is worth about a million:

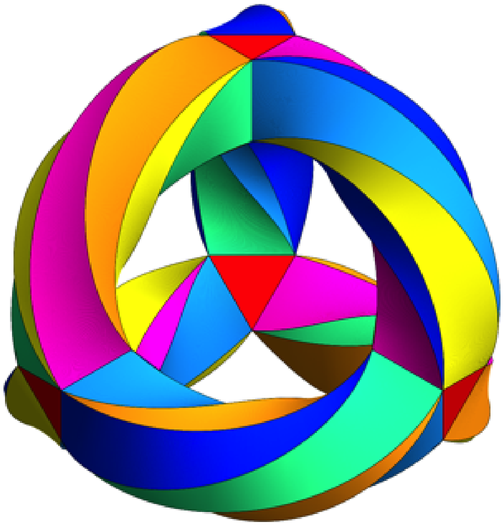
2) Greg Egan, "Turning Klein's quartic curve inside out", `http://math.ucr.edu/home/baez/mathematical/KleinDualInsideOut.gif`

So, we get a total of $7 \times 24 = 168$ symmetries.

Even better, if you watch carefully, you'll see that the tetrahedron in Egan's movie gets *reflected* as it turns inside out. More precisely, if you follow the four corners of the tetrahedron, you'll see that two come back to where they were, while the other two get switched. So, this symmetry acts as a reflection, or odd permutation, of the 4 corners. The rotations act as even permutations of the corners.

This means that the Klein quartic has 24 symmetries forming a group isomorphic to the rotation/reflection symmetry group of a tetrahedron. Algebraically speaking, this group is S_4: the permutations of 4 things.

This group is also the rotational symmetry group of a cube. In fact, Egan was able to spot a hidden cube lurking in his picture! Can you?



If you look carefully, you'll see each corner of his tetrahedral gadget is made of a little triangular prism with one triangle facing out and one facing in: for example, the pink triangle staring you right in the face, or the light blue one on top. Since $4 \times 2 = 8$, there are 8 of these triangles. Abstractly, we can think of these as the 8 corners of a cube! They aren't really, but we can pretend. The way these 8 triangles come in pairs corresponds to how the vertices of a cube come in diagonally opposite pairs.

Using this, you can see that the group S_4 acts on these 8 triangles in precisely the same way it acts via rotations on the vertices of a cube.

In fact, you can even draw a *picture* of a cube on the Klein quartic by drawing suitable curves that connect the centers of these 8 triangles! It's horribly distorted, but topologically correct. Part of the distortion is caused by embedding the Klein quartic in ordinary 3d Euclidean space. If we gave the Klein quartic the metric it inherits from the hyperbolic plane, the edges of the cube would be geodesics.

This remark also helps us see something else. The Klein quartic is tiled by 56 triangles. 8 of them give the cube we've just been discussing. In Egan's picture these triangles look special, since they lie at the corners of his tetrahedral gadget. But this is just an illusion caused by embedding the Klein quartic in 3d space. In reality, the Klein quartic is perfectly symmetrical: every triangle is just like every other. So in fact there are lots of these cubes, and every triangle lies in some cube.

But this is where it gets really cool. In fact, each triangle lies in just *one* cube. So, there's precisely one way to take the 56 triangles and divide them into 7 bunches of 8 so that each bunch forms a cube.

So: the symmetry group of the Klein quartic acts on the set of cubes, which has 7 elements.

But as I explained last week, this symmetry group also acts on the Fano plane, which has 7 points.

This suggests that cubes in the Klein quartic naturally correspond to points of the Fano plane. And Egan showed this is true!

He showed this by showing more. The Fano plane also has 7 lines. What 7 things in the Klein quartic do these lines correspond to?

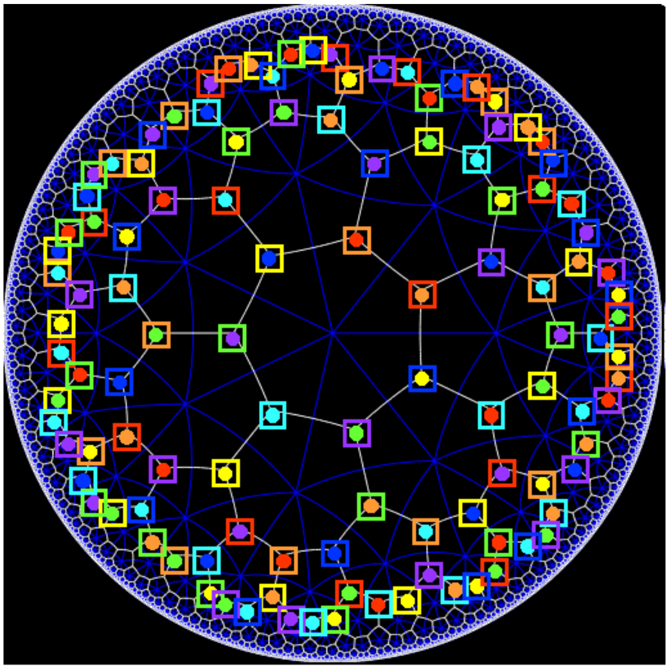
Anticubes!

You see, the cubes in the Klein quartic have an inherent handedness to them. You can go between the 8 triangles of a given cube by following certain driving directions, but these driving directions involve some left and right turns. If you follow the mirror-image driving directions with "left" and "right" switched, you'll get an *anticube*.

Apart from having the opposite handedness, anticubes are just like cubes. In particular, there's precisely one way to take the 56 triangles and divide them into 7 bunches of 8 so that each bunch forms an anticube.

Here's a picture:

3) Greg Egan, Cubes and anticubes in the Klein quartic curve, `http://math.ucr.edu/home/baez/KleinFigures.gif`



Each triangle has a colored circle and a colored square on it. There are 7 colors. The colored circle says which of the 7 *cubes* the triangle belongs to. The colored square says which of the 7 *anticubes* it belongs to.

If you stare at this picture for a few hours, you'll see that each cube is completely disjoint from precisely 3 anticubes. Similarly, each anticube is completely disjoint from precisely 3 cubes.

This is just like the Fano plane, where each point lies on 3 lines, and each line contains 3 points!

So, we get a vivid way of seeing how every figure in the Fano plane corresponds to some figure in the Klein quartic curve. This is why they have the same symmetry group.

This is an excellent example of Klein's Erlangen program for reducing geometry to group theory, which I discussed in "Week 213". Here we are beginning to see how two superficially different geometries are secretly the same:

Fano plane	Klein's quartic curve
7 points	7 cubes
7 lines	7 anticubes
incidence of points and lines	disjointness of cubes and anticubes

However, we're only half done! We've seen how to translate simple figures and incidence relations in the Fano plane to complicated ones in Klein's quartic curve. But, we haven't figured out translate back!

Klein's quartic curve	Fano plane
24 vertices	???
84 edges	???
56 triangular faces	???
incidence of vertices and edges	???
incidence of edges and faces	???

Here I'm talking about the tiling of Klein's quartic curve by 56 equilateral triangles. We could equally well talk about its tiling by 24 regular heptagons, which is the Poincare dual. Either way, the puzzle is to fill in the question marks. I don't know the answer!

To conclude — at least for now — I want to give the driving directions that define a "cube" or an "anticube" in Klein's quartic curve. Say you're on some triangle and you want to get to a nearby triangle that belongs to the same cube. Here's what you do:

> *hop across any edge,*
> *turn right,*
> *hop across the edge in front of you,*
> *turn left,*
> *then hop across the edge in front of you.*

Or, suppose you're on some triangle and you want to get to another that's in the same anticube. Here's what you do:

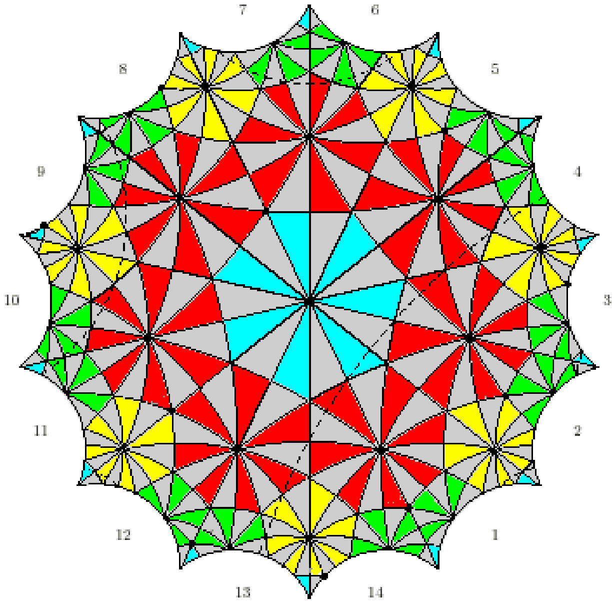
> *hop across any edge,*
> *turn left,*
> *hop across the edge in front of you,*
> *turn right,*
> *then hop across the edge in front of you.*

(If you don't understand this stuff, look at the picture above and see how to get from any circle or square to any other circle or square of the same color.)

You'll notice that these instructions are mirror-image versions of each other. They're also both $1/4$ of the "driving directions from hell" that I described last time. In other words, if you go LRLRLRLR or RLRLRLRL, you wind up at the same triangle you started from. You'll have circled around one face of a cube or anticube!

In fact, your path will be a closed geodesic on the Klein quartic curve... like the long dashed line in Klein and Fricke's original picture:

4) Klein and Fricke, Klein's quartic curve with geodesic, `http://math.ucr.edu/home/baez/Klein168.gif`



Next, a little about geodesics and prime numbers. I've just been talking a little about geodesics in the Klein quartic, which is the quotient

$$H/G$$

of the hyperbolic plane H by a certain group G which I explained last week. This group, usually called $\Gamma(7)$, is a nice example of a "Fuchsian group" — that is, a discrete subgroup of the isometries of the hyperbolic plane.

Darin Brown and his thesis advisor Jeff Stopple at U. C. Santa Barbara have been thinking about geodesics in H/G for other Fuchsian groups G, and their relation to number theory:

5) Jeff Stopple, "A reciprocity law for prime geodesics", *J. Number Theory* **29** (1988), 224–230.

6) Darin Brown, *Lifting properties of prime geodesics on hyperbolic surfaces*, Ph.D. thesis, U. C. Santa Barbara, 2004.

I'd really like to learn about this, because it connects all sorts of stuff I dream of understanding someday, especially quantum chaos ("Week 190"), zeta functions in physics and number theory ("Week 199"), and Galois theory as a theory of covering spaces ("Week 205"). Also, it involves a big mysterious analogy, and I always like those!

I don't understand this stuff well enough to try a full-fledged explanation yet, so I'll just give a vague sketch. A "prime geodesic" in a Riemannian manifold X is a closed geodesic

$$f\colon S^1 \to X$$

that cycles around just once. In other words, f should be one-to-one.

We say a closed geodesic is the "nth power" of a prime one if it's just like the prime one but it cycles around n times. Every closed geodesic is the nth power of a prime one in a unique way.

If we have a Fuchsian group G, H/G is a surface with a Riemannian metric. It looks locally like the hyperbolic plane, so it's called a "hyperbolic surface". And, we can look at prime geodesics in it.

If G' is a subgroup of G, we get a covering map

$$H/G' \to H/G$$

so we can ask about lifting prime geodesics in H/G to closed geodesics in H/G'. There can be a bunch of ways to do this, so we say a prime geodesic in H/G "splits" into powers of prime geodesics up in H/G'.

If you know any number theory — reading "Week 205" should be enough — this should remind you of how a prime ideal in some algebraic number field can "split" into prime ideals in an extension of this field, and/or "ramify" into powers of prime ideals.

And indeed, Darin Brown has found a big mysterious analogy that goes like this:

Number field K	Hyperbolic surface H/G
Field extension K' of K	Covering $p\colon H/G' \to H/G$
Galois group $\mathrm{Gal}(K'/K)$	Deck transformation group $\mathrm{Aut}(p)$
Prime ideal Q of K	Prime geodesic f in H/G
Prime ideal Q' lying over Q	Prime geodesic f' lying over f
Splitting of prime ideal Q of K'	Lifting of prime geodesic f to H/G'
Norm $N(Q)$ of ideal Q	Norm $N(f)$ of closed geodesic f
Frobenius conjugacy class of Q	Frobenius conjugacy class of f
Artin L-function	Selberg zeta function

(Here by "prime ideal of K" we mean a prime ideal in the ring of algebraic integers of K.)

But this is more than an analogy: there's even a way to associate number fields to certain hyperbolic surfaces! The reason is that often Fuchsian groups will consist of matrices whose entries lie in some number field.

I would like to understand the Selberg zeta function and its relation to quantum mechanics. The Selberg zeta function is related to closed geodesics, which are periodic classical trajectories, while the zeta function of a Laplacian is related to periodic *quantum* trajectories (namely eigenfunctions of the Laplacian). So, the two are related. I know there's a lot of cool stuff going on here — especially since the motion of a particle on a hyperbolic surface tends to be chaotic, so "quantum chaos" rears its ugly head. But, I don't understand any of the details.

In some notes on quantum chaos, Gutzwiller wrote:

> *The classical periodic orbits are a crucial stepping stone in the understanding of quantum mechanics, in particular when then classical system is chaotic. This situation is very satisfying when one thinks of Poincar who emphasized the importance of periodic orbits in classical mechanics, but could not have had any*

> *idea of what they could mean for quantum mechanics. The set of energy levels and the set of periodic orbits are complementary to each other since they are essentially related through a Fourier transform. Such a relation had been found earlier by the mathematicians in the study of the Laplacian operator on Riemannian surfaces with constant negative curvature. This led to Selberg's trace formula in 1956 which has exactly the same form, but happens to be exact. The mathematical proof, however, is based on the high degree of symmetry of these surfaces which can be compared to the sphere, although the negative curvature allows for many more different shapes.*

When I get serious, I'll read these:

7) M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics*, Springer, Berlin, 1990.

8) Predrag Cvitanovic, Roberto Artuso, Per Dahlqvist, Ronnie Mainieri, Gregor Tanner, Gabor Vattay, Niall Whelan and Andreas Wirzba, *Chaos: Classical and Quantum*, available at `http://www.nbi.dk/ChaosBook/`

9) Svetlana Katok, *Fuchsian Groups*, U. Chicago Press, Chicago, 1992.

10) J. Elstrodt, F. Grunewald, and J. Mennicke, *Groups Acting on Hyperbolic Space*, Springer, Berlin, 1998.

11) Peter Sarnak, "Quantum chaos, symmetry and zeta functions", in *Current Developments in Mathematics*, 1997, eds R. Bott et al., International Press, Boston, 1999, pp. 127–159.

12) C. Schmit, "Quantum and classical properties of some billiards on the hyperbolic plane", in *Chaos and Quantum Physics*, eds. M.-J. Giannoni et al., Elsevier, New York, 1991, pp. 333–369.

For a nice pop treatment of quantum chaos and the Riemann hypothesis, try this:

13) Martin Gutzwiller, "Quantum chaos", *Scientific American*, January 1992. Also available at `http://www.maths.ex.ac.uk/~mwatkins/zeta/quantumchaos.html`

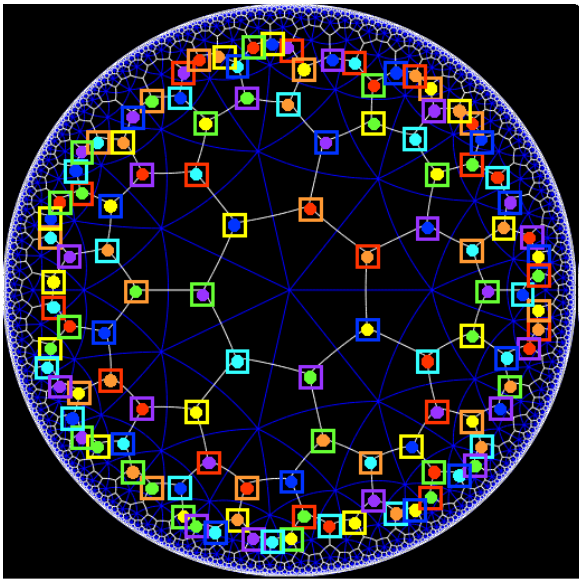
Addendum: Here is some email I got from Greg Egan and Toby Bartels, and a post from Darin Brown which corrects some mistakes and answers some questions raised by my pal Squark.

Greg Egan wrote me the following after I suggested a relation between the Klein quartic curve and 3d Minkowski spacetime over the field $\mathbb{Z}/7$ — a relation that he later exploited in some fascinating ways.

> *Hi*
>
> *Thanks for all the Lorentz group stuff! This will take me a while to digest.*
>
> *In the meantime, here are some more translations between the geometries.*
>
> *Every cube intersects 4 anticubes, and any pair of cubes, between them, intersect 6 anticubes (two of the 4 for each will always be shared). So together the pair*

of cubes single out one anticube: the 7th one that neither of them intersect. This is analogous to the fact that any two Fano points single out a Fano line.

I'll write $\mathrm{anti}(\{c_1, c_2\})$ *for the anticube singled out by a pair of cubes, and similarly* $\mathrm{cube}(\{a_1, a_2\})$ *for the cube singled out by a pair of anticubes. In the scheme used in this diagram:*

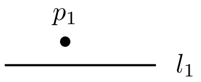


both functions have identical outputs for the same input colours:

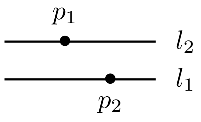
Table 16: $\mathrm{anti}(\{c_1, c_2\})$ *and* $\mathrm{cube}(\{a_1, a_2\})$

	R	***O***	***Y***	***G***	***LB***	***P***	***DB***
R	—	*DB*	*R*	*DB*	*Y*	*Y*	*R*
O	*DB*	—	*P*	*DB*	*P*	*O*	*O*
Y	*R*	*P*	—	*LB*	*P*	*LB*	*R*
G	*DB*	*DB*	*LB*	—	*G*	*LB*	*G*
LB	*Y*	*P*	*P*	*G*	—	*Y*	*G*
P	*Y*	*O*	*LB*	*LB*	*Y*	—	*O*
DB	*R*	*O*	*R*	*G*	*G*	*O*	—

Now for some actual translations.

Klein's Quartic Curve	*Fano plane*
28 pairs of opposite triangular faces	*28 choices of a point and a non-incident line,* $\{p, l\}$*.*  $7 \times 4 = 28$

In Klein's quartic curve, we specify a pair of opposite triangular faces by picking one of seven cubes, then one of four anticubes that intersect it. The intersection is a pair of triangular faces which are diagonally opposite each other both on the cube and on the anticube. The 56 order-3 *elements of* G *preserve these pairs of triangular faces, and consist of rotations by* $1/3$ *and* $2/3$ *turns for each such pair.*

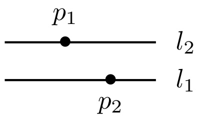
Triangular faces sharing an edge	*Pairings of points and non-incident lines* $\{p_1, l_1\}$ *and* $\{p_1, l_1\}$ *having* p_1 *incident on* l_2 *and* p_2 *incident on* l_1*.* 

In Klein's quartic curve, whenever two triangular faces share an edge, the cube each face belongs to will be disjoint from the anticube that the other face belongs to. This can be checked by noting that the colour of the anticube appears in the row for $\mathrm{anti}(c, \cdot)$*.*

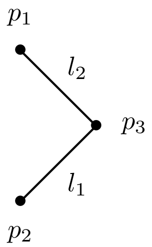
If you inspect a triangle and the three neighbours that share edges with it, the neighbours will always belong to the three anticubes disjoint from the cube the central triangle belongs to, i.e. they will have exactly the three colours appearing in the row for $\mathrm{anti}(c, -)$

84 edges

84 choices of $\{p_1, l_1\}$ and $\{p_2, l_2\}$ non-incident, but $\{p_1, l_2\}$ and $\{p_2, l_1\}$ incident.



or equivalently: 84 choices of 3 non-colinear points with one point singled out. In this definition, the special 3rd point is the one point shared by l_1 and l_2 of the previous definition.



We can count this as $\binom{7}{3}$ triples, minus 7 triples that are colinear, times three for three choices of distinguished point:

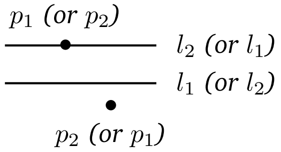
$$(\binom{7}{3} - 7) \times 3 = 28 \times 3 = 84$$

In Klein's quartic curve, we specify an edge by picking a pair of cubes $\{c_1, c_2\}$ and then a distinguished third one, c_3, so that the three aren't all disjoint from any one anticube. This means that, between them, they must intersect all seven anticubes. So the third cube must be one that intersects $\mathrm{anti}(\{c_1, c_2\})$. There are exactly 4 of these (and c_1 and c_2 aren't among them, by definition). So another way of counting the total is $((\binom{7}{2})) \times 4 = 21 \times 4 = 84$ choices.

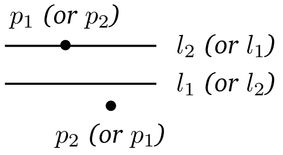
To identify the particular edge, suppose we've chosen $\{\{c_1, c_2\}, c_3\}$ as our cubes. Then $\{c_1, \mathrm{anti}(\{c_2, c_3\})\}$ is a cube and an anticube that intersects it, which specifies a pair of diagonally opposite triangular faces, and the same is true of $\{c_2, \mathrm{anti}(\{c_1, c_3\})\}$. There is a unique edge where two of these triangles meet.

For example, if we pick $\{\{$red, orange$\}$, yellow$\}$ then we have $\{$red, anti-purple$\}$ and $\{$orange, anti-red$\}$. Both cube/anticube choices specify two triangles each, *but there is only one edge that belongs to both a* $\{$*red*, *anti-purple*$\}$ *and an* $\{$*orange*, *anti-red*$\}$ *triangle.*

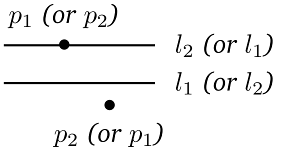
To reverse the process, if we look at the cube/anticube markings on the triangles either side of some edge, and they are $\{c_1, a_1\}$ *and* $\{c_2, a_2\}$*, then we can describe this edge by* $\{\{c_1, c_2\}, \mathrm{cube}(\{a_1, a_2\})\}$.

Triangular faces each sharing an edge with a common neighbour, but not each other. (This is sufficient, but not necessary, for them to share a vertex.)	*Pairings of points and non-incident lines* $\{p_1, l_1\}$ *and* $\{p_2, l_2\}$ *having* either p_1 *incident on* l_2 *or* p_2 *incident on* l_1*, but* not *both.*  p_1 (or p_2) l_2 (or l_1) l_1 (or l_2) p_2 (or p_1)

In Klein's quartic curve, as you go around a triangle anticlockwise and look at its three (edge-sharing) neighbours, the cube a triangle belongs to will be disjoint from the anticube of the triangle that follows, but the anticube it belongs to will intersect the cube of the triangle that follows. (But what the sense of the rotation means in the Fano plane depends on whether we map cubes to points and anticubes to lines or vice versa!)

24 vertices	*168 pairings of points and non-incident lines* $\{p_1, l_1\}$ *and* $\{p_2, l_2\}$ *having* either p_1 *incident on* l_2 *or* p_2 *incident on* l_1*, but* not *both.*  p_1 (or p_2) l_2 (or l_1) l_1 (or l_2) p_2 (or p_1) *There are: (*28 *choices for* $\{p_1, l_1\}$*)* × *(*3 *choices for* l_2 *passing through* p_1*)* × *(*$7 - 5 = 2$ *choices for* p_2 *not in* l_1 *or* l_2*). This count identifies each vertex as shared by common neighbours of a particular triangle, so we expect to count each vertex 7 times for the seven triangles. We could double this to count for swapping the role of* p_1 *and* p_2*, and the we'd be counting each vertex twice as often: once going anticlockwise between each pair of neighbours, and once going clockwise*

This is all a bit strange and inconvenient! I can pin down an edge, but I haven't really pinned down a single face, or a way to count a vertex just once. I guess the answer for a vertex is to talk about an equivalence class of the structures:



where we mod out by $\mathbb{Z}/7$ *and "gauge fix"* l_1*. Every vertex is surrounded by 7 triangular faces encompassing all seven cubes and all seven anticubes, so these equivalence classes do fix a single vertex.*

Best wishes

Greg

Toby Bartels wrote:

In Week 215, you wrote:

> *We say a closed geodesic is the "*n*th power" of a prime one if it's just like the prime one but it cycles around* n *times. Every closed geodesic is the* n*th power of a prime one in a unique way.*

> *The latter sentence is not quite true; you've forgotten $n = 0$ again!*
>
> *Some manifolds, like the real line, have no prime geodesics, but every (pointed) manifold has a unique unit closed geodesic, which is the geodesic that just sits at the basepoint the whole time. Given any prime geodesic f, this unit geodesic is f^0.*
>
> *Thinking about this, I noticed that multiplication of closed geodesics, which involves (the often technically tricky) concatenation of paths, has a unique definition that's associative on the nose. (Parametrise by arclength, concatenate, then parametrise to unit length; since the paths are geodesics, the last step is also unique.)*
>
> *Unfortunately, this gives no way to define multiplication of closed geodesics that are (positive) powers of different primes. We could generalise to piecewise geodesics that may turn corners at the basepoint, but this seems somewhat artificial, and it doesn't have very nice properties.*
>
> *– Toby*

Darin Brown wrote, in response to some questions by Squark on `sci.physics.research`:

> > *Squark wrote:*
> >
> > > *John Baez wrote:*
> > > *If G' is a subgroup of G, we get a covering map*
> > >
> > > $$H/G' \to H/G$$
> > >
> > > *so we can ask about lifting prime geodesics in H/G to closed geodesics in H/G'. There can be a bunch of ways to do this, so we say a prime geodesic in H/G "splits" into powers of prime geodesics up in H/G'.*
> >
> > *I don't quite understand how can the lift be a power, rather than just a prime.*
>
> *Quite true. When you lift a geodesic, once you get back to the starting basepoint, you've gone around once up above, corr. to a prime above, so it doesn't make sense to go around more than once! (I think this is what the author of this comment meant.) In fact, I think it's true (I can ask Jeff) that in a sense, there are no "ramified primes" in the geodesic context. (There are only finitely many in the number theory context. Actually, ramified primes are bad behaviour in a sense.) It is true, when you lift a prime, the geodesic above has length a multiple of the prime below, this is the analogue of the* inertial degree*, not the ramification degree. It seems all the ramification degrees are 1, and the magic equation reduces to degree of extension = sum(inertia degrees).*

Norm $N(Q)$ of ideal Q	*Norm $N(f)$ of closed geodesic f*

>> *What is a norm of a geodesic? The length or the energy or… ?*
>
> *Explicitly, the length of a geodesic is the (natural) log of the norm, or equivalently, the norm is $\exp($length$)$. For closed geodesics on $\Gamma\backslash H$, you find the norm explicitly as follows: consider the associated hyp. conj. class $\{\gamma\}$, take an eigenvalue ε of an element of this conj. class, then the norm is ε^2. The length of the geodesic is then $2\log(\varepsilon)$. This is independent of the choice of γ in the conj. class.*
>
> *This is why I now like to think of the norm of an ideal as a kind of "length function on ideals".*

Frobenius conjugacy class of Q	*Frobenius conjugacy class of f*

>> *Again, what is the Frobenius on the right side here?*
>
> *I can give 2 answers. The first answer is a cop-out, because it would just give the concrete definition given in Jeff's paper or my thesis, e.g. Namely, you take the associated matrix γ, and reduce entries mod the prime Q, where Q determines the covering surface $\Gamma(Q)\backslash H$. This is a very concrete definition that doesn't hint at the connection to number theory. Remember, secretly, $\mathrm{PSL}(2,q)$, $q = \mathrm{norm}(Q)$ is really (isomorphic to) the deck transformation group of $\Gamma(Q)\backslash H$ over $\Gamma(H)$, and the Frob conj. class of a geodesic f should be a conj. class in this deck transformation group. Conceptually, it should be an element of the decomposition group, those deck transformations that fix the prime geodesic above. Choosing different primes above the prime below should give elements of the deck transformation group which are conjugate to each other. At least, that should be the idea.*
>
> *darin*

Darin's description of the Frobenius associated to a prime geodesic in H/G is a bit technical. Here's my guess as to a simpler description:

We have a covering space of a Riemannian manifold. A geodesic down below gives an element of the fundamental group of the base. This acts as deck transformations of the cover. So, it acts on the set of prime geodesics in the cover! Indeed, it acts on the set of prime geodesics which are lifts of the geodesic down below. This is the "Frobenius automorphism" associated to the geodesic.

It's just a guess, but I feel sure it's right, or at least close. It's just like the Frobenius automorphisms in number theory — at least if we realize that a Galois group is secretly a fundamental group, as explained in "Week 213".

Wherever there is number, there is beauty.

— *Proclus*

Week 216

May 23, 2005

There are lots of different things called "zeta functions" in mathematics and physics. The grand-daddy of them all is the Riemann zeta function:

$$\zeta(s) = \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \dots$$

This is deeply related to prime numbers, thanks to Euler's product formula

$$\zeta(s) = \prod \frac{1}{1 - p^{-s}}$$

where we take a product over all primes. This formula is fun to prove: just use the geometric series to expand each factor, multiply them out and see what happens!

Using this, Riemann and von Mangoldt derived an explicit formula for how many primes are less than a given number as a sum over the *zeros* of the Riemann zeta function. Instead of showing you this formula, I'll just urge you to watch a *movie* of how it works:

1) Matthew Watkins, Animation: the prime counting function $\pi(x)$, `http://www.maths.ex.ac.uk/~mwatkins/zeta/pianim.htm`

Thanks to this formula, information about the Riemann zeta function is secretly information about the distribution of primes!

For example, the Riemann Hypothesis says that when we analytically continue the zeta function to the complex plane, the only zeros occur at negative even integers and numbers with real part equal to $1/2$. And, knowing this would be equivalent to knowing that the number of primes less than x differs from

$$\mathrm{Li}(x) = \int_2^x \frac{dt}{\ln t}$$

by less than some constant times $\ln(x)\sqrt{x}$. Everyone feels sure these facts are true. But, despite over a century of hard work and a million-dollar prize offered by the Clay Mathematics Institute, nobody has come close to proving them!

It's known that apart from the negative even integers, the only place the Riemann ζ function can vanish is in the strip where

$$0 < \Re(s) < 1$$

But, nobody has been able to show that all the zeros in this "critical strip" lie on the line

$$\Re(s) = \frac{1}{2}$$

Of course, this can be checked in special cases. The current record may belong to Xavier Gourdon, who on October 12th of 2004 claimed to have shown — with the help of a computer — that the first *ten trillion* zeros in the critical strip lie on the line Re(s) = 1/2.

2) Xavier Gourdon, Computation of zeros of the Riemann zeta function, `http://numbers.computation.free.fr/Constants/Miscellaneous/zetazeroscompute.html`

Alas, such monster calculations don't seem helpful for proving the Riemann hypothesis. They're more useful when it comes to formulating and testing conjectures about the *statistical properties* of the zeros.

The most famous of these traces its way back to a teatime conversation between Hugh Montgomery and Freeman Dyson. . . you can read the story here:

3) K. Sabbagh, *Dr. Riemann's Zeros*, Atlantic Books, 2002, pp. 134=-136. Story about Hugh Montgomery and Freeman Dyson also available at `http://www.maths.ex.ac.uk/~mwatkins/zeta/dyson.htm`

The upshot is that the distribution of spacings between Riemann zeros closely mimics the spacings between eigenvalues of a large randomly chosen self-adjoint matrix.

This suggests fascinating relations between the Riemann zeta function and quantum physics. In fact, one popular dream for proving the Riemann zeta function is to find a chaotic classical system whose quantized version has energy levels related to the Riemann zeta zeros!

I would like to understand this stuff, but it all seems a bit intimidating — especially since the coolest aspects are the ones *nobody* understands.

Luckily, the Riemann zeta function has spawned a lot of other functions called zeta functions and $L$-functions, and many of these are *simpler* than the original one — or at least raise fascinating questions that are easier to solve. Many of these are listed here:

4) Matthew R. Watkins, A directory of all known zeta functions, `http://www.maths.ex.ac.uk/~mwatkins/zeta/directoryofzetafunctions.htm`

   Matthew R. Watkins, A directory of all known $L$-functions, `http://www.maths.ex.ac.uk/~mwatkins/zeta/directoryof$L$-functions.htm`

Lately I've been talking about zeta functions with James Dolan and also Darin Brown, whose work I mentioned last week. I feel some things are starting to make sense, so I'd like to explain them before it turns out they don't.

I'll just list some zeta functions, so you can see what we're dealing with:

A) The zeta function of a number field. A "number field" is something you get by taking the rational numbers and throwing in some algebraic numbers. One can define "algebraic integers" for any number field, and they act a lot like the ordinary integers. So, one can define a zeta function for any number field.

Technically, we do this by summing over all nonzero ideals I in the ring A of algebraic integers in our number field:

$$\zeta(s) = \sum |I|^{-s}$$

where $|I|$ is a number called the "norm" of I, which is just the cardinality of A/I. We also have an Euler product formula:

$$\zeta(s) = \prod \frac{1}{1 - |P|^{-s}}$$

where we take the product over prime ideals P.

For example, if our number field is the rational numbers, its algebraic integers are the ordinary integers. So, each ideal consists of multiples of some number $n = 1, 2, 3, \ldots$, and its norm is just n, so we get:

$$\zeta(s) = \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \ldots$$

A more fun example is the number field $\mathbb{Q}[i]$, where we take the rational numbers and throw in a square root of -1. Here the algebraic integers are the so-called "Gaussian integers" $\mathbb{Z}[i]$, namely guys like

$$a + bi$$

where a and b are ordinary integers.

In this example it's easiest to work out the zeta function using the Euler product formula. If you ask one of your number theory pals about prime ideals in the Gaussian integers, they'll say:

> *”Well, the Gaussian integers are a principal ideal domain, so every ideal is generated by a single element. So, we can actually talk about* prime numbers *in the Gaussian integers. And there are 3 cases:*
>
> - ***INERT**: An ordinary prime number of the form* $4n + 3$ *is also prime in the Gaussian integers: for example, 3.*
> - ***SPLIT**: An ordinary prime numbers of the form* $4n + 1$ *is the product of two complex conjugate primes in the Gaussian integers: for example,* $5 = (2 + i)(2 - i)$.
> - ***RAMIFIED**: The ordinary prime 2 equals* $(1 + i)(1 - i)$*, but here the two factors give the same prime ideal, since* $(1-i) = i(1+i)$*, and* i *is invertible in the Gaussian integers.*
>
> *We get all primes in the Gaussian integers this way.”*

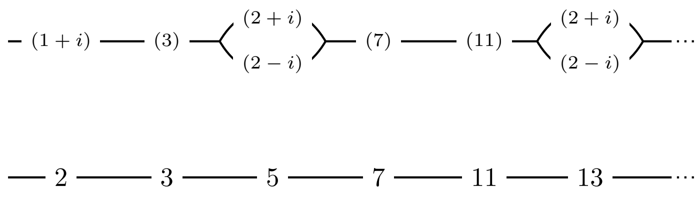
So, the zeta function of the Gaussian integers goes like this:

$$\zeta(s) = \frac{1}{1 - 2^{-s}} \frac{1}{1 - 3^{-2s}} \frac{1}{1 - 5^{-s}} \frac{1}{1 - 5^{-s}} \cdots$$

I went just far enough to show you what happens for each kind of prime. As you might expect, we get two factors for each prime that splits in two.

I should explain the other details, but number theory is best absorbed in small doses, especially if you're a physicist. The main lesson to take home is this:

A number field is like a funny sort of "branched covering space" of the set of ordinary primes. Sitting over each ordinary prime there are one or more prime ideals in our number field:



And, the ζ function records the details of how this works!

For more on this covering space philosophy see "Week 205" and "Week 213". This geometrical metaphor lies behind a lot of the really cool work on number theory.

B) The zeta function of a discrete dynamical system. A "discrete dynamical system" consists of a set X together with a one-to-one and onto function

$$f\colon X \to X$$

Here X is the set of "states" of some physical system, and f describes one step of time evolution. For each integer n we get a function

$$f^n\colon X \to X$$

Since the integers are called $\mathbb{Z}$, mathematicians would call our discrete dynamical system a "$\mathbb{Z}$-set".

Whatever you call it, its zeta function is defined to be:

$$\zeta(s) = \prod \frac{1}{1 - |P|^{-s}}$$

where P ranges over all periodic orbits and $|P|$ is the *exponential* of the size of this periodic orbit.

This is like an Euler product formula with the periodic orbits being the "primes". Just as every natural number can be uniquely factored into primes, every discrete dynamical system can be uniquely written as a disjoint union of periodic orbits. This explains the exponential in the definition of $|P|$ above: primes like to multiply, while sizes of orbits like to add.

One nice thing about this zeta function is that when we take the disjoint union of two discrete dynamical systems, their zeta functions multiply. Another nice thing is that the zeta function of a dynamical system completely describes it up to isomorphism, at least when the set X is finite. Decategorification at work!

We can also rewrite this zeta function as a sum:

$$\zeta(s) = \sum |I|^{-s}$$

where I ranges over all formal products of periodic orbits, and we define $|P_1 \ldots P_n| = |P_1| \ldots |P_n|$.

Even better, examples A) and B) overlap. I'll explain how later, but the key is to associate to any number field and any prime p a discrete dynamical system

$$f\colon X \to X$$

called the "Frobenius automorphism". This gives a zeta function. It works best if we take the exponential of the size of each periodic orbit using the base p instead of base e. Then, if we multiply all these zeta functions, one for each ordinary prime p, we get the zeta function of our number field!

C) The zeta function of a continuous dynamical system. Now suppose X is some topological space and we have a time evolution map

$$f_t\colon X \to X$$

for each real number t. We can define a zeta function

$$\zeta(s) = \prod \frac{1}{1 - |P|^{-s}}$$

where P ranges over all periodic orbits and $|P|$ is the exponential of the "period" of P — meaning the time it takes for points on this orbit to loop around back to where they started.

A famous example is when we have a Riemannian manifold. A free particle moving around on such a space will trace out geodesics, giving us a dynamical system. The analogue of primes are now "prime geodesics": periodic geodesics that loop around just once.

The "covering space" philosophy described in example A) can now be taken literally! If the Riemannian manifold M' is a covering space of M, any prime geodesic P in M defines a deck transformation of M'. This transformation acts on the set X of prime geodesics sitting over P, so we get a one-to-one and onto map

$$f\colon X \to X$$

This is exactly like the "Frobenius automorphism" in number theory!

All this is particularly interesting when our manifold is a quotient of the upper half-plane by a discrete group — see "Week 215" for more on this. The reason is that some of these quotients are related to number theory. So, we get some direct interactions with example A).

D) The zeta function of a graph. We can take the idea of "periodic geodesic on a Riemannian manifold" and vastly simplify it by looking at closed loops in a graph with finitely many edges and vertices. We get a ζ function

$$\zeta(s) = \prod \frac{1}{1 - |P|^{-s}}$$

where P ranges over all "prime loops" in our graph: loops that don't backtrack or loop around more than once. Now $|P|$ is the exponential of the length of the loop.

The "covering space" philosophy still applies, since we can define what it means for a graph G' to be a covering space of a graph G. Any prime loop P in G defines a deck transformation of G'. This acts on the set X of prime loops sitting over P, so we get a one-to-one and onto map

$$f\colon X \to X$$

which again deserves to be called the "Frobenius automorphism".

E) The zeta function of an affine scheme. Given a commutative ring, we can think of it as the ring of functions on some space. The ζ function of the commutative ring then counts the points of this space.

To make this precise, we cleverly invent a kind of space called an "affine scheme", which is secretly *just another name for a commutative ring!* So, any commutative ring R gives an affine scheme called $\mathrm{Spec}(R)$, purely by our fiendish definition.

If we take a function and evaluate it at a point, we should get a number. This should give a homomorphism from functions to numbers. But in algebraic geometry, "numbers" can be elements of any field k. So, let's say the "k-points" of $\mathrm{Spec}(R)$ are the homomorphisms from R to k.

(This is a bit nontraditional, but I really need this here. For a more traditional alternative, see "Week 205".)

In particular, for any prime p we can take k to be the algebraic closure of the field with p elements. Let X be the set of k-points of some affine scheme $\mathrm{Spec}(R)$. Then comes something wonderful: if x is a k-point, so is x^p, since "raising to the pth power" is an automorphism of k. So, we get a map

$$f\colon X \to X$$

sending x to x^p. This is called the "Frobenius automorphism"!

Since f is a discrete dynamical system, we can define its zeta function as in example B):

$$\zeta(s) = \prod \frac{1}{1 - |P|^{-s}}$$

where P ranges over all periodic orbits, and $|P|$ is the exponential of the size of the periodic orbit, defined using the base p.

So far, this is the zeta function of our affine scheme "localized at p". If we multiply a bunch of factors like this, one for each ordinary prime p, we get the zeta function of our affine scheme. For example, the affine scheme $\mathrm{Spec}(\mathbb{Z})$ gives the Riemann zeta function.

In fact, all of example A) fits neatly into this one as a sub-example. And if you know about schemes that aren't affine — like projective varieties, such as elliptic curves and other curves — you'll see this definition works for them too.

If you know someone else's definition of the zeta function of a scheme, it may not look like what I wrote here! But don't panic. The reason is that people like to express the zeta function of a discrete dynamical system $f\colon X \to X$ in terms of the number of fixed points of f^n. When f is the Frobenius automorphism, these are usually called "points defined over the field with p^n elements". So, you'll see lots of formulas for zeta functions phrased in terms of these. . . .

Okay. Enough examples.

There are a lot more, but I think these are the simplest. I hope you see how all these examples are just different expressions of the same idea. To go further, I would tell you how there are versions of the Riemann Hypothesis in all these examples, and also things called "L-functions", and lots of theorems and conjectures concerning them, too!

It's a wonderful example of the unity of mathematics. But, it's also a wonderful example of how this unity is obscured by people who zoom in on their own favorite special case and its particular technical complexities while never discussing the big picture. You wouldn't believe how hard it's been for me to figure out what I just told you! It's like trying to learn English by reading the US legal code, or learning basic chord progressions by listening to Schoenberg.

If you're just trying to get started, here's one of the more readable introductions:

5) Anton Deitman, "Panorama of zeta functions", available as `math.NT/0210060`.

Audrey Terras has a lot of nice slide presentations about the ζ functions and L-functions of graphs:

6) Audrey Terras, Artin L-functions of graph coverings, available at `http://math.ucsd.edu/~aterras/artin1.pdf`

"More on L-functions", available at `http://math.ucsd.edu/~aterras/artin2.pdf`

Here's a paper written in broken English but making a very serious attempt to explain things to the nonexpert:

7) David Zywina, "The zeta function of a graph", available at `http://math.berkeley.edu/~zywina/zeta.pdf`

He gives a characterization of graphs whose zeta functions satisfy an analogue of the Riemann Hypothesis. Strangely, this analogue involves *poles* of the ζ function in the critical strip

$$0 < \Re(s) < 1$$

Is this a real difference or the result of some difference in conventions?

Finally, I should explain some more technical stuff about ζ functions and fixed points, just so I don't forget it. Suppose we have a discrete dynamical system

$$f\colon X \to X$$

and let

$$|\mathrm{fix}(f^n)|$$

be the number of fixed points of the nth iterate of f.

We can define a weird function like this:

$$Z(u) = \exp\left(\sum_{n>0} \frac{|\mathrm{fix}(f^n)| u^n}{n}\right)$$

You'll often see formulas like this running around, especially when f is some sort of "Frobenius automorphism". Sometimes people even call these guys zeta functions. But what in the world do they have to do with zeta functions???

Let's see. Suppose first that X consists of a single cycle of length k. Then f^n has k fixed points when n is a multiple of k, but none otherwise, so:

$$\begin{aligned} Z(u) &= \exp\left(\frac{ku^k}{k} + \frac{ku^{2k}}{2k} + \frac{ku^{3k}}{3k} + \dots\right) \\ &= \exp\left(u^k + \frac{u^{2k}}{2} + \frac{u^{3k}}{3} + \dots\right) \\ &= \exp\left(\ln \frac{1}{1-u^k}\right) \\ &= \frac{1}{1-u^k} \end{aligned}$$

This is starting to look more like the zeta functions we know and love. It looks even better if we pick some prime number p and define

$$\zeta(s) = Z(p^{-s})$$

Then we get

$$\zeta(s) = \frac{1}{1 - p^{-ks}}$$

which is precisely what we'd get using the definition in example B).

Furthermore, for both that old definition and our new one, the zeta function of a disjoint union of discrete dynamical systems is the *product* of the zeta functions of the parts. Since every discrete dynamical system is a disjoint union of cycles, we conclude that the definitions *always* agree. In other words,

$$\zeta(s) = Z(p^{-s})$$

with

$$Z(u) = \exp\left(\sum_{n>0} \frac{|\mathrm{fix}(f^n)| u^n}{n}\right)$$

is always equal to the zeta function defined in example B).

So, don't let anyone fool you — there aren't lots of completely different kinds of zeta functions! There's just a few kinds, and we could probably boil them down to just ONE kind with some work.

> Some decades ago I made — somewhat in jest — the suggestion that one should get accepted a non-proliferation treaty of zeta functions. There was becoming such an overwhelming variety of these objects.
>
> — *Atle Selberg*

Week 217

May 30, 2005

Last week I described lots of different zeta functions, but didn't say much about what they're good for. This week I'd like to get started on fixing that problem.

People have made lots of big conjectures related to zeta functions. So far they've just proved just a few. . . but it's still a big deal.

For example, Andrew Wiles' proof of Fermat's Last Theorem was just a tiny spin-off of his work on something much bigger called the Taniyama-Shimura conjecture. Now, personally, I think Fermat's Last Theorem is a ridiculous thing. The last thing I'd ever want to know is whether this equation:

$$x^n + y^n = z^n$$

has nontrivial integer solutions for $n > 2$. But the Taniyama-Shimura Conjecture is really interesting! It's all about the connection between geometry, complex analysis and arithmetic, and it ties together some big ideas in an unexpected way. This is how it usually works in number theory: cute but goofy puzzles get solved as a side-effect of deep and interesting results related to zeta functions and L-functions — sort of like how the powdered drink "Tang" was invented as a spinoff of going to the moon.

For a good popular book on Fermat's Last Theorem and the Taniyama-Shimura Conjecture, try:

1) Simon Singh, *Fermat's Enigma: The Epic Quest to Solve the World's Greatest Mathematical Problem*, Walker, New York, 1997.

Despite the "world's greatest mathematical problem" baloney, this book does a great job of telling the story without drowning the reader in math.

But you read This Week's Finds because you *want* to be drowned in math, and I wouldn't want to disappoint you. So, let me list a few of the big conjectures and theorems related to zeta functions.

Here goes:

> ***A) The Riemann Hypothesis — the zeros of the Riemann zeta function in the critical strip*** $0 \leqslant \Re(s) \leqslant 1$ ***actually lie on the line*** $\Re(s) = 1/2$.

First stated in 1859 by Bernhard Riemann; still open.

This implies a good estimate on the number of primes less than a given number, as described in "Week 216".

> ***B) The Generalized Riemann Hypothesis — the zeros of any Dirichlet*** L***-function that lie in the critical strip actually lie on the line*** $\Re(s) = 1/2$.

Still open, since the Riemann Hypothesis is a special case.

A "Dirichlet L-function" is a function like this:

$$L(\chi, s) = \sum_{n>0} \frac{\chi(n)}{n^s}$$

where χ is any "Dirichlet character", meaning a periodic complex function on the positive integers such that

$$\chi(nm) = \chi(n)\chi(m)$$

If we take $\chi = 1$ we get back the Riemann zeta function.

Dirichlet used these L-functions to prove that there are infinitely many primes equal to $k \mod n$ as long as k is relatively prime to n. The Generalized Riemann Hypothesis would give a good estimate on the number of such primes less than a given number, just as the Riemann Hypothesis does for plain old primes.

Erich Hecke established the basic properties of Dirichlet L-functions in 1936, including a special symmetry called the "functional equation" which Riemann had already shown for his zeta function. So I bet Hecke must have dreamt of the Generalized Riemann Hypothesis, even if he didn't dare state it.

> ***C) The Extended Riemann Hypothesis — for any number field, the zeros of its zeta function in the critical strip actually lie on the line*** $\Re(s) = 1/2$***.***

Still open, since the Riemann Hypothesis is a special case.

I described the zeta functions of number fields in "Week 216". These are usually called "Dedekind zeta functions". Hecke also proved a functional equation for these back in 1936.

> ***D) The Grand Riemann Hypothesis — for any automorphic*** L***-function, the zeros in the critical strip actually lie on the line*** $\Re(s) = 1/2$***.***

This is still open too, since it includes A)–C) as special cases!

I don't want to tell you what "automorphic L-functions" are yet. For now, you can just think of them as grand generalizations of both Dirichlet L-functions and zeta functions of number fields.

> ***E) The Weil Conjectures — The zeros and poles of the zeta function of any smooth algebraic variety over a finite field lie on the lines*** $\Re(s) = 1/2, 1, 3/2, \ldots, d/2$ ***where*** d ***is the dimension of the variety.*** *The zeros lie on the lines* $1/2, 3/2, \ldots$ *while the poles lie on the lines* $1, 2, \ldots$*. Also: such zeta functions are quotients of polynomials, they satisfy a functional equation, and a lot of information about their zeroes and poles can be recovered from the topology of the corresponding* complex *algebraic varieties.*

First stated in 1949 by Andre Weil; proof completed by Pierre Deligne in 1974 based on much work by Michael Artin, J.-L. Verdier, and especially Alexander Grothendieck. Grothendieck invented topos theory as part of the attack on this problem!

> ***F) The Taniyama-Shimura Conjecture — every elliptic curve over the rational numbers is a modular curve.*** *Or, equivalently: every* L*-function of an elliptic curve is the* L*-function of a modular curve.*

This was first conjectured in 1955 by Yukata Taniyama, who worked on it with Goro Shimura until committing suicide in 1958. Around 1982 Gerhard Frey suggested that this conjecture would imply Fermat's Last Theorem; this was proved in 1986 by Ken

Ribet. In 1995 Andrew Wiles and Richard Taylor proved a big enough special case of the Taniyama-Shimura Conjecture to get Fermat's Last Theorem. The full conjecture was shown in 1999 by Breuil, Conrad, Diamond, and Taylor.

I don't want to say what L-functions of curves are yet... but they are a lot like Dirichlet L-functions.

> ***G) The Langlands Conjectures — any automorphic representation*** π ***of a connected reductive group*** G***, together with a finite-dimensional representation of its*** L***-group, gives an automorphic*** L***-function*** $L(s, \pi)$***.*** *Also: these* L*-functions all satisfy functional equations. Furthermore, they depend functorially on the group* G*, its* L*-group, and their representations.*

Zounds! Don't worry if this sounds like complete gobbledygook! I only mention it to show how scary math can get. As Stephen Gelbart once wrote:

> *The conjectures of Langlands just alluded to amount (roughly) to the assertion that the other zeta-functions arising in number theory are but special realizations of these* $L(s, \pi)$*.*
>
> *Herein lies the agony as well as the ecstacy of Langlands' program. To merely state the conjectures correctly requires much of the machinery of class field theory, the structure theory of algebraic groups, the representation theory of real and* p*-adic groups, and (at least) the language of algebraic geometry. In other words, though the promised rewards are great, the initiation process is forbidding.*

I hope someday I'll understand this stuff well enough to say something more helpful! Lately I've been catching little glimpses of what it's about....

But, right now I think it's best if I talk about the "functional equation" satisfied by the Riemann zeta function, since this gives the quickest way to see some of the strange things that are going on.

The Riemann zeta function starts out life as a sum:

$$\zeta(s) = 1^{-s} + 2^{-s} + 3^{-s} + 4^{-s} + \dots$$

This only converges for $\Re(s) > 1$. It blows up as we approach $s = 1$, since then we get the series

$$\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots$$

which diverges. However, in 1859 Riemann showed that we can analytically continue the zeta function to the whole complex plane except for this pole at $s = 1$.

He also showed that the zeta function has an unexpected symmetry: its value at any complex number s is closely related to its value at $1-s$. It's not true that $\zeta(s) = \zeta(1-s)$, but something similar is true, where we multiply the zeta function by an extra fudge factor.

To be precise: if we form the function

$$\pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\zeta(s)$$

then this function is unchanged by the transformation

$$s \mapsto 1 - s$$

This symmetry maps the line

$$\Re(s) = \frac{1}{2}$$

to itself, and the Riemann Hypothesis says all the ζ zeros in the critical strip actually lie on this magic line.

This symmetry is called the "functional equation". It's the tiny tip of a peninsula of a vast and mysterious continent which mathematicians are still struggling to explore. Riemann gave two proofs of this equation. You can find a precise statement and a version of Riemann's second proof here:

2) Daniel Bump, Zeta Function, lecture notes on "the functional equation" available at `http://math.stanford.edu/~bump/zeta.html` and `http://www.maths.ex.ac.uk/~mwatkins/zeta/fnleqn.htm`

This proof is a beautiful application of Fourier analysis. Everyone should learn it, but let me try to sketch the essential idea here.

I will deliberately be VERY rough, and use some simplified nonstandard definitions, since the precise details have a way of distracting your eye just as the magician pulls the rabbit out of the hat.

We start with the function $\zeta(2s)$:

$$1^{-s} + 4^{-s} + 9^{-s} + 25^{-s} + \ldots.$$

Then we apply a curious thing called the "inverse Mellin transform", which turns this function into

$$z^1 + z^4 + z^9 + z^{25} + \ldots$$

Weird, huh? This is almost the "theta function"

$$\theta(t) = \sum_{n \in \mathbb{Z}} e^{\pi i n^2 t}$$

Indeed, it's easy to see that

$$\frac{\theta(t) - 1}{2} = z^1 + z^4 + z^9 + z^2 5^+ \ldots$$

when

$$z = e^{\pi i t}$$

The theta function transforms in a very simple way when we replace t by $-1/t$, as one can show using Fourier analysis.

Unravelling the consequences, this implies that the zeta function transforms in a simple way when we replace s by $1 - s$. You have to go through the calculation to see precisely how this works. . . but the basic idea is: a symmetry in the theta function yields a symmetry in the zeta function.

Hmm, I'm not sure that explained anything! But I hope at least the mystery is more evident now. A bunch of weird tricks, and then *presto* — the functional equation! To make progress on understanding the Riemann Hypothesis and its descendants, we need to get what's going on here.

I feel I *do* get the inverse Mellin transform; I'll say more about that later. But for now, note that the theta function transforms in a simple way, not just when we do this:

$$t \mapsto -\frac{1}{t}$$

but also when we do this:

$$t \mapsto t+2$$

Indeed, it doesn't change at all when we add 2 to t, since $e^{2\pi i} = 1$.

Now, the maps

$$t \mapsto -\frac{1}{t}$$

and

$$t \mapsto t+1$$

generate the group of all maps

$$t \mapsto \frac{at+b}{ct+d}$$

where a, b, c, d form a 2×2 matrix of integers with determinant 1. These maps form a group called $\mathrm{PSL}(2, \mathbb{Z})$, or the "modular group".

A function that transforms simply under this group and doesn't blow up in nasty ways is called a "modular form". In "Week 197" I gave the precise definition of what counts as transforming simply and not blowing up in nasty ways. I also explained how modular forms are related to elliptic curves and string theory. So, please either reread "Week 197" or take my word for it: modular forms are cool!

The theta function is almost a modular form, but not quite. It doesn't blow up in nasty ways. However, it only transforms simply under a subgroup of $\mathrm{PSL}(2, \mathbb{Z})$, namely that generated by

$$t \mapsto -\frac{1}{t}$$

and

$$t \mapsto t+2$$

So, the theta function isn't a full-fledged modular form. But since it comes close, we call it an "automorphic form".

Indeed, for any discrete subgroup G of $\mathrm{PSL}(2, \mathbb{Z})$, functions that transform nicely under G and don't blow up in nasty ways are called "automorphic forms" for G. They act a lot like modular forms, and people know vast amounts about them. It's the power of automorphic forms that makes number theory what it is today!

We can summarize everything so far in this slogan:

> *THE FUNCTIONAL EQUATION FOR THE RIEMANN ZETA FUNCTION SAYS:*
> *"THE THETA FUNCTION IS AN AUTOMORPHIC FORM"*

But before you start printing out bumper stickers, I should explain....

The point of this slogan is this. We *thought* we were interested in the Riemann zeta function for its own sake, or what it could tell us about prime numbers. But with the wisdom of hindsight, the first thing we should do is hit this function with the inverse Mellin transform and repackage all its information into an automorphic form — the theta function.

ζ is dead, long live θ!

The Riemann zeta function is just like all the fancier zeta functions and L-functions in this respect. The fact that they satisfy a "functional equation" is just another way of saying their inverse Mellin transforms are automorphic forms... and it's these automorphic forms that exhibit the deeper aspects of what's going on.

Now let me say a little bit about the inverse Mellin transform.

Ignoring various fudge factors, the inverse Mellin transform is basically just the linear map that sends any function of s like this:

$$n^{-s}$$

to this function of z:

$$z^n$$

In other words, it basically just turns things upside down, replacing the base by the exponent and vice versa. The minus sign is just a matter of convention; don't worry about that too much.

So, the inverse Mellin transform basically sends any function like this, called a "Dirichlet series":

$$a_1 1^{-s} + a_2 2^{-s} + a_3 3^{-s} + a_4 4^{-s} + \dots$$

to this function, called a "Taylor series":

$$a_1 z^1 + a_2 z^2 + a_3 z^3 + a_4 z^4 + \dots$$

Now, why would we want to do this?

The reason is that multiplying Taylor series is closely related to *addition* of natural numbers:

$$z^n z^m = z^{n+m}$$

while multiplying Dirichlet series is closely related to *multiplication* of natural numbers:

$$n^{-s} m^{-s} = (nm)^{-s}$$

The Mellin transform (and its inverse) are how we switch between these two pleasant setups!

Indeed, it's all about algebra — at least at first. We can add natural numbers and multiply them, so $\mathbb{N}$ becomes a monoid in two ways. A "monoid", recall, is a set with a binary associative product and unit. So, we have two closely related monoids:

$$(\mathbb{N}, +, 0)$$

and

$$(\mathbb{N}, \times, 1)$$

Given a monoid, we can form something called its "monoid algebra" by taking formal complex linear combinations of monoid elements. We multiply these in the obvious way, using the product in our monoid.

If we take the monoid algebra of $(\mathbb{N}, +, 0)$, we get the algebra of Taylor series! If we take the monoid algebra of $(\mathbb{N}, \times, 1)$, we get the algebra of Dirichlet series!

(Actually, this is only true if we allow ourselves to use *infinite* linear combinations of monoid elements in our monoid algebra. So, let's do that. If we used finite linear combinations, as people often do, $(\mathbb{N}, +, 0)$ would give us the algebra of polynomials, while $(\mathbb{N}, \times, 0)$ would give us the algebra of "Dirichlet polynomials".)

Of course, algebraically we can combine these structures. $(\mathbb{N}, +, \times, 0, 1)$ is a rig, and by taking formal complex linear combinations of natural numbers we get a "rig algebra" with two products: the usual product of Taylor series, and the usual product of Dirichlet series. They're compatible, too, since one distributes over the other. They both distribute over addition.

However, if we're trying to get an algebra of functions on the complex plane, with pointwise multiplication as the product, we need to make up our mind: either Taylor series or Dirichlet series! We then need the Mellin transform to translate between the two.

So, what seems to be going on is that people take a puzzle, like

> *"what is the sum of the squares of the divisors of n?"*

or

> *"how many ideals of order n are there in this number field?"*

and they call the answer a_n.

Then they encode this sequence as either a Dirichlet series:

$$a_1 1^{-s} + a_2 2^{-s} + a_3 3^{-s} + a_4 4^{-s} + \dots$$

or a Taylor series:

$$a_1 z^1 + a_2 z^2 + a_3 z^3 + a_4 z^4 + \dots$$

The first format is nice because it gets along well with multiplication of natural numbers. For example, in the puzzle about ideals, every ideal is a product of prime ideals, and its norm is the product of the norms of those prime ideals... so our Dirichlet series will have an Euler product formula.

The second format is nice *if* our Taylor series is an automorphic form. This will happen precisely when our Dirichlet series satisfies a functional equation.

(For experts: I'm ignoring some fudge factors involving the gamma function.)

I still need to say more about *which* puzzles give automorphic forms, what it really means when they *do*. But, not this week! I'm tired, and I bet you are too.

For now, let me just give some references. There's a vast amount of material on all these subjects, and I've already referred to lots of it. But right now I want to focus on stuff that's free online, especially stuff that's readable by anyone with a solid math background — not journal articles for experts, but not fluff, either.

Here's some information on the Riemann Hypothesis provided by the Clay Mathematics Institute, which is offering a million dollars for its solution:

3) Clay Mathematics Institute, "Problems of the Millennium: the Riemann Hypothesis", `http://www.claymath.org/millennium/`

The official problem description by Enrico Bombieri talks about evidence for the Riemann Hypothesis, including the Weil Conjectures. The article by Peter Sarnak describes generalizations leading up to the Grand Riemann Hypothesis. In particular, he gives a super-rapid introduction to automorphic L-functions.

Here's a nice webpage that sketches Wiles and Taylor's proof of Fermat's last theorem:

4) Charles Daney, "The Mathematics of Fermat's Last Theorem", `http://www.mbay.net/~cgd/flt/fltmain.htm`

I like the quick introductions to "Elliptic curves and elliptic functions", "Elliptic curves and modular functions", "Zeta and L-functions", and "Galois Representations" — they're neither too detailed nor too vague, at least for me.

Here's a nice little intro to the Weil Conjectures:

5) Runar Ile, "Introduction to the Weil Conjectures", `http://folk.uio.no/~ile/WeilA4.pdf`

James Milne goes a lot deeper — his course notes on etale cohomology include a proof of the Weil Conjectures:

6) James Milne, "Lectures on Etale Cohomology", `http://www.jmilne.org/math/CourseNotes/math732.html`

while his course notes on elliptic curves sketch the proof of Fermat's Last Theorem:

7) James Milne, "Elliptic Curves", `http://www.jmilne.org/math/CourseNotes/math679.html`

Here's a nice history of what I've been calling the Taniyama-Shimura Conjecture, which explains why some people call it the Taniyama-Shimura-Weil conjecture, or other things:

8) Serge Lang, "Some history of the Shimura-Taniyama Conjecture", *AMS Notices* **42** (November 1995), 1301–1307. Available at `http://www.ams.org/notices/199511/forum.pdf`

Here's a quick introduction to the proof of this conjecture, whatever it's called:

9) Henri Darmon, "A proof of the full Shimura-Taniyama-Weil Conjecture is announced", *AMS Notices* **46** (December 1999), 1397–1401. Available at `http://www.ams.org/notices/199911/comm-darmon.pdf`

I won't give any references to the Langlands Conjectures, since I hope to talk a lot more about those some other time.

And, I hope to keep on understanding this stuff better and better!

I thank James Borger and Kevin Buzzard for help with this issue of This Week's Finds.

Addendum: Here's part of an email exchange I had with Kevin Buzzard of Imperial College after he read this Week's Finds:

I wrote:

> *What I* REALLY *want to know is: what combinatorial properties of an integer sequence an are we being told when we're told that the Dirichlet series*
>
> $$\sum_n a^n n^{-s}$$
>
> *comes from an automorphic form?*

He replied:

> *Yeah, that's a really key question. I'm not sure that there is an elementary answer. Here is another question: given a sequence of complex numbers $a_1, a_2, \ldots, a_n, \ldots$, with*
>
> $$a_n = \mathcal{O}(n^r),$$
>
> *what is a neat easy-to-understand property of this sequence which implies (or is implied by, or is equivalent to) the statement that*
>
> $$\sum \frac{a_n}{n^s}$$
>
> *has analytic (or meromorphic) continuation beyond $\Re(s) > r$? Maybe even this is hard—or maybe there is no such elementary criterion.*
>
> > *I'll be happy to assume for starters that an is multiplicative.*
>
> *This might not "logically speaking" be necessary, but on the other hand probably the most interesting cases have this property. Here is an example. Take a sequence of complex numbers $a_1, a_2, \ldots$ which is periodic with prime period p (primality probably isn't necessary but it simplifies the combinatorics). Then the associated L-function*
>
> $$\sum \frac{a_n}{n^s}$$
>
> *has meromorphic continuation with at worst a simple pole at $s = 1$ and no other poles, and one could even argue that the a_i "came from an automorphic form".*
>
> *On the other hand, this is not the kind of automorphic form that people usually think about because it's not an eigenform. What is happening is that "there are enough Dirichlet characters": consider the trivial character $\chi(n) = 1$ for all n, and then the $p - 1$ Dirichlet characters of level p, those defined by group homomorphisms*
>
> $$\chi\colon (\mathbb{Z}/p\mathbb{Z})^* \to \mathbb{C}^*$$
>
> *and extended to functions on $\mathbb{Z}$ by $\chi(n) = 0$ for n a multiple of p. These p functions on $\mathbb{Z}$ form a basis of the vector space of periodic functions on $\mathbb{Z}$ with period p.*

The Dirichlet characters give automorphic forms, but automorphic forms are a vector space so you can add them together and get an automorphic form for any periodic function. However the Dirichlet characters are the most interesting such forms because these are the ones which are eigenforms. The eigenforms give multiplicative a_n, but it's certainly not true in general that a periodic function is multiplicative.

But I can't really enlighten you much further. I know that the L-function of an automorphic form has meromorphic continuation and that we understand the poles (but we only conjecturally understand the zeroes). However if someone put some a_n in front of me I would demand that they told me where the a_n had come from before I put my money on whether there was an automorphic form.

The example where I actually proved something was in the case where the a_n came from a finite-dimensional complex representation of $\mathrm{Gal}(K/\mathbb{Q})$ with K a number field, Galois over $\mathbb{Q}$. (In fact my only contribution was in the 2-dimensional case, the 1-dimensional case having been understood for some time.) The a_n are then related to traces of the representation. Artin conjectured in the 1930s that

$$\sum a_n/n^s$$

should have analytic continuation to all of the complex plane if the representation was irreducible and not the trivial 1-dimensional representation. Langlands conjectured much later that the a_n should come from an automorphic form, and we knew by then that Langlands' conjecture implied Artin's. But none of the analytic guys know how to prove Artin's conjecture without essentially deducing it from Langlands'! I did something with some other Brits in the 2-dimensional case. Ironically, to deduce our analytic continuation results, we proved some p-adic analytic continuation results first :-) We constructed a modular form using p-adic techniques (and all of Wiles' machinery).

> *I get the feeling that nobody knows the answer, except perhaps for specific cases like modular forms, where we know they're all linear combinations of products of Eisenstein series, so that an is built out of sequences like $\sigma_k(n)$ — sums of kth powers of divisors.*

But unfortunately this is only true for "level 1" modular forms: you can build all modular forms of level 1 from the Eisenstein series E_4 and E_6. There is no neat analogous result for modular forms for the group $\Gamma_0(N)$ for general N. In particular you will never see the a_n for an elliptic curve built up from Eisenstein series in this way :-(

> *What I'd like is a really "conceptual" answer... or else for someone to admit that nobody knows such an answer yet.*

I think I will freely admit that, although that's just a personal opinion.

Here is why Dirichlet characters are the same as 1*-dimensional complex representations of* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$*. It's called the Kronecker-Weber theorem, it pre-dates class field theory (although it is now a special case), and you can* just about *prove it at the end of an introductory undergraduate course on number fields, as the last starred question on the example sheet, and only for people that have done the Galois theory course too. Let* K *be a number field and assume* K *is Galois over* $\mathbb{Q}$ *(equivalently, that there is a polynomial* f *with rational coefficients such that* K *is the smallest subfield of the complex numbers containing all the roots of* f*;* K *is called the "splitting field" of* f*).*

Then K *has a Galois group* $\mathrm{Gal}(K/\mathbb{Q})$*, which is the field automorphisms of* K *that fix* $\mathbb{Q}$*. Say* f *has degree* n *and* n *roots* $z_1, z_2, \ldots, z_n$*. Then any automorphism of* K *fixes* f*, so it permutes the roots of* f*. So we get an embedding*

$$\mathrm{Gal}(K/\mathbb{Q}) \to S_n$$

where S_n *is the symmetric group on the set* $z_1, z_2, \ldots, z_n$*. Then any automorphism of* K *fixes* f*, so it permutes the roots of* f*. So we get an embedding*

"Generically" this map is an isomorphism. But certainly not always — if there are subtle relations amongst the z_i *with rational coefficients then these subtle relations must also be preserved by the Galois group. One of my favourite examples is the equation* $x^4 - 2$*. This is an irreducible polynomial of degree 4, the four roots are* z*,* iz*,* $-z$ *and* $-iz$ *where* z *is the real 4th root of* 2 *and* i *is* $\sqrt{-1}$*. But* $z + (-z) = 0$*, so if* K *is the field generated by these roots and* σ *is an automorphism of* K *then* $\sigma(z) + \sigma(-z)$ *will also be zero, and the Galois group cannot possibly be all of* S_4*. In fact one can check that the Galois group is* D_8*, the dihedral group with 8 elements (the four roots form a square in the complex plane and it's the symmetries of this square).*

So there's Galois theory. Now here's a question: can we classify all the number fields K*, Galois over* $\mathbb{Q}$*, with* $\mathrm{Gal}(K/\mathbb{Q})$ *an abelian group?*

Here are some examples: $K = \mathbb{Q}(\sqrt{d})$*, the splitting field of* $x^2 - d$*. If* d *is the square of a rational then* $K = \mathbb{Q}$ *and if not then it's an extension of degree 2, with Galois group* S_2 *which is abelian.*

The splitting field of $x^n - 1$*, called the* n*th cyclotomic field, also turns out to have abelian Galois group; if* $z = \exp(2\pi i n)$ *then any automorphism of* K *must send* z *to another* n*th root of unity and furthermore the* n*th root of unity must have "exact order* n*", i.e. its* n*th power must be 1 but none of its* m*th powers can be 1 for* $0 < m < n$*. So* z *must get sent to* z^a *with* $0 < a < n$ *coprime to* n*. This gives us an injection*

$$\mathrm{Gal}(K/\mathbb{Q}) \to (\mathbb{Z}/n\mathbb{Z})^*,$$

where $(\mathbb{Z}/n\mathbb{Z})^*$ *is the units in the ring* $\mathbb{Z}/n\mathbb{Z}$ *(send* σ *to* a*), and it's tricky but true that this is in fact an isomorphism (use the fundamental theorem of Galois theory, and cyclotomic polynomials, or try and get a "trick" proof that uses as little of this as possible, but it's still some work). In any case* $\mathrm{Gal}(K/\mathbb{Q})$ *is certainly abelian.*

Next example: any subfield of any previous example, because this is how Galois theory works: if K contains L contains $\mathbb{Q}$, and both K and L are Galois over Q, then the obvious restriction map $\mathrm{Gal}(K/\mathbb{Q}) \to \mathrm{Gal}(L/\mathbb{Q})$ is a surjection.

So we now have quite a general example of a number field K with $\mathrm{Gal}(K/\mathbb{Q})$ abelian: any subfield of a cyclotomic field. The hard theorem (not really too hard, but quite messy) is that the converse is true: $\mathrm{Gal}(K/\mathbb{Q})$ abelian implies that K is contained within a cyclotomic field. For example $\mathbb{Q}(\sqrt{5})$ is in the 5th cyclotomic field because $\cos(72^\circ) = (\sqrt{5} - 1)/4$.

Any Dirichlet character gives a group homomorphism

$$(\mathbb{Z}/n\mathbb{Z})^* \to \mathbb{C}^*,$$

so a map

$$\mathrm{Gal}(K_n/\mathbb{Q}) \to \mathbb{C}^*,$$

with K_n the nth cyclotomic field, so a continuous group homomorphism

$$\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathbb{C}^*.$$

Conversely any continuous group homomorphism

$$\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathbb{C}^*$$

factors through a compact discrete quotient of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, which is just $\mathrm{Gal}(K/\mathbb{Q})$ for some number field K, and we get an injection

$$\mathrm{Gal}(K/\mathbb{Q}) \to \mathbb{C}^*,$$

so $\mathrm{Gal}(K/\mathbb{Q})$ is abelian, so K is contained in a K^n for some n, so we get a map

$$\mathrm{Gal}(K_n/\mathbb{Q}) \to \mathbb{C}^*$$

so it's a Dirichlet character.

Amazingly it might have been Langlands who really sold this "duality" point, 100 years after it was understood: people always used to state Kronecker–Weber as "any abelian number field is contained in a cyclo field" rather than the dual "any 1-dimensional rep of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ comes from a Dirichlet character". It was perhaps Langlands who realised that the correct generalisation of this statement was "any n-dimensional rep of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ comes from an automorphic form", rather than a statement about non-abelian extensions of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. Perhaps you will like the reason that people find the representation-theoretic approach appealing: I have been talking about number fields as subfields of the complexes, but really a number field is an abstract object which is a field and happens to have finite dimension over $\mathbb{Q}$, but it does not have a preferred embedding into the complexes. As a result of this sort of thinking, one realises that $\overline{\mathbb{Q}}$ is unique, but only up to highly non-unique isomorphism, and hence $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ is a "group only defined up to inner automorphism"! Hence it almost makes no sense to study this group—we cannot make any serious sense of an element of this group, because it's only the conjugacy classes that are well-defined. Hence we should *study the representations of the group on abstract vector spaces (i.e. ones without preferred bases), because these are well-defined up to isomorphism. The reason there was so much mileage in the abelian case was that this subtlety goes away: an abelian group up to inner automorphism is an abelian group.*

I know some facts about the sequence a_n coming from, say, an elliptic curve over the rationals, but the killer, and one that is really hard to "read off" from the a_n, is that the a_n are related to the traces of Frobenius elements on a 2-dimensional p-adic Galois representation (the Tate module of the curve). The moment I see a Galois representation I think that this must be something to do with automorphic forms, so that's why I believe that the L-function of an elliptic curve should come from a modular form. And it does! On the other hand, if you give me any finite set of primes p, and any integers a_p with p running through the set, such that $|a_p|^2 \leqslant 4p$, then I can concoct an elliptic curve with these a_p, so at the very least one has to look at infinitely many of the a_n before one can begin to guess whether the a_n come from an automorphic form.

Kevin

If I were to awaken after having slept for a thousand years, my first question would be: "Has the Riemann hypothesis been proven?"

— *David Hilbert*

Week 218

June 5, 2005

Classes are over! Summer is here! Now I can finally get some work done! I'll be travelling to Sydney, Canberra, Beijing, Chengdu and Calgary, but mainly I want to finish writing some papers.

First, though, I need to recover from a hard quarter. I need to goof off a bit! I spent most of yesterday lying in bed reading. Now I want to talk some more about number theory.

Let's see, where were we? I had just begun to introduce the theme of L-functions and their corresponding automorphic forms. My ultimate goal is to understand the Langlands Conjectures well enough to give a decent explanation of what they say. Instead of simply stating them, I'd like to really make them plausible, and this will take quite an elaborate warmup. So, this Week I want to talk about some background.

Actually, this reminds me of something Feynman wrote: whenever he worked on a problem, he needed the feeling he had some "inside track" — some insight or trick up his sleeve that nobody else had. Most of us will never be as good as Feynman at choosing an "inside track". But I think we all need one to convert what would otherwise be a dutiful and doomed struggle to catch up with the experts into something more hopeful and exciting: a personal quest!

For anyone with a physics background, a good "inside track" on almost any math problem is to convert it into some kind of crazy physics problem. It doesn't need to be realistic physics, just anything you can apply physics intuition to! This is part of why string theorists have been so successful in cracking math problems. It also underlies Alain Connes' attempt to prove the Riemann Hypothesis:

1) Alain Connes, "Noncommutative Geometry, Trace Formulas, and the Zeros of the Riemann Zeta Function". Ohio State course notes and videos at `http://www.math.ohio-state.edu/lectures/connes/Connes.html`

   Alain Connes, "Trace formula in noncommutative geometry and the zeros of the Riemann zeta function", available as `math.NT/9811068`.

2) Mathilde Marcolli, "Noncommutative geometry and number theory", available at `http://www.math.fsu.edu/~marcolli/ncgntE.pdf`

Of course, Connes also has another "inside track", namely his theory of noncommutative geometry.

By the way: a number theorist I know says he thinks Connes has essentially proved the Riemann Hypothesis, in the same way that Riemann "essentially" proved the Prime Number Theorem. Namely, he has reduced it to some facts that seem obviously true! Of course, it took about 40 years, from 1859 to 1896, for Riemann's plan to be fulfilled by Hadamard and De La Vallee Poussin. So, even if Connes' insights are correct, it may be a while before the Riemann Hypothesis is actually proved.

For anyone with a background in geometry, a good "inside track" on almost any math problem is to convert it into a geometry problem. In the case of number theory this trick is old news, but still very much worth knowing. It's based on an analogy which I began discussing in "Week 198".

The analogy starts out like this:

Number theory	Complex geometry
Integers, $\mathbb{Z}$	Polynomial functions on the complex plane, $\mathbb{C}[z]$
Rational numbers, $\mathbb{Q}$	Rational functions on the complex plane, $\mathbb{C}(z)$
Prime numbers, $\mathbb{P}$	Points in the complex plane, $\mathbb{C}$

Why is this analogy good? Well, for starters:

Every rational number is a ratio of integers.

Every rational function is a ratio of polynomials.

Better yet:

Every integer can be uniquely factored into primes
(modulo invertible integers, namely $+1$ and -1).

Every complex polynomial can be uniquely factored into linear polynomials
(modulo invertible polynomials, namely nonzero constants).

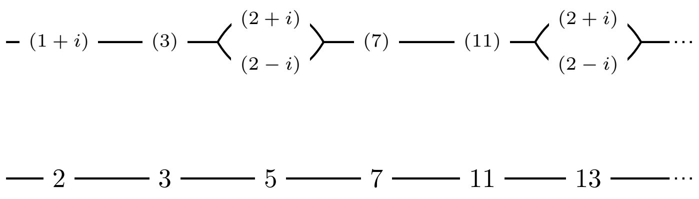
There's one linear polynomial $z - a$ for each point a in the complex plane, so *primes* are like *points* in the complex plane.

We can make this precise using the concept of "spectrum", which I defined in "Week 199". Ignoring a certain little sublety which is discussed there:

The spectrum of $\mathbb{Z}$ is the set of prime numbers.

The spectrum of $\mathbb{C}[z]$ is the complex plane.

This way of thinking lets us treat the spectrum of any algebraic extension of the integers, like the Gaussian integers, as a "covering space" of the set of prime numbers. I've already drawn this picture:



But, now I'm saying that the "line" down below really acts like the complex *plane*. Taking this strange idea seriously leads to all sorts of amazing insights.

For example, if you poke a hole in this "plane" at some prime, there's something like a little *loop* that goes around this hole! In other words, there's a sense in which the spectrum of $\mathbb{Z}$ has a nontrivial "fundamental group", which contains an element for each prime. Technically this group is called the Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, and we get an element in it for each prime, called the "Frobenius automorphism" for that prime.

Another cool thing is that we can study integers "locally", one prime at a time, just like we study complex functions locally. We can analyze functions at a point using Taylor series and Laurent series. And, we can stretch our analogy to include these concepts:

Number theory	Complex geometry
Integers, $\mathbb{Z}$	Polynomial functions on the complex plane, $\mathbb{C}[z]$
Rational numbers, $\mathbb{Q}$	Rational functions on the complex plane, $\mathbb{C}(z)$
Prime numbers, $\mathbb{P}$	Points in the complex plane, $\mathbb{C}$
Integers $\mod p^n$, $\mathbb{Z}/p^n$	$(n-1)$st-order Taylor series, $\mathbb{C}[z]/(z-a)^n$
p-adic integers, $\mathbb{Z}_p$	Taylor series, $\mathbb{C}[[z-a]]$
p-adic numbers, $\mathbb{Q}_p$	Laurent series, $\mathbb{C}((z-a))$

All the weird symbols are just the standard notations for these gadgets. The analogy goes as follows:

> *To study a polynomial "at a point" a in the complex plane,*
> *we can look at its value modulo $(z-a)$, or more generally mod $(z-a)^n$.*
>
> *To study an integer "at a prime" p,*
> *we can look at its value modulo p, or more generally $\mod p^n$.*

This is nice because the value of a polynomial modulo $(z-a)^n$ is just its Taylor series at the point a, where we keep terms up to order $n-1$.

We can also also take the limit as $n \to \infty$. If we do this to the integers $\mod p^n$ we get a ring called the "p-adic integers". For example, a typical 3-adic integer, written in base 3, looks like this:

$$\ldots 2100110202011010201210 2201$$

They're just like natural numbers in base 3, except they go on forever to the left! We add and multiply them in the obvious way, for example:

$$\begin{aligned} &\ldots 21001102020110102012102201 \\ +&\ldots 10201101012201201122010012 \\ =&\ldots 01202210110012010211112220 \end{aligned}$$

If we take the same sort of limit for Taylor series, we get Taylor series that go on forever — in other words, formal power series.

We can also ratios of p-adic integers, which are called p-adic numbers, and ratios of Taylor series, which are called Laurent series. A typical 3-adic number, written in base 3, looks like this:

$$\ldots 121010010012121201201201011.21021$$

They have to stop at some finite stage at the right, just as Laurent series have to stop at some finite stage: they can't have arbitrarily large negative powers of $z - a$.

Laurent series can be used to describe functions that have a pole at some point, like rational functions. Similarly, p-adic numbers can be used to describe rational numbers. Using more math jargon:

> *For any point a in $\mathbb{C}$, there's a homomorphism*
> *from the field of rational functions*
> *to the field of Laurent series,*
> *which sends polynomials to Taylor series.*

> *For any prime p, there's a homomorphism*
> *from the field of rational numbers*
> *to the field of p-adic numbers,*
> *which sends integers to p-adic integers.*

This lets us study rational numbers "locally" at the prime p using p-adic numbers, just as we can study a rational function locally at a point using its Laurent series. This technique can be quite useful. For example, a polynomial equation can have rational solutions only if it has p-adic solution for all primes p.

We might hope for the converse, but then we would be ignoring a funny extra "prime" besides the usual ones... something called the "real prime"!

The point is, besides being able to embed the rational numbers in the p-adics for any prime p, we can also embed them in the real numbers! This embedding is a bit different than the rest: it's based on a weird thing called an "Archimedean valuation", while the usual primes correspond to non-Archimedean valuations.

I'm sort of joking here, since if you're more used to real numbers than p-adics, you'll probably find Archimedean valuations to be *less* weird than non-Archimedean ones. The Archimedean valuation on the rational numbers is just the usual absolute value, while the non-Archimedean ones are other concepts of "absolute value", one for each prime p. If we take limits of rational numbers that converge using the usual distance function $|x - y|$, we get real numbers; if we take limits that converge using one of the non-Archimedean versions of this distance function, we get p-adic numbers. But from the viewpoint of number theory, it's the Archimedean valuation that's the odd man out! It indeed does act very weird and different than all the rest. That's why someone wrote this book:

3) M. J. Shai Haran, *The Mysteries of the Real Prime*, Oxford U. Press, Oxford, 2001.

... which you will see is deeply connected to mathematical physics.

If we take this weird "real prime" into account, things work better. We sometimes get results saying that some kind of polynomial equations have a rational solution if they have p-adic solutions for all primes p and also a real solution. For example, Hasse proved this was true for systems of quadratic equations in many variables.

Results like this are called "local-to-global" results, since they're analogous to constructing a function from local information, like its Laurent series at all different points.

In 1950, in his famous PhD thesis, John Tate came up with a clever way to formalize this "Laurent series at all different points" idea in the context of number theory. To do this, he formed a ring called the "adeles".

Indeed, this is what my whole discussion so far has been leading up to! Adeles are a really nice formalism, and you pretty much need to understand them to follow what people are doing in work on the Langlands Conjectures, or even simpler things, like class field theory. But, adeles seem like an arbitrary construction until you see them as an inevitable outgrowth of our desire to study integers "locally" at all different primes, including the real prime.

The definition is simple. An adele consists of a p-adic number for each prime p, together with a real number. . . but where all but finitely many of the p-adic numbers are p-adic integers!

This is the number-theoretic analogue of a Laurent series for each point in the complex plane, including the point at infinity. . . but with poles at only finitely many points! We could call such a thing an "adele for the rational functions".

Any rational function gives such a thing, just as any rational number gives an adele. And, we don't lose any information this way:

There's a one-to-one (but not onto) homomorphism
from the rational functions to the adeles for the rational functions.

There's a one-to-one (but not onto) homomorphism
from the rational numbers to the adeles for the rational numbers.

So, our table now looks like this. For good measure, I'll combine it with the related table in "Week 205":

Number theory	Complex geometry
Integers	Polynomial functions on the complex plane
Rational numbers	Rational functions on the complex plane
Prime numbers	Points in the complex plane
Integers $\mod p^n$	$(n-1)$st-order Taylor series
p-adic integers	Taylor series
p-adic numbers	Laurent series
Adeles for the rationals	Adeles for the rational functions
Fields	One-point spaces
Homomorphisms to fields	Maps from one-point spaces
Algebraic number fields	Branched covering spaces of the complex plane

There's a *lot* more to say about this analogy, but I think this is enough for now. Again, one of my secret goals was to start getting you comfy with adeles and the idea of studying number theory "locally".

For more on the geometrical side of number theory, I again recommend these:

4) Juergen Neukirch, *Algebraic Number Theory*, trans. Norbert Schappacher, Springer, Berlin, 1986.

5) Dino Lorenzini, *An Invitation to Arithmetic Geometry*, American Mathematical Society, Providence, Rhode Island, 1996.

But now, back to the subject of "inside tracks" — sneaky ways to get the beneficial feeling that you have secret insights into some problem.

For anyone with a background in categories, a good "inside track" on almost any math problem is to categorify it: to see that people are using sets where they could, and therefore *should*, be using categories or n-categories.

I've already hinted that zeta functions are an example of "decategorification". Now I'd like to make this more precise.

Let's think about the zeta function of a set X equipped with a one-to-one and onto function

$$f\colon X \to X$$

If you're a physicist, you might call this a "discrete dynamical system", with f describing one step of "time evolution". If you're a mathematician, you might call this a "$\mathbb{Z}$-set". After all, for any group G, a "G-set" is a set equipped with an action of G. If $G = \mathbb{Z}$ (the additive group of integers), this amounts to a one-to-one and onto function from some set to itself.

No matter what you call them, these are fundamental things. So, let's look at the *category* of $\mathbb{Z}$-sets! Here the objects are $\mathbb{Z}$-sets and the morphisms are functions that commute with time evolution.

As explained near the end of "Week 216", we can define a kind of zeta function for a $\mathbb{Z}$-set as follows:

$$Z(x) = \exp\left(\sum_{n>0} \frac{|\mathrm{fix}(f^n)| x^n}{n}\right)$$

where $|\mathrm{fix}(f^n)|$ is the number of fixed points of f^n. Of course, this only makes sense if all these numbers are finite; henceforth I'll assume my $\mathbb{Z}$-sets are "finite" in this special sense.

It turns out that you know a finite $\mathbb{Z}$-set up to isomorphism if you know its zeta function. So, a zeta function is just a sneaky way of talking about an *isomorphism class* of finite $\mathbb{Z}$-sets.

This is a fancy example of something we all learn as kids: counting! When we "count" a finite set, assigning a natural number to it, we are really determining its isomorphism class. Two finite sets are isomorphic if and only if they have the same number of elements. Operations on finite sets, like disjoint union and Cartesian product, are what give rise to operations on natural numbers, like addition and multiplication.

Summarizing this, we have the following motto, suitable for making into a bumper sticker:

*THE SET OF NATURAL NUMBERS IS THE DECATEGORIFICATION OF
THE CATEGORY OF FINITE SETS*

Similarly, this is what we're seeing now:

*THE SET OF ZETA FUNCTIONS IS THE DECATEGORIFICATION OF
THE CATEGORY OF FINITE $\mathbb{Z}$-SETS*

Beware: here I'm only talking about zeta functions of the above form. There are lots of other things people call zeta functions. So, don't read too much into this statement. But don't read too little into it, either! With an extra twist we can get most of the zeta functions showing up in number theory. In number theory, we typically get a $\mathbb{Z}$-set for each prime p, coming from the "Frobenius" for that prime. We thus get a bunch of "local" zeta functions $Z_p(x)$, one for each prime. We then multiply these to get one big fat "global" zeta function:

$$\zeta(s) = \prod_p Z(p^{-s})$$

Each local zeta function is a formal power series, while this global zeta function is a Dirichlet series. As I mentioned in "Week 217", formal power series live in the monoid algebra of $(\mathbb{N}, +, 0)$, while Dirichlet series live in the monoid algebra of $(\mathbb{N}, \times, 1)$. $(\mathbb{N}, +, 0)$ is the free commutative monoid on one generator, while $(\mathbb{N}, \times, 1)$ is the free commutative monoid on countably many generators — the primes! Everything fits together sweetly.

So, it's a good first step to think about the zeta function of a single $\mathbb{Z}$-set.

Now, there's another motto along the lines of the above two, which I've talked about before:

> *THE SET OF GENERATING FUNCTIONS IS A DECATEGORIFICATION OF THE CATEGORY OF FINITE STRUCTURE TYPES*

I explained this in "Week 185", "Week 190", and "Week 202". I've even taught a whole course on structure types (also known as "species") and the combinatorics of Feynman diagrams. The course notes by Derek Wise are available online:

6) John Baez and Derek Wise, "Quantization and Categorification", available at:
`http://math.ucr.edu/home/baez/qg-fall2003/`
`http://math.ucr.edu/home/baez/qg-winter2004/`
`http://math.ucr.edu/home/baez/qg-spring2004/`

So, I think this third example of decategorification is great. But, I'm not going to explain it in much detail here — just enough to say how it's related to zeta functions!

A stucture type F is a gadget that gives a set F_n for each $n = 0, 1, 2, \ldots$. We think of the elements of F_n as "structures of type F" on an n-element set — for example, orderings, or cyclic orderings, or n-colorings, or whatever type of structure you like. We only require that permutations of the n-element set act on this set of structures.

Let's say a structure type is "finite" if all the sets F_n are finite. Any finite structure type has a "generating function", which is a formal power series $|F|$ given by

$$|F|(x) = \sum \frac{|F_n|}{n!} x^n$$

Isomorphic structure types have the same generating function. However, structure types with the same generating function can fail to be isomorphic. This is why I said generating functions are "a" decategorification of finite structure types, instead of "the" decategorification.

Despite this defect, generating functions are still very useful in combinatorics. So, when we see a zeta function like

$$Z(x) = \exp\left(\sum_{n>0} \frac{|\mathrm{fix}(f^n)| x^n}{n}\right)$$

as a trick for decategorifying $\mathbb{Z}$-sets, we should instantly wonder if it's a generating function in disguise. And of course, it is!

Actually it's easiest to leave out the exponential at first. This power series:

$$\sum_{n>0} \frac{|\mathrm{fix}(f^n)| x^n}{n}$$

is the generating function for the structure type "being cyclically ordered and equipped with a morphism to the $\mathbb{Z}$-set X".

Huh?

We "cyclically order" a finite set by drawing it as a little circle of dots with arrows pointing clockwise from each dot to the next. A cyclically ordered set is automatically a $\mathbb{Z}$-set in an obvious way. So, here's a type of structure you can put on a finite set: cyclically ordering it and equipping the resulting $\mathbb{Z}$-set with a morphism to the $\mathbb{Z}$-set X.

And, if you work out the generating function of this structure type, you get

$$\sum_{n>0} \frac{|\mathrm{fix}(f^n)| x^n}{n}$$

Check it and see!

What about the exponential? Luckily, there's a standard way to take the exponential of a structure type: to put an $\exp(F)$-structure on a finite set S, we chop S into disjoint parts and put an F-structure on each part. So, the zeta function

$$Z(x) = \exp\left(\sum_{n>0} \frac{|\mathrm{fix}(f^n)| x^n}{n}\right)$$

is the generating function for "being chopped up into cyclically ordered parts, each equipped with a morphism to the $\mathbb{Z}$-set X".

But this is just a long way of saying: "being made into a $\mathbb{Z}$-set and equipped with a morphism to the $\mathbb{Z}$-set X".

Or, in category theory jargon, "being a $\mathbb{Z}$-set over X".

So:

> *THE ZETA FUNCTION OF THE $\mathbb{Z} - SET\ X$ IS THE GENERATING FUNCTION*
> *OF*
> *"BEING A $\mathbb{Z} - SET$ OVER X"*

By the way, this is the kind of thing you could do with *any* structure type F. Given an F-structured set X, we get a new structure type "being an F-structured set equipped with a morphism to X". Or, in category theory jargon, "being an F-structured set over X". The generating function of this could be called the "zeta function" of our F-structured set X. I have no idea how important this is...

... but I want to keep gnawing away on the connection between zeta functions and the generating functions of combinatorics, because to understand number theory, I need all the "inside tracks" I can get!

Addendum: After reading this Week's Finds, Kevin Buzzard emailed me the following remarks. He begins by talking about adeles for any algebraic number field K — a fairly straightforward generalization of the case I discussed above, where K is the rational numbers:

> *The adeles were used in a really powerful way in the theorems and proofs of global class field theory (you don't want to read the proofs. I did this precisely once in my life and they are very unilluminating). But the theorem — if K is a number field then the abelianisation of $\mathrm{Gal}(\overline{K}/K)$ is canonically isomorphic to*
>
> $$K^* \setminus \mathrm{Adeles}_{K^*}/(K^*_\infty)^0$$
>
> *(the last term being the connected component of the product of the infinite completions of K) — is incredibly important.*
>
> *Much easier going is Tate's thesis (in the book by Cassels and Froehlich). Tate observes that Fourier analysis works on any locally compact abelian group (Haar measure is the replacement for "usual" measure), and then gives a very short proof of the analytic continuation and functional equation of all Hecke's L-functions by simply pushing through an analogue of the proof you know of the functional equation of the zeta function in this much more general context. I think this is an amazingly powerful use of the adeles. Tate's approach explains the fudge factors, the factors at infinity, everything.*
>
> *A word on analogies. If you want to say that the p-adic integers are analogous to the formal power series ring $\mathbb{C}[[z-a]]$ (call it $\mathbb{C}[[z]]$ for simplicity) then in fact some people would say that this was* not *an analogy—this was simply two instances of the same thing, namely a complete discrete valuation ring. Similarly, you might say that $\mathbb{Z}$ is analogous to $\mathbb{C}[z]$, but again some people would just tell you to go and get a book on commutative algebra and look up the word "Dedekind domain" — both of these are examples. A geometer might even go and tell you to go and find out what a regular 1-dimensional scheme was!*
>
> *One thing I didn't realise when I was learning all this stuff however, was that there is some stuff that just goes through for all Dedekind domains (e.g. construction of adeles, existence of class group etc), and there is some stuff that actually requires more. Tate's thesis for example requires more — it doesn't just work for all Dedekind domains because Tate needs a Haar measure and so he needs completions to be locally compact, which is basically the same as demanding that* all residue fields are finite*. Here's something you can do for $\mathbb{Z}_p$ which you can't do for $\mathbb{C}[[z]]$: let's define the measure of $a + p^n\mathbb{Z}_p$ to be p^{-n}. Then this*

is finitely additive, because $\mathbb{Z}_p$ is the disjoint union $p\mathbb{Z} \cup 1+p\mathbb{Z} \cup \ldots \cup (p-1)+p\mathbb{Z}$, and $1/p + 1/p + \ldots + 1/p$ (p times) is 1. But you can't do this for $\mathbb{C}[[z]]$ because the cardinality of $\mathbb{C}$ is infinite. This naive measure on Z_p is exactly what you need to define p-adic L-functions, by the way! But they are another (related) story.

When you move from Discrete Valuation Rings to Dedekind Domains the same care needs to be applied: it's a famous theorem that the ideal class group of (the integers of) a number field is finite. But it's not true that the class group of a Dedekind domain is finite: the class group of $\mathbb{C}[z]$ is finite as $\mathbb{C}[z]$ is a principal ideal domain, but the class group of $\mathbb{C}[x,y]/(y^2 - x^3 - 1)$ is infinite (the class group is essentially the underlying elliptic curve, which is an infinite group). Again you have to demand that residue fields are finite. So this stops you thinking about $\mathbb{C}[z]$ and its finite extensions, it forces you to start thinking about $k[z]$ where k is a finite *field. Of course algebraic geometers aren't scared of finite fields (well, at least, the ones I talk to the most aren't), so after a while your analogy is going to break because $\mathbb{C}$ is infinite. Langlands' philosophy is (or at least, was—it has been generalised in various directions now) about global fields, which means either number fields or finite extensions of $k(z)$ where k is a finite field. Of course Lafforgue recently proved everything in the function field case, hence the Fields Medal.*

Kevin

I replied:

> *The adeles were used in a really powerful way in the theorems and proofs of global class field theory (you don't want to read the proofs. I did this precisely once in my life and they are very unilluminating).*

Then I think there must exist nicer proofs! There can't possibly be such important and beautiful results where the best possible proof is unilluminating. So, someone needs to work on this more... perhaps me, if everyone else is too busy. :-)

> *But the theorem — if K is a number field then the abelianisation of $\mathrm{Gal}(\overline{K}/K)$ is canonically isomorphic to $K^* \setminus \mathrm{Adeles}_{K^*}/(K^*_\infty)^0$ (the last term being the connected component of the product of the infinite completions of K) — is incredibly important.*

It's beautiful, too!

> *Much easier going is Tate's thesis (in the book by Cassels and Froehlich). Tate observes that Fourier analysis works on any locally compact group (Haar measure is the replacement for "usual" measure), and then gives a very short proof of the analytic continuation and functional equation of all Hecke's L-functions by simply pushing through an analogue of the proof you know of the functional equation of the zeta function in this much more general context. I think this is an amazingly powerful use of the adeles. Tate's approach explains the fudge factors, the factors at infinity, everything.*

This sounds great. I've always heard people rave about Tate's thesis, and now it's time for me to read it... or at least the book you mention — but I get the feeling the actual thesis is good.

> *A word on analogies.If you want to say that the p-adic integers are analogous to the formal power series ring $\mathbb{C}[[z-a]]$ (call it $\mathbb{C}[[z]]$ for simplicity) then in fact some people would say that this was not an analogy — this was simply two instances of the same thing, namely a complete discrete valuation ring.*

Yes, but I don't want to intimidate my readers with concepts like "complete discrete valuation ring" — I'd rather lure them in with the charm of a mysterious analogy! I think think this is how things went historically, too... judging from Weil's remarks:

7) *Martin H. Krieger, "A 1940 letter of Andre Weil on analogy in mathematics",* AMS Notices ***52*** *(March 2005), 334–341. Available at* `http://www.ams.org/notices/200503/200503-toc.html`

He even talks about how the charm of an analogy evaporates when you find a generalization that encompasses both terms:

> *"The day dawns when the illusion vanishes; intuition turns to certitude; the twin theories reveal their common source before disappearing; as the Gita teaches us, knowledge and indifference are attained at the same moment. Metaphysics has become mathematics, ready to form the material for a treatise whose icy beauty no longer has the power to move us."*

Wouldn't want that!

> *Similarly, you might say that $\mathbb{Z}$ is analogous to $\mathbb{C}[z]$, but again some people would just tell you to go and get a book on commutative algebra and look up the word "Dedekind domain" — both of these are examples. A geometer might even go and tell you to go and find out what a regular 1-dimensional scheme was!*

I've read lots of theorems about Dedekind domains, and on good days I can even remember the definition...

But, I really want to keep things a bit vague and misty for my readers — most importantly because This Week's Finds is supposed to be fun, but also because a lot of the coolest stuff happens when you extend vague analogies in shocking ways.

For example, thinking of $\mathrm{Spec}(\mathbb{Z})$ as a plane that gets a fundamental group when you poke a hole in it and remove a prime is nice for visualizing an individual Frobenius generator, but deeper results suggest that it's good to think of $\mathrm{Spec}(\mathbb{Z})$ as 3-dimensional! This leads to the extensive analogy between $\mathrm{Spec}(\mathbb{Z})$ and knot theory discussed here...

> > 8) *Adam Sikora, "Analogies between group actions on 3-manifolds and number fields", available as math.GT/0107210.*
> >
> > 9) *Christopher Deninger, "A note on arithmetic topology and dynamical systems", available as math.NT/0204274.*
> >
> > > *(Actually I think there is even a kind of Langlands philosophy for $\mathbb{C}(z)$ and its finite extensions nowadays worked out recently by Beilinson and Drinfeld. I saw Beilinson give several lectures on it, more than once, and still didn't really get it, I am too number-theoretic.)*
>
> *Is this the "geometric Langlands program" stuff? Physicists are getting interested in that...*
>
> *Best,*
> *jb*

To understand Kevin's reply, recall that any algebraic number field K has a "maximal abelian extension" K^{ab}. This is the biggest algebraic extension of K whose Galois group is *abelian*. When $K = \mathbb{Q}$, the Kronecker-Weber theorem says this is obtained by throwing in all the roots of unity. Since a field obtained from $\mathbb{Q}$ by throwing in a root of unity is called a "cyclotomic field", people sometimes call this $\mathbb{Q}^{\mathrm{cyc}}$.

In "Week 201" I described the Galois group $\mathrm{Gal}(\mathbb{Q}^{\mathrm{cyc}}/\mathbb{Q})$. Unsurprisingly, this is the abelianization of the big bad Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$: the Galois group of the algebraic closure of $\mathbb{Q}$. In what follows, Kevin more generally discusses $\mathrm{Gal}(K^{\mathrm{ab}}/K)$, which is the abelianization of $\mathrm{Gal}(\overline{K}/K)$.

Understanding groups like $\mathrm{Gal}(\overline{K}/K)$ is one of the great unfulfilled dreams of number theory. Understanding its abelianization is one of the great triumphs of late nineteenth to mid twentieth century mathematics. This is called *class field theory*.

> > > *The adeles were used in a really powerful way in the theorems and proofs of global class field theory (you don't want to read the proofs. I did this precisely once in my life and they are very unilluminating).*
>
> > *Then I think there must exist nicer proofs!*
>
> *This is related to one of Hilbert's problems! (the 12th one). So you must be thinking along the right lines :-)*
>
> *Abstract theorem: if K is a number field then the abelianisation of $\mathrm{Gal}(\overline{K}/K)$ is isomorphic to $K^* \setminus \mathrm{Adeles}_{K^*}/(K^*_\infty)^0$.*
>
> *Remark: the right hand group is very "concrete", in the sense that one can write down explicit finite quotients of it. (Why quotients? Because quotients of Galois groups are again Galois groups.) For example, I can just write down a big subgroup e.g. K^*_∞ times the product of $O^*_{K_v}$, where v runs through all the finite places of K, and the quotient of the big group by the big subgroup can be checked to be compact and discrete, so it's finite. We have hence "explicitly" written down a finite quotient of $\mathrm{Gal}(\overline{K}/K)$, corresponding to a finite extension H of K. The objects on the right hand side are rather abstract, but this is as*

explicit a quotient group of the right hand side as you could possibly ask for — we understand exactly what's going on at every place, for example. Hence this is as explicit a finite extension of K as you could possibly ask for, if you admit the isomorphism of class field theory. Indeed, if you know a bit more about the isomorphism, you will know that this quotient H is unramified at all the primes of K, and is indeed the largest abelian extension of K with this property. H is (by definition) the Hilbert Class Field of K. In your analogy, given a curve, a natural thing to think of would be the universal covering space of the curve. Unfortunately number theorists aren't good enough to understand all of $\mathrm{Gal}(\overline{K}/K)$, they have to abelianise first, so H corresponds to the maximal unramified cover of the curve with abelian covering group.

Great! So we have all this machinery, this beautiful isomorphism, this completely canonical description of the Galois group, and we make a very explicit and natural construction on the right hand side, so now let me give you a number field like $\mathbb{Q}(\sqrt{10})$ and ask you what H is!

Now suddenly you see a big disadvantage of the glorious proof of the isomorphism which goes via all this cohomological chasing around — it shows the existence of H but doesn't tell us what it is. At all. And at the end of the day, there are a lot of number theorists out there that are actually interested in numbers*, rather than in abstract results which hold for all number fields or whatever.*

Hilbert's question was: "well, this is all well and good, but can anyone actually write down *the isomorphism, rather than actually prove its existence? Can anyone write it down sufficiently concretely so that people can just read off the Hilbert Class Group of a number field, given the field?" And, to be honest, although a lot is known, Hilbert would probably say that the answer is still "no". If you were to find a "better" proof of global class field theory then perhaps the answer would change. In fact the Hilbert Class Field is just the tip of the iceberg — global class field theory gives us a description of the abelianisation of $\mathrm{Gal}(\overline{K}/K)$, and this abelianisation corresponds to a field K^{ab}, of infinite degree over K, but Galois, with infinite abelian Galois group. If I give you K, can you tell me K^{ab}? Hilbert even wanted to know this (his questions are maximally greedy, I guess).*

Hilbert's question was not totally out of the blue. It can be done for $K = \mathbb{Q}$, indeed it had been done 50 years before Hilbert's question. Kronecker and Weber knew not just the Hilbert class field of $\mathbb{Q}$, they even knew $\mathbb{Q}^{\mathrm{ab}}$, it's just the union of $\mathbb{Q}(1^{\frac{1}{n}})$, where $1^{\frac{1}{n}}$ is $\exp(2\pi i/n)$. Let me labour a point which experts in the theory feel is highly important: the exponential function is transcendental — it doesn't belong in algebraic geometry because it's not in $\mathbb{C}(z)$. On the other hand, this transcendental function, when evaluated at certain places, gives algebraic numbers out, and it is these algebraic numbers which explain all the class field theory of the rationals.

*Now a **TOTALLY AMAZING GENERALISATION**: let K be an imaginary quadratic field, so $\mathbb{Q}(\sqrt{d})$ for some integer $d < 0$. Let L be the lattice in the complex numbers with basis 1 and $\sqrt{d}$ (this is a lattice as $d < 0$). Quotient out the complex numbers by this lattice. You get a 1-dimensional complex torus, so an elliptic* *curve. The curve has a j-invariant, which is going to be a "random" complex number. One can compute this number to as many decimal places as one likes nowadays (in practice). For example, if $d = -5$ then my computer instantly tells me that the j-invariant of the corresponding elliptic curve is*

$$1264538.909475140509320227047410703421481442121566908396881751412781728159444422249946349547842$$

Equally quickly, my computer spots that this (to 100 decimal places, at least) looks awfully like one of the roots of

$$x^2 - 1264000x - 681472000$$

(it agrees with it to 100 decimal places, despite the fact that the j-function is again "transcendental" — we have put an algebraic number in and appear to have got an algebraic number out).

The awesome fact is that the splitting field of this polynomial over $\mathbb{Q}(\sqrt{-5})$ (i.e. the field you get by adjoining all of the roots of this polynomial to $\mathbb{Q}(\sqrt{-5})$) is the Hilbert Class Field of $\mathbb{Q}(\sqrt{5})$! Even better*: I can even tell you K^{ab}, if K is $\mathbb{Q}(\sqrt{-5})$: write down a model for the elliptic curve in the form $y^2 = f(x)$ with f a cubic with coefficients in K (use the Weierstrass $\wp$-function and its derivative), and now look at all the points of finite order on this elliptic curve. The x and y coordinates of all these points are algebraic numbers, and they generate K^{ab}.*

I am proud now to give you a genuine analogy :-)

Rational field	*Imaginary quadratic field*
Group $\mathbb{C}/\mathbb{Z}$	*Elliptic curve* $\mathbb{C}/($*integers of the field*$)$
Element of finite order in the group	*Element of finite order in the group*
function $z \mapsto \exp(2\pi i z)$	*Weierstrass $\wp$-function (and its derivative)*

In both cases, the function maps the group to an algebraic variety ($\mathbb{C}^*$ in the first case, y^2 = cubic in the second), and evaluating the function at complex numbers which give torsion points of the group (rational numbers in the first case, elements of the imaginary quadratic field in the second) gives numbers which by all rights should be random complex numbers, but turn out to be not only algebraic, but to generate the maximal abelian extension of the number field.

This really is an analogy because no-one has (as far as I know) a clue how to do this more generally. Note that the rationals and the imaginary quadratic fields are the only fields with exactly one infinite place. Is this why they are the only fields we can "do"? This technique, of "explicitly" generating abelian extensions of a number field, is called "explicit class field theory" and, other than the (non-trivial) contribution by Shimura and Taniyama where they used higher-dimensional abelian varieties to push the analogy slightly further, it's still a big mystery.

> *There can't possibly be such important and beautiful results where the best possible proof is unilluminating.*

In the case of local class field theory, there are now some really neat proofs, where you in some sense really do write down the maximal abelian extension of an arbitrary finite extension of Q_p, again using torsion points in groups (formal groups). But people have spent a century looking for more illuminating proofs, motivated by Hilbert's question. Until then, we just have to rely on known algorithms for computing Hilbert Class Fields (there are algorithms that work in lots of cases, and they rely on known abstract theorems, but one might argue that none of them are really "explicit", they just go, I think, by essentially looking at lots of fields until one finds the one that works, rather than working out which one is the right one by pure thought).

You should talk about special values of L-functions! Do you know the analytic class number formula? The degree of H over K is called the class number of K and, totally amazingly, it is related to the special value of an L-function. The Birch-Swinnerton-Dyer conjecture is just a natural generalisation of this formula to elliptic curves over the rationals, but again, what used to be an analogy has turned into two instances of a more general piece of mathematics (ranks of K-groups, Beilinson's conjectures etc.).

Sorry to go on! I just get quite enthusiastic about this stuff.

> *This sounds great. I've always heard people rave about Tate's thesis, and now it's time for me to read it... or at least the book you mention — but I get the feeling the actual thesis is good.*

The thesis was never published, Tate I guess wasn't happy that he just reproved a known theorem or something? Cassels–Froehlich is the canonical reference. Serre's article in there talks about the links to elliptic curves in the im quad case too. A nice book!

Thanks for the link to the letter of Weil — interesting stuff! I have pity on his sister. I have heard that (Andre) Weil's house in Paris had a plaque on it saying "Simone Weil used to live here" (because she did). Funny that a genius had to live in the shadows of his sister (who by all accounts might also have been a genius).

Another funny piece of gossip. I think it was Chevalley who originally started thinking about ideles (the ideles are the group of units of the adeles). I am no historian so might have this wrong. Chevalley(?) wrote a book on algebraic number theory where he talked about ideals and also about these ideles, which he referred to as "ideal elements" and which he abbreviated as "id.eles". Later on the period was dropped, so they became ideles. I think it was Serre who saw that the ideles were the units of a ring, and christened the ring with the name of "adeles". If you get Serre's collected works and look at his CV at the beginning, you will see that his mother was called Adele. Coincidence? :-)

> *(Actually I think there is even a kind of Langlands philosophy for $\mathbb{C}(z)$ and its finite extensions nowadays worked out*

> > *recently by Beilinson and Drinfeld. I saw Beilinson give several lectures on it, more than once, and still didn't really get it, I am too number-theoretic.)*
>
> *Is this the "geometric Langlands program" stuff? Physicists are getting interested in that. . .*

Yes.

Kevin

The scientific life of mathematicians can be pictured as a trip inside the geography of the "mathematical reality" which they unveil gradually in their own private mental frame.

It often begins by an act of rebellion with respect to the existing dogmatic description of that reality that one will find in existing books. The young "to be mathematician" realize in their own mind that their perception of the mathematical world captures some features which do not fit with the existing dogma. This first act is often due in most cases to ignorance but it allows one to free oneself from the reverence to authority by relying on one's intuition provided it is backed by actual proofs. Once mathematicians get to really know, in an original and "personal" manner, a small part of the mathematical world, as esoteric as it can look at first, their trip can really start. It is of course vital not to break the "fil d'arianne" which allows one to constantly keep a fresh eye on whatever one will encounter along the way, and also to go back to the source if one feels lost at times. . .

It is also vital to always keep moving. The risk otherwise is to confine oneself in a relatively small area of extreme technical specialization, thus shrinking one's perception of the mathematical world and its bewildering diversity.

The really fundamental point in that respect is that while so many mathematicians have been spending their entire life exploring that world they all agree on its contours and on its connexity: whatever the origin of one's itinerary, one day or another if one walks long enough, one is bound to reach a well known town i.e. for instance to meet elliptic functions, modular forms, zeta functions. "All roads lead to Rome" and the mathematical world is "connected".

In other words there is just "one" mathematical world, whose exploration is the task of all mathematicians, and they are all in the same boat somehow.

— *Alain Connes*

Week 219

July 4, 2005

I'm about to head to Sydney and Canberra to help celebrate the 60th birthday of Ross Street . . . the world's best n-category theorist!

1) Categories in Algebra, Geometry and Mathematical Physics, conference in honor of the 60th birthday of Ross Street, http://streetfest.maths.mq.edu.au/

Lots of people will be talking about the interface of higher-dimensional algebra and physics. Some will be talking about higher gauge theory, which generalizes ordinary gauge theory from point particles to strings, loops, or higher-dimensional "branes" by replacing groups with n-groups.

Ezra Getzler will be speaking on how to get an n-group from a Lie n-algebra, and Mikhail Kapranov will be speaking on higher-dimensional holonomies. Alissa Crans, Danny Stevenson and I will also be talking about this stuff.

There will eventually be a conference proceedings, but for now you can see the abstracts of people's talks on the website. You can see my talks here:

2) John Baez, "Higher gauge theory", http://math.ucr.edu/home/baez/street/

Anyway, before I take off, here's a roundup of stuff I've been meaning to mention: the centennial of Einstein's "annus mirabilis", the Pioneer anomaly, silicon photonics, a company that plans to sell quantum computers, and some relationships between Klein's quartic curve, the Fano plane, and special relativity over the integers $\bmod 7$. I want to leave you lots of stuff to think about. :-)

Okay. . . let's start with Einstein's "annus mirabilis".

1905 was indeed a miraculous year for Albert Einstein. He published four earth-shaking papers, all in the same journal:

3) Albert Einstein, "On a heuristic viewpoint concerning the production and transformation of light", *Annalen der Physik* **17** (1905), 132–148. Available at http://dbserv.ihep.su/~elan/src/einstein05/eng.pdf

   "On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat", *Annalen der Physik* **17** (1905), 549–560. Available at http://lorentz.phl.jhu.edu/AnnusMirabilis/AeReserveArticles/eins_brownian.pdf

   "On the electrodynamics of moving bodies", *Annalen der Physik* **17** (1905), 891–921. Available at http://dbserv.ihep.su/~elan/src/einstein05b/eng.pdf

   "Does the inertia of a body depend upon its energy content?", *Annalen der Physik* 18 (1905), 639–641. Available at http://dbserv.ihep.su/~elan/src/einstein05c/eng.pdf

In the first of these papers, Einstein explained the photoelectric effect by assuming that light consisted of particles each carrying an energy equal to Planck's constant times its frequency. This was an important step towards quantum mechanics — a theory he would later fight against.

In the second, he showed that Brownian motion was explained by the existence of atoms. His calculations were later used to measure Boltzmann's constant.

In the third, Einstein derived the formula for Lorentz transformations from two simple assumptions: only relative motion can be detected, and every unaccelerated observer measures light to have the same speed.

In the fourth, he derived a relation between mass, momentum, and energy, including as a special case the famous formula $E = mc^2$ relating the mass of a body to its energy at rest.

The most impressive thing about these papers is how simple and readable they are: they go from clearly stated assumptions to world-shaking conclusions with clear logic and only a little math.

Which makes one wonder: will there ever be another Einstein? Can one person ever again make such revolutions in physics? Or are all the remaining discoveries yet to be made in physics too complicated? Why has fundamental physics been stuck ever since the completion of the Standard Model? There are lots of glorious theories, but none of them have gotten any experimental confirmation. There have also been some big empirical discoveries — dark matter, dark energy, neutrino oscillations, and more — but these weren't predicted, and our theories still don't shed much light on them.

So, are the times ripe for a good new idea... a really smart person who will fit the jigsaw puzzle together? Or are there still too many missing pieces?

Lee Smolin has some interesting thoughts about this:

4) Lee Smolin, 'Why no "new Einsteins"?', *Physics Today*, June 2005, 56–57.

I think the mention of Einstein is mainly just a trick to get people to read the article. He focuses on institutional pressures that push physicists to conform: to follow "research programs" in big teams rather than strike out on their own.

This is certainly a big problem. But I'm not sure it's the only problem.

Fundamental physics is in a tough situation, partly a victim of its own success. Pure thought may not be enough. To make real progress, it helps to have lots of experimental results that don't fit the current theories. Not just a few numbers here or there, like neutrino masses. We want *piles* of unexplained data! — enough to draw lots of graphs. Then we could cook up theories that fit the data.

Barring this, we need to assemble all the mysteries we can get our hands on, and see if they fit together somehow.

Luckily, right now several spacecraft are leaving the the Solar System... and they're leaving us with the best of gifts: a mystery!

Pioneer 10 was launched in 1972. It was the first spacecraft to travel through the asteroid belt, and the first to take closeups photos of Jupiter, and study the intense magnetic field of this planet. I remember getting excited about this when I was a kid.

Pioneer 10 was also the first spacecraft to explore the outer solar system. In 1983 it passed the orbit of Neptune. On February 7th, 2003, Pioneer 10's radioactive power source weakened to the point where its radio signals became too feeble to detect. When this happened, it was 80 astronomical units from the Sun: in other words, 80 times as far from the Sun as we are, or about twice as far out as Pluto.

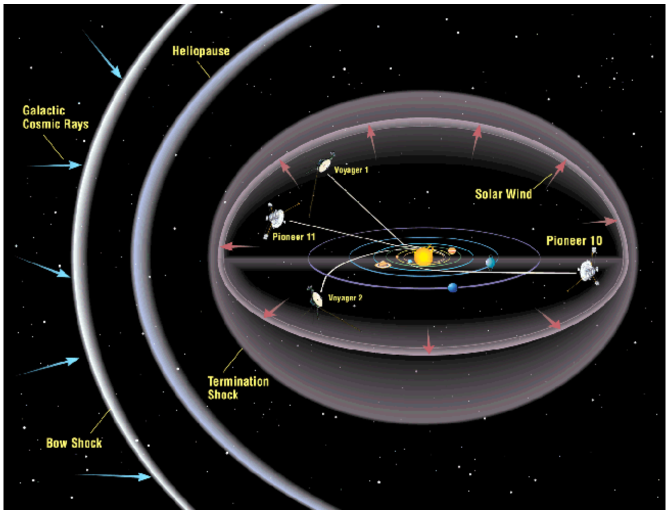
But this was just the beginning of its cold dark journey. It will continue to coast through deep space in the general direction of the red giant Aldebaran, 68 light years from us. A light year is about 63,000 astronomical units, so this journey will take a while: over 2 million years! We'll probably be long gone by then, for better or worse.

Pioneer 11 was launched in 1973. It followed its sister ship to Jupiter, then swung past Saturn, and then sailed out into the night, studying the solar wind as it went. On September 30, 1995 its power source became too weak to run any more experiments, and NASA stopped monitoring it. It was 45 astronomical units from the Sun, moving out at 2 AU per year.

We are no longer in the beam of its radio signal. Its antenna cannot be rotated. It is heading towards Aquila — the Eagle — and it will pass one of the stars in this constellation in 4 million years.

Here's a picture of what these spacecraft are doing:

5) NASA, Pioneer path, `http://spaceprojects.arc.nasa.gov/Space_Projects/pioneer/path.html`



See how they'll eventually pass through a bubble called the "termination shock"? That's where the solar wind drops below supersonic speed as it crashes into the gas of the Milky Way. You can see a pattern vaguely reminscient of the wave formed by a boat sailing through the sea — that's because we're moving through the Milky Way.

There are also two other spacecraft in this picture: Voyager 1 and 2! These craft are still transmitting, and Voyager 1 is now farther than Pioneer 10. It crossed the termination shock in December 2004, zipping along at 3.6 AU per year. It's sending back information about this region of space — called the "heliosheath" — and its batteries should be good until 2020.

But what's the mystery?

Well, the Pioneer missions yielded the most precise information we have about navigation in deep space. However, analysis of their radio tracking data indicates a small unexplained acceleration towards the Sun! The magnitude of this acceleration is roughly 10^{-9} meters per second per second. It's called the "Pioneer anomaly".

This anomaly has also been seen in the Ulysses spacecraft, and possibly also in the Galileo spacecraft, though the data is much more noisy, since these were Jupiter probes, hence much closer to the Sun, where there is a lot more pressure from solar radiation. The Voyagers are not set up to provide good data on this issue, since they are able to rotate themselves and this messes things up. The Viking mission to Mars did *not* detect the Pioneer anomaly — and it would have, had an acceleration of this magnitude been present, because its radio tracking was incredibly accurate — good to about 12 meters.

Many physicists and astronomers have tried to explain the Pioneer anomaly using conventional physics, but so far nobody seems to have succeeded. Radiation pressure from the Sun, the solar wind, the back-reaction from the radio emissions: all these point the wrong way! Other explanations also seem to fail, like gravity from the Kuiper belt, small amounts of gas venting from the spacecraft, and thermal radiation from the craft,

As conventional explanations seemed to fail, people started trying to explain the anomaly using new physics — for example, modified theories of gravity, or dark matter. But it's hard to get these explanations to work, either. For example, explaining the Pioneer anomaly by the gravitational attraction of dark matter would require more than .0003 solar masses of dark matter within 50 AU units of the Sun. But this is in conflict with our highly successful calculations of planetary orbits — even a millionth of a solar mass of dark matter in this region would be enough to throw those off!

So, we may have an interesting clue on our hands.

For more information on the Pioneer anomaly, see these sources and the many references therein:

6) Wikipedia, "Pioneer anomaly", `http://en.wikipedia.org/wiki/Pioneer_anomaly`

7) Chris P. Duif, "Pioneer anomaly — literature and links", `http://www.space-time.info/pioneer/pioanomlit.html`

I got interested in the Pioneer anomaly when I read this recent proposal for a mission to study it:

8) The Pioneer Collaboration, "A mission to explore the Pioneer anomaly", available as gr-qc/0506139.

Finally, here's a fun book on the Pioneer missions:

9) Mark Wolverton, *The Depths of Space: The Story of the Pioneer Planetary Probes*, Joseph Henry Press, 2004. Available at `http://www.nap.edu/books/0309090504/html/`

Next: silicon lasers. I don't want to say much about these, just that Intel thinks they could be the next big thing. You've probably heard about "Moore's Law", how the number of components in an integrated circuit doubles every 2 years. Or more vaguely: how computers keeping getting more powerful, really fast! And you've probably heard that this exponential growth is in danger of running into a brick wall someday. One of the problems is that copper wire carries too little data, too slowly.

For a long time some people have touted "photonics" as the way around this: transmitting information as pulses of light, rather than clumps of electrons. One problem is that substances that emit such pulses tend to be expensive, like gallium arsenide and indium phosphide. But now they've figured out to make silicon function as a laser! Researchers at Intel have created a silicon laser that emits a continuous beam of light. They've also developed a modulator that chops the beam into 10 billion pulses per second — a 10 gigahertz signal.

For details, try this:

10) Intel, "Silicon photonics", `http://www.intel.com/technology/silicon/sp/`

11) Robert Service, "Intel's breakthrough", *Technology Review*, July 2005, 62–65. Also available at `http://www.technologyreview.com/articles/05/07/issue/feature_intel.asp`

While we're talking about high tech, how about quantum computers? Most of them don't work very well — yet, say the optimists. Interaction with the environment ruins the coherence of the quantum state. But there's a company called D-Wave Systems that aims to build a different breed of quantum computer — one that doesn't mind a fair amount of noise. It's an analog chip made of lots of superconductors, which is supposed to quantum tunnel to a lowest-energy state, The idea is that you can get it to solve lots of minimization problems this way, like the travelling salesman problem.

I don't see why it'll work better than other analogue computers, or methods like simulated annealing. It could get stuck when there are lots of "almost-minima" to sift through. . . just like glass gets stuck when it tries to find *its* lowest energy state. Physicists call this "frustration": the poor glass in your window is "frustrated", trying to crystallize but unable to decide how to do it. In principle it could do it by quantum tunnelling, but in practice that takes forever. Why will D-Wave Systems' computer be better?

By the way, they admit their computer *can't* do cool stuff like rapidly factor large numbers via Shor's algorithm, the way full-fledged quantum computers should. True devotees of quantum computation probably wouldn't even call it a quantum computer! But I'll be happy if it works.

They've certain managed to convince some investors. Here's some more info:

12) Erika Jonietz, "Quantum calculation", *Technology Review*, July 2005, 24–25. Also available at `http://www.technologyreview.com/articles/05/07/issue/forward_quantum.asp`

But enough practical stuff! Now let me say a bit more about the Klein quartic.

I wrote about this gadget in "Week 214" and "Week 215". Simply put, this is a "Platonic surface": a 3-holed torus tiled by regular heptagons, with 3 heptagons meeting at each vertex.

It's perfectly symmetrical, but you can't stuff it into 3-dimensional Euclidean space without warping it. So, its charms are a bit more esoteric than those of, say, a dodecahedron.

Of course, this is the kind of challenge that some people just can't resist. Mike Stay and Gerard Westendorp bravely tried making paper models of it:

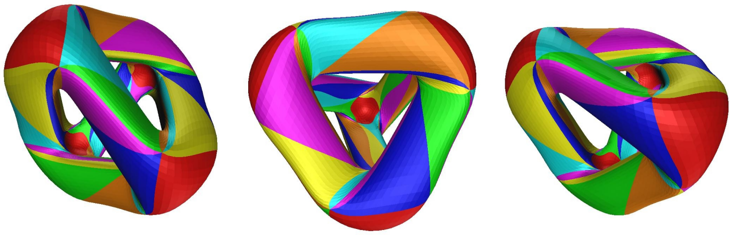
13) Mike Stay, "Klein quartic", `http://math.ucr.edu/~mike/klein/`

14) Gerard Westendorp, "Geometry", `http://www.xs4all.nl/~westy31/Geometry/Geometry.html`

Mike wisely stopped just short of the final step, which would create a nasty crumpled mess. Gerard succeeded in completing the task by switching to *pastry* instead of paper. Check it out!

There might be a small niche market for Klein quartic birthday cakes... but computer graphics are probably better if you just want to visualize this surface instead of actually eat it. About 10 or 12 years ago, Joe Christy made the following pictures using a program called Geomview, which makes virtual 3d objects:

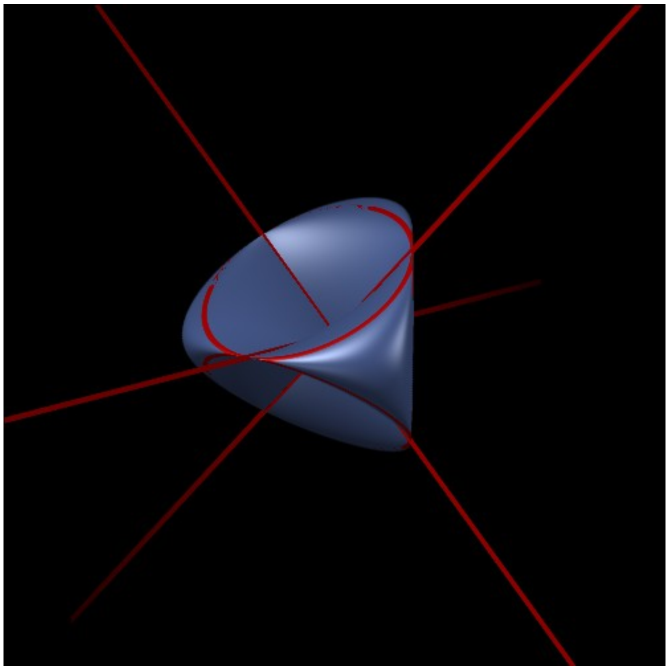
15) Joe Christy, Klein quartic pictures:
`http://math.ucr.edu/home/baez/pentacontihexahedron.jpg`
`http://math.ucr.edu/home/baez/pentacontihexahedron2.jpg`
`http://math.ucr.edu/home/baez/pentacontihexahedron3.jpg`



It took about a day on the fastest Linux machine they had at the time. The advantage of Geomview is that once the virtual object is done, you can quickly create different views. These pictures show the dual version of Klein's quartic, which has 7 regular triangles meeting at each quarter. The funky name "pentacontihexahedron" refers to the fact that there are 56 of these triangles.

While I'm at it, I can't resist showing you another beautiful picture:

16) Joe Christy, "Fano plane on Roman surface", `http://math.ucr.edu/home/baez/roman.jpg`



The blue thing is called the "Roman surface" because it was discovered by the mathematician Jakob Steiner while he was visiting Rome. It's a self-intersecting immersion of the projective plane in 3d space. On it, the 7 lines of the Fano plane are visible in red, with four of them drawn as circles.

Mathematically, one nice thing about this picture is that it exhibits the tetrahedral symmetry of the Fano plane!

You can see all these pictures and much more on my Klein quartic website:

17) John Baez, "Klein's quartic curve", `http://math.ucr.edu/home/baez/klein.html`

You may think I'm digressing, but the relation between Klein's quartic curve and the Fano plane underlies what I want to talk about today. Greg Egan and I realized that this relation is just part of a bigger picture involving special relativity in 3-dimensional spacetime... over the integers $\bmod 7$.

Huh?

Well, these days so-called physicists have no shame studying physics in all sorts of dimensions, but they usually confine themselves to building their spacetimes out of the real numbers.

That makes sense if they're trying to claim some relevance to real-world physics, however slight. But mathematically, there's no reason not to try other number systems, like finite fields, just for the fun of it.

And this sheds new light on the Klein quartic. Why? Because the symmetries of the Klein quartic and the Fano plane also act as *Lorentz transformations* in 3 dimensional spacetime if you work using the integers $\bmod 7$. This lets us see the Klein quartic and Fano plane as being closely related to special relativity in this funny context.

Let's see how this goes.

First, recall that a "field" is a number system where you can add, subtract, multiply and divide to your hearts content, with all the basic laws holding that hold for real numbers. A "finite field" is one with finitely many elements, like the integers mod any prime number p. This example is called $\mathbb{Z}/p$, or $\mathbb{F}_p$ if you really want to emphasize that you're thinking of it as a field.

So, let's do some 3d special relativity with $\mathbb{Z}/7$, and see what it has to say about the Klein quartic.

First, some basic stuff about finite fields.

The concepts of "positive" and "negative" make sense in any finite field! Say a nonzero element of the field is "positive" if it's of the form x^2 and "negative" otherwise. (Number theorists call the positive elements "quadratic residues", just to intimidate outsiders.)

Then multiplication works nicely:

- if you multiply two positive elements you get a positive one;
- if you multiply two negative elements you get a positive one;
- if you multiply a positive and a negative element you get a negative one.

There are finite fields whose cardinality is any prime power, but if we focus on those whose cardinality is a prime, namely the fields $\mathbb{Z}/p$, there are three possibilities: the good, the bad, and the ugly.

- GOOD: If the field is $\mathbb{Z}/p$ for $p = 4n + 3$ then -1 is negative,
  so we can switch the "sign" of a number by multiplying it by -1.
- BAD: If the field is $\mathbb{Z}/p$ for $p = 4n + 1$ then -1 is positive,
  so we can't do this.

In both these cases there are as many positive as negative elements. Then there's

- UGLY: If the field is $\mathbb{Z}/2$ then every element is positive.

Luckily, $p = 7$ is good. But beware: addition doesn't get along with positivity very well. In fields like $\mathbb{Z}/p$, *every* element is a sum of positive elements.

Next, let's ponder the peculiarities of special relativity over a finite field.

We can define Minkowski spacetime of any dimension over any field $\mathbb{F}$: it's just $\mathbb{F}^{n+1}$ with the quadratic form

$$x^2 = x_0^2 - x_1^2 - \ldots - x_n^2$$

We define $\mathrm{O}(n, 1)$ to be the group of transformations of $\mathbb{F}^{n+1}$ that preserve the above quadratic form, and define the "Lorentz group" $\mathrm{SO}(n, 1)$ to be the subgroup consisting of transformations that also have determinant 1.

As usual, we say that a vector x in Minkowski spacetime is:

- timelike if $x^2 > 0$
- lightlike if $x^2 = 0$

- spacelike if $x^2 < 0$

We define a “ray” to be a line through the origin. We say a ray is timelike, spacelike or lightlike if any vector on it — hence all! — is of that type.

The lightlike rays are usually called “light rays”, both because it sounds cool (we went into physics because we liked things like X-rays and rayguns) and because it's accurate. The light rays going through a given point — the origin — are precisely like this.

Next, let's ponder the peculiarities of 3-dimensional spacetime.

For any field $\mathbb{F}$, 2×2 matrices with determinant 1 act as Lorentz transformations of 3d Minkowski spacetime. I touched upon this idea when discussing Trautman's “Pythagorean spinors” in “Week 196”. Here's how it works:

We can think of 3d Minkowski spacetime as consisting of all 2×2 matrices that are equal to their own transpose:

$$x = \begin{pmatrix} x_0 + x_1 & x_2 \\ x_2 & x_0 - x_1 \end{pmatrix}$$

since the determinant of such a matrix is just

$$x^2 = x_0^2 - x_1^2 - x_2^2$$

In this picture, the group $\mathrm{SL}(2, \mathbb{F})$ consisting of 2×2 matrices with determinant 1 acts as Lorentz transformations:

$$g \colon x \mapsto gxg^*$$

where g^* is the transpose of g. So, we get a homomorphism from $\mathrm{SL}(2, \mathbb{F})$ to the 3d Lorentz group:

$$\mathrm{SL}(2, \mathbb{F}) \to \mathrm{SO}(2, 1)$$

This is two-to-one, since it sends both 1 and -1 to the identity Lorentz transformation. People typically cure this by defining

$$\mathrm{PSL}(2, \mathbb{F}) = \mathrm{SL}(2, \mathbb{F})/\{\pm 1\}$$

We then get a one-to-one homomorphism

$$\mathrm{PSL}(2, \mathbb{F}) \to \mathrm{SO}(2, 1)$$

Alas, this homomorphism is not onto: it's only “half-onto”. In the traditional case where $\mathbb{F}$ is the real numbers, its range is just one of the two connected components of $\mathrm{SO}(2, 1)$. In the case we're interested in here, where $\mathbb{F} = \mathbb{Z}/7$, the group $\mathrm{PSL}(2, \mathbb{Z}/7)$ has 168 elements but $\mathrm{SO}(2, 1)$ has twice as many.

Next, let's bring the hyperbolic plane into the game!

Special relativity in 3 dimensions is closely related to the hyperbolic plane. The reason is that the set of timelike rays in $(n + 1)$-dimensional Minkowski spacetime forms a copy of hyperbolic n-space: physicists call this the “mass hyperboloid”. So, for $n = 2$, we get the hyperbolic plane. This is most familiar for special relativity based on the real numbers, but the same idea applies to other fields.

So, let's make some definitions:

- the hyperbolic plane H_+ is the set of timelike rays in 3d Minkowski spacetime

- the heavenly circle L is the set of light rays in 3d Minkowski spacetime
- the hyperbolic cylinder H_- is the set of spacelike rays in 3d Minkowski spacetime

You may think I'm being silly to call the set of light rays "the heavenly circle", but in 4-dimensional spacetime the analogous thing is often called "the heavenly sphere", and we're studying things one dimension down.

Why "heavenly sphere"? Well, when you look at the stars at night, they seem to be lying on a sphere. That's the heavenly sphere: the set of light rays entering your eye — the set of directions you can look!

One dimension down, flatlanders get to enjoy the "heavenly circle". Mathematicians call this the projective line, since it's a line with one extra point added on.

Now for something fun: points on the hyperbolic plane give lines on the hyperbolic cylinder and vice versa!

This is basically by definition of "line". We define a "line" in H_+ to consist of all points that are orthogonal to a given point in H_-, and vice versa. Note a point in either of these spaces is really a ray in Minkowski spacetime, but it makes sense to say that two rays are orthogonal.

This definition isn't arbitrary: it reduces to a standard notion of "line" in the hyperbolic plane — namely a geodesic — when our field is the real numbers.

Finally, let's dive into the case we're really interested in: 3-dimensional Minkowski spacetime over $\mathbb{F} = \mathbb{Z}/7$.

In this case the positive numbers are $1, 2, 4$, and the negative numbers are $3, 5, 6$.

It turns out that:

- The hyperbolic plane over $\mathbb{Z}/7$, namely H_+, has size 21.
- The heavenly circle over $\mathbb{Z}/7$, namely L, has size 8.
- The hyperbolic cylinder over $\mathbb{Z}/7$, namely H_-, has size 28.

So, H_+ is a nice finite version of the hyperbolic plane with 21 points and 28 lines! A little calculation shows there are 3 points on each line and 4 lines through each point.

We know that $\mathrm{PSL}(2, \mathbb{Z}/7)$ acts on everything in sight here: H_+, H_-, and L. It also acts on the Fano plane and Klein's quartic curve. So, we can try to match up various features of 3d special relativity with features in the Fano plane or Klein's quartic curve!

Greg Egan found the following correspondence:

Hyperbolic plane over $\mathbb{Z}/7$	Fano plane
7 triads	7 points
7 antitriads	7 lines
21 points	21 flags
28 lines	28 apartments

As usual with these correspondences, simple things in the Fano plane correspond to subtle things in the hyperbolic plane, and simple things in the hyperbolic plane correspond to subtle things in the Fano plane.

First of all, points and lines in the Fano plane correspond to "triads" and "antitriads" in H_+.

Huh?

Well, for starters, a "triad" or "antitriad" is an unordered triple of orthogonal points in H_+.

Huh?

Well, remember that a point in H_+ is a timelike ray in 3d Minkowski spacetime. You can't have three timelike rays that are orthogonal in ordinary special relativity, but you can over $\mathbb{Z}/7$, because the sum of postive numbers can be zero, or even negative. For example, these three vectors are timelike and orthogonal:

$$(1,0,0), \quad (0,4,2), \quad (0,-2,4)$$

We call the corresponding triple of rays a "triad", and we get a total of 7 triads by applying elements of $\mathrm{PSL}(2,\mathbb{Z}/7)$ to it. But, there are triples of orthogonal timelike rays that aren't among these 7. For example, we get one from these three vectors:

$$(1,0,0), \quad (0,2,4), \quad (0,-4,2)$$

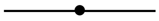
We call the corresponding triple of rays an "antitriad", and we get 7 antitriads by applying elements of $\mathrm{PSL}(2,\mathbb{Z}/7)$.

Each line in the Fano plane contains 3 points, and each point lies on 3 lines. This incidence relation can also be seen in terms of triads and antitriads: each triad has nonempty intersection with 3 antitriads, and each antitriad has nonempty intersection with 3 triads!

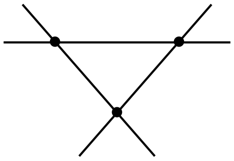
We can also go backwards: points and lines in the hyperbolic plane correspond to "flags" and "apartments" in the Fano plane.

Huh?

Flags and apartments are standard concepts in the theory of "buildings" which I began to explain in "Week 186". But, I don't want or need to explain this general theory here. In the Fano plane, a "flag" consists of a point lying on a line:



An "apartment" consists of 3 distinct points lying on 3 distinct lines like this:



Each apartment in the Fano plane contains 3 flags, and each flag is contained in 4 apartments. This incidence can also been seen in terms of points and lines in the hyperbolic plane: each line contains 3 points, and each point lies on 4 lines!

So, there's an interesting but complicated relation between hyperbolic geometry over $\mathbb{Z}/7$ and the Fano plane. How does the Klein quartic curve fit in? There's more to this side of the story than I've managed to absorb, so I'll just say a few words — probably more than you want to hear. For more detail, try my Klein quartic curve webpage.

There are 48 nonzero lightlike vectors in 3d Minkowksi spacetime, but if you take one of them and apply elements of $\mathrm{PSL}(2, \mathbb{Z}/7)$ to it, you get an orbit consisting of only 24. These 24 guys correspond to the 24 heptagons in the heptagonal tiling of the Klein quartic curve! In other words, $\mathrm{PSL}(2, \mathbb{Z}/7)$ acts in precisely the same way.

You may ask what the point of all this stuff is, and the answer is — I'm not sure yet, except that it's fun! Apparently the coincidence

$$\mathrm{PSL}(2, \mathbb{Z}/7) = \mathrm{PSL}(3, \mathbb{Z}/2)$$

is the only coincidence among classical groups over finite fields, not counting the ones we already know over the real numbers. So, it's got to be good for something! And, I haven't even begun to exploit the fact that the Klein quartic curve is a quotient space of the real hyperbolic plane: this has got to be related to the hyperbolic plane over $\mathbb{Z}/7$. So, I think something interesting should emerge, though I'm not sure what.

Addendum: Regarding the Pioneer anomaly, the cosmologist Ned Wright writes:

> *John,*
>
> *I don't think the obvious possibilities have been ruled out. In this I disagree with Anderson but that's science at the fringes.*
>
> *I have a Web page on this at:*
> *http://www.astro.ucla.edu/~wright/PioneerAA.html*

Just for the record, that's:

17) Ned Wright, "Pioneer anomalous acceleration", http://www.astro.ucla.edu/~wright/PioneerAA.html

For a contrasting viewpoint see:

18) Slava G. Turyshev, Michael Martin Nieto, and John D. Anderson, "The Pioneer anomaly and its implications", available as gr-qc/0510081.

> Space isn't remote at all. It's only an hour's drive away if your car could go straight upwards.
>
> — *Sir Fred Hoyle*

Week 220

August 31, 2005

Work on quantum gravity has seemed stagnant and stuck for the last couple of years, which is why I've been turning more towards pure math.

Over in string theory they're contemplating a vast "landscape" of possible universes, each with their own laws of physics — one or more of which might be ours. Each one is supposed to correspond to a different "vacuum" or "background" for the marvelous unifying M-theory that we don't completely understand yet. They can't choose the right vacuum except by the good old method of fitting the experimental data. But these days, this time-honored method gets a lot less airplay than the "anthropic principle":

1) Leonard Susskind, "The anthropic landscape of string theory", available as `hep-th/0302219`.

Perhaps this is because it's more grandiose to imagine choosing one theory out of a multitude by discovering that it's among the few that supports intelligent life, than by noticing that it correctly predicts experimental results. Or, perhaps it's because nobody really knows how to get string theory to predict experimental results! Even after you chose a vacuum, you'd need to see how supersymmetry gets broken, and this remain quite obscure.

There's still tons of beautiful math coming out of string theory, mind you: right now I'm just talking about physics.

What about loop quantum gravity? This line of research has always been less ambitious than string theory. Instead of finding the correct theory of everything, its goal has merely been to find *any* theory that combines gravity and quantum mechanics in a background-free way. But, it has major problems of its own: nobody knows how it can successfully mimic general relativity at large length scales, as it must to be realistic! Old-fashioned perturbative quantum gravity failed on this score because it wasn't renormalizable. Loop quantum gravity may get around this somehow. . . but it's about time to see exactly how.

Loop quantum gravity follows two main approaches: the so-called "Hamiltonian" or "spin network" approach, which focuses on the geometry of space at a given time, and the so-called "Lagrangian" or "spin foam" approach, which focuses on the geometry of spacetime.

In the last couple of years, the most interesting new work in the Hamiltonian approach has focussed on problems with extra symmetry, like black holes and the big bang. Here's a nontechnical introduction:

2) Abhay Ashtekar, "Gravity and the quantum", available as `gr-qc/0410054`.

and here's some new work that treats the information loss puzzle:

3) Abhay Ashtekar and Martin Bojowald, "Black hole evaporation: a paradigm", *Class. Quant. Grav.* **22** (2005) 3349–3362. Also available as `gr-qc/0504029`.

However, by focusing on solutions with extra symmetry, one puts off facing the hardest aspects of renormalization, or whatever its equivalent might be in loop quantum gravity.

The other approach — the spin foam approach — got stalled when the most popular model seemed to give spacetimes made mostly of squashed-flat "degenerate 4-simplexes". Various papers have found an effect like this: see "Week 198" for more details. So, there's definitely a real phenomenon going on here. However, its physical significance remains a bit obscure. The devil is in the details.

In particular, even though the *amplitude* for a single large 4-simplex in the Barrett-Crane model is dominated by degenerate geometries, certain *second derivatives* of the amplitude might not — and this may be what really matters. Carlo Rovelli has recently come out with a paper on this:

4) Carlo Rovelli, "Graviton propagator from background-independent quantum gravity", available as gr-qc/0508124.

If the idea holds up, I'll be pretty excited. If not, I'll be bummed. But luckily, I've already gone through the withdrawal pains of switching my focus away from quantum gravity. When you do theoretical physics, sometimes you feel the high of discovering hidden truths about the physical universe. Sometimes you feel the agony of suspecting that those "hidden truths" were probably just a bunch of baloney... or, realizing that you may never know. Ultimately nature has the last word.

Math is, at least for me, a less nerve-racking pursuit, since the truths we find can be confirmed simply by discussing them: we don't need to wait for experiment. Math is just as grand as physics, or more so. But it's more wispy and ethereal, since it's about pure pattern in general — not the particular magic patterns that became the world we see. So, the stakes are lower, but the odds are higher.

Speaking of math, I really want to talk about the Streetfest — the conference in honor of Ross Street's 60th birthday. It was a real blast: over sixty talks in two weeks in two cities, Sydney and Canberra. However, I accidentally left my notes from those talks at home before zipping off to Calgary for a summer school on homotopy theory:

5) *Topics in Homotopy Theory*, graduate summer school at the Pacific Institute of Mathematics run by Kristine Bauer and Laura Scull. Recommended reading material available at http://www.pims.math.ca/science/2005/05homotopy/reading.html

So, I'll say a bit about what I learned at this school.

Dan Dugger spoke about motivic homotopy theory, which was *great*, because I've been trying to understand stuff from number theory and algebraic geometry like the Weil conjectures, etale cohomology, motives, and Voevodsky's proof of the Milnor conjecture... and thanks to his wonderfully pedagogical lectures, it's all starting to make some sense!

I hope to talk about this someday, but not now.

Alejandro Adem spoke about orbifolds and group cohomology. Purely personally, the most exciting thing here was seeing that orbifolds can also be seen as certain kinds of topological groupoids, or stacks, or topoi... so that various versions of "categorified topology" are actually different faces of the same thing!

I may talk about this someday, too, but not now.

I spoke about higher gauge theory and its relation to Eilenberg-Mac Lane spaces. I may talk about that too someday, but not now.

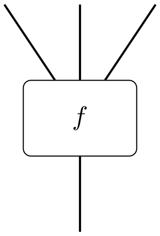
Dev Sinha spoke about operads, and besides explaining the basics, he said a couple of things that really blew me away. So, I want to talk about this now.

For one, the homology of the little k-cubes operad is a graded version of the Poisson operad! For two, the little 2-cubes operad acts on the space of thickened long knots!

But for this to thrill you like it thrills me, I'd better say a word about operads — and especially little k-cubes operads.

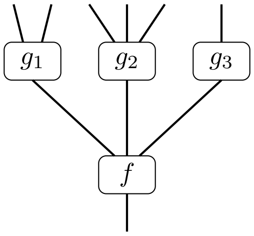
Operads, and especially the little k-cubes operads, were invented by Peter May in the early 1970s to formalize the algebraic structures lurking in "infinite loop spaces". In "Week 149" I explained what infinite loop spaces are, and how they give generalized cohomology theories, but let's not get bogged down in this motivation now, since operads are actually quite simple.

In its simplest form, an operad is a gizmo that has for each $n = 0, 1, 2, \ldots$ a set $\mathcal{O}(n)$ whose elements are thought of as n-ary operations — operations with n inputs. It's good to draw such operations as black boxes with n input wires and one output:



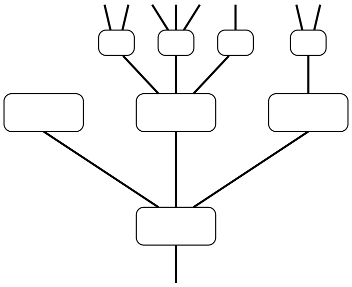
For starters these operations are purely abstract things that don't actually operate on anything. Only when we consider a "representation" or "action" of an operad do they get incarnated as actual n-ary operations on some set. The point of operads is to study their actions.

But, for completeness, let me sketch the definition of an operad. An operad tells us how to compose its operations, like this:



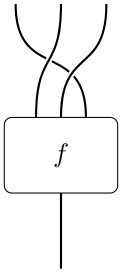
Here we are composing f with g_1, g_2, and g_3 to get an operation with 6 inputs called $f \circ (g_1, g_2, g_3)$.

An operad needs to have a unary operation serving as the identity for composition. It also needs to satisfy an "associative law" that makes a composite of composites like this well-defined:

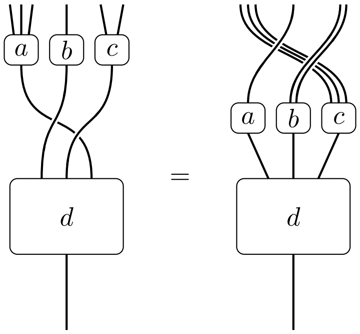


(This picture has a 0-ary operation in it, just to emphasize that this is allowed.)

That's the complete definition of a "planar operad". In a full-fledged operad we can do more: we can permute the inputs of any operation and get a new operation:



This gives actions of the permutation groups on the sets $\mathcal{O}(n)$. We also demand that these actions be compatible with composition, in a way that's supposed to be obvious from the pictures. For example:



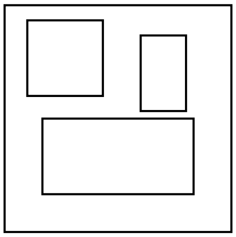
and similarly for permuting the inputs of the black boxes on top.

Voil!

Now, operads make sense in various contexts. So far we've been talking about operads that have a *set* $\mathcal{O}(n)$ of n-ary operations for each n. These have actions on *sets*, where each guy in $\mathcal{O}(n)$ gets incarnated as a *function* that eats n elements of some set and spits out an element of that set.

But historically, Peter May started by inventing operads that have a *topological space* of n-ary operations for each n. These like to act on *topological spaces*, with the operations getting incarnated as *continuous maps*.

Most importantly, he invented an operad called the "little k-cubes operad". Here $\mathcal{O}(n)$ is the space of ways of putting n nonoverlapping little k-dimensional cubes in a big one. We don't demand that the little cubes are actually cubes: they can be rectangular boxes. We do demand that their walls are nicely lined up with the walls of the big cube:



typical 3-ary operation
in the little 2-cubes operad

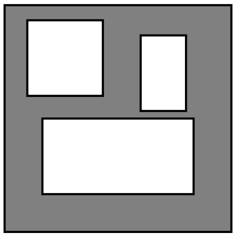
This is an operation in O(3), where $\mathcal{O}$ is the little 2-cubes operad. Or, at least it would be if I labelled each of the 3 little 2-cubes — we need that extra information.

We compose operations by sticking pictures like this into each of the little k-cubes in another picture like this! I should draw you an example, but I'm too lazy. So, figure it out yourself and check the associative law.

The reason this example is so important is that we get an action of the little k-cubes operad whenever we have a "k-fold loop space".

Starting from a space S equipped with a chosen point $*$, the k-fold loop space $\Omega^k(S)$ is the space of all maps from a k-sphere into S that send the north pole to the point $*$. But this is also the space of all maps from a k-cube into S sending the boundary of the k-cube to the point $*$.

So, given n such such maps, we can glom them together using an n-ary operation in the little k-cubes operad:



where we map all the shaded stuff to the point $*$. We get another map from the k-cube to S sending the boundary to $*$. So:

> *ANY k-FOLD LOOP SPACE HAS AN ACTION OF*
> *THE LITTLE k-CUBES OPERAD*

But the really cool part is the converse:

> *ANY CONNECTED POINTED SPACE WITH AN ACTION OF*
> *THE LITTLE k-CUBES OPERAD IS*
> *HOMOTOPY EQUIVALENT TO A k-FOLD LOOP SPACE*

This is too technical to make a good bumper sticker, so if you want people in your neighborhood to get interested in operads, I suggest combining both the above slogans into one:

> *A k-FOLD LOOP SPACE IS THE SAME AS*
> *AN ACTION OF THE LITTLE k-CUBES OPERAD*

Like any good slogan, this leaves out some important fine print, but it gets the basic idea across. Modulo some details, being a k-fold loop space amounts to having a bunch of operations: one for each way of stuffing little k-cubes in a big one!

By the way:

Speaking of bumper stickers, I'm in Montreal now, and there's a funky hangout on the Boulevard Saint-Laurent called Cafe π where people play chess — and they sell T-shirts, key rings, baseball caps and coffee mugs decorated with the Greek letter π! The T-shirts are great if you're going for a kind of math-nerd/punk look; I got one to wow the students in my undergraduate courses. I don't usually provide links to commercial websites, but I made an exception for Acme Klein Bottles, and I'll make an exception for Cafe π:

6) $Cafe\pi$, `http://www.cafepi.ca/`

Unfortunately they don't sell bumper stickers.

But where were we? Ah yes — the little k-cubes operad.

The little k-cubes operad sits in the little $(k+1)$-cubes operad in an obvious way. Indeed, it's a "sub-operad". So, we can take the limit as k goes to ∞ and form the "little ∞-cubes operad". Any infinite loop space gets an action of this... and that's why Peter May invented operads!

You can read more about these ideas in May's book:

7) J. Peter May, *The Geometry of Iterated Loop Spaces*, Lecture Notes in Mathematics **271**, Springer, Berlin, 1972.

or for a more gentle treatment, try this expository article:

8) J. Peter May, "Infinite loop space theory", *Bull. Amer. Math. Soc.* **83** (1977), 456–494.

But Dev Sinha told us about some subsequent work by Fred Cohen, who computed the homology and cohomology of the little k-cubes operad.

For this, we need to think about operads in the world of linear algebra. Here we consider operads that have a *vector space* of n-ary operations for each n, which get incarnated as *multilinear maps* when they act on some *vector space*. These are sometimes called "linear operads".

An example is the operad for Lie algebras. This one is called "Lie". $\mathrm{Lie}(n)$ is the vector space of n-ary operations that one can do whenever one has a Lie algebra. In this example:

- $\mathrm{Lie}(0)$ is zero-dimensional, since there are no nullary operations (constants) built into the definition of Lie algebra, except zero.

- $\mathrm{Lie}(1)$ is one-dimensional, since the only unary operations are multiples of the identity operation:
$$a \to a$$
- $\mathrm{Lie}(2)$ is one-dimensional, since the only binary operations are multiples of the Lie bracket:
$$(a, b) \to [a, b]$$
You might think we need a second guy in $\mathrm{Lie}(2)$, namely
$$(a, b) \to [b, a]$$
but the antisymmetry of the Lie bracket says this is linearly dependent on the first one:
$$[b, a] = -[a, b]$$
- $\mathrm{Lie}(3)$ is two-dimensional, since the only ternary operations are linear combinations of these two:
$$(a, b, c) \to [a, b], c]$$
$$(a, b, c) \to [b, [a, c]$$
You might think we need a third guy in $\mathrm{Lie}(3)$, for example
$$(a, b, c) \to [a, [b, c]$$
but the Jacobi identity says this is linearly dependent on the first two:
$$[a, [b, c] = [a, b], c] + [b, [a, c]$$

You may enjoy trying to show that the dimension of $\mathrm{Lie}(n)$ is $(n-1)!$, at least for $n > 0$. There's an incredibly beautiful conceptual proof, and probably lots of obnoxious brute-force proofs.

There's a lot more to say about the Lie operad, but right now I want to talk about the Poisson operad. A "Poisson algebra" is a commutative associative algebra that has a bracket operation $\{a, b\}$ making it into a Lie algebra, with the property that
$$\{a, bc\} = \{a, b\}c + b\{a, c\}$$
So, bracketing with any element is like taking a derivative: it satisfies the product rule.

For this reason, Poisson algebras arise naturally as algebras of observables in classical mechanics — the Poisson bracket of any observable A with an observable H called the "Hamiltonian" tells you the time derivative of A:
$$\frac{dA}{dt} = \{H, A\}$$
This is the beginning of a nice big story.

But, what's got me excited now is how Poisson algebras show up in topology!

To understand this, we need to note that there's a linear operad whose algebras are Poisson algebras. That's not surprising. But, we can get a very similar operad in a rather shocking way, as follows.

Take the little k-cubes operad. This has a space $\mathcal{O}(n)$ of n-ary operations for each n. Now take the homology of these spaces $\mathcal{O}(n)$, using coefficients in your favorite field, and get vector spaces $H(\mathcal{O}(n))$. By functorial abstract nonsense these form a linear operad. And this is the operad for Poisson algebras!

Alas, we actually have to be a bit more careful. The homology of each space $\mathcal{O}(n)$ with coefficients in some field is really a *graded* vector space over that field. So, taking the homology of the little k-cubes operad gives an operad in the category of graded vector spaces. And, it's the operad whose algebras are graded Poisson algebras with a bracket of degree $k-1$.

What are those? Well, they're like Poisson algebras, but if a is an element of degree $|a|$ and b is an element of degree $|b|$, then:

- ab has degree $|a| + |b|$ (we've got a graded algebra)
- $\{a, b\}$ has degree $|a| + |b| + k - 1$ (with a bracket of degree $k - 1$)

and the usual axioms for a Poisson algebra hold, but sprinkled with minus signs according to the usual yoga of graded vector spaces.

So: whenever we have a k-fold loop space, its homology is a graded Poisson algebra with a bracket of degree $k - 1$.

To get an idea of this works, let me sketch how the product and the bracket work. Suppose we have an space X with an action of the little k-cubes operad:

- The product on homology corresponds to sticking two little cubes side by side. Given two points in X, this gives another point in X. More generally, given two homology classes a and b in X, we get a homology class of degree $|a| + |b|$ in X.
- The bracket comes from taking one little cube and moving it around to trace out a sphere surrounding the other little cube. Given two points in X, this gives a $(k-1)$-sphere in X. More generally, given a homology class a in X, and a homology class b in X, we get a homology class $\{a, b\}$ of degree $|a| + |b| + k - 1$.

The equation

$$\{a, bc\} = \{a, b\}c + b\{a, c\}$$

then says "moving a around b and c is like moving a around b while c stands by, plus moving a around c while b stands by".

I guess this result can be found here:

9) Frederick Cohen, "Homology of $\Omega^{n+1}\Sigma^{n+1}X$ and $C_{n+1}X$, $n > 0$", *Bull. Amer. Math. Soc.* **79** (1973), 1236–1241.

10) Frederick Cohen, Tom Lada and J. Peter May, *The homology of iterated loop spaces*, Lecture Notes in Mathematics **533**, Springer, Berlin, 1976.

But, I don't think these old papers talk about graded Poisson operads! Dev Sinha has a paper where he takes these ideas and distills them all into the combinatorics of graphs and trees:

11) Dev Sinha, "A pairing between graphs and trees", available as math.QA/0502547.

However, what I really like is how he gets these graphs and trees starting from the homology and cohomology (respectively) of the little k-cubes operad! He first wrote about it here:

12) Dev Sinha, "Manifold theoretic compactifications of configuration spaces", available as math.GT/0306385.

    Dev Sinha, "The homology of the little disks operad", available as math/0610236.

I have a vague feeling that this relation between the little k-cubes operad and the Poisson operad is part of a big picture involving braids and quantization. Another hint in this direction is Deligne's Conjecture, now proved in many ways, which says that the operad of singular chains coming from the little 2-disks operad acts on the Hochschild cochain complex of any associative algebra. Since Hochschild cohomology classifies the ways you can deform an associative algebra, this result is related to quantization and Poisson algebras. But, I don't get the big picture! This might help:

13) Maxim Kontsevich, "Operads and motives in deformation quantization", *Lett. Math. Phys.* **48** (1999) 35–72. Also available as math.QA/9904055.

I'd like to ponder this now! But I'm getting tired, and I still need to say how the little 2-cubes operad acts on the space of thickened long knots.

What's a thickened long knot? In k dimensions, it's an embedding of a little k-cube in a big one:

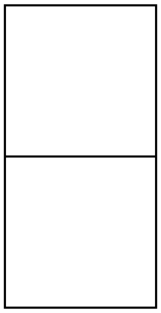
$$f\colon [0,1]^k \to [0,1]^k$$

subject to the condition that the top and bottom of the little cube get mapped to the top and bottom of the big one via the identity map. So, you should imagine a thickened long knot as a fat square rope going from the ceiling to the floor, all tied up in knots.

There are two ways to "compose" thickened long knots.

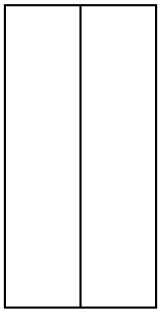
If you're a knot theorist, the obvious way is to stick one on top of the other — just like the usual composition of tangles. But if you just think of thickened long knots as functions, you can also compose them just by composing functions! This amounts to stuffing one knot inside another... a little hard to visualize, but fun.

Anyway, it turns out that the whole little 2-cubes operad acts on the space of thickened long knots, with the two operations I just mentioned corresponding to this:



sticking one thickened long
knot on top of another

and this:



sticking one thickened long
knot inside another

This isn't supposed to make obvious sense, but the point is, there are lots of binary operations interpolating between these two — one for each binary operation in the little 2-cubes operad!

This gives a new proof that the operation of "sticking one thickened long knot on top of another" is commutative up to homotopy.

And, using these ideas, Ryan Budney has managed to figure out a lot of information about the homotopy type of the space of long knots. Check out these papers:

14) Ryan Budney, "Little cubes and long knots", available as math.GT/0309427.

15) Ryan Budney and Frederick Cohen, "On the homology of the space of long knots", available as math.GT/0504206.

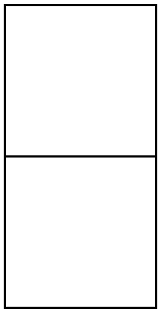
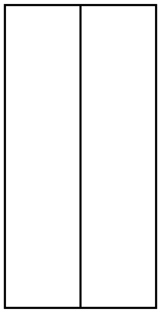
16) Ryan Budney, "Topology of spaces of knots in dimension 3", available as math.GT/0506524.

The paper by Budney and Cohen combines the two ideas I just described — the action of the little 2-cubes operad on thickened long knots and its relation to the Poisson operad. Using these, they show that the rational homology of the space of thickened long knots in 3 dimensions is a free Poisson algebra! They also show that the mod-p homology of this space is a free "restricted Poisson" algebra.

Addendum: Jesse McKeown had a question about the two operations on long thickened knots. Here's what he asked:

> *Perhaps I'm being too imaginative, but I don't feel very convinced the two operations described towards the end of "Week 220" are fundamentally different.*
>
> *VagueSpecifically, in map-composition, can't one stretch all the knottednes of the first composand into an "upper", essentially unknotty portion of the second composand, and similarly squish the knottedness of the second composand into a "lower" section of the big all-encompassing box?*

Here's my reply:

> *Right! That's exactly what having an action of the little* 2*-cubes operad says! There's nothing "fundamentally different" between this:*
>
> 
>
> *A: sticking one thickened long knot on top of another*
>
> *and this:*
>
> 
>
> *B: sticking one thickened long knot inside another*
>
> *because there is a continuous family of operations interpolating between these — one for each way of sticking two little squares in a big one.*
>
> *But, the* process *of moving from operation A to operation B is itself nontrivial. If you loop all the way around from A to B to A — moving the two little squares around each other in the big one — you can get a noncontractible loop in the space of long thickened knots!*
>
> *And, this is what gives the bracket operation on the homology of the space of thickened long knots.*
>
> *Operads were born to deal with issues like this.*

On another note — in the summer of 2007, Urs Schreiber posted a 3-part article on Batalin–Vilkovisky quantization over at the n-Category Cafe. The third part has links to the other two:

18) Urs Schreiber, "Lyakhovich and Sharapov on QFT (On BV-Quantization, Part III)", `http://golem.ph.utexas.edu/category/2007/08/lyakhonov_and_sharapov_on_qft.html`.

and in addition to providing lots of references, it led me back to puzzling about Poisson algebras and the little disks operad. Here's what I wrote, roughly:

> *Try these:*
>
> 19) *Ezra Getzler, "Batalin-Vilkovisky algebras and two-dimensional topological field theories",* Comm. Math. Phys. ***159*** *(1994), 265–285. Available at* `http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.cmp/1104254599`
>
> 20) *Takashi Kimura, Jim Stasheff and Alexander A. Voronov, "On operad structures of moduli spaces and string theory",* Comm. Math. Phys. ***171*** *(1995), 1–25. "Section 3.7: the Batalin-Vilkovisky (BV) algebra". Available at* `http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid` `cmp/1104273401`
>
> *I don't understand this stuff very well!*
>
> *More precisely:*
>
> *If you take the space of multivector fields V on a manifold M, and think of V equipped with its wedge product and Schouten bracket, you get the easiest example of a Gerstenhaber algebra.*
>
> *A Gerstenhaber algebra is an associative supercommutative graded algebra A together with a bracket of degree -1 which makes A into a kind of "graded Poisson algebra with bracket of degree -1". All the usual Poisson algebra axioms hold, but sprinkled with minus signs according to the usual conventions.*
>
> *If your manifold M is a Poisson manifold, then the space V of multivector fields comes equipped with a differential given by taking the Schouten bracket with the Poisson bivector field Π in V.*
>
> *Axiomatizing this mess, we get the definition of a Batalin–Vilkovisky algebra: a Gerstenhaber algebra with differential that's compatible with the other structure in a certain way.*
>
> *There are also lots of Batalin–Vilkovisky algebras that don't come from Poisson manifolds. But just like Poisson manifolds, we can still think of these as describing phase spaces in classical mechanics — in a clever algebraic way. And, that's what BV quantization is all about: figuring out how to treat these Batalin-Vilkovisky algebras as classical phase spaces and quantize them!*
>
> *All this makes some sense to me. But then it gets weird and mystical. . .*
>
> *First, thanks to an old result of Fred Cohen, a Gerstenhaber algebra is the same as an algebra of the operad $H(D)$ — the homology of the little disks operad!*
>
> *Did I just hear some of you say "Huh?"*
>
> *Well, let me sketch what that means. The little disks operad is a gadget with a bunch of n-ary operations corresponding to ways of sticking n little disks in a big one. For each n there's a topological space of these n-ary operations. Taking the homology of this topological space, we get a graded vector space. These are the n-ary operations of the operad I'm calling $H(D)$.*

While I roughly follow how this works, I don't understand the deep inner meaning. It seems amazing: there's a mystical relation between ways of sticking little 2d disks in bigger ones, and operations you can do on the space of multivector fields on a manifold!

I don't know if the connections to 2d topological and conformal field theory (described in the articles I cite) actually explain *this mystical relation, or merely* exploit *it.*

Now, as I said, a Batalin–Vilovisky algebra is a Gerstenhaber algebra with an extra operation. And, Getzler showed that this extra operation corresponds to our ability to twist *a little disk* 360°. *(Until we started twisting like this, we could equally have used little* 2*-cubes.)*

More precisely, Getzler showed that a Batalin–Vilkovisky algebra is the same as an algebra of the homology of the framed *little discs operad.*

This extra twist of the knife only makes me more curious to know what's really going on here.

Here's a clue that could help. As I explained to Urs a couple days ago, this business of "taking homology" is really some sort of procedure for turning weak ∞*-groupoids (i.e. spaces) into stable strict* ∞*-groupoids (i.e. chain complexes) — followed by taking the homology of the chain complex, which in principle loses even* more *information, but doesn't in this particular example. That suggests that these Gerstenhaber (and Batalin–Vilkovisky) algebras are really just watered-down chain complex versions of* spaces *equipped with* n*-ary operations corresponding to ways of sticking* n *(framed) little disks into a big disk.*

But still: what's really going on? What do classical phase spaces have to do with little 2-dimensional disks???

As far as I'm concerned, the Rosetta Stone on the third page of Getzler's paper only serves to heighten the mystery further!

One gets the impression that some physicists have gone for so long without any experimental data that might resolve the quantum-gravity debates that they are going a little crazy.

— *Jaron Lanier*

Week 221

September 18, 2005

After going to the Streetfest this summer, I wandered around China. I began by going to a big conference in Beijing, the 22nd International Congress on the History of Science. I learned some interesting stuff. For example:

You may have heard of Andalusia, that fascinating melting-pot of cultures that formed when southern Spain was invaded by Muslims. The eleventh century was the golden age of Andalusian astronomy and mathematics, with a lot of innovation in astrolabes. During the Caliphate (929–1031), three quarters of all mathematical manuscripts were produced in Cordoba, most of the rest in Sevilla, and only a few in Granada and Toledo.

I didn't understand the mathematical predominance of Cordoba when I first heard about it, but the underlying reason is simple. The first great Muslim dynasty were the Ummayyads, who ruled from Damascus. They were massacred by the Abbasids in 750, who then moved the capital to Baghdad. When Abd ar-Rahman fled Damascus in 750 as the only Ummayyad survivor of this massacre, he went to Spain, which had already been invaded by Muslim Berbers in 711.

Abd ar-Rahman made Cordoba his capital. And, by enforcing a certain level of religious tolerance, he made this city into *the place to be* for Muslims, Jews and Christians — the "ornament of the world", and a beacon of learning — until it was sacked by Berber troops in 1009.

Other cities in Andalusia became important later. The great philosopher Ibn Rushd — known to Westerners by the Latin name "Averroes" — was born in Cordoba in 1128. He later became a judge there. He studied mathematics, medicine, and astronomy, and wrote detailed line-by-line commentaries on the works of Aristotle. It was through these commentaries that most of Aristotle's works, including his Physics, found their way into Western Europe! By 1177, the bishop of Paris had banned the teaching of many of these new ideas — but to little effect.

Toledo seems to have only gained real prominence after Alfonso VI made it his capital upon capturing it in 1085 as part of the Christian "reconquista". By the 1200s, it became a lively center for translating Arabic and Hebrew texts into Latin.

Mathematics also passed from the Arabs to Western Europe in other ways. Fibonacci (1170-1250) studied Arabic accounting methods in North Africa where his father was a diplomat. His book *Liber Abaci* was important in transmitting the Indian system of numerals (including zero) from the Arabs to Europe. However, he wasn't the first to bring these numbers to Europe. They'd been around for over 200 years!

For example: Gerbert d'Aurillac (940–1003) spent years studying mathematics in various Andalusian cities including Cordoba. On his return to France, he wrote a book about a cumbersome sort of "abacus" labelled by a Western form of the Indian numerals — close to what we now call "Arabic numerals". This remained popular in intellectual circles until the mid-12th century.

Amusingly, Arabic numerals were also called "dust numerals" since they were used in calculations on an easily erasable "dust board". Their use was described in the *Liber Pulveris*, or "book of dust".

I want to learn more about Andalusian science! I found this book a great place to start — it's really fascinating:

1) Maria Rose Menocal, *The Ornament of the World: How Muslims, Jews and Christians Created a Culture of Tolerance in Medieval Spain*, Little, Brown and Co., 2002.

For something quick and pretty, try this:

2) Steve Edwards, "Tilings from the Alhambra", `http://www2.spsu.edu/math/tile/grammar/moor.htm`

Apparently 13 of the 17 planar symmetry groups can be found in tile patterns in the Alhambra, a Moorish palace built in Granada in the 1300s.

To dig deeper into the splendors of Arabic mathematics, try these:

3) John J. O'Connor and Edmund F. Robertson, "Arabic mathematics: forgotten brilliance?", `http://www-groups.dcs.st-and.ac.uk/~history/HistTopics/Arabic_mathematics.html`

   John J. O'Connor and Edmund F. Robertson, "Biographies of Arab/Islamic mathematicians", `http://www-groups.dcs.st-and.ac.uk/~history/Indexes/Arabs.html`

For more on Fibonacci and Arabic mathematics, try this paper by Charles Burnett, who spoke about the history of "Arabic numerals" in Beijing:

4) Charles Burnett, "Leonard of Pisa and Arabic Arithmetic", `http://muslimheritage.com/topics/default.cfm?ArticleID=472`

Another interesting talk in Beijing was about the role of the Syriac language in the transmission of Greek science to Europe. Many important texts didn't get translated directly from Greek to Arabic! Instead, they were first translated into *Syriac*.

I don't understand the details yet, but luckily there's a great book on the subject, available free online:

5) De Lacy O'Leary, *How Greek Science Passed to the Arabs*, Routledge & Kegan Paul Ltd, 1949. Also available at `http://www.aina.org/books/hgsptta.htm`

So, medieval Europe learned a lot of Greek science by reading Latin translations of Arab translations of Syriac translations of second-hand copies of the original Greek texts!

George Baloglu recommends this book:

6) Dimitri Gutas, *Greek Thought, Arabic Culture: The Graeco–Arabic Translation Movement in Baghdad and Early 'Abbasid Society (2nd–4th/8th–10th Centuries)*, Routledge, 1998.

I want to read this book, too:

7) Scott L. Montgomery, *Science in Translation: Movements of Knowledge through Cultures and Time*, U. of Chicago Press, 2000. Review by William R. Everdell available at MAA Online, `http://www.maa.org/publications/maa-reviews/science-in-translation-movements-`

The historian of science John Stachel, famous for his studies of Einstein, says this book "strikes a blow at one of the founding myths of 'Western Civilization' " — namely, that Renaissance Europeans single-handedly picked up doing science where the Greeks left off. As Everdell writes in his review:

> *Perhaps the best of the book's many delightful challenges to conventional wisdom comes in the first section on the translations of Greek science. Here we learn why it is ridiculous to use a phrase like "the Renaissance recovery of the Greek classics"; that in fact the Renaissance recovered very little from the original Greek and that it was long before the Renaissance that Aristotle and Ptolemy, to name the two most important examples, were finally translated into Latin. What the Renaissance did was to create a myth by eliminating all the intermediate steps in the transmission. To assume that Greek was translated into Arabic "still essentially erases centuries of history" (p. 93). What was translated into Arabic was usually Syriac, and the translators were neither Arabs (as the great Muslim historian Ibn Khaldun admitted) nor Muslims. The real story involves Sanskrit compilers of ancient Babylonian astronomy, Nestorian Christian Syriac-speaking scholars of Greek in the Persian city of Jundishapur, and Arabic- and Pahlavi-speaking Muslim scholars of Syriac, including the Nestorian Hunayn Ibn Ishak (809–873) of Baghdad, "the greatest of all translators during this era" (p. 98).*

And now for something completely different: the Langlands program! I want to keep going on my gradual quest to understand and explain this profoundly difficult hunk of mathematics, which connects number theory to representations of algebraic groups. I've found this introduction to be really helpful:

8) Stephen Gelbart: "An elementary introduction to the Langlands program", *Bulletin of the AMS* **10** (1984), 177–219.

There are a lot of more detailed sources of information on the Langlands program, but the problem for the beginner (me) is that the overall goal gets swamped in a mass of technicalities. Gelbart's introduction does the best at avoiding this problem.

I've also found parts of this article to be helpful:

9) Edward Frenkel, "Recent advances in the Langlands program", available at `math.AG/0303074`.

It focuses on the "geometric Langlands program", which I'd rather not talk about now. But, it starts with a pretty clear introduction to the basic Langlands stuff... at least, clear to me after I've battered my head on this for about a year!

If you know some number theory or you followed recent issues This Week's Finds (especially "Week 217" and "Week 218") it should make sense, so I'll quote it:

> *The Langlands Program has emerged in the late 60's in the form of a series of far-reaching conjectures tying together seemingly unrelated objects in number theory, algebraic geometry, and the theory of automorphic forms. To motivate it, recall the classical Kronecker-Weber theorem which describes the maximal*

abelian extension $\mathbb{Q}^{\mathrm{ab}}$ *of the field* $\mathbb{Q}$ *of rational numbers (i.e., the maximal extension of* $\mathbb{Q}$ *whose Galois group is abelian). This theorem states that* $\mathbb{Q}^{\mathrm{ab}}$ *is obtained by adjoining to* $\mathbb{Q}$ *all roots of unity; in other words,* $\mathbb{Q}^{\mathrm{ab}}$ *is the union of all cyclotomic fields* $\mathbb{Q}(1^{\frac{1}{N}})$ *obtained by adjoining to* $\mathbb{Q}$ *a primitive* N*th root of unity*

$$1^{\frac{1}{N}}$$

The Galois group $\mathrm{Gal}(Q(1^{\frac{1}{N}})/\mathbb{Q})$ *of automorphisms of* $\mathbb{Q}(1^{\frac{1}{N}})$ *preserving* $\mathbb{Q}$ *is isomorphic to the group* $(\mathbb{Z}/N)^*$ *of units of the ring* $\mathbb{Z}/N$*. Indeed, each element* m *in* $(\mathbb{Z}/N)^*$*, viewed as an integer relatively prime to* N*, gives rise to an automorphism of* $\mathbb{Q}(1^{\frac{1}{N}})$ *which sends*

$$1^{\frac{1}{N}}$$

to

$$1^{\frac{m}{N}}.$$

Therefore we obtain that the Galois group $\mathrm{Gal}(\mathbb{Q}^{\mathrm{ab}}/\mathbb{Q})$*, or, equivalently, the maximal abelian quotient of* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$*, where* $\overline{\mathbb{Q}}$ *is an algebraic closure of* $\mathbb{Q}$*, is isomorphic to the projective limit of the groups* $(\mathbb{Z}/N)^*$ *with respect to the system of surjections*

$$(\mathbb{Z}/N)^* \to (\mathbb{Z}/M)^*$$

for M *dividing* N*. This projective limit is nothing but the direct product of the multiplicative groups of the rings of* p*-adic integers,* $\mathbb{Z}_p^*$*, where* p *runs over the set of all primes. Thus, we obtain that*

$$\mathrm{Gal}(\mathbb{Q}^{\mathrm{ab}}/\mathbb{Q}) = \prod \mathbb{Z}_p^*.$$

The abelian class field theory gives a similar description for the maximal abelian quotient $\mathrm{Gal}(\mathbb{F}^{\mathrm{ab}}/\mathbb{F})$ *of the Galois group* $\mathrm{Gal}(\overline{\mathbb{F}}/\mathbb{F})$*, where* $\mathbb{F}$ *is an arbitrary global field, i.e., a finite extension of* $\mathbb{Q}$ *(number field), or the field of rational functions on a smooth projective curve defined over a finite field (function field). Namely,* $\mathrm{Gal}(\mathbb{F}^{\mathrm{ab}}/\mathbb{F})$ *is almost isomorphic to the quotient* $A(\mathbb{F})^*/\mathbb{F}^*$*, where* $A(\mathbb{F})$ *is the ring of adeles of* $\mathbb{F}$*, a subring in the direct product of all completions of* $\mathbb{F}$*. Here we use the word "almost" because we need to take the group of components of this quotient if* $\mathbb{F}$ *is a number field, or its profinite completion if* $\mathbb{F}$ *is a function field.*

When $\mathbb{F} = \mathbb{Q}$ *the ring* $A(\mathbb{Q})$ *is a subring of the direct product of the fields* $\mathbb{Q}_p$ *of* p*-adic numbers and the field* $\mathbb{R}$ *of real numbers, and the quotient* $A(\mathbb{F})^*/\mathbb{F}^*$ *is isomorphic to*

$$\mathbb{R}^+ \times \prod_p \mathbb{Z}_p^*.$$

where $\mathbb{R}^+$ *is the multiplicative group of positive real numbers. Hence the group of its components is*

$$\prod_p \mathbb{Z}_p^*$$

in agreement with the Kronecker-Weber theorem.

One can obtain complete information about the maximal abelian quotient of a group by considering its one-dimensional representations. The above statement of the abelian class field theory may then be reformulated as saying that one-dimensional representations of $\mathrm{Gal}(\overline{\mathbb{F}}/\mathbb{F})$ *are essentially in bijection with one-dimensional representations of the abelian group*

$$A(\mathbb{F})^* = \mathrm{GL}(1, A(\mathbb{F}))$$

which occur in the space of functions on

$$A(\mathbb{F})^*/\mathbb{F}^* = \mathrm{GL}(1, A(\mathbb{F}))/\mathrm{GL}(1, \mathbb{F})$$

A marvelous insight of Robert Langlands was to conjecture that there exists a similar description of n-dimensional representations *of* $\mathrm{Gal}(\overline{\mathbb{F}}/\mathbb{F})$. *Namely, he proposed that those may be related to irreducible representations of the group* $\mathrm{GL}(n, A(\mathbb{F}))$ *which are* automorphic*, that is those occurring in the space of functions on the quotient*

$$\mathrm{GL}(n, A(\mathbb{F}))/\mathrm{GL}(n, \mathbb{F})$$

This relation is now called the Langlands correspondence.

At this point one might ask a legitimate question: why is it important to know what the n*-dimensional representations of the Galois group look like, and why is it useful to relate them to things like automorphic representations? There are indeed many reasons for that. First of all, it should be remarked that according to the Tannakian philosophy, one can reconstruct a group from the category of its finite-dimensional representations, equipped with the structure of the tensor product. Therefore looking at* n*-dimensional representations of the Galois group is a natural step towards understanding its structure. But even more importantly, one finds many interesting representations of Galois groups in "nature".*

For example, the group $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ *will act on the geometric invariants (such as the etale cohomologies) of an algebraic variety defined over* $\mathbb{Q}$*. Thus, if we take an elliptic curve* E *over* $\mathbb{Q}$*, then we will obtain a two-dimensional Galois representation on its first etale cohomology. This representation contains a lot of important information about the curve* E*, such as the number of points of* E *over* $\mathbb{Z}/p$ *for various primes* p.

The point is that the Langlands correspondence is supposed to relate n*-dimensional Galois representations to automorphic representations of* $\mathrm{GL}(n, A(\mathbb{F}))$ *in such a way that the data on the Galois side, such as the number of points of* E *over* $\mathbb{Z}/p$*, are translated into something more tractable on the automorphic side, such as the coefficients in the* q*-expansion of the modular forms that encapsulate automorphic representations of* $\mathrm{GL}(2, A(\mathbb{Q}))$.

More precisely, one asks that under the Langlands correspondence certain natural invariants attached to the Galois representations and to the automorphic representations be matched. These invariants are the Frobenius conjugacy classes *on the Galois side and the* Hecke eigenvalues *on the automorphic side.*

Since I haven't talked about Hecke operators yet, I'll stop here!

But, someday I should really explain the ideas behind the baby "abelian" case of the Langlands philosophy in simpler terms than Frenkel does here. The abelian case goes back way before Langlands: it's called "class field theory". And, it's all about exploiting this analogy, which I last mentioned in "Week 218":

Number theory	Complex geometry
Integers	Polynomial functions on the complex plane
Rational numbers	Rational functions on the complex plane
Prime numbers	Points in the complex plane
Integers $\mod p^n$	$(n-1)$st-order Taylor series
p-adic integers	Taylor series
p-adic numbers	Laurent series
Adeles for the rationals	Adeles for the rational functions
Fields	One-point spaces
Homomorphisms to fields	Maps from one-point spaces
Algebraic number fields	Branched covering spaces of the complex plane

Addendum: I thank Fabien Besnard for some suggestions on how to improve this Week's Finds. Bruce Smith, Noam Elkies, and Miguel Carrin-lvarez had some things to say about the history of science. In response to this comment of mine:

> *So, medieval Europe learned a lot of Greek science by reading Latin translations of Arab translations of Syriac translations of second-hand copies of the original Greek texts!*

my friend Bruce wrote:

> *This all seems so precarious a process that it makes me wonder whether there was ten times as much valuable ancient math and philosophy as we know about, most of which got* completely *lost.*

Something like this almost certainly true.

Like Plato, Aristotle is believed to have written dialogs which presented his ideas in a polished form. They were all lost. His extant writings are just "lecture notesquot; for courses he taught!

Euripides wrote at least 75 plays, of which only 19 survive in their full form. We have fragments or excerpts of some more. This isn't philosophy or math, but it's still incredibly tragic (pardon the pun).

The mathematician Apollonius wrote a book on *Tangencies* which is lost. Only four of his eight books on *Conics* survive in Greek. Luckily, the first seven survive in Arabic.

The burning of the library of Alexandria is partially to blame for these losses.

There's some good news, though:

Archimedes did more work on calculus than previously believed! We know this now because a manuscript of his on mechanics that had been erased and written over has recently been read with the help of a synchrotron X-ray beam! This is a great example of modern science helping the history of science.

This manuscript, called the Archimedes Palimpsest, also reveals for the first time that he did work on combinatorics:

10) Nova, "The Archimedes Palimpsest", `http://www.pbs.org/wgbh/nova/archimedes/palimpsest.html`

11) Heather Rock Woods, "Placed under X-ray gaze, Archimedes manuscript yields secrets lost to time", *Stanford Report*, May 19, 2005, `http://news-service.stanford.edu/news/2005/may25/archimedes-052505.html`

12) Erica Klarreich, "Glimpses of genius: mathematicians and historians piece together a puzzle that Archimedes pondered", *Science News* **165** (2004), 314. Also available at `http://www.sciencenews.org/articles/20040515/bob9.asp`

Also: a team using "multispectral imaging" has recently been able to read parts of a Roman library in the town of Herculaneum. The books in this library were "roasted in place" — heavily carbonized — during the eruption of Vesuvius that destroyed Pompeii. By distinguishing between different shades of black, researchers were able to reconstruct the entire book *On Piety* by one Philodemus:

13) Julie Walker, "A library of mud and ashes", *BYU Magazine*, Spring 2001, `http://magazine.byu.edu/?act=view&a=43`

I can't resist quoting a bit:

> *A sister city to Pompeii that was also buried in the volcanic eruption of A.D. 79, Herculaneum was a seaside town that sat between Vesuvius' fertile foot and the gleaming Bay of Naples. The collection of 2,000 carbonized Greek and Latin scrolls, primarily Epicurean philosophical writings, was found in a luxurious Herculaneum house known as the Villa of the Papyri, which was discovered in 1752.*
>
> *The scrolls have endured a destructive path through history: first, rain soaked the papyri, then a 570-degree swell of molasses-thick mud engulfed the villa and charred the scrolls. They would remain buried under 65 feet of mud for hundreds of years.*
>
> *As a result, many of the fragile scroll cylinders are pressed into trapezoidal columns; some are bowed and snaked into half-moons, others folded into v-shapes.*
>
> *After their discovery the mortality rate for the scrolls continued to climb as would-be conservators struggled to find a way to unroll the fragile manuscripts. Some scrolls were turned to mush when they were painted with mercury; many were sliced down the middle and cut into fragments. Early transcribers would copy the visible outer layer of a scroll, then scrape it off and discard it to read the next layer.*
>
> *Even today, scholars use metaphors of near impossibility to describe the scroll unrolling process. It is like "flattening out a potato chip" without destroying it,*

> *or like "separating (burned) layers of two-ply tissue," says Jeffrey Fish of Baylor University.*
>
> *The current unrolling methoddeveloped by a team of Norwegian conservators involves applying a gelatin-based adhesive to the scroll's outer surface. As the adhesive dries, the outer shell — which bears the text on its interior — can be slowly peeled off. It can take days to remove a single fragment, months or years to process a complete scroll. Some 300 of the library's scrolls have yet to be unrolled, and many more scrolls are in various stages of conservation and repair.*
>
> *On the Herculaneum project, CPART researchers Steve and Susan Booras conducted multispectral imaging (MSI) on 3,100 trays of papyrus fragments and photographed them with a high-quality digital camera. The images will be used to create a digital library that can be accessed by scholars worldwide. Developed for NASA scientists, the imaging technique has only recently been applied to the study of ancient texts. Rather than focusing on light that is seen at wave lengths visible to the eye, MSI uses filters to focus on nonvisible portions of the light spectrum. In the nonvisible infrared spectrum, the black ink on a blackened scroll can be clearly differentiated. In some cases clear, legible writings have been found on fragments that researchers believed were completely blank.*

The same team is now studying over 400,000 fragments of papyrus found in an ancient garbage dump in the old Egyptian town of Oxyrhynchus. They've pieced together new fragments of plays by Euripides, Sophocles and Menander, lost lines from the poets Sappho, Hesiod, and Archilocus, and most of a book by Hesiod:

14) Oxyrhynchus Online, "multispectral imaging", `http://www.papyrology.ox.ac.uk/multi/procedure.html`

If you just want to look at a nice "before and after" movie of what multispectral imaging can do, try this link.

Finally, in response to this remark of mine:

> *Amusingly, Arabic numerals were also called "dust numerals" since they were used in calculations on an easily erasable "dust board". Their use was described in the Liber Pulveris, or "book of dust".*

Noam Elkies wrote:

> *This is even more amusing than you may realize: the word "abacus" comes from a Greek word "abax, abak-" for "counting board", which conjecturally might come from the Hebrew word (or a cognate word in another semitic language) for "dust"! See for instance:*
>
> *`http://education.yahoo.com/reference/dictionary/entry/abacus`*
>
> *So these "dust numerals" replaced a reckoning device whose name may also originate with calculation a dust board. . .*

Interesting! While "calculus" refers back to pebbles.

My erstwhile student Miguel Carrin-lvarez clarified the issue somewhat:

The first abaci were drawn in the sand with sticks. The next step was to carve grooves in a board (wooden, or clay: think cuneiform tablets) and place beads in them. Pierced beads moving on beams (wood, later metal) must have been a pretty recent development, relatively speaking.

Remember that Archimedes was studying geometry by drawing figures in the sand when he was slain. If a sand abacus is the precursor of the modern calculator, Archimedes' sandbox is the precursor of GUI geometry software.

One of Archimedes' most fanciful works is "The Sand Reckoner". Here the reckoner can be understood to be himself, as he is counting the grains of sand which fit inside the sphere of fixed stars, but it can also refer to a sand abacus (reckoner = calculator). In fact, romance translations of this title that I've seen (French: L'arenaire, Spanish: El arenario, etc.) unambiguously refer to an object, not a person. It is easy to imagine Archimedes inventing his positional number system on a sand abacus, and using the counting of grains of sand as an excuse to write about it.

We avail ourselves of what our predecessors may have said. That they were or were not our coreligionists is of no account.... Whatever accords with the truth, we shall happily and gratefully accept, and whatever conflicts, we shall scrupulously but generously point out.

— *Averroes*

Week 222

October 17, 2005

Last week there was a big conference on quantum gravity at the Albert Einstein Institute near Berlin:

1) Loops '05, `http://loops05.aei.mpg.de`

The focus was loop quantum gravity and spin foams, but there were also talks about other approaches, so it was much bigger than last year's get-together in Marseille. Last year about 100 people attended; this time about 160 did! It was strange seeing old pals like Ashtekar, Lewandowski, Loll, Rovelli and Smolin almost lost in a sea of new faces. But, it was great to talk to everyone, both old and new.

I'll say more about this conference, but first let's talk about γ ray bursters, a black hole without a host galaxy, the newly discovered moon of planet Xena, and lots of other transneptunian objects.

Actually, just for fun, let's start with this science fiction novel I picked up in Heathrow en route to Berlin:

2) Charles Stross, *Accelerando*, Ace Books, New York. Also available at `http://www.accelerando.org/book/`

This is one of the few tales I've read that does a good job of fleshing out Verner Vinge's "Singularity" scenario, where the accelerating development of technology soars past human comprehension and undergoes a phase transition to a thoroughly different world. This is a real possibility, and it's been discussed a lot:

3) Wikipedia, "Technological singularity", `http://en.wikipedia.org/wiki/Technological_singularity`

   Ray Kurzweil, "The Singularity", `http://www.kurzweilai.net/meme/frame.html?m=1`

   Anders Sandberg, "The Singularity", `http://www.aleph.se/Trans/Global/Singularity/`

However, it's not an easy subject for fiction — at least not for mere human readers! Stross makes it gripping: sometimes goofy, sometimes thrilling, and sometimes rather sad. Characters include a robot cat with ever-growing powers and some space-faring uploaded lobsters.

The hero, Manfred Macx, starts out as a freeware developer, futurist and all-purpose wheeler-dealer. Here's a scene from the beginning of the book, before all hell breaks loose:

> *Manfred's mood of dynamic optimism is gone, broken by the knowledge that his vivisectionist stalker has followed him to Amsterdam to say nothing of Pamela, his dominatrix, source of so much yearning and so many morning-after weals. He slips his glasses on, takes the universe off hold, and tells it to take him for a long walk while he catches up on the latest on the tensor-mode gravitational waves in the cosmic background radiation (which, it is theorized, may be waste*

> *heat generated by irreversible computational processes back during the inflationary epoch; the present-day universe being merely the data left behind by a really huge calculation). And then there's the weirdness beyond M31: according to the more conservative cosmologists, an alien superpower — maybe a collective of Kardashev Type Three galaxy-spanning civilizations — is running a timing channel attack on the computational ultrastructure of space-time itself, trying to break through to whatever's underneath. The tofu-Alzheimer's link can wait.*

An idea a minute — and the book is free online: what more could you want?

But right now, the big news in astronomy is *not* about a type III civilization lurking beyond M31 (otherwise known as the Andromeda Galaxy). It's some evidence that short γ ray bursts are caused by collisions involving neutron stars and black holes!

γ ray bursts are among the most energetic events known in the heavens. They happen in galaxies throughout the universe; we see about one a day, and each releases somewhere between 10^{45} and 10^{47} joules of energy. The larger figure is what you'd get by turning the entire mass of the Sun into energy.

There could be several kinds of γ ray bursts, but there seem to be at least two: short and long. Short bursts last between 40 milliseconds and 10 seconds — imagine the whole Sun turning into energy that fast! Long ones last between 10 and 100 seconds. The two kinds seem to be qualitatively different: for example, the short ones consist of higher-frequency γ rays. The big news is that they happen in different kinds of galaxies!

In "Week 204", I described how people caught a long γ ray burst in the act in March 2003. A γ ray detector aboard a satellite relayed information to telescopes in Australia and Japan, allowing them to spot a visible afterglow right after the burst. The details of this glow fit the "hypernova" theory of long γ ray bursts.

The hypernova theory says that when a star more than 25 times heavier than the Sun runs out of fuel and collapses, it forms a black hole that sucks down the star's iron core before a normal supernova explosion can occur. In just a few seconds, about a solar mass of iron spirals into the black hole, forming a pancake-shaped disk as it goes down. In the process, this disk becomes incredibly hot and shoots out jets of radiation in the transverse directions. As they plow through the star's outer layers, these jets create beams of γ rays.

The short bursts have been harder to catch. By the time a telescope on Earth could be aimed at the spot where the γ rays were seen, no afterglow could be seen!

So, in October 2004 NASA launched Swift: a γ-ray detecting satellite equipped with an X-ray telescope and an ultraviolet/optical telescope that can respond quickly whenever a burst is seen:

4) Official NASA Swift homepage, `http://swift.gsfc.nasa.gov/docs/swift/swiftsc.html`

5) Gamma-ray burst real-time sky map, `http://grb.sonoma.edu/`

On May 9th, 2005, Swift detected a short burst and caught 11 photons of the burst's X-ray afterglow. Another short burst detected by HETE-II had its X-ray afterglow caught by the Chandra X-ray satellite. Analysis of these and two more short bursts has convinced some scientists that they're caused by collisions between neutron stars and/or black holes:

6) D. B. Fox et al, "The afterglow of GRB050709 and the nature of the short-hard γ-ray bursts", *Nature* **437** (October 2005), 849–850. Also available at `http://www.nasa.gov/pdf/135397main_nature_fox_final.pdf`

Despite what the news media are saying, I don't see that this paper "proves" the short γ-ray bursts are caused by such collisions. Instead, I see some good pieces of evidence.

The faintness of the afterglows suggests some mechanism other than a hypernova. But as far as I can tell, the best evidence is that short γ ray bursts tend to happen near the edges of old galaxies, while the long ones happen near the centers of young galaxies.

The center of a young galaxy is where you'd expect to find a really huge Wolf-Rayet star, the sort that dies in a hypernova. The edge of an old galaxy is where you'd expect to see black holes and neutron stars collide. Why? Because such collisions can only happen long after stars are first formed. First you need an orbiting pair of giant stars to go supernova and collapse into neutron stars and/or black holes. Then you need plenty more time for this pair to spiral down thanks to gravitational radiation, and eventually collide. By then the pair may sail off to the edge of the galaxy, thanks to the "kick" delivered by the supernova explosions.

I hope astronomers can clinch the case for the collision theory of short γ ray bursts. After all, these collisions involving neutron stars and black holes are precisely what gravitational wave detectors like LIGO and VIRGO are hoping to see! If we know to look for gravitational waves precisely when we see short γ ray bursts, and we know where they're coming from, we'll have a better chance of finding them.

(Of course, we'll also have a better chance of *fooling* ourselves into *thinking* we found them, until we do some double-blind tests.)

By the way, LIGO is already analysing data to look for gravitational waves. I talked about this in "Week 189", but here's something new: now you can help them by running a cool screensaver called Einstein@Home on your computer! Check it out:

7) Einstein@Home, `http://einstein.phys.uwm.edu/`

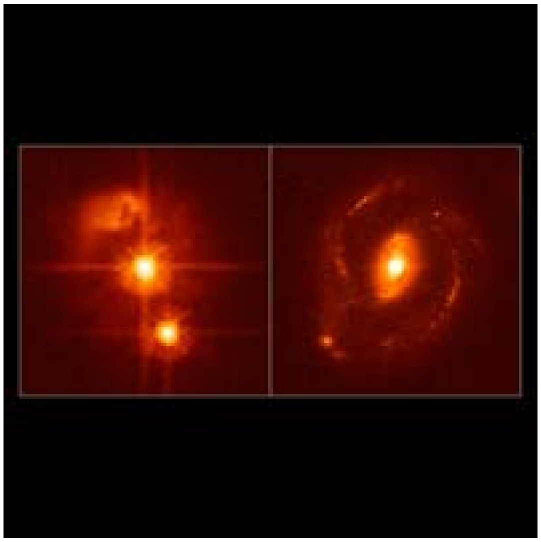
Speaking of black holes, last month the Hubble Space Telescope and the Very Large Telescope in Chile detected a quasar that seems to have no host galaxy:

8) European Southern Observatory, "Black hole in search of a home", `http://www.eso.org/outreach/press-rel/pr-2005/pr-23-05.html`

HubbleSite, "Quasar without host galaxy compared with normal quasar", `http://hubblesite.org/newscenter/newsdesk/archive/releases/2005/13/image/a`

Quasars are thought to be super massive black holes; they're usually found in the centers of galaxies, where they devour stars and shoot out enormously powerful jets of radiation. However, the quasar HE0450-2958 is surrounded only by a blob of ionized gas. Nearby, a wildly disturbed spiral galaxy can be seen.

Compare HE0450-2958 (at left) with a normal quasar (at right):

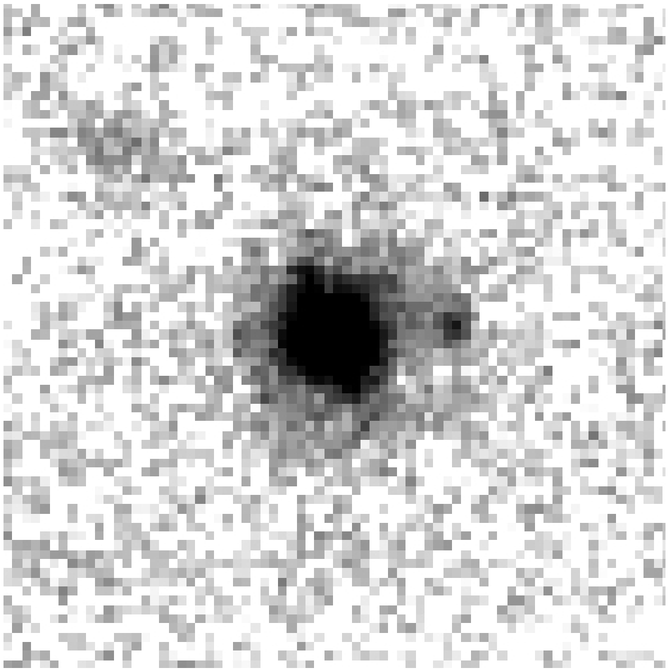


The quasar HE0450-2958 is in the middle of the left-hand picture; the disturbed galaxy is above and a completely irrelevant foreground star is below. For more details on what this image means, click on it.

Did this quasar begin life in the middle of a galaxy and then get kicked out when that galaxy collided with something containing a super-massive black hole? What could that something be?

Puzzles, puzzles, in the sky. . . .

Closer to home, astronomers at the Keck Observatory in Hawaii have discovered that planet Xena has a moon!



They nicknamed it Gabrielle, after this famous TV character's sidekick:

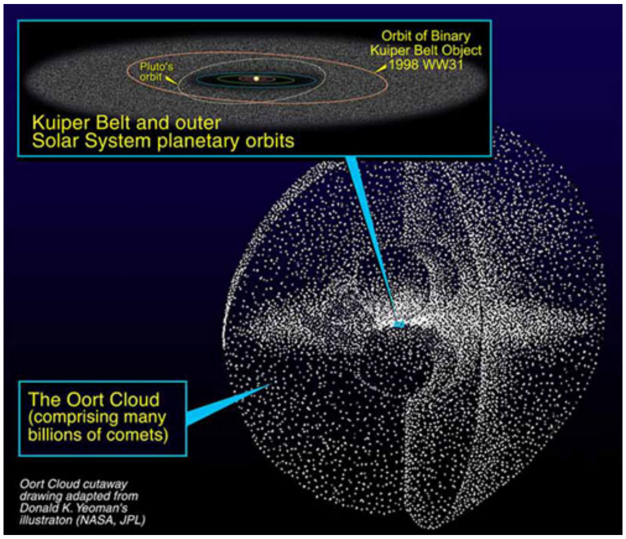
9) Michael E. Brown, "The moon of the 10th planet", `http://www.gps.caltech.edu/~mbrown/planetlila/moon/index.html`

If you hadn't heard about planet Xena, or you don't like the idea of naming a planet after a TV character — even a "warrior princess" - don't get worked up just yet. Xena's official name is currently 2003 UB$_{313}$, and though she's larger than Pluto, the International Astronomical Union has not decided whether she'll officially be considered a planet.

If Xena becomes a planet, she'll probably be renamed Persephone, after the reluctant queen of the underworld in Greek mythology. But, she may have to settle for the status of a mere "transneptunian object", like Quaoar and Sedna. Indeed, if Pluto had been discovered more recently, folks probably wouldn't have called him a planet either.

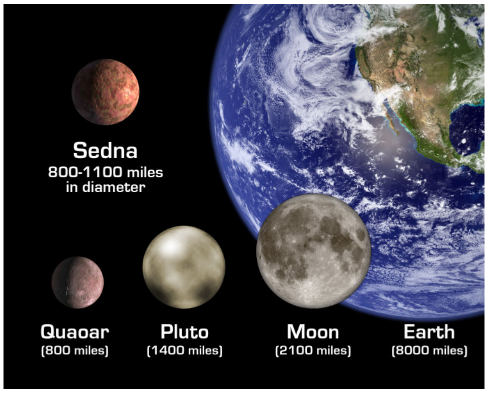
If you haven't even heard of Quaoar and Sedna... well, you must be too absorbed by mundane concerns to keep track of the burgeoning population of our Solar System. But it's not too late to mend your ways! Impress your friends by casually dropping some of this jargon:

- **Transneptunian object** — any object that orbits the Sun at an average distance greater than that of Neptune. Neptune is about 30 AU from the Sun, meaning it's 30 times farther from the Sun than we are. Transneptunian objects can be roughly divided into three kinds: Kuiper belt objects, scattered disc objects, and Oort cloud objects.



- **Kuiper belt object** — any object whose orbit lies in the Kuiper belt. This is the region in the ecliptic (the plane of the planets' orbits) between 30 and 50 AU from the Sun. There are a bunch of planetoids in this belt. Beyond 50 AU there seems to be a sharp dropoff in their density. Three main kinds of Kuiper belt objects have been found so far: cubewanos, plutinos and twotinos.

  ∗ **Cubewano** — A cubewano is a Kuiper belt object whose orbit is not in resonance with any of the outer planets. The curious name comes from "QB1", since the first example was named 1992 QB_1.
  One of the biggest cubewanos is Quaoar, with a diameter of about 1200 kilometers. This is about half the diameter of Pluto, or a third the size of the Moon: much bigger than anything in the asteroid belt!



  Folks believe Quaoar is a mixture of ice and rock. It's very dark in color, but last year crystalline water ice was detected on its surface, using infrared spectroscopy. This came as a surprise, because cosmic rays and solar wind should convert exposed ice crystals to the amorphous form of ice within about 10 million years. Could there have been liquid water volcanos active on Quaoar this recently?? Or maybe meteor impacts melt amorphous ice and then it crystallizes?

Other big cubewanos include Chaos, Varuna, and Deucalion. 2003 EL61 and 2005 FY9 are even bigger, but they haven't got nice names yet.

∗ **Plutino** — A plutino is a Kuiper belt object whose orbit is in 3:2 resonance with Neptune: they go around the Sun twice while Neptune goes around three times. About a quarter of Kuiper belt objects are plutinos.

The most famous plutino is Pluto itself, though some pedants argue that Pluto can't be a "little Pluto". Pluto is quite different than anything else we call a planet: it has an eccentric orbit that ranges between 30 and 50 AU, and its orbit is tilted 17 degrees to the ecliptic. Its surface is light brown, consisting mainly of frozen nitrogen and carbon monoxide. When it comes near the sun, as it recently did, it also gets a thin atmosphere made of these gases.

Other plutinos include Ixion, Orcus, Rhadamanthus, and Pluto's moon Charon. If you know Greek mythology, you'll know these guys are all named after deities of the underworld.

∗ **Twotino** — A twotino is a Kuiper belt object whose orbit is in $2:1$ resonance with Neptune. These are rare compared to plutinos, and they're smaller, so they're stuck with boring names like 1996 TR66. There are also a couple of Kuiper belt objects in $4:3$ and $5:3$ resonances with Neptune.

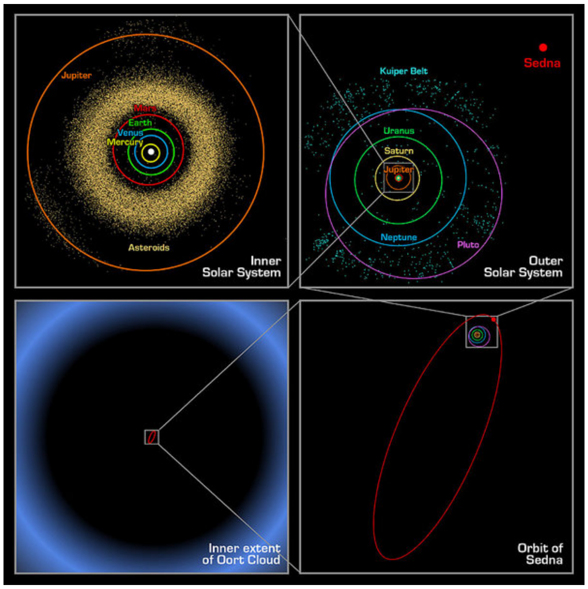
- **Scattered disc object** — A scattered disc object is a Kuiper belt object that has been perturbed by interactions with Neptune into an orbit that is more eccentric or more tilted from the ecliptic.

  Xena (or more properly 2003 UB_{313}) is a highly eccentric scattered disc object whose orbit carries it between 40 to 100 AU from the sun. Its orbit is inclined a whopping 44 degrees, and it's locked in a complicated $17:5$ resonance with Neptune. It's probably larger than Pluto — a reasonable rough guess is 2900 kilometers in diameter, as compared with 2400 for Pluto. Its surface has methane ice, and we now know it has a moon.

  It's quite possible that scattered disc objects are related to centaurs, which are planetoids orbiting the Sun between Jupiter and Neptune. The centaurs may be Kuiper belt objects that got knocked towards the Sun instead of away from it! Centaurs have chaotic orbits and will probably either collide with something or be ejected from the Solar System.

- **Oort cloud object** — the Oort cloud is a hypothesized spherical cloud of comets, perhaps between 50,000 and 100,000 AU from the Sun. The idea is this cloud consists of leftovers from the original nebula that collapsed to form our Solar system, and comets come from this region when they are perturbed from their orbits by the gravity of other stars.

  Nobody has seen a certified Oort cloud object. The best candidate so far is Sedna, an object roughly 1500 kilometers in diameter with a wildly eccentric orbit taking it between 80 to 930 AU from the Sun.



Sedna was discovered in 2004 when it was 90 AU from the Sun. It's redder than Mars, its temperature never rises above 23 Kelvin, and its year lasts 11,250 years. It's the farthest known object in our Solar System, but still much closer than the Oort cloud was supposed to be. Maybe it's a drastic example of a scattered disc object, maybe it's part of an "inner Oort cloud"... or maybe the Oort cloud isn't as far out as people thought.

The closest people have come to seeing the Oort cloud is seeing a "Bok globule":

Pre-Collapse Black Cloud B68 (visual view)
(VLT ANTU + FORS 1)

ESO PR Photo 02a/01 (10 January 2001)

© European Southern Observatory

A Bok globule is a cloud of dust and gas that's collapsing to form a star. This one is about 12,500 AU across. The scientists who observed it say it's about the size of the Oort cloud. This just goes to show how little we know about the Oort cloud!

For a great introduction to the Kuiper belt and related topics, try this:

10) David C. Jewitt, "Kuiper belt", `http://www.ifa.hawaii.edu/faculty/jewitt/kb.html`

For transneptunian objects in general, try:

11) William Robert Johnston, "Transneptunian objects", `http://www.johnstonsarchive.net/astro/tnos.html`

Also check out this newsletter:

12) *Distant EKOs: the Kuiper Belt Electronic Newsletter*, `http://www.boulder.swri.edu/ekonews/`

Quaoar was discovered in 2002 by Chad Trujillo and Mike Brown of Caltech:

13) Chad Trujillo, "Quaoar", `http://www.gps.caltech.edu/~chad/quaoar/`

For evidence of crystalline water ice on Quaoar, see:

14) David C. Jewitt and Jane Luu, "Crystalline water ice on the Kuiper belt object (50000) Quaoar", *Nature* **432** (2004), 731–733. Also available at `http://www.ifa.hawaii.edu/faculty/jewitt/quaoar.html`

Xena was discovered in 2003 by Trujillo, Brown and a colleague of theirs at Yale University:

15) Michael E. Brown, Chad A. Trujillo and David L. Rabinowitz, "Discovery of a planetary-sized object in the scattered Kuiper belt", submitted to *ApJ Letters*, available at `http://www.gps.caltech.edu/%7Embrown/papers/ps/xena.pdf`

Brown has a nice webpage about Xena and Gabrielle:

16) Michael E. Brown, "The discovery of UB313, the 10th planet", `http://www.gps.caltech.edu/~mbrown/planetlila/`

The same gang of three also discovered Sedna in 2003:

17) Michael E. Brown, Chad A. Trujillo and David L. Rabinowitz, "Discovery of a candidate inner Oort cloud planetoid", to appear in *ApJ Letters*, available at `http://www.gps.caltech.edu/%7Embrown/papers/ps/sedna.pdf`

. . . and Brown has a fun Sedna webpage too:

18) Michael E. Brown, "Sedna (2003 VB12)", `http://www.gps.caltech.edu/~mbrown/sedna/`

How all these transneptunian objects got where they are is a wonderful puzzle in celestial mechanics, but you can read more about that in the references above, especially Jewitt's Kuiper belt webpage.

Now I want to talk about Loops '05!

Instead of trying to review all the talks — a hopeless task, since there were 86 — I'll just mention the *two* strands of work I find most exciting.

First, there's new evidence that a quantum theory of pure gravity (meaning gravity without matter) makes sense in 4-dimensional spacetime.

To understand why this is exciting, you have to realize that in some quarters, the conventional wisdom says a quantum theory of pure gravity can't possibly make sense, except as a crude approximation at large distance scales, because this theory is "perturbatively nonrenormalizable".

Very roughly, this means that as we zoom in and look at the theory at shorter and shorter distance scales, it looks less and less like a "free field theory" where gravitons zip about without interacting. Instead, the interactions get stronger and more complicated!

So, in the jargon of the trade, we don't get a "Gaussian ultraviolet fixed point".

Huh?

Well, roughly, an "ultraviolet fixed point" is a quantum field theory that keeps looking the same as you keep viewing it on shorter and shorter distance scales. A "Gaussian" ultraviolet fixed point is one that's also a free quantum field theory: one where particles don't interact.

If quantum gravity approached a Gaussian ultraviolet fixed point as we zoomed in, we could calculate what gravitons do at arbitrarily high energies (at least perturbatively, as power series in Newton's constant — no guarantee that these series converge). Particle physicists would then be happy and say the theory was "perturbatively renormalizable".

But, it's not.

The conventional wisdom concludes that to save quantum gravity, we must include matter of precisely the right sort to *make* it perturbatively renormalizable. This is the quest that led people first to supergravity and ultimately to superstring theory — see "Week 195" for more of this story.

But, as far back as 1979, the particle physicist Weinberg raised the possibility that pure quantum gravity is "nonperturbatively renormalizable", or "asymptotically safe". This means that as we zoom in and look at the theory at shorter and shorter distance scales, it approaches some theory *other than* that of noninteracting gravitons.

In other words, Weinberg was suggesting that pure quantum gravity approaches a non-obvious ultraviolet fixed point — possibly a "non-Gaussian" one.

The big news is that this seems to be true!

Even cooler, in this theory spacetime seems to act 2-*dimensional* at very short distance scales.

This idea has been brewing for a long time — I talked about it extensively back in "Week 139". But now there's more solid evidence for it, coming from two quite different approaches.

First, people doing numerical quantum gravity in the "causal dynamical triangulations" approach are seeing this effect in their computer calculations. This is what Renate Loll explained at Loops '05. The best place to read the details is here:

19) Jan Ambjrn, J. Jurkiewicz and Renate Loll, "Reconstructing the universe", *Phys. Rev.* **D72** (2005) 064014. Also available as hep-th/0505154.

but if you need something less technical, try this:

20) Jan Ambjrn, J. Jurkiewicz and Renate Loll, "The universe from scratch", available as hep-th/0509010.

The titles of their papers are a bit grandiose, but their calculations are solid stuff — truly magnificent. I described their basic strategy in my report on the Marseille conference in week206. So, I won't explain that again. I'll just mention their big new result: in pure quantum gravity, spacetime has a spectral dimension of 4.02 ± 0.1 on large distance scales, but 1.80 ± 0.25 in the limit of very short distance scales!

Zounds! What does that mean?

The "spectral dimension" of a spacetime is the dimension as measured by watching heat spread out on this spacetime: the short-time behavior of the heat equation probes the spacetime at short distance scales, while its large-time behavior probes large distance scales. Spectral dimensions don't need to be integers — for fractals they're typically not. But, Loll and company believe they're seeing spacetimes that are *exactly* 2-dimensional in the limit of very small distance scales, *exactly* 4-dimensional in the limit of very large scales, with a continuous change in dimension in between. The error bars in the above figures come from doing Monte Carlo simulations. They're just using ordinary computers, not supercomputers. So, with more work one could shrink their error bars and test their result.

My main worry about their work is that it uses a fixed slicing of spactime by timelike slices. So, there's a danger that their procedure breaks Lorentz-invariance, even in the continuum limit which they are attempting to compute. I would like to find a way around this problem!

Luckily, some other people are getting similar results from a second procedure that definitely does *not* break Lorentz invariance:

21) Oliver Lauscher and Martin Reuter, "Fractal spacetime structure in asymptotically safe gravity", available as `hep-th/0508202`.

Reuter spoke about all this work at Loops '05. The idea is to investigate Weinberg's original idea in excruciatingly precise detail using "renormalization group flow" ideas. The above paper is a review of lots of others, and you need to read a bunch to get what's really going on. The upshot, however, is that they find evidence for a non-Gaussian ultraviolet fixed point in pure quantum gravity. Moreover, the spectral dimension of spacetime approaches 2 in the limit of very short distance scales.

Suppose this is all true. What does it mean?

Nobody knows yet; there are lots of attitudes one could take.

Ambjrn, Jurkiewicz and Loll could probably just plunge ahead and use computers to calculate lots of things about quantum gravity. (Right now they want to test their results in lots of ways.) One good thing would be to include matter of various sorts and see how it affects the conclusions.

Similarly, Lauscher and Reuter could just plunge ahead and compute, if they wanted.

This is excellent. But personally, I'd like to find a beautiful theory in which spacetime is 2-dimensional at short distance scales, which reduces to general relativity at large scales. In other words, to redo all these calculations "from the bottom up".

Unsurprisingly, I hope this beautiful theory is a spin foam model, since spin foams are 2-dimensional and I like them a lot. I presented some rough ideas on how one might invent such a model:

22) John Baez, "Towards a spin foam model of quantum gravity", talk at *Loops '05*, available at `http://math.ucr.edu/home/baez/loops05/`

But, these ideas are very tentative and only time will tell if they amount to anything. What's more important is that pure quantum gravity seems to exist — as a theory, that is — and people seem to be learning actual facts about it, instead of just arguing endlessly about it. That's progress!

The second most exciting thing at Loops '05, in my biased opinion, was the work of John Barrett, Laurent Freidel, Karim Noui and others on "matter without matter" in 3d quantum gravity. Simply by carving a Feynman-diagram-shaped hole in 3d spacetime and doing quantum gravity on the spacetime that's left over, you get a good theory of quantum gravity coupled to matter! You can even take the limit as Newton's gravitational constant goes to zero and get ordinary quantum field theory on flat spacetime!

Check these out:

23) John Barrett, "Feynman diagams coupled to three-dimensional quantum gravity", available as gr-qc/0502048.

    John Barrett, "Feynman loops and three-dimensional quantum gravity", *Mod. Phys. Lett.* **A20** (2005) 1271. Also available as gr-qc/0412107.

24) Laurent Freidel and David Louapre, "Ponzano-Regge model revisited I: gauge fixing, observables and interacting spinning particles", *Class. Quant. Grav.* **21** (2004) 5685–5726. Also available as hep-th/0401076.

    Laurent Freidel and David Louapre, "Ponzano-Regge model revisited II: equivalence with Chern-Simons", available as gr-qc/0410141

    Laurent Freidel and Etera R. Livine, "Ponzano-Regge model revisited III: Feynman diagrams and effective field theory", available as hep-th/0502106.

25) Laurent Freidel, Daniele Oriti, and James Ryan, "A group field theory for 3d quantum gravity coupled to a scalar field", available as gr-qc/0506067.

26) Karin Noui and Alejandro Perez, "Three dimensional loop quantum gravity: coupling to point particles", available as gr-qc/0402111.

This is mindblowingly beautiful, especially because lots of it is already mathematically rigorous, and we can easily make more so. It's even related to n-categories: my student Jeffrey Morton presented a poster on this aspect.

Together with my student Derek Wise, Jeffrey Morton and I plan to have a lot of fun studying this stuff. So, I won't talk about it more now — I'll probably get around to saying more someday, especially about how the whole story generalizes to 4 dimensions.

There's a lot more to say about Loops '05, but this will have to do. In a while, a bunch of the talks should be visible on the conference homepage.... that should give you a better idea of what happened.

Addendum: Here are some comments on this Week's Finds by Gene Partlow, Phillip Helbig and Robert Helling, and my replies — as well as a replies by Jonathan Thornburg and Arnold Neumaier.

Gene Partlow writes:

> *In a recent John Baez post he mentions discovery of probably the first known quasar found without a host galaxy. He says:*
>
>> *Did this quasar begin life in this galaxy and then get kicked out when the galaxy collided with something containing a super-massive black hole?*
>
> *I suggest that a fairly ordinary explanation may be that the nearby "wildly disturbed" galaxy may have contained several supermassive black holes which interacted via a gravitational slingshot scenario. This would be like a larger version of the effect where smaller mass stars can be flung out of globular clusters when encountering larger mass stars near the cluster center.*

Sounds like a possibility worth exploring. I'm no expert, so I can't tell how likely this is. I agree that a collision with some other object containing a super-massive black hole sounds a little odd, given that this object — most plausibly another galaxy — has *not been seen*. I wouldn't have ventured such a guess myself. But, it's mentioned on the European Southern Observatory webpage I cite above. To quote:

> *The absence of a massive host galaxy, combined with the existence of the blob and the star-forming galaxy, lead us to believe that we have uncovered a really exotic quasar, says team member Frederic Courbin (Ecole Polytechnique Federale de Lausanne, Switzerland). "There is little doubt that a burst in the formation of stars in the companion galaxy and the quasar itself have been ignited by a collision that must haven taken place about 100 million years ago. What happened to the putative quasar host remains unknown."*
>
> *HE0450-2958 constitutes a challenging case of interpretation. The astronomers propose several possible explanations, that will need to be further investigated and confronted. Has the host galaxy been completely disrupted as a result of the collision? It is hard to imagine how that could happen. Has an isolated black hole captured gas while crossing the disc of a spiral galaxy? This would require very special conditions and would probably not have caused such a tremendous perturbation as is observed in the neighbouring galaxy.*
>
> *Another intriguing hypothesis is that the galaxy harbouring the black hole was almost exclusively made of dark matter. "Whatever the solution of this riddle, the strong observable fact is that the quasar host galaxy, if any, is much too faint", says team member Knud Jahnke (Astrophysikalisches Institut Potsdam, Germany).*

Phillip Helbig writes:

> *John Baez writes:*
>
>> *Pluto is quite different than anything else we call a planet: it has an eccentric orbit that ranges between 30 and 50 AU, and its orbit is tilted 17 degrees to the ecliptic.*

> *Also, for a period of several years during each orbit, it is closer to the Sun than Neptune ever is. Until relatively recently, in fact, it was in this phase and Neptune was farther from the Sun than Pluto.*

Yeah! The US may still send a mission to Pluto before its atmosphere freezes, despite delays and indecision. The "Pluto Express" satellite was scheduled for launch in December 2004, but in 2000 NASA ordered a stop-work order on the project, due to lack of money and rising cost estimates. In 2003, Congress gave NASA money to proceed with the project:

27) The Planetary Society, "Pluto and Europa Campaign Page", `http://www.planetary.org/html/UPDATES/Pluto/pluto_europa_action.html`

and I guess it's now called "New Horizons":

28) "New Horizons: NASA's Pluto-Kuiper Belt Mission", `http://pluto.jhuapl.edu/`

This webpage gives a timetable of:

- launch in January 2005
- slingshot off Jupiter in February 2007
- arrival at Pluto and Charon in July 2015
- exploration of Kuiper belt during 2016-2020

I wonder if they have any plans to study the Pioneer anomaly?
Robert Helling writes:

> *John Baez wrote:*
>
> > - ***Twotino*** — *A twotino is a Kuiper belt object whose orbit is in* $2:1$ *resonance with Neptune. These are rare compared to plutinos, and they're smaller, so they're stuck with boring names like 1996 TR66. There are also a couple of Kuiper belt objects in* $4:3$ *and* $5:3$ *resonances with Neptune.*
>
> *Is there an easy way to see why these resonance orbits come about? Why do three body systems with a large central object, an intermediate planet and a small probe happen to get the probe in resonance with the planet? Is this just "frequency locking happens in chaotic systems" or is there an easy but more quantitative way to understand this?*

I'm shamefully ignorant of this, so ten minutes' research on the web was able to double my knowledge. I got ahold of this paper online:

29) B. Garfinkel, "On resonance in celestial mechanics: a survey", *Celestial Mech.* **28** (1982), 275–290, also available at `http://adsabs.harvard.edu/cgi-bin/nph-bib_query?bibcode=1982Ce.28..275G`

and while not easy to understand — I guess there's a huge body of work on this subject — it uses Hamiltonian perturbation theory and continued fractions to study resonance, and talks about a difference between "shallow" and "deep" resonances.

The orbits of Jupiter and Saturn are almost in $5:2$ resonance, but this is a "shallow" resonance, and Saturn wiggles back and forth around this resonance with a period of about 880 years — an effect called the "Great Inequality". The first person to study this was Laplace. I read elsewhere that:

> *The dynamics of the Sun-Jupiter-Saturn system was recognized as problematic from the beginnings of perturbation theory. The problems are due to the so-called Great Inequality (GI), which is the Jupiter-Saturn $2:5$ mean-motion near-commensurability.*

This is from:

30) F. Varadi, M. Ghil, and W. M. Kaula, "The great inequality in a planetary Hamiltonian theory", available as `chao-dyn/9311011`.

This shallow $5:2$ resonance is related to the continued fraction

$$\frac{1}{2+\frac{1}{2+\frac{1}{14+\frac{1}{2+\dots}}}}$$

which is close to $2/5$.

The Pluto-Neptune $3:2$ resonance, on the other hand, is a "deep resonance" and related to the continued fraction

$$\frac{1}{2-\frac{1}{2+\frac{1}{10+\dots}}}$$

which starts out close to $2/3$.

I wish I understood the connection between continued fractions and dynamical systems better! I know it gives rise to the role of the Golden Ratio in chaos theory, which I tried to explain in "Week 203". But, I don't understand it very deeply.

Robert Helling also asked:

> *There are people doing numerical long term stability analysis of the solar system. From what I know, they are not just taking $F=ma$ and Newton's law of gravity, replacing dt by Δt and then integrating but use much fancier spectral methods. Could somebody please point me to an introduction into these methods?*

I referred him to the work of Gerald Sussman, Jack Wisdom and others in "Week 107", but Jonathan Thornburg posted this reply:

> *I don't do this sort of work myself, but the buzzwords you want are "symplectic ODE integrator". The basic idea is to use an ODE integration scheme which conserves energy, angular momentum, and maybe other nice things, up to floating-point roundoff error, rather than just up to finite differencing error like a standard ODE integrator would do.*

prompting Arnold Neumaier to give a nice list of references, which I will take the liberty of numbering:

> 31) *Tetsuharu Fuse, "Planetary perturbations on the* 2 : 3 *mean motion resonance with Neptune",* Publ. Astron. Soc. Japan ***54*** *(2002), 494–499. Also available at* `http://astronomy.nju.edu.cn/~xswan/reference/Fuse_PASJ54_493.pdf`
>
> *uses symplectic integration to study* 2 : 3 *resonances numerically.*
>
> *The thesis:*
>
> 32) *Luz Vianey Vela-Arevalo,* Time-frequency analysis based on wavelets for Hamiltonian systems*, Caltech, 2002. Also available at* `http://www.cds.caltech.edu/~luzvela/th2s.pdf`
>
> *contains in Chapter 4 interesting numerical information about chaos, resonances, and stability in the restricted 3-body problem. Other interesting papers include:*
>
> 33) *George Voyatzis and John D. Hadjidemetriou, "Symmetric and asymmetric librations in planetary and satellite systems at the* 2/1 *resonance", available at* `http://users.auth.gr/~hadjidem/Asymmetric1.pdf`
>
> *Mihailo Cubrovic, "Regimes of stability and scaling relations for the removal time in the asteroid belt", available as* `astro-ph/0501004`*.*
>
> *Ryszard Gabryszewski and Ireneusz Wlodarczyk, "The resonant dynamical evolution of small body orbits among giant planets", available as* `astro-ph/0203182`*.*
>
> *Luz V. Vela-Arevalo and Jerrold E. Marsden, "Time-frequency analysis of the restricted three-body problem: transport and resonance transitions",* Class. Quant. Grav. ***21*** *(2004), S351–S375. Also available at* `http://cns.physics.gatech.edu/~luzvela/VelaArevaloMarsdenCQG_2004.pdf`
>
> *Harry Varvogli, Kleomenis Tsiganis, and John D. Hadjidemetriou, "The 'third' integral in the restricted three-body problem revisited", available at* `http://www.astro.auth.gr/~varvogli/varv5.ps`

So many worlds, so much to do,
So little done, such things to be.

— *Tennyson, In Memoriam*

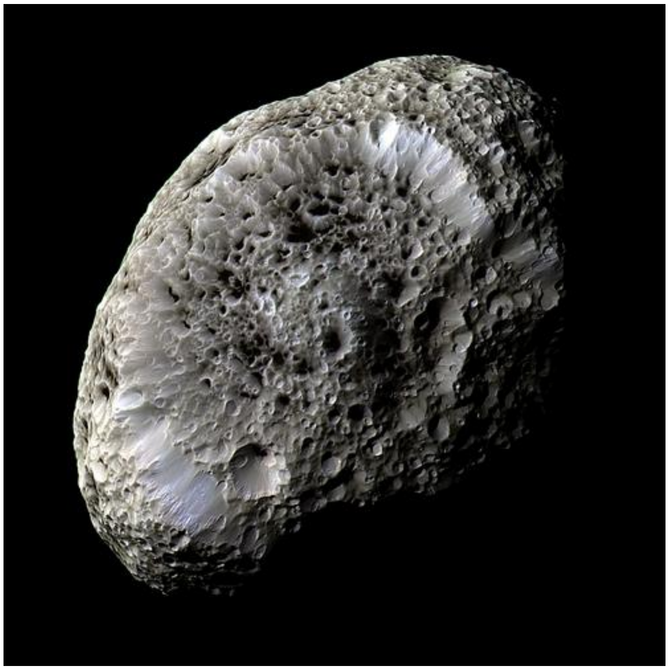
Week 223

November 14, 2005

This week I'd like to talk about two aspects of higher gauge theory: p-form electromagnetism and nonabelian cohomology. Lurking behind both of these is the mathematics of n-categories, but I'll do my best to hide that until the end, to build up the suspense.

But first, some cool pictures. Astronomy is booming these days, and it's a great way to see beautiful complexity emerging from simple laws in this wonderful universe of ours. So, I'd like the freedom to occasionally start This Week's Finds with some pictures from the skies. Think of it as an appetizer before the main course. Sometimes I'll explicitly relate these pictures to math and physics; other times not.

Here's Saturn's moon Hyperion, photographed up close by the Cassini probe:



1) Cassini-Huyghens Mission, "Hyperion: Odd World", `http://saturn.jpl.nasa.gov/multimedia/images/image-details.cfm?imageID=1762`

It seems to be a huge pile of rubble loosely held together by gravity and heavily cratered by meteor bombardments.

Hyperion is interesting because it's the only known moon that tumbles chaotically on a short time scale, thanks to its eccentric shape and gravitational interactions with Saturn and Titan.

This leads to some interesting math. We can think of Hyperion's angular momentum vector as a point on a sphere. If we started out knowing this point lay inside some small disk, time evolution would warp this disk into an ever more complicated region as time passed. This region would always have the same area, thanks to the wonders of symplectic geometry. But it would sprout ever more complicated tendrils, with its perimeter growing by a factor of e about every 100 days or so!

That's chaos for you.

Indeed, only quantum mechanics would stop the intricacy from growing forever, by blurring it out. After about 37 years, the area of a typical tendril would equal Planck's constant. At this point, classical mechanics would no longer be accurate. You'd really need to describe Hyperion's spin state using quantum theory: for example, a holomorphic section of some line bundle on the sphere.

Well. . . at least you would if it weren't for decoherence caused by the interaction of Hyperion with its environment, for example solar radiation! For an explanation of how this changes the story, try:

2) Michael Berry, "Chaos and the semiclassical limit of quantum mechanics (is the moon there when somebody looks?)", in *Quantum Mechanics: Scientific Perspectives on Divine Action*, CTNS Publications, Vatican Observatory, 2001. Also available at `http://www.phy.bris.ac.uk/people/berry_mv/the_papers/berry337.pdf`

Here's another great picture:

3) The Hubble Heritage Project, "Cat's Eye Nebula — NGC 6543", `http://heritage.stsci.edu/2004/27/index.html`

This is a star about the size of the Sun, nearing the end of its life, emitting pulses of gas and dust. Astronomers call such a thing a "planetary nebula", though it has nothing to do with planets. It's in our galaxy, about 3000 light years from us. When it's done shedding its outer layers, all that's left of this star will be a white dwarf.

Our own Sun will become a planetary nebula in about 6.9 billion years, after two separate stages of being a red giant — one as it runs out hydrogen, and one as it runs out of helium. When the helium is all gone, the Sun will start to pulsate every 100,000 years, ejecting more and more mass in each pulse, eventually throwing off all but a hot inner core made of heavier elements. The astronomer Bruce Balick has written eloquently on what this will mean for the Earth:

> *Here on Earth, we'll feel the wind of the ejected gases sweeping past, slowly at first (a mere 5 miles per second!), and then picking up speed as the spasms continue — eventually to reach 1000 miles per second!! The remnant Sun will rise as a dot of intense light, no larger than Venus, more brilliant than 100 present Suns, and an intensely hot blue-white color hotter than any welder's torch. Light from the fiendish blue "pinprick" will braise the Earth and tear apart its surface molecules and atoms. A new but very thin "atmosphere" of free electrons will form as the Earth's surface turns to dust.*

So, don't keep procrastinating — enjoy life now!

For other pictures of planetary nebulae, try Balick's webpage:

4) Bruce Balick, "Hubble Space Telescope images of planetary nebulae", `http://www.astro.washington.edu/balick/WFPC2/index.html`

For a timeline of the universe, including the future life of our Sun, try:

5) John Baez, "A brief history of the universe", `http://math.ucr.edu/home/baez/timeline.html`

Now. . . on to p-form electromagnetism!

In ordinary electromagnetism, the secret star of the show turns out to be not the electromagnetic field but the "vector potential", A. At least locally, we can think of this as a 1-form. A 1-form is just a gadget that you can integrate along a path and get a number. In the case of the vector potential, this number describes the change in phase that a charged particle acquires as it moves along this path.

The 1-form A gives rise to a 2-form F called the "electromagnetic field". A 2-form is a gadget you can integrate over a surface and get a number. Here's how we get F from A. Suppose we move a charged particle around a loop that's the boundary of some surface. Then the integral of F over this surface is defined to be the integral of A around the loop! We summarize this by saying that F is the "exterior derivative" of A, and writing

$$F = dA.$$

F is called the electromagnetic field because. . . that's what it is! It contains both the electric and magnetic fields in a single neat package. In 4d spacetime, the magnetic field describes the change in a phase of a charged particle that loops around a surface in the xy, yz or zx planes. The electric field describes the change in phase of a charged particle that loops around a surface in the xt, yt or zt planes.

If you don't know this stuff, you're missing some of the best fun life has to offer. For an easy introduction with lots of gorgeous pictures, see:

6) Derek Wise, "Electricity, magnetism and hypercubes", available at `http://math.ucr.edu/~derek/talks/050916bw.pdf`

The idea of p-form electromagnetism is to replace point particles by strings or higher-dimensional membranes. To see how this goes, it's enough to look at 2-form electromagnetism.

In 2-form electromagnetism, the star of the show is a 2-form, A. As already mentioned, a 2-form is a gadget you can integrate over a surface and get a number. In 2-form electromagnetism, this number describes the change in phase that a charged string acquires as it moves along, tracing out a surface in spacetime.

The 2-form A gives rise to a 3-form, F. A 3-form is a gadget you can integrate over a 3-dimensional region and get a number. Suppose we move a charged string and let it trace out a surface that's the boundary of some 3-dimensional region. Then the integral of F over this region is defined to be the integral of A over the surface! Again we write this as:

$$F = dA.$$

So, we're just adding one to the dimensions of things. This makes it easy to keep on going. In fact, for any integer p, we can write down a generalization of Maxwell's equations.

It goes like this. We start with a p-form A. We define a $(p+1)$-form

$$F = dA$$

This automatically implies some of Maxwell's equations:

$$dF = 0$$

but the nontrivial Maxwell equations say that

$$*d*F = J$$

where $*$ is the Hodge star operator and J is a p-form called the "current", which is produced by charged matter.

What does this mean, physically? The idea is that we have charged matter consisting of $(p-1)$-dimensional membranes. These trace out p-dimensional surfaces in spacetime as time passes. The current J is a p-form that's concentrated on these surfaces. The current affects the A field in a manner governed by Maxwell's equations. Conversely, the A field affects the motion of the membranes. Classically, we just integrate the A field over the surface traced out by a membrane and add the result to the *action* for the membrane. In the path integral approach to quantum mechanics, this number gives a change in phase, as already mentioned.

Maxwell's equations and their p-form generalization make sense when spacetime is any Lorentzian manifold. However, to get a theory where initial data determine a unique global solution, we want our spacetime to be "globally hyperbolic", which means that it has a "Cauchy surface": roughly, a spacelike surface that any sufficiently long timelike curve hits precisely once. To get a good *quantum* theory of p-form electromagnetism with a Hilbert space of states on which time evolution acts as unitary operators, we need more: our spacetime should be "stationary", meaning that it has time translation symmetry. Otherwise there's no way to define energy and the vacuum state — which is defined to be the least-energy state.

My student Miguel Carrion-Alvarez tackled an important special case in his thesis, namely "static" globally hyperbolic spacetimes:

7) Miguel Carrion-Alvarez, "Loop quantization versus Fock quantization of p-form electromagnetism on static spacetimes", available as math-ph/0412032.

There's a lot of interesting analysis involved, especially when space (the Cauchy surface) is noncompact. When it's compact, we can use "Hodge's theorem" to relate its deRham cohomology to its topology, and this turns out to be crucial for understanding p-form electromagnetism — especially issues like the p-form Bohm-Aharonov effect. When it's noncompact we need something called "twisted Lˆ2 cohomology" instead, and Miguel proved a generalization of Hodge's theorem for this.

With the analysis under control, Miguel was able to set up a very beautiful approach to "loop quantum electromagnetism" and its p-form generalization. Here the idea is to write Maxwell's equations in terms of the integrals of A around all possible loops in space — or more generally, over all p-dimensional surfaces. People interested in loop quantum gravity should like this.

As you can guess, either from seeing all the "d" operators or seeing all the buzzwords I'm throwing around, p-form electromagnetism is really just cohomology incarnated as physics! My student Derek Wise made this very precise for a version of the theory where spacetime is *discrete* — so-called "lattice p-form electromagnetism":

8) Derek Wise, "Lattice p-form electromagnetism and chain field theory", available as gr-qc/0510033. Version with better graphics and related material at http://math.ucr.edu/~derek/pform/index.html

In this paper, he shows lattice p-form electromagnetism is a "chain field theory": something like a topological quantum field theory, but where what matters is not spacetime itself so much as the cochain complex of differential forms *on* spacetime, equipped with just enough extra geometrical structure to write down the p-form version of Maxwell's equations.

Both Miguel's thesis and Derek's papers are great if you want to learn lots of math and physics. I seem to attract students who enjoy explaining things.

Speaking of which. . . .

Next I want to explain some stuff Danny Stevenson told me at a mall in the little town of Cabazon while we were recovering from a hike in the desert followed by pancakes at the Wheel Inn — a roadside restaurant famous for its enormous statues of dinosaurs.

Danny works on gerbes, stacks, and higher gauge theory. Last year we wrote a paper with Alissa Crans and Urs Schreiber constructing 2-groups (categorified groups) from the math of string theory — more precisely, from central extensions of loop groups. Since then I've been spending a lot of time writing a paper with Urs on higher gauge theory, where we set up a theory of parallel transport along surfaces. 2-form electromagnetism is the simplest case of this theory. Meanwhile, Danny has been thinking about connections on 2-vector bundles and their relation to the cohomology of Lie 2-algebras.

This has led him to generalize Schreier theory in some interesting ways. So, let me tell you about Schreier theory!

Schreier theory is a way to classify short exact sequences of groups. I'll say what I mean by that in a minute... but what makes Schreier theory special is that avoids some simplifying assumptions you might have seen if you've studied short exact sequences before.

Normally people water down their short exact sequences by assuming some of the groups in question are *abelian*. This lets them use "cohomology theory" to do the classification. See "Week 210" for a nice book that takes this approach.

This standard approach is great — I'm not knocking it — but Schreier theory is more general: it's really a branch of "nonabelian cohomology theory". It's not all that hard to explain, either. So, I'll explain it and then talk about various simplifying assumptions people make.

The goal of Schreier theory is to classify short exact sequences of groups:

$$1 \to F \to E \to B \to 1$$

for a given choice of F and B. "Exact" means that the arrows stand for homomorphisms and the image of each arrow is the kernel of the next. Here this just means that F is a normal subgroup of E and B is the quotient group E/F. Such a short exact sequence is also called an "extension of B by F", since E is bigger than B and contains F. The simplest choice is to let E be the direct sum of F and B. Usually there are other more interesting extensions as well.

To classify these, the trick is to use the analogy between group theory and topology.

As I explained in "Week 213", you can think of a group as a watered-down version of a connected space with a chosen point. The reason is that given such a space, we get a group consisting of homotopy classes of loops based at the chosen point. This is called the "fundamental group" of our space. There's a lot more information in our space than this group. But pretty much anything you can do for groups, you can do for such spaces. It's usually harder, but it's completely analogous!

In particular, classifying short exact sequences is a lot like classifying "fibrations":

$$1 \to F \to E \to B \to 1$$

where now the letters stand for connected spaces with a chosen point, and the arrows stand for continuous maps. If you're a physicist or geometer you may prefer fiber bundles to "fibrations" — but luckily, they're so similar we can ignore the difference in a vague discussion like this. The idea is basically just that E maps onto B, and sitting over each point of B we have a copy of F. We call B the "base space", E the "total space" and F the "fiber".

If we want to classify such fibrations we can consider carrying the fiber F around a loop in B and see how it twists around. For example, if all our spaces are smooth manifolds, we can pick a connection on the total space E and see what parallel transport around a loop in the base space B does to points in the fiber F. This gives a kind of homomorphism

$$\Omega B \to \mathrm{Aut}(F)$$

sending loops in B to invertible maps from F to itself. And, the cool thing is: this homomorphism lets us classify the fibration!

Here I say "kind of homomorphism" since ΩB, the space of loops in B based at the chosen point, is only "kind of" a topological group: the group laws only hold up to homotopy. But let's not worry about this technicality — especially since I'm being vague about all sorts of other equally important issues!

The reason I can get away with not worrying about these issues is that I'm trying to explain a very robust powerful principle — one that can easily survive a dose of vagueness that would kill a lesser idea. Namely, if B is a connected space with a chosen basepoint,

> *FIBRATIONS OVER THE BASE SPACE B WITH FIBER F*
> *ARE "THE SAME" AS*
> *HOMOMORPHISMS SENDING LOOPS IN B TO AUTOMORPHISMS OF F.*

This could be called "the basic principle of Galois theory", for reasons explained in "Week 213". There I explained the special case where the fiber is discrete. Then our fibration called a "covering space", and the basic principle of Galois theory boils down to this:

> *COVERING SPACES OVER B WITH FIBER F*
> *ARE "THE SAME" AS*
> *HOMOMORPHISMS FROM THE FUNDAMENTAL GROUP OF B TO AUTOMORPHISMS OF F.*

Okay. Now let's use the same principle to classify extensions of a group B by a group F:

$$1 \to F \to E \to B \to 1$$

The group B here acts like "loops in the base". But what acts like "automorphisms of the fiber"?

You might guess it's the group of automorphisms of F. But, it's actually the 2-*group* of automorphisms of F!

A 2-group is a categorified version of a group where all the usual group laws hold up to natural isomorphism. They play a role in higher gauge theory like that of groups in ordinary gauge theory. In higher gauge theory, parallel transport along a path is described by an *object* in a 2-group, while parallel transport along a path-of-paths is described by a *morphism*. In 2-form electromagnetism we use a very simple "abelian" 2-group, which has one object and either the real line or the circle as morphism. But there are other more interesting "nonabelian" examples.

If you want to learn more about 2-form electromagnetism from this perspective, try "Week 210". For 2-groups in general, try this paper:

9) John Baez and Aaron Lauda, "Higher-dimensional algebra V: 2-groups", *Theory and Applications of Categories* **12** (2004), 423–491. Available online at `http://www.tac.mta.ca/tac/volumes/12/14/12-14abs.html` or as `math.QA/0307200`.

Anyway: it turns out that any group F gives a 2-group $\mathrm{AUT}(F)$ where the objects are automorphisms of F and the morphisms are "conjugations" — elements of F acting to conjugate one automorphism and yield another. And, extensions

$$1 \to F \to E \to B \to 1$$

are classified by homomorphisms

$$B \to \mathrm{AUT}(F)$$

where we think of B as a 2-group with only identity morphisms. More precisely:

> *EXTENSIONS OF THE GROUP B BY THE GROUP F*
> *ARE "THE SAME" AS*
> *HOMOMORPHISMS FROM B TO THE 2-GROUP $\mathrm{AUT}(F)$*

It's fun to work out the details, but it's probably not a good use of our time together grinding through them here. So, I'll just sketch how it works.

Starting with our extension

$$1 \to F \xrightarrow{i} E \xrightarrow{p} B \to 1$$

we pick a "section"

$$E \xleftarrow{s} B$$

meaning a function with

$$p(s(b)) = b$$

for all b in B. We can find a section because p is onto. However, the section usually *isn't* a homomorphism.

Given the section s, we get a function

$$\alpha \colon B \to \mathrm{Aut}(F)$$

where $\mathrm{Aut}(F)$ is the group of automorphisms of F. Here's how:

$$\alpha(b)f = s(b)fs(b)^{-1}$$

However, usually α *isn't* a homomorphism.

So far this seems a bit sad: functions between groups want to be homomorphisms. But, we can measure how much s fails to be a homomorphism using the function

$$g \colon B^2 \to F$$

defined by

$$g(b,b') = s(bb')s(b')^{-1}s(b)^{-1}$$

Note that $g = 1$ iff s is a homomorphism.

We can then use this function g to save α. The sad thing about α is that it's not a homomorphism… but the good thing is, it's a homomorphism *up to conjugation by g!* In other words:

$$\alpha(bb')f = g(b,b')[\alpha(b)\alpha(b')f]g(b,b')^{-1}$$

Taken together, α and g satisfy some equations ("cocycle conditions") which say precisely that they form a homomorphism from B to the 2-group $\mathrm{AUT}(F)$. Conversely, any such homomorphism gives an extension of B by F.

In fact, isomorphism classes of extensions of B by F correspond in a 1-1 way with isomorphism classes of homomorphisms from B to $\mathrm{AUT}(F)$. So, we've classified these extensions!

In fact, something even better is true! It's evil to "decategorify" by taking isomorphism classes as we did in the previous paragraph. To avoid this, we can form a groupoid whose objects are extensions of B by F, and a groupoid whose objects are homomorphisms $B \to \mathrm{AUT}(F)$. I'm pretty sure that if you form these groupoids in the obvious way, they're equivalent. And that's what this slogan really means:

> *EXTENSIONS OF THE GROUP B BY THE GROUP F*
> *ARE "THE SAME" AS*
> *HOMOMORPHISMS FROM B TO THE 2-GROUP* $\mathrm{AUT}(F)$

Next, let me say how Schreier theory reduces to more familiar ideas in two special cases.

People have thought a lot about the special case where F is abelian and lies in the center of E. These are called "central extensions". This is just the case where $\alpha = 1$. The set of isomorphism classes of central extensions is called $H^2(B, F)$ — the "second cohomology" of B with coefficients in F.

People have also thought about "abelian extensions". That's an even more special case where all three groups are abelian. The set of isomorphism classes of such extensions is called $\mathrm{Ext}(B, F)$.

Since we don't make any simplifying assumptions like this in Schreier theory, it's part of a subject called "nonabelian cohomology". It was actually worked out by Dedecker in the 1960's, based on much earlier work by Schreier:

10) O. Schreier, "Ueber die Erweiterung von Gruppen I", *Monatschefte fur Mathematik and Physick* **34** (1926), 165–180. "Ueber die Erweiterung von Gruppen II", *Abh. Math. Sem. Hamburg* **4** (1926), 321–346.

11) P. Dedecker, "Les foncteuers Ext_Π, H^2_Π and H^2_Π non abeliens", *C. R. Acad. Sci. Paris* **258** (1964), 4891–4895.

More recently, Schreier theory was pushed one step up the categorical ladder by Larry Breen. As far as I can tell, he essentially classified the extensions of a 2-group B by a 2-group F in terms of homomorphisms $B \to \mathrm{AUT}(F)$, where $\mathrm{AUT}(F)$ is the *3-group* of automorphisms of F:

12) Lawrence Breen, "Theorie de Schreier superieure", *Ann. Sci. Ecole Norm. Sup.* **25** (1992), 465–514. Also available at `http://www.numdam.org/numdam-bin/feuilleter?id=ASENS_1992_4_25_5`

We can keep pushing Schreier theory upwards like this, but we can also expand it "sideways" by replacing groups with groupoids. You should have been annoyed by how I kept assuming my topological spaces were connected and equipped with a specified point. I did this to make them analogous to groups. For example, it's for only spaces like this that the fundamental group is powerful enough to classify covering spaces. For more general spaces, we must use the fundamental *groupoid*. And, we can set up a Schreier theory for extensions of groupoids:

13) V. Blanco, M. Bullejos and E. Faro, "Categorical non abelian cohomology, and the Schreier theory of groupoids", available as math.CT/0410202.

In fact, these authors note that Grothendieck did something similar back in 1971: he classified *all* groupoids fibered over a groupoid B in terms of weak 2-functors from B to Gpd, which is the 2-groupoid of groupoids! The point here is that Gpd contains $\mathrm{AUT}(F)$ for any fixed groupoid F:

14) Alexander Grothendieck, *Revłtements tales et Groupe Fondamental (SGA1)*, chapter VI: "Catgories fibres et descente", Lecture Notes in Mathematics **224**, Springer, Berlin, 1971. Also available as math.AG/0206203.

Having extended the idea "sideways" like this, one can then continue marching "upwards". I don't know how much work has been done on this, but the slogan should be something like this:

n-GROUPOIDS FIBERED OVER AN n-GROUPOID B
ARE "THE SAME" AS
WEAK $(n+1)$-FUNCTORS FROM B TO THE $(n+1)$-GROUPOID $n\mathsf{Gpd}$

Grothendieck also studied this kind of thing with categories replacing groupoids, so there should also be an n-category version, I think... but it's more delicate to define "fibrations" for categories than for groupoids, so I'm a bit scared to state a slogan suitable for n-categories.

However, I'm not scared to go from n-groupoids to ω-groupoids, which are basically the same as spaces. In terms of spaces, the slogan goes like this:

SPACES FIBERED OVER THE SPACE B
ARE "THE SAME" AS
MAPS FROM B TO THE SPACE OF ALL SPACES

This is how James Dolan taught it to me. Most mortals are scared of "the space of all spaces" — both for fear of Russell's paradox, and because we really need a *space* of all spaces, not just a mere set of them. To avoid these terrors, you can water down Jim's slogan by choosing a specific space F to be the fiber:

FIBRATIONS WITH FIBER F OVER THE SPACE B
ARE "THE SAME" AS
MAPS FROM B TO THE CLASSIFYING SPACE OF $\mathrm{AUT}(F)$

where $\mathrm{AUT}(F)$ is the topological group of homotopy self-equivalences of F. The fearsome "space of all spaces" is then the disjoint union of the classifying spaces of all these topological groups $\mathrm{AUT}(F)$. It's too large to be a space unless you pass to a larger universe of sets, but otherwise it's perfectly fine. Grothendieck invented the notion of a "Grothendieck universe" for precisely this purpose:

15) Wikipedia, "Grothendieck universe", http://en.wikipedia.org/wiki/Grothendieck_universe

Addendum: I'd like to thank Leo Alonso for pointing out that Grothendieck's famous "SGA1" is now available on the arXiv in TeX form, courtesy of the Socit Mathmatique de France. SGA2 is also available on the arXiv, and more are coming.

Here are further addenda thanks to Kevin Buzzard, Toby Bartels, David Corfield, Peter May, Jim Stasheff, and Ronnie Brown.

First, an email I sent in reply to Kevin Buzzard, who was curious about how we classify extensions of the group B by the group F using homomorphisms from B to the 2-group $\mathrm{AUT}(F)$. In particular, he wanted to know the definition of "2-group" and "homomorphism between 2-groups", and how $\mathrm{AUT}(F)$ is defined:

> *Dear Kevin -*
>
> *You write:*
>
> > *I'm just checking some of the details (extensions of groups are morphisms of 2-groups) and I find that you've not given me quite enough information to do it (in the sense that I'm not knowledgeable enough about standard facts about 2-groups to fill in some of the details that you allude to).*
>
> *Sorry. I'm glad you care enough to want to know this stuff: 2-groups and homomorphisms between them are defined in loving detail in my paper with Aaron Lauda, but I'll answer your questions here and append this to "Week 223" to help out anyone else who cares.*
>
> > *Say we're in the "classical" case where F is an abelian group, B is a group, and we're classifying extensions $1 \to F \to E \to B \to 1$ where F lies in the centre of E. We know the answer here: such E's are classified by $H^2(B,F)$ which, for me, means 2-cocycles over 2-coboundaries. Recall that a 2-cocycle is $g\colon B^2 \to F$ satisfying $g(a,bc)+g(b,c)=g(a,b)+g(ab,c)$.*
>
> *Right.*
>
> > *You want the answer to be homomorphisms $B \to \mathrm{AUT}(F)$. You don't quite give the definitions of these things.*
>
> *True. Let me start by saying what a 2-group is, and then how the group B becomes a 2-group, and then how $\mathrm{AUT}(F)$ is defined.*
>
> *A group is just a category with one object and with all morphisms invertible. Slick! But, we usually prefer to think of a group as a set: the morphisms of our category get called "elements" of this set. This set then has a multiplication function*
>
> $$m\colon G\times G\to G$$
>
> *and an identity element*
>
> $$1\in G$$
>
> *satisfying the associative and unit laws, and such that every element has an inverse.*

All this categorifies painlessly:

A 2-group is just a 2-category with one object and with all morphisms and 2-morphisms invertible. Slick! But, we sometimes prefer to think of a 2-group as a category: the morphisms of our 2-category get called "objects" of this category, and the 2-morphisms get called "morphisms". This category then has a multiplication functor

$$m\colon G \times G \to G$$

and an identity object

$$1 \in G$$

satisfying the associative and unit laws, and such that every object and morphism has an inverse.

In short: a 2-group is just like a group, but with the word "category" replacing the word "set", the word "object" replacing the word "element", and so on!

Now, how does a group become a 2-group? Simple: we take the elements *of our group and make them the* objects *of our 2-group; then we say the only morphisms of our 2-group are identity morphisms. The 2-group multiplication $m\colon G \times G \to G$ comes from the multiplication in our group, and so on.*

> *Let's stick to F abelian. You think of $\mathrm{AUT}(F)$ as being the 2-category with objects $\mathrm{Aut}(F)$, 1-morphisms $\mathrm{Hom}(\varphi, \psi) = F$ if $\varphi = \psi$ and empty otherwise (because F is abelian). What are the 2-morphisms? Is $\mathrm{Hom}(\varphi, \psi)$ supposed to be a category with objects F?*

Alas, you're one dimension down: thought of as a 2-category, we want our 2-group $\mathrm{AUT}(F)$ to have one object, the usual group $\mathrm{Aut}(F)$ as morphisms, and conjugations between these as 2-morphisms.

Here's how we get this. Think of our group F as a category. Then let $\mathrm{AUT}(F)$ have F as its only object, invertible functors

$$a\colon F \to F$$

as its morphisms, and natural isomorphisms between these as its 2-morphisms.

That's very slick. But let me say this in a different way using the other viewpoint, where we think of a 2-group as a "category with multiplication and inverses". Given a group F, our 2-group $\mathrm{AUT}(F)$ will be the category where the objects are automorphisms

$$a\colon F \to F$$

and where a morphism $f\colon a \to a'$ is an element f of F that conjugates a to give a':

$$f a(g) f^{-1} = a'(g) \quad \textit{for all } g \in F$$

This viewpoint requires some extra work to check that $\mathrm{AUT}(F)$ is a 2-group. The 2-category viewpoint is actually much more efficient.

> *Now what is a homomorphism $B \to \mathrm{AUT}(F)$?*

> *Here's where the subtlety comes in: weakening! So far we haven't weakened anything: all the equations in the definition of a group became equations in the definition of a* 2*-group. We're really just dealing with "strict"* 2*-groups here. But we need to* weaken *the definition of homomorphism, replacing some equations by isomorphisms, to get things to work out well now.*
>
> *If we think of* B *and* $\mathrm{AUT}(F)$ *as* 2*-categories, a homomorphism* $B \to \mathrm{AUT}(F)$ *is just a weak* 2*-functor. Slick! But, you may not enjoy this definition as much as I do.*
>
> *So, let's think of* B *and* $\mathrm{AUT}(F)$ *as "categories with multiplication and inverses". Then a homomorphism*
>
> $$\alpha\colon B \to \mathrm{AUT}(F)$$
>
> *is a functor that preserves multiplication of objects* up to a specified isomorphism*, which must satisfy some laws of its own.*
>
> *Quite roughly, this means that given two objects* b *and* b' *of* B*, we don't have an equation*
>
> $$\alpha(bb') = \alpha(b)\alpha(b')$$
>
> *Instead, we have an isomorphism*
>
> $$g(b, b')\colon \alpha(bb') \to \alpha(b)\alpha(b')$$
>
> *This needs to satisfy some equations. I can tell you these if you want, but for starters you can check that in our application, this* $g(b, b')$ *thing will be the* 2*-cocycle familiar from group cohomology!*
>
> *And, the laws* g *must satisfy say precisely that* g *is a* 2*-cocycle.*
>
> *(Indeed, for a* central *extension* $\alpha = 1$*, so all we really need to think about is this* 2*-cocycle* g*. Schreier theory goes on to consider more general extensions, where* $\alpha \neq 1$*.)*
>
>> *I'm sure I could just go and read Breen's book, but these questions are so trivial that I'm sure you can instantly answer them, and you also get the confirmation that there are still people out there reading TWF.*
>
> *That's worth a lot! If you ever want to learn more about* 2*-groups and homomorphisms between them, I think* my paper with Aaron Lauda *could be easier than* Breen's opus*. Breen's opus concerns "higher Schreier" theory — classifying extensions of* 2*-groups with the help of* 3*-groups!*
>
> *Best,*
> *jb*

It's important to note that it's *isomorphism classes* of extensions that correspond to *isomorphism classes* of homomorphisms $B \to \mathrm{Aut}(F)$. For this, one needs to know what a "2-isomorphism" between homomorphisms of 2-groups is. Again, this is explained in my paper with Lauda. It's a special case of a weak natural isomorphism between weak 2-functors between 2-categories, but we say what this means in terms that working mathematicians can understand.

Also, Toby Bartels had some comments on the dinosaurs of Cabazon and size issues in category theory:

> *John wrote in part:*
>
> > *Next I want to explain some stuff Danny Stevenson told me at a mall in the little town of Cabazon while we were recovering from a hike in the desert followed by pancakes at the Wheel Inn — a roadside restaurant famous for its enormous statues of dinosaurs.*
>
> *Did you see the creationist sign by the dinosaurs?*
>
> > *SPACES FIBERED OVER THE SPACE B ARE "THE SAME" AS MAPS FROM B TO THE SPACE OF ALL SPACES*
> >
> > *The fearsome "space of all spaces" is then the disjoint union of the classifying spaces of all these topological groups $\mathrm{AUT}(F)$. It's too large to be a space unless you pass to a larger universe of sets, but otherwise it's perfectly fine.*
>
> *So if you want your slogan to treat size issues carefully:*
>
> > *SPACES FIBERED OVER THE SPACE B*
> > *ARE "THE SAME" AS*
> > *MAPS FROM B TO THE SPACE OF ALL SMALL SPACES*
>
> *But you were secretly doing this all along! After all, when you wrote:*
>
> > *n-GROUPOIDS FIBERED OVER AN n-GROUPOID B ARE "THE SAME" AS WEAK $(n+1)$-FUNCTORS FROM B TO THE $(n+1)$-GROUPOID $n\mathsf{Gpd}$*
>
> *you simply used "$n\mathsf{Gpd}$" to abbreviate "OF ALL SMALL n-GROUPOIDS". So there are really no new size issues at the ω level.*
>
> *(Exercise for those that like this sort of thing: Do we need to state that B and F are small?)*
>
> *– Toby*

Thanks for clearing this up. I prefer not to distract people with size issues, so I didn't mention them until "the space of all spaces" walked in the door, at which point I figured alarm bells would start ringing for lots of people. But, it was already hiding in the "the $(n+1)$-groupoid of all n-groupoids". I prefer to use a new Grothendieck universe for each level of the n-categorical hierarchy, to justify such expressions. I guess ω-categories then require an ω-th universe.

Yes, I saw that silly sign in front of the dinosaurs, though I didn't understand its full meaning until now. It wasn't there the last time I visited. Apparently the new owners have decided to enlist this nice roadside attraction as part of the propaganda campaign for creationism. They actually believe Adam and Eve walked with dinosaurs in Eden — as one biologist put it, "they think The Flintstones is a documentary".

It's sad how just as the magnificent history of the universe is becoming vividly clear, some want to truncate it to a pitifully human scale — and then claim *that* was God's work.

Next, David Corfield had some questions about the "space of all spaces", which I answered in this email:

> *Dear David -*
>
> *You wrote:*
>
> > *Hi,*
> >
> > > *SPACES FIBERED OVER THE SPACE B ARE "THE SAME" AS MAPS FROM B TO THE SPACE OF ALL SMALL SPACES*
> >
> > *Is there another handle on this, other than*
> >
> > > *OMEGA-GROUPOIDS FIBERED OVER THE OMEGA-GROUPOID B ARE "THE SAME" AS WEAK OMEGA-FUNCTORS FROM B TO THE OMEGA-GROUPOID OF ALL SMALL OMEGA-GROUPOIDS ?*
> >
> > *Presumably B must be small, and the spaces fibered over it.*
>
> *It suffices for the fibers to be small, so if you want a really nitpicky motto:*
>
> > *SMALL ω-GROUPOIDS FIBERED OVER THE ω-GROUPOID B*
> > *ARE "THE SAME" AS*
> > *WEAK ω-FUNCTORS FROM B TO THE ω-GROUPOID OF ALL*
> > *SMALL ω-GROUPOIDS*
>
> > *Do you and Jim have other intuitions about THE SPACE OF ALL SMALL SPACES?*
>
> *One can describe it in a completely precise and rigorous way. It's the disjoint union over all homotopy types of small spaces F of the classifying spaces $B(\mathrm{Aut}(F))$. Here $\mathrm{Aut}(F)$ is the topological group of homotopy self-equivalences of F.*
>
> *Note: the "largeness" of this space is solely due to it being a disjoint union of a proper class of connected components. When we map any small space to it, the map can only hit a set's worth of components. So, it's not really scary.*
>
> *And, if we map a connected small space X to it, we get a map*
>
> $$X \to B(\mathrm{Aut}(F))$$
>
> *for some F, which is just what you need to get an F-bundle over X.*

> > *Like, is it one of your FREE SUCH-AND-SUCHES?*
>
> *I don't know a description like that offhand, since "free" suggests a left universal property, and the space of all (small) spaces mainly has a right universal property, which describes maps* into *it.*
>
> *Namely: maps from X into the space of all spaces are "the same as" fibrations over X.*
>
> *I could make this completely precise, but it's probably not worth bothering here; one just needs suitable equivalence relations.*

Next, Peter May wrote:

> *In his posting today, John Baez advertised the slogan:*
>
> > *FIBRATIONS OVER THE BASE SPACE B WITH FIBER F*
> > *ARE "THE SAME" AS*
> > *HOMOMORPHISMS SENDING LOOPS IN B TO AUTOMORPHISMS OF F.*
>
> *He hedged it with a "dose of vagueness", but in fact I proved a completely precise and general version of exactly this result in:*
>
> 16) *Peter May, "Classifying spaces and fibrations",* Memoirs AMS ***155****, Jan. 1975.*
>
> *Using Moore loops on B, LB, one has a topological monoid, and one also has the topological monoid $\mathrm{Aut}(F)$ of homotopy equivalences of F. A "transport" is a homomorphism of topological monoids from LB to $\mathrm{Aut}(F)$. Allowing F to vary by a homotopy equivalence, one can define an equivalence relation on transports such that the equivalence classes are in natural bijective correspondence with the equivalence classes of "fibrations over the base space B with fiber F". One can generalize the context by allowing fibers in some nice category and prove the same result. See opus cit, Theorem 14.2, page 83. That was over 30 years ago, so naturally I wasn't thinking about categorification, but I would imagine that the methods categorify.*

Here and below I've taken the liberty of numbering the references to papers, so it's easier to find them in my table of contents for This Week's Finds.

Jim Stasheff wrote:

> *Even more ancient:*
>
> 17) *James Stasheff, "Parallel transport in fibre spaces",* Bol. Soc. Mat. Mexicana *(1968), 68–86.*
>
> *Unfortunately that's a hard paper to get a hold of.*
>
> *Somewhat related is:*
>
> 18) *James Stasheff, "Associated fibre spaces",* Michigan Math. Journal ***15*** *(1968), 457–470.*

and at the survey level:

> 19) *James Stasheff, "H-spaces and classifying spaces, I-IV",* AMS Proc. Symp. Pure Math. ***22*** *(1971), 247–272.*

Of course, as you might expect, I describe things in terms of A_∞-morphisms from the space of loops into $\mathrm{Aut}(F)$ of homotopy equivalences of F.

Now that some of us are comfortable with A_∞-cats, categorification should proceed, perhaps with some technical details.

jim

Ronnie Brown wrote:

John Baez gave an interesting account of nonabelian cohomology and extension theory. Here are a few more references with which I have been involved, all using crossed complexes and free crossed resolutions:

> 20) *Ronald Brown and P. J. Higgins, "Crossed complexes and non-abelian extensions",* Category theory proceedings, Gummersbach, 1981, *(ed. K.H. Kamps et al) Lecture Notes in Math.* ***962*** *Springer, Berlin, 1982, pp. 39–50.*

This generalises classical Schreier theory to extensions of groupoids.

> 21) *Ronald Brown and O. Mucuk, "Covering groups of non-connected topological groups revisited",* Math. Proc. Camb. Phil. Soc. ***115*** *(1994) 97–110. Also available as* `math.AT/0009021`.

This relates the theory of covering topological groups of non connected topological groups to the classical theory of extensions and obstructions to a Q-kernel with an invariant in $H3$. It uses the properties of the internal hom in crossed complexes $\mathrm{CRS}(F,C)$, and exact sequences derived from a fibration $C \to D$ and the induced fibration on $\mathrm{CRS}(F,-)$.

> 22) *Ronald Brown and Timothy Porter, "On the Schreier theory of non-abelian extensions: generalisations and computations",* Proceedings Royal Irish Academy ***96A*** *(1996), 213–227. Also available at* `http://www.informatics.bangor.ac.uk/public/math/research/ftp/algtop/rb/nonabex4.ps.gz`

This establishes a generalisation of the Schreier theory in two ways (but only for groups). One is using coefficients in a crossed module, following Dedecker's key ideas, as in the references in John's account. Second it shows how to compute with such extensions

$$A \to E \to G$$

in terms of presentations of the group G. This involves identities among relations for the presentation, as shown originally by Turing in

23) *Alan Turing, "The extensions of a group",* Compositio Mathematica ***5*** *(1938), 357–367.*

The advantage of this method is that one can actually do sums, even when the group G may be infinite. The example given by us is G = trefoil group with two generators x, y and relation $x^3 = y^2$. This presentation has no identities among relations, and so the calculation is especially simple. Equivalence of extensions is described in terms of homotopies of morphisms of crossed complexes, and this relates the ideas to other forms of homological or homotopical algebra.

An advantage of this approach is the ability to calculate some small free crossed resolutions of some groups: this is one reason for using crossed complexes. Note that a convenient monoidal closed structure on the category of crossed complexes has been explicitly written down, and this allows convenient calculation and representations of homotopies, using the 'unit interval' groupid, as a crossed complex.

One of the problems I have with the globular approach is the difficulty of writing down homotopies, and higher homotopies. For example, Ilhan Icen and I found it difficult to rewrite in terms of group-groupoids the well known Whitehead theory of automorphisms of crossed modules, explained for the crossed modules of groupoids case in:

24) *Ronald Brown and Ilhan Icen, "Homotopies and automorphisms of crossed modules of groupoids",* Applied Categorical Structures ***11*** *(2003) 185–206. Also available as* `math.CT/0008117`*.*

It looks as if it would be better expressed in terms of automorphisms of 2-groupoids: good marks to anyone who writes it down in that way!

One knows such homotopies of globular ∞-groupoids exist because globular ∞-groupoids are equivalent to crossed complexes:

25) *Ronald Brown and P. J. Higgins, "The equivalence of ∞-groupoids and crossed complexes",* Cah. Top. Geom. Diff. ***22*** *(1981) 371–386.*

(This paper contains an early definition of a (strict, globular) ∞ category.)

This raises a question: what is the crossed complex associated to the free globular groupoid on one generator of dimension n? I have a round-about sketch proof, using also cubical theory, and a van Kampen theorem, that it is the fundamental crossed complex of the n-globe. Does anyone have a purely algebraic proof?

The idea of singular nonabelian cohomology of a space X with coefficients in a crossed complex C is given in:

26) *Ronald Brown and P. J. Higgins, "The classifying space of a crossed complex",* Math. Proc. Camb. Phil. Soc. ***110*** *(1991) 95–120.*

This cohomology is given by $[\Pi SX, C]$, homotopy classes of maps from the fundamental crossed complex of the singular complex of X, to C. There is also a Cech version (current work with Jim Glazebrook and Tim Porter).

> *An interesting problem is to classify extensions of crossed complexes!*
>
> *There is an interesting account of extensions of principal bundles and transitive Lie groupoids by Androulidakis, developing work of Mackenzie, at:*
>
> *27) Iakovos Androulidakis, "Classification of extensions of principal bundles and transitive Lie groupoids with prescribed kernel and cokernel",* J. Math. Phys. ***45*** *(2004), 3095–4012. Also available as* `math.DG/0402007`*.*
>
> *(not using crossed complexes).*
>
> *Ronnie Brown*
> `http:www.bangor.ac.uk/r.brown`

Finally, here's my reply to a bemused comment by Jim Stasheff:

> *Jim Stasheff wrote:*
>
>> *John and anyone else who cares to weigh in, here are some comments from the purely topological or rather homotopy theory side:*
>>
>> *For both bundles and fibrations (e.g. over a paracompact base), your last slogan is the oldest:*
>>
>> *FIBRATIONS WITH FIBER F OVER THE SPACE B ARE "THE SAME" AS MAPS FROM B TO THE CLASSIFYING SPACE OF $\mathrm{AUT}(F)$*
>>
>> *"the same as" referring to homotopy classes.*
>
> *It's certainly old, but I mentioned another that may be older:*
>
>> *COVERING SPACES OF B WITH FIBER F*
>> *ARE "THE SAME" AS*
>> *HOMOMORPHISMS FROM THE FUNDAMENTAL GROUP OF B TO AUTOMORPHISMS OF F*
>
> *although one usually sees this special case (which I didn't bother to mention):*
>
>> *CONNECTED COVERING SPACES OF B WITH FIBER F*
>> *ARE "THE SAME" AS*
>> *TRANSITIVE ACTIONS OF THE FUNDAMENTAL GROUP OF B ON F*
>
> *which is usually disguised as follows:*
>
>> *CONNECTED COVERING SPACES OF B*
>> *ARE "THE SAME" AS*
>> *SUBGROUPS OF THE FUNDAMENTAL GROUP OF B*

Anyway, I wasn't trying to present things in historical order. I was trying present them roughly in order of increasing "dimension", starting with extensions of groups, then going up to 2-groups, then expanding out to groupoids, then going up to n-groupoids, and finally ω-groupoids. . . which are the same as homotopy types!

And here, as usual, the n-category theorists meet up with the topologists — and find that the topologists have already done everything there is to do with ω-groupoids . . . but usually by thinking of them of them as spaces*, rather than ω-groupoids!*

It's sort of like climbing a mountain, surmounting steep cliffs with the help of ropes and other equipment, and then finding a Holiday Inn on top and realizing there was a 4-lane highway going up the other side.

So, as usual, the main point of calling homotopy types "ω-groupoids" instead of "spaces" is not to reinvent topology, but to see how ideas from topology generalize to n-category theory. Think of spaces as ω-groupoids, but use those as a springboard for ω-categories — or at least n-categories, perhaps just for low values of n if one is feeling tired.

In the case at hand, the ω-groupoidal slogan:

> *FIBRATIONS OF ω-GROUPOIDS WITH FIBER F AND BASE B*
> *ARE "THE SAME" AS*
> *WEAK ω-FUNCTORS FROM B TO $\mathrm{AUT}(F)$*

is just a reformulation of:

> *FIBRATIONS WITH FIBER F OVER THE SPACE B*
> *ARE "THE SAME" AS*
> *MAPS FROM B TO THE CLASSIFYING SPACE OF $\mathrm{AUT}(F)$*

but it suggests a grandiose generalization:

> *FIBRATIONS OF ω-CATEGORIES WITH BASE B*
> *ARE "THE SAME" AS*
> *WEAK ω-FUNCTORS FROM B^{op} TO THE ω-CATEGORY OF ω-CATEGORIES!*

I guess we can thank Grothendieck for making precise and proving a version of this with ω replaced by $n = 1$:

> *FIBRATIONS OF CATEGORIES WITH BASE B*
> *ARE "THE SAME" AS*
> *WEAK 2-FUNCTORS FROM B^{op} TO THE 2-CATEGORY OF CATEGORIES.*

More recently people have been thinking about the $n = 2$ case, especially Claudio Hermida:

28) Claudio Hermida, "Descent on 2-fibrations and strongly 2-regular 2-categories", Applied Categorical Structures ***12** (2004), 427–459. Also available at* `http://maggie.cs.queensu.ca/chermida/papers/2-descent.pdf`

He states something that hints at this:

> *FIBRATIONS OF 2-CATEGORIES WITH BASE B*
> *ARE "THE SAME" AS*
> *WEAK 3-FUNCTORS FROM B^{op} TO THE WEAK 3-CATEGORY OF 2-CATEGORIES.*

Here I'm using B^{op} to mean B with the directions of both 1-morphisms and 2-morphisms reversed. Hermida follows tradition and calls this B^{coop} — "op" for reversing 1-morphisms and "co" for reversing 2-morphisms. But, it looks like we'll be needing to reverse all kinds of morphisms in n-category case, so we'll need a short name for that.

Best,
jb

"Hah, what a fantastic night," Gunn said. "Arcturus is absolutely steady." He leaned back, put his elbows on the rail of the lift, and looked up at the sky. His glasses glinted faintly in the starlight. "Astronomy is not terribly important," he said. He fell silent for a moment. "Although it is one of the more important things we do as a species," he said. He did not see any contradiction there.

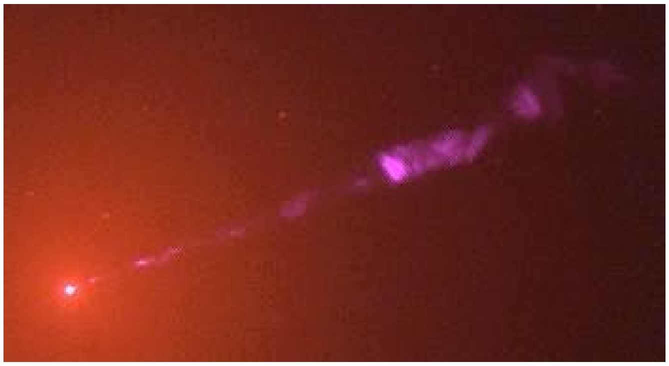
— *Richard Preston,* First Light

Week 224

December 14, 2005

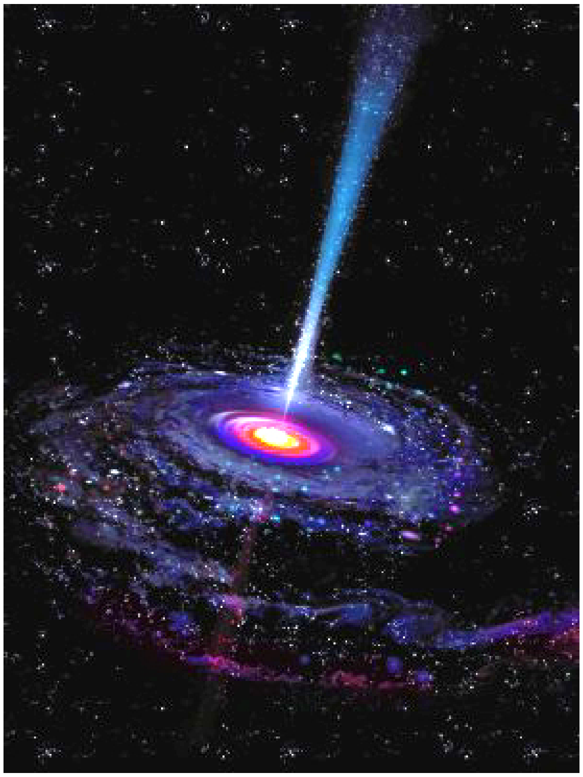
This week I want to mention a couple of papers lying on the interface of physics, topology, and higher-dimensional algebra. But first, some astronomy pictures... and a bit about the mathematical physicist Hamilton!

I like this photo of a jet emanating from the black hole in the center of galaxy M87:



1) NASA and John Biretta, "M87", `http://hubblesite.org/newscenter/newsdesk/archive/releases/2005/12/image/o`

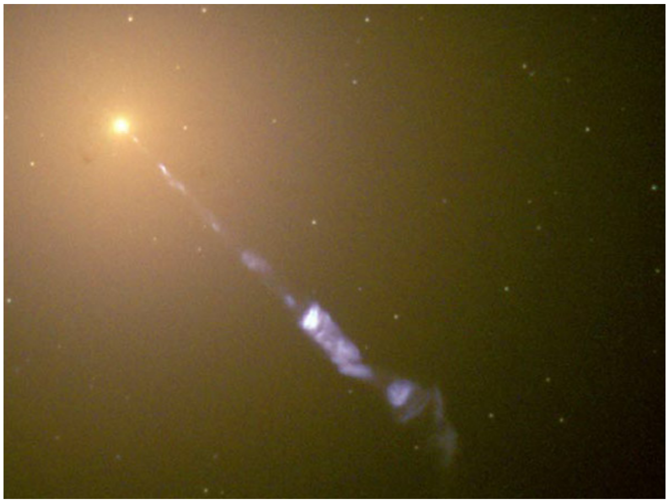
M87 is a giant elliptical galaxy. It's long been known as a powerful radio source, and now we know why: there's a supermassive black hole in the center, about 3 billion times the mass of our Sun. As matter spirals into this huge black hole, it forms an "accretion disk", and some gets so hot that it shoots out in a jet, as envisioned here:



2) NASA, MAXIM: "Micro-Arcsecond X-ray Imaging Mission", `http://maxim.gsfc.nasa.gov/docs/science/science.html`

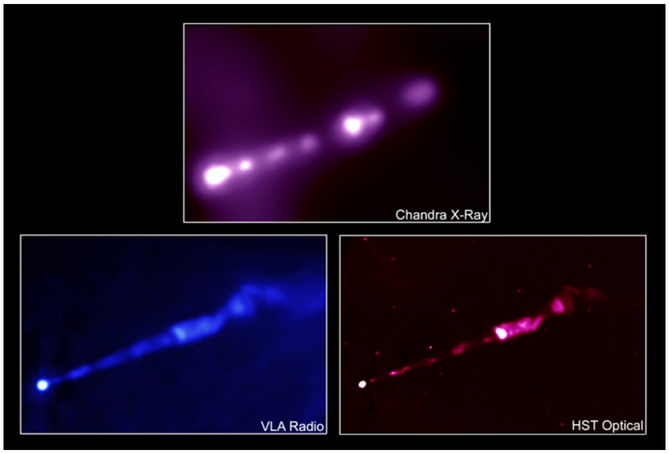
Accretion disks and jets are common at many different scales in our universe. They're just nature's way of letting a bunch of matter fall in under its own gravitation while losing angular momentum and energy. We see them when dust clouds collapse to form stars, we see them when black holes sucks in mass from companion stars, and they're probably also responsible for slow γ ray bursts as huge stars collapse when they run out of fuel — see "Week 204" for that story.

But, among the biggest accretion disks and jets are those surrounding supermassive black holes in the middle of galaxies. These are probably responsible for all the "active galactic nuclei" or "quasars" that we see. In the case of M87 the jet is enormous: 5000 light years long! To get a sense of the scale, look at the small white specks away from the jet in the next picture. These are globular clusters: clusters containing between ten thousand and a million stars.



3) A jet from galaxy M87, "Astronomy Picture of the Day, July 6, 2000", `http://antwrp.gsfc.nasa.gov/apod/ap000706.html`

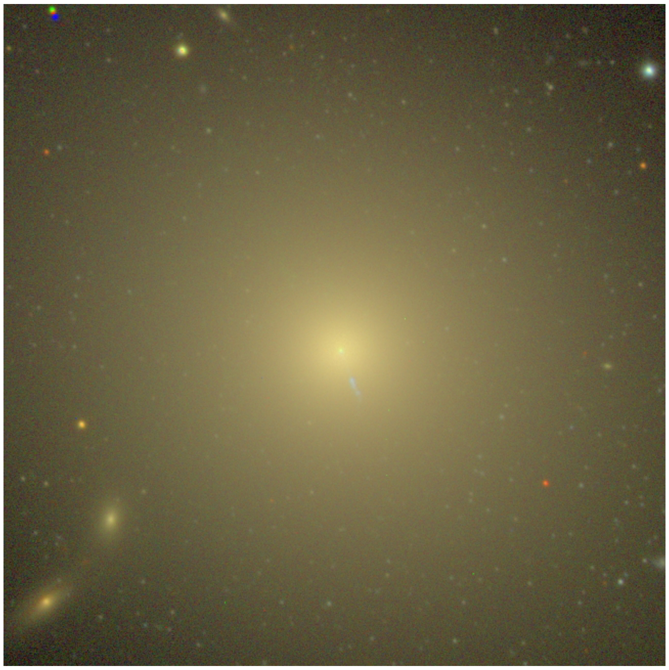
The jet in M87 is so hot that it emits not just radio waves and visible light, but even X-rays, as seen by the Chandra X-ray telescope:



4) "M87: Chandra sheds light on the knotty problem of the M87 jet", `http://chandra.harvard.edu/photo/2001/0134/`

It seems the jet consists mainly of electrons moving at relativistic speeds, focused by the magnetic field of the accretion disk. They come in blobs called "knots". People can actually see these blobs moving out, getting brighter and dimmer.

In fact, many galaxies have super-massive black holes at their centers with jets like this one. The special thing about M87 is that it's fairly nearby, hence easy to see. M87 is the biggest galaxy in the Virgo Cluster. This is the closest galaxy cluster to us, about 50 million light years away. That sounds pretty far, but it's only 1000 times the radius of the Milky Way. If the Milky Way were a pebble, M87 would be only a stone's throw away. So, even amateur astronomers — really good ones, at least — can take photos of M87 that show the jet. But here's a high-quality picture produced by Robert Lupton using data from the Sloan Digital Sky Survey — you can see the jet in light blue:

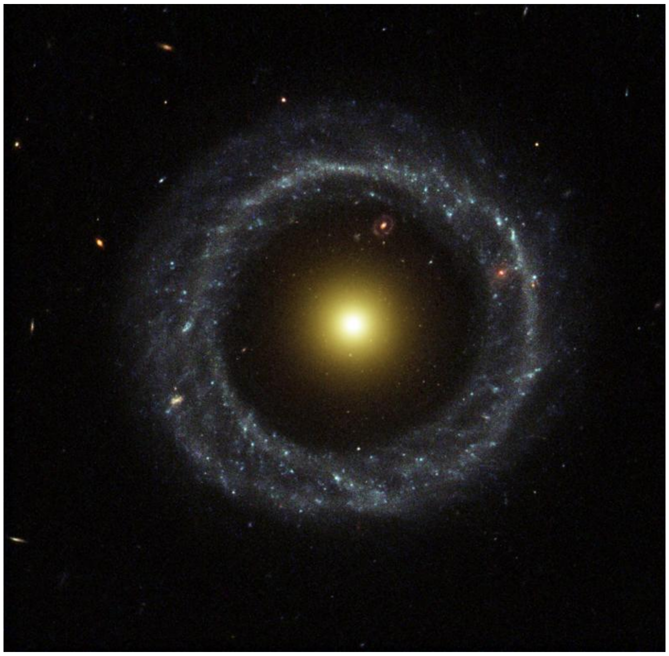


5) Robert Lupton and the Sloan Digital Sky Survey Consortium, "The central regions of M87", `http://www.astro.princeton.edu/~rhl/PrettyPictures/`

Backing off a bit further, let's take a look at the Virgo Cluster. It contains over a thousand galaxies, but we can tell it's fairly new as clusters go, since it consists of a bunch of "subclusters" that haven't merged yet. Our galaxy, and indeed the whole Local Group to which it belongs, is being pulled towards the Virgo Cluster and will eventually join it. Here's a nice closeup of part of the Virgo Cluster:

6) Chris Mihos, Paul Harding, John Feldmeier and Heather Morrison, "Deep imaging of the Virgo Cluster", `http://burro.astr.cwru.edu/Schmidt/Virgo/`

Finally, just for fun, something unrelated — and more mysterious. It's called "Hoag's object":



7) The Hubble Heritage Project, "Hoag's Object", `http://heritage.stsci.edu/2002/21/`

It's a ring-shaped galaxy full of hot young blue stars surrounding a ball of yellower stars. Nobody knows how it formed: perhaps by a collision of two galaxies? Such collisions are fairly common, but they don't typically create this sort of structure.

The weirdest part is that inside the ring, in the upper right, you can see *another* ring galaxy in the distance! Maybe an advanced civilization over there enjoys this form of art? Probably not, but if it turns out to be true, you heard it here first.

Anyway... back here on Earth, in the summer of 2004, I visited Dublin for a conference on general relativity called GR17. As recounted in "Week 207", this was where Hawking admitted defeat in his famous bet with John Preskill about information loss due to black hole evaporation. In August of this year, Hawking finally came out with a short paper on the subject:

8) Stephen W. Hawking, "Information loss in black holes", available as hep-th/0507171.

I spent a lot of time talking to physicists, but I also wandered around Dublin a bit. Besides listening to some great music at a pub called Cobblestones — Kevin Rowsome plays a mean uilleann pipe! — and tracking down some sites mentioned in James Joyce's novel "Ulysses", I went with Tevian Dray on a pilgrimage to Brougham Bridge.

Tevian Dray is an expert on the octonions, and Brougham Bridge is where Hamilton carved his famous formula defining the quaternions! Now there is a plaque commemorating this event, which reads:

> *Here as he walked by*
> *on the 16th of October 1843*
> *Sir William Rowan Hamilton*
> *in a flash of genius discovered*
> *the fundamental formula for*
> *quaternion multiplication*
> $i^2 = j^2 = k^2 = ijk = -1$
> *& cut it on a stone of this bridge*

It does't mention that Hamilton had been racking his brain for the entire month of October trying to solve this problem: "flashes of genius" favor the prepared mind. But it's a nice story and a nice place. My friend Tevian Dray took some photos, which you can see here:

9) John Baez, "Dublin", `http://math.ucr.edu/home/baez/dublin/`

It was a bit of a challenge finding Brougham Bridge, since nobody at the main bus station gave us correct information about which bus went there — except the bus driver who finally took us there. So, to ease your way in case you want to make your own pilgrimage, the above page includes directions. And now, thanks to Dirk Schlimm, it also includes a link to a map showing the bridge!

Speaking of Hamilton, Theron Stanford recently sent me an answer to one of life's persistent questions: why is momentum denoted by the letter p?

Since momentum and position play fundamental roles in Hamiltonian mechanics, and they're denoted by p and q, one wonders: could this notation be related to Hamilton's alcoholism in later life? After all, some claim the saying *mind your p's and q's* began as a friendly Irish warning not to imbibe too many pints and quarts! So, maybe he used these letters in his work on physics as a secret plea for help.

Umm... probably not. Just kidding. But in the absence of hard facts, speculation runs rampant. So, I'm glad Stanford provided some of the former, to squelch the latter.

He sent me this email:

> *While Googling various subjects, I came across the following from your Quantum Gravity Seminar notes from 2001:*
>
> > *Again Oz was overcome with curiosity, so mimicking Toby's voice, he asked, "Why do we call the momentum p?"*
> >
> > *The Wiz glared at Toby. "Because m is already taken – it stands for mass! Seriously, I don't know why people call position q and momentum p. All I know is that if you use any other letters, people can tell you're not a physicist. So I urge you to follow tradition on this point."*
>
> *Well, I have an answer. Hamilton, the first physicist to actually understand the importance of the concept of momentum, chose ϖ to stand for momentum (it's not the usual π, but what TeX calls `varpi`, a lower-case ω with a top, kinda like the top of a lower-case τ). Jacobi changed this to p in one of his seminal papers on the subject; he also used q in the same paper to stand for position. In the 1800s (I want to say 1850s, though it might have been a decade or two later) Cayley presented a paper to the Royal Academy in which he says (and I paraphrase), "Well, it seems that p and q are pretty well established now, so that's what I'm going to use."*

So, now the question is why Hamilton chose the letter ϖ for momentum. This variant of π was fairly common in the mathematical literature of the day, so there may be no special explanation. For some further detective work, see:

10) "Hamilton: two mysteries solved", `http://groups.google.com/group/sci.physics/browse_thread/thread/d1b7b4a998682bbb/3a868ae8218a4bca#3a868ae8218a4bca`

Also see equation 12 in this paper for one of the first uses of "ϖ" to mean momentum:

11) William Rowan Hamilton, "Second essay on a general method in dynamics", ed. David R. Wilkins, available at `http://www.maths.tcd.ie/pub/HistMath/People/Hamilton/Dynamics/SecEssay.pdf`

He doesn't say why he chose this letter — it may have been completely random!

Before I turn to higher-dimensional algebra, maybe this is a good time to mention a paper related to the octonions:

12) Jakob Palmkvist, "A realization of the Lie algebra associated to a Kantor triple system", available as `math.RA/0504544`.

In "Week 193" I mentioned how 3-graded Lie algebras come from "Jordan triple systems", and vaguely hinted that 5-graded Lie algebras come from "Kantor triple systems". I explained how the exceptional Lie algebra E_8 gets to be 5-graded, but I didn't really say anything about Kantor triple systems because my understanding of them was so poor. This paper by Palmkvist explains them very clearly! And even better, he shows how the "magic square" Lie algebras F_4, E_6, E_7, and E_8 can be systematically obtained from the octonions, bioctonions, quateroctonions and octooctonions by means of Kantor triple systems.

Now for some mathematical physics that touches on higher-dimensional algebra. If you still don't get why topological field theory and n-categories are so cool, read this thesis:

13) Bruce H. Bartlett, *Categorical aspects of topological quantum field theories*, M.Sc. Thesis, Utrecht University, 2005. Available as `math.QA/0512103`.

It's a great explanation of the big picture! I can't wait to see what Bartlett does for his Ph.D..

If you're a bit deeper into this stuff, you'll enjoy this:

14) Aaron Lauda and Hendryk Pfeiffer, "Open-closed strings: two-dimensional extended TQFTs and Frobenius algebras", available as `math.AT/0510664`.

This paper gives a purely algebraic description of the topology of open and closed strings, making precise and proving some famous results stated without proof by Moore and Segal, which can be seen here:

15) Greg Moore, "Lectures on branes, K-theory and RR charges", *Clay Math Institute Lecture Notes* (2002), available at `http://www.physics.rutgers.edu/~gmoore/clay1/clay1.html`

Lauda and Pfeiffer's paper makes heavy use of Frobenius algebras, developing more deeply some of the themes I mentioned in "Week 174". In a related piece of work, Lauda has figured out how to *categorify* the concept of a Frobenius algebra, and has applied this to 3d topology:

16) Aaron Lauda, "Frobenius algebras and ambidextrous adjunctions", *Theory and Applications of Categories* **16** (2006) 84–122. Also available as `math.CT/0502550`.

Aaron Lauda, "Frobenius algebras and planar open string topological field theories", available as `math.QA/0508349`.

The basic idea behind all this work is a "periodic table" of categorified Frobenius algebras, which are related to topology in different dimensions. For example, in "Week 174" I explained how Frobenius algebras formalize the idea of paint drips on a sheet of rubber. As you move your gaze down a sheet of rubber covered with drips of paint, you'll notice that drips can merge:

```
\ \         / /
 \ \       / /
  \ \     / /
   \ \   / /
    \ \_/ /
     \   /
      | |
      | |
      | |
      | |
      | |
```

but also split:

```
      | |
      | |
      | |
      | |
      | |
     / _ \
    / / \ \
   / /   \ \
  / /     \ \
 / /       \ \
/ /         \ \
```

In addition, drips can start:

```
 _
| |
| |
| |
| |
| |
| |
| |
| |
| |
```

but also end:

```
| |
| |
| |
| |
| |
| |
| |
| |
|_|
```

In a Frobenius algebra, these four pictures correspond to four operations called "multiplication" (merging), "comultiplication" (splitting), the "unit" (starting) and the "counit" (ending). Moreover, these operations satisfy precisely the relations that you can prove by warping the piece of rubber and seeing how the pictures change. For example, there's the associative law:

```
\ \    / /    / /      \ \    \ \    / /
 \ \  / /    / /        \ \    \ \  / /
  \ \/ /    / /          \ \    \ \/ /
\   /   / /            \ \    \  /
 \ \   / /              \ \   / /
  \ \_/ /                \ \_/ /
   \   /                  \   /
    | |                    | |
    | |                    | |
    | |          =         | |
    | |                    | |
    | |                    | |
    | |                    | |
    | |                    | |
    | |                    | |
```

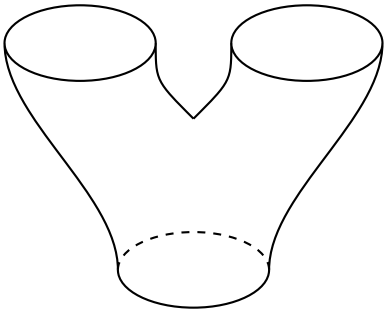
The idea here is that if you draw the picture on the left-hand side on a sheet of rubber, you can warp the rubber until it looks like the right-hand side! There's also the "coassociative law", which is just an upside-down version of the above picture. But the most interesting laws are the "$I = N$" equation:

```
\ \     / /                | |        | |
 \ \   / /                 | |        | |
  \ \_/ /                  | |        | |
   \   /                   |  \       | |
    | |                    |   \      | |
    | |                    | |\ \     | |
    | |                    | | \ \    | |
    | |                    | |  \ \   | |
    | |          =         | |   \ \  | |
    | |                    | |    \ \ | |
    | |                    | |     \ | |
    | |                    | |      \   |
   / _ \                   | |       \  |
  / / \ \                  | |        | |
 / /   \ \                 | |        | |
/ /     \ \                | |        | |
```

and its mirror-image version.

So, the concept of Frobenius algebra captures the topology of regions in the plane! Aaron Lauda makes this fact into a precise theorem in his paper on planar open string field theories, and then generalizes it to consider "categorified" Frobenius algebras where the above equations are replaced by isomorphisms, which describe the *process* of warping the sheet of rubber until the left side looks like the right. You should look at his paper even if you don't understand the math, because it's full of cool pictures.

Lauda and Pfeiffer's paper goes still further, by considering these paint stripes as "open strings", not living in the plane anymore, but zipping around in some spacetime of high dimension, where they might as well be abstract 2-manifolds with corners. Following Moore and Segal, they also bring "closed strings" into the game, which form a Frobenius algebra of their own, where the multiplication looks like an upside-down pair of pants:



These topological closed strings are the subject of Joachim Kock's book mentioned in "Week 202"; they correspond to *commutative* Frobenius algebras. The fun new stuff comes from letting the open strings and closed strings interact.

You can read more about Lauda and Pfeiffer's work at Urs Schreiber's blog:

17) Urs Schreiber, "Lauda and Pfeiffer on open-closed topological strings", `http://golem.ph.utexas.edu/string/archives/000680.html`

In fact, I recommend Schreiber's blog quite generally to anyone interested in higher categories and/or the math of string theory!

Addendum: Bruce Smith, David Rusin and Robert Lupton had some comments about the astronomy section; Urs Schreiber had more to say about the role of Frobenius algebras in string theory.

Bruce Smith picked up on my comment about accretion disks being common at many different scales, and wondered what the smallest accretion disks are. We talked about it and agreed that hurricanes, tornados, dust devils and whirlpools are *related* phenomena, but not true accretion disks.

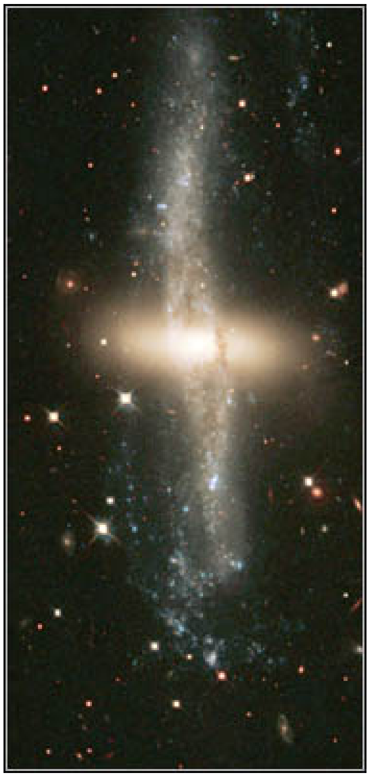
Given this, the smallest accretion disks I know are those that led to the formation of planets in our Solar System, and perhaps even some moons. These probably began as eddies in the bigger accretion disk that became our Sun. The Sun is about 300,00 times heavier than the Earth, and the super-massive black hole in M87 is about 3 billion times heavier than the Sun, so we're seeing accretion disks that differ in mass by a factor of a trillion!

David Rusin's reaction to Hoag's object was:

> *Cool. But what are the chances that there would be not just one but TWO fascinating objects which have a significant plane of symmetry, which "just happens" to be perpendicular to our line of sight?*

He asked how many ring galaxies are known!

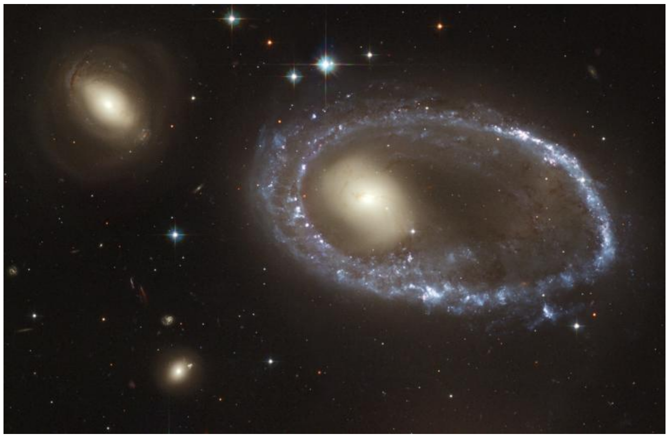
I checked and read there are 100 known "polar-ring galaxies". Here's a nice one called NGC 4650:



18) "Ring around a galaxy", *HubbleSite News Archive*, May 6, 1999, `http://hubblesite.org/newscenter/newsdesk/archive/releases/1999/16/image/a`

I can imagine this thing looking like Hoag's object if we viewed it head-on.

Here's another ring galaxy, called AM 0644-741:



19) "The lure of the rings", *Hubblesite News Archive*, April 22, 2004, `http://hubblesite.org/newscenter/newsdesk/archive/releases/2004/15/image/a`

It's the result of a collision involving a galaxy that's not in this picture. So, maybe Hoag's object is just a specially pretty case of a galaxy collision!

Robert Lupton referred me to a picture that covers more of the Virgo Cluster — but the file is huge, so I won't include it here:

20) Doug Finkbeiner and the Sloan Digital Sky Survey Consortium, "Some pretty objects as observed by the SDSS: Virgo Cluster", `http://www.astro.princeton.edu/~rhl/dfink`

See the lower right corner for the picture called "virgobig".

Here's what Urs Schreiber had to say about Frobenius algebras and string theory:

> *John Baez wrote:*
>
> > *[. . .] Following Moore and Segal, they also bring "closed strings" into the game, which form a Frobenius algebra of their own, where the multiplication looks like an upside-down pair of pants: [. . .]*
>
> *I would like to make the following general comment on the meaning of Frobenius algebras in 2-dimensional quantum field theory.*
>
> *Interestingly,* non-*commutative Frobenius algebras play a role even for closed strings, and even if the worldhseet theory is not purely topological.*
>
> *The archetypical example for this is the class of 2D TFTs invented by Fukuma, Hosono and Kawai. There one has a non-commutative Frobenius algebra which describes not the splitting/joining of the entire worldsheet, but rather the splitting/joining of edges in any one of its dual triangulations. It is the* center *of (the Morita class of) the noncommutative Frobenius algebra decorating dual triangulations which is the commutative Frobenius algebra describing the closed 2D TFT.*
>
> *One might wonder if it has any value to remember a non-commutative Frobenius algebra when only its center matters (in the closed case). The point is that the details of the non-commutative Frobenius algebra acting in the "interior" of the world sheet affects the nature of "bulk field insertions" that one can consider and hence affects the (available notions of) n-point correlators of the theory, for $n > 0$.*
>
> *This aspect, however, is pronounced only when one switches from 2D topological field theories to conformal ones.*
>
> *The fascinating thing is that even 2D* conformal *field theories are governed by Frobenius algebras. The difference lies in different categorical internalization. The Frobenius algebras relevant for CFT don't live in* Vect*, but in some other (modular) tensor category, usually that of representations of some chiral vertex operator algebra. It is that ambient tensor category which "knows" if the Frobenius algebra describes a topological or a conformal field theory (in 2D) — and which one.*

Of course what I am referring to here is the work by Fjelstad, Froehlich, Fuchs, Runkel, Schweigert and others. I can recommend their most recent review which will appear in the Streetfest proceedings. It is available as `math.CT/0512076`*.*

The main result is, roughly, that given any modular tensor category with certain properties, and given any (symmetric and special) Frobenius algebra object internal to that category, one can construct functions on surfaces that satisfy all the properties that one would demand of an n-point function of a 2D (conformal) field theory.

If we define a field theory to be something not given by an ill-defined path integral, but something given by its set of correlation functions, then this amounts to constructing a (conformal) field theory.

This result is achieved by first defining a somewhat involved procedure for generating certain classes of functions on marked surfaces, and then proving that the functions generated by this procedure do indeed satisfy all the required properties.

In broad terms, the prescription is to choose a dual triangulation of the marked worldsheet whose correlation function is to be computed, to decorate its edges with symmetric special Frobenius algebra objects in some modular tensor category, to decorate its vertices by product and coproduct morphisms of this algebra, to embed the whole thing in a certain 3-manifold in a certain way and for every boundary or bulk field insertion to add one or two threads labeled by simple objects of the tensor category which connect edges of the chosen triangulation with the boundary of that 3-manifold. Then you are to hit the resulting extended 3-manifold with the functor of a 3D TFT and hence obtain a vector in a certain vector space. This vector, finally, is claimed to encode the correlation function.

This procedure is deeply rooted in well-known relations between 3-(!)-dimensional topological field theory, modular functors and modular tensor categories and may seem very natural to people who have thought long enough about it. It is already indicated in Witten's paper on the Jone's polynomial, that 3D TFT (Chern-Simons field theory in that case) computes conformal blocks of conformal field theories on the boundaries of these 3-manifolds. To others, like me in the beginning, it may seem like a miracle that an involved and superficially ad hoc procedure like this has anything to do with correlations functions of conformal field theory in the end.

In trying to understand the deeper "meaning" of it all I played around with the idea that this prescription is really, to some extent at least, the "dual" incarnation of the application of a certain 2-functor to the worldsheet. Namely a good part of the rough structure appearing here automatically drops out when a 2-functor applied to some 2-category of surfaces is "locally trivialized". I claim that any local trivialization of a 2-functor on some sort of 2-category of surface elements gives rise to a dual triangulation of the surface whose edges are labeled by (possibly a generalization of) a Frobenius algebra object and whose vertices are labeled by (possibly a generalization of) product and coproduct operations. There is more data in a locally trivialized 2-functor, and it seems to correctly

> *reproduce the main structure of bulk field insertions as appearing above. But of course there is a limit to what a 2-functor can know about a structure that is inherently 3-dimensional.*
>
> *I have begun outlining some of the details that I have in mind here:*
>
> *http://golem.ph.utexas.edu/string/archives/000697.html*
>
> *This has grown out of a description of gerbes with connective structure in terms of transport 2-functors. Note that in what is called a* bundle *gerbe we also do have a certain product operation playing a decisive role. Bundle gerbes can be understood as "pre-trivializations" of 2-functors to* Vect*:*
>
> *http://golem.ph.utexas.edu/string/archives/000686.html*
>
> *and the product appearing is one of the Frobenius products mentioned above. For a bundle gerbe the coproduct is simply the inverse of the product, since this happens to be an isomorphism. The claim is that 2-functors to* Vect *more generally give rise to non-trivial Frobenius algebras when locally trivialized.*
>
> *This is work in progress and will need to be refined. I thought I'd mention it here as a comment to John's general statements about how Frobenius algebras know about 2-dimensional physics. I am grateful for all kinds of comments.*

Here's the paper Urs refers to:

21) Ingo Runkel, Jens Fjelstad, Jurgen Fuchs and Christoph Schweigert, "Topological and conformal field theory as Frobenius algebras", available as `math.CT/0512076`.

Here's how you do it:
First you're obtuse,
Then you intuit,
Then you deduce!

— *Garrison Keillor*

Week 225

December 24, 2005

Happy holidays! I'll start with some gift suggestions for people who put off their Christmas shopping a bit too long, before moving on to this week's astronomy pictures and then some mathematical physics: minimal surfaces.

Back in 2000 I listed some gift ideas in "Week 162". I decided to do it again this year. After all, where else can you read about quantum gravity, nonabelian cohomology, higher categories... and also get helpful shopping tips? But, I put off writing this Week's Finds a bit too late. Oh well.

I just saw this book in a local store, and it's *great*:

1) Robert Dinwiddie, Philip Eales, David Hughes, Ian Nicholson, Ian Ridpath, Giles Sparrow, Pam Spence, Carole Stott, Kevin Tildsley, and Martin Rees, *Universe*, DK Publishing, New York, 2005.

If you like the astronomy pictures you've seen here lately, you'll love this book, because it's *full* of them — all as part of a well-organized, clearly written, information-packed but nontechnical introduction to astronomy. It starts with the Solar System and sails out through the Oort Cloud to the Milky Way to the Local Group to the Virgo Supercluster ... and all the way out and back to the Big Bang!

The only thing this book seems to lack — though I could have missed it — is a 3d map showing the relative scales of our Solar System, Galaxy, and so on. I recommended a wall chart like this back in "Week 162", and my friend Danny Stevenson just bought me one. I'll probably put it up near my office in the math department... gotta keep the kids thinking big!

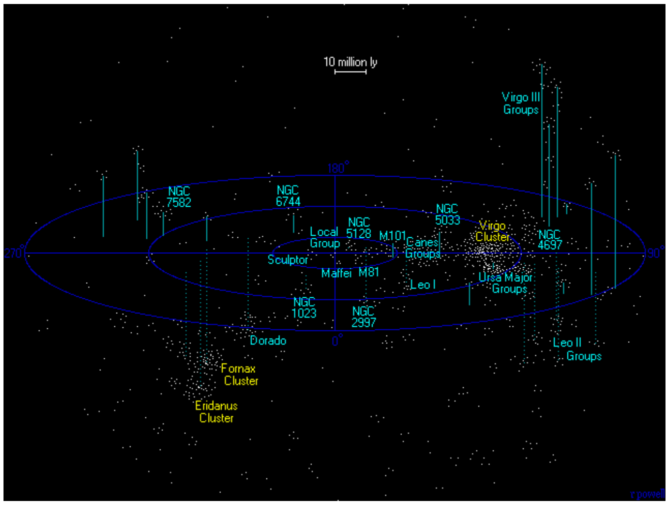
You don't really need to buy a chart like this. You can just look at this website:

2) Richard Powell, "An Atlas of the Universe", `http://www.anzwers.org/free/universe/`

It has nine maps, starting with the stars within 12.5 light years and zooming out repeatedly by factors of 10 until it reaches the limits of the observable universe, roughly 14 billion light years away. Or more precisely, 14 billion years ago!

(The farther we look, the older things we see, since light takes time to travel. The most distant thing we see is light released when hot gas from the Big Bang cooled down just enough to let light through! If we calculate how far this gas would be *now*, thanks to the expansion of the universe, we get a figure of roughly 78 billion light years. But of course we can't see what that gas looks like *now* unless we wait a lot longer. It's a bit confusing until you think about it for a while.)

For example, here are the clusters of galaxies within 100 million lightyears of us:



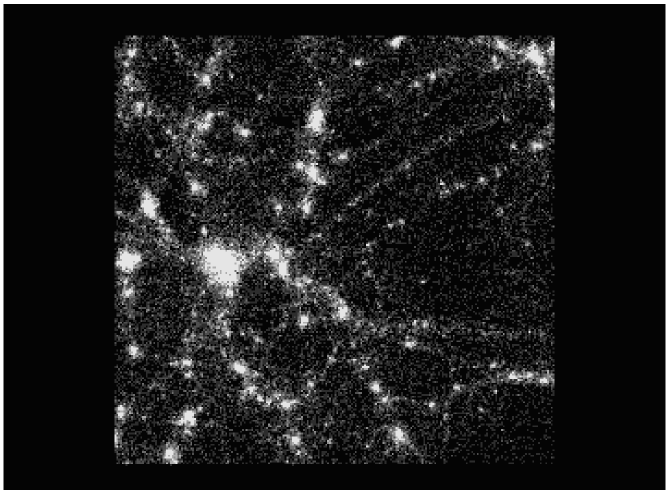
The biggest of these is the Virgo cluster, which I discussed in "Week 224". This contains about 2000 galaxies. The second biggest is the Fornax cluster. The whole agglomeration shown here is called the Virgo Supercluster. Superclusters are among the biggest structures in the universe.

This atlas is fun to browse when you're at your computer. But, if someone you know wants to contemplate the universe in a more relaxing way, try getting them one of these:

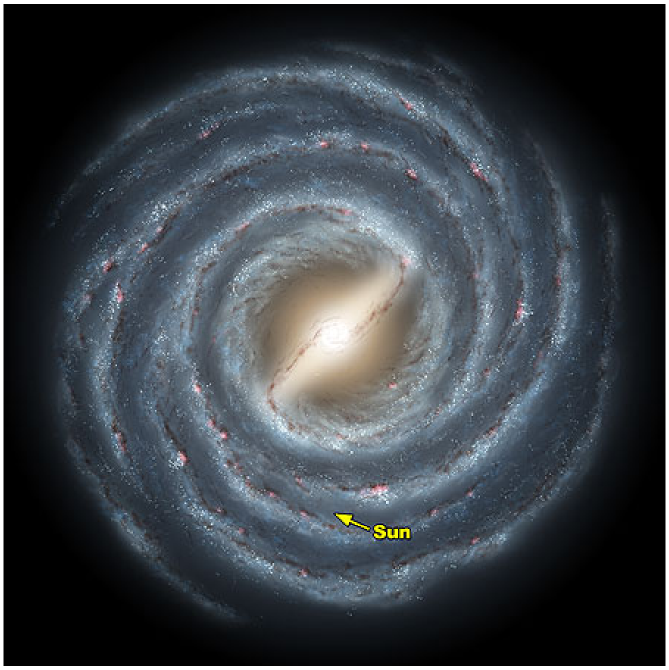
3) Bathsheba Grossman, "Crystal model of a typical 100-megaparsec cube of the universe", `http://www.bathsheba.com/crystal/largescale/`

   "Crystal model of the Milky Way", `http://www.bathsheba.com/crystal/galaxy/`

I found out about these from David Scharffenberg, who owns the Riverside Computer Center nearby — a cool little shop that's decorated with archaic technology ranging from a mammoth slide rule to a gizmo that computes square roots using air pressure. He gave me the 100-megaparsec cube as a present, and it's great! It's lit up from below, and it shows the filaments, sheets and superclusters of galaxies that reign supreme at this distance scale. 100 megaparsecs is about 300 million light years, so this view is a bit bigger than the previous picture:

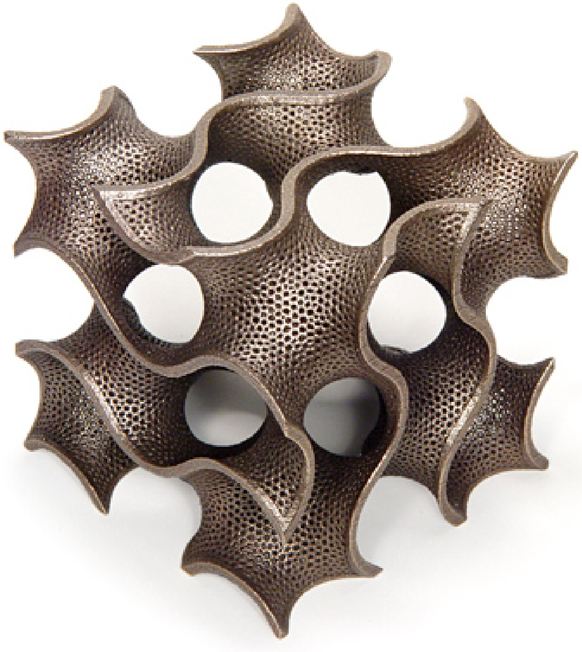


David says Grossman's model of the Milky Way is also nice: it takes into account the latest research, which shows our galaxy is a "barred" spiral! You can see the bar in the middle here:



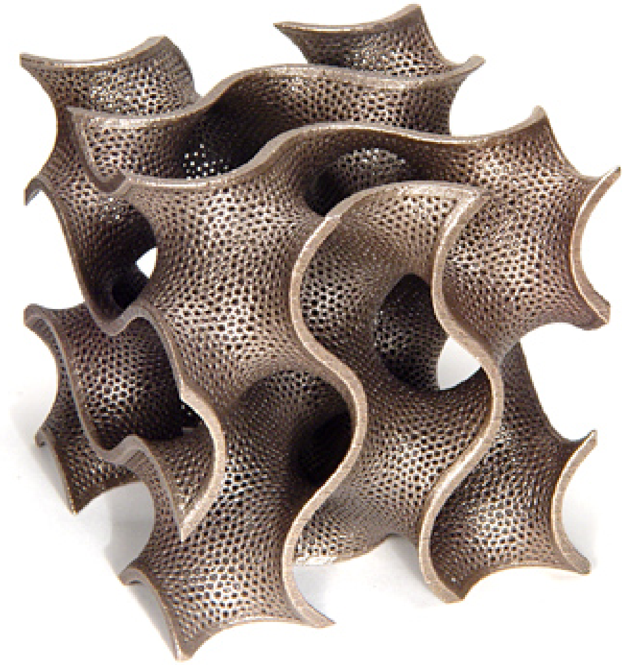
4) R. Hurt, NASA/JPL-Caltech, "Milky Way Bar", `http://www.spitzer.caltech.edu/Media/mediaimages/sig/sig05-010.shtml`

If you really have money to burn, Grossman has also made nice sculptures of mathematical objects like the 24-cell, the 600-cell and Schoen's gyroid — a triply periodic minimal surface that chops 3-space into two parts:



5) Bathsheba Grossman, Mathematical models, `http://www.bathsheba.com/math/`

However, the great thing about the web is that lots of beautiful stuff is free — like these *pictures* of the gyroid:



I explained the 24-cell and 600-cell in “Week 155”. So, let me explain the gyroid — then I need to start cooking up a Christmas eve dinner!

A “minimal surface” is a surface in ordinary 3d space that can’t reduce its area by changing shape slightly. You can create a minimal surface by building a wire frame and then creating a soap film on it. As long as the soap film doesn’t actually enclose any air, it will try to minimize its area — so it will end up being a minimal surface.

If you make a minimal surface this way, it will have edges along the wire frame. A minimal surface without edges is called “complete”. For a long time, the only known complete minimal surfaces that didn’t intersect themselves were the plane, the catenoid, and the helicoid. You get a “catenoid” by taking an infinitely long chain and letting it hang to form a curve called a “catenary”; then you use this curve to form a surface of revolution, which is the catenoid:

6) Eric Weisstein, “Catenoid”, from *Mathworld*, `http://mathworld.wolfram.com/Catenoid.html`

In cylindrical coordinates the catenoid is given by the equation

$$r = c\cosh\left(\frac{z}{c}\right)$$

for your favorite constant c.

A “helicoid” is like a spiral staircase; in cylindrical coordinates it’s given by the equation

$$z = c\theta$$

for some constant c. You can see a helicoid here — and see how it can continuously deform into a catenoid:

7) Eric Weisstein, "Helicoid", from *Mathworld*, `http://mathworld.wolfram.com/Helicoid.html`

In 1987 a fellow named Hoffman discovered a bunch more complete non-self-intersecting minimal surfaces with the help of a computer:

8) D. Hoffman, "The computer-aided discovery of new embedded minimal surfaces", *Mathematical Intelligencer* **9** (1987), 8–21.

Since then people have gotten good at inventing minimal surfaces. You can see a bunch here:

9) GRAPE (Graphics Programming Environment), "Surface overview", `http://www-sfb256.iam.uni-bonn.de/grape/EXAMPLES/AMANDUS/bmandus.html`

10) GANG (Geometry Analysis Numerics Graphics), "Gallery of minimal surfaces", `http://www.gang.umass.edu/gallery/min/`

As you can see, people who work on mininal surfaces like goofy acronyms. If you look at the pictures, you can also see that a minimal surface needs to be locally saddle-shaped. More precisely, it has "zero mean curvature": at any point, if it curves one way along one principal axis of curvature, it has to curve an equal and opposite amount along the perpendicular axis. Supposedly this was proved by Euler.

If we write this requirement as an equation, we get a second-order nonlinear differential equation called "Lagrange's equation" — a special case of the Euler-Lagrange equation we get from any problem in the variational calculus. So, finding new minimal surfaces amounts to finding new solutions of this equation. Soap films solve this equation automatically, but only with the help of a wire frame; it's a lot more work to find minimal surfaces that are complete.

For the theoretical physicist, minimal surfaces also go by another name: *strings!* The "worldsheet" of a bosonic string is just a 2-dimensional surface in spacetime. The equation governing the string's motion just says that the area of this surface can't be reduced by wiggling it slightly. In other words, it's just Lagrange's equation. There's a big difference between string theory and the theory of minimal surfaces, though: in string theory we need to take quantum mechanics into account! (Another big difference is that spacetime is a Lorentzian rather than Riemannian manifold, unless we do a trick called "Wick rotation".)

So, bosonic string theory is about the quantum version of soap films - and "D-branes" serve as the wire frames. But if this reminds you of "spin foams", I should warn you: there are a few big differences. The main thing is that spin foams are background-free: they don't live in spacetime, they *are* spacetime. So, it doesn't make any obvious sense for them to minimize area, though Smolin has suggested it might make an *unobvious* kind of sense. All the fun must happen when the "bubbles" of a spin foam meet along their edges. . . but we don't really know how this should work, to create a foam with the right consistency at large scales.

Anyway. . . .

There are a lot of minimal surfaces that have periodic symmetry in 3 directions, like a crystal lattice. You can learn about them here:

11) Elke Koch, "3-periodic minimal surfaces", `http://staff-www.uni-marburg.de/~kochelke/minsurfs.htm`

In fact, they have interesting relations to crystallography:

12) Elke Koch and Werner Fischer, "Mathematical crystallography", `http://www.staff.uni-marburg.de/~kochelke/mathcryst.htm#minsurf`

I guess people can figure out which of the 230 crystal symmetry groups (or "space groups") can arise as symmetries of triply periodic minimal surfaces, and use this to help classify these rascals. A cool mixture of group theory and differential geometry! I don't get the impression that they have completed the classification, though.

Anyway, Schoen's "gyroid" is one of these triply periodic minimal surfaces. Schoen discovered it before the computer revolution kicked in. He was working for NASA, and his idea was to use it for building ultra-light, super-strong structures:

13) A. H. Schoen, "Infinite periodic minimal surfaces without selfintersections", *NASA Tech. Note No.* **D-5541**, Washington, DC, 1970.

You can learn more about the gyroid here:

14) Eric Weisstein, "Gyroid", from *Mathworld*, `http://mathworld.wolfram.com/Gyroid.html`

Apparently it's the only triply periodic non-self-intersecting minimal surface with "triple junctions". I'm not quite sure what that means mathematically, but I can see them in the picture!

I said that soap films weren't good at creating *complete* minimal surfaces. But actually, people have seen at least approximate gyroids in nature, made from soap-like films:

15) P. Garstecki and R. Holyst, "Scattering patterns of self-assembled gyroid cubic phases in amphiphilic systems", *J. Chem. Phys.* **115** (2001), 1095–1099.

An "amphiphilic" molecule is one that's attracted by water at one end and repelled by water at the other. For example, the stuff in soap. Mixed with water and oil, such molecules form membranes, and really complicated patterns can emerge, verging on the biological. Sometimes the membranes make a gyroid pattern, with oil on one side and water on other! It's a great example of how any sufficiently beautiful mathematical pattern tends to show up in nature somewhere... as Plato hinted in his theory of "forms".

People have fun simulating these "ternary amphiphilic fluids" on computers:

16) Nelido Gonzalez-Segredo and Peter V. Coveney, "Coarsening dynamics of ternary amphiphilic fluids and the self-assembly of the gyroid and sponge mesophases: lattice-Boltzmann simulations", available as `cond-mat/0311002`.

17) Pittsburgh Supercomputing Center, “Ketchup on the grid with joysticks”, `http://www.psc.edu/science/2004/teragyroid/`

The second site above describes the “TeraGyroid Project”, in which people used 17 teraflops of computing power at 6 different facilities to simulate the gyroidal phase of oil/water/amphiphile mixtures and study how “defects” move around in what's otherwise a regular pattern. The reference to ketchup comes from some supposed relationship between these ternary amphiphilic fluids and how ketchup gets stuck in the bottle. I'm not sure ketchup actually *is* a ternary amphiphilic fluid, though!

Hmm. I just noticed a pattern to the websites I've been referring to: first one about a “Milky Way bar”, then one about a “GRAPE”, and now one about ketchup! I think it's time to cook that dinner.

Daydreaming admiring being
Quietly, open the world
I hear the time of the starry sky
Turning over at midnight

— *Massive Attack*

Week 226

February 10, 2006

This month I'm hanging out at CIRM, the "Centre International de Recontres Mathematiques" near Marseille. It's like a little hotel with a classroom, library and computers, set at the edge of a forest that borders the region of limestone hills and cliffs called the Calanques on the coast of the Mediterranean. All they do here is hold conferences: guests come in, listen to math talks, and eat nice French meals prepared by the overworked staff.

This February they're having a conference on logic and computation, with a strong slant towards the use of diagrams:

1) *Geometry of Computation 2006* (Geocal06), `http://iml.univ-mrs.fr/geocal06/`

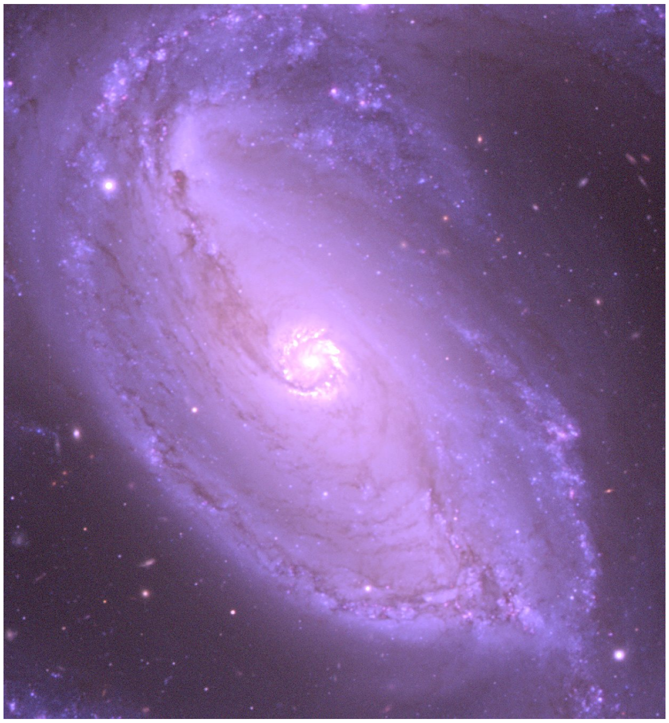
The first week they had lots of talks on "higher-dimensional rewriting", where you use diagrams to draw ways of rewriting ways of rewriting ways of... rewriting strings of symbols. This is inherently connected to n-categories, since n-categories show up whenever you consider "processes that take a process and turn it into another process" — and iterate this notion until your eyes start bugging out.

The first week was pretty cool, and I hope to talk about it someday. The second week I've been recovering from the first week. But today I'll start with some astronomy pictures and some gossip I heard about cryptography, random sequences, and logic.

Let's zoom in on NGC 1097. Here's a photograph of this galaxy taken in ultraviolet light. It's a barred spiral galaxy that's colliding with a smaller elliptical galaxy called NGC 1097A:

2) "NGC 1097", photograph taken by the Galaxy Evolution Explorer, NASA/JPL-Caltech/SSC, `http://www.galex.caltech.edu/MEDIA/2005-02/images.html`

NGC 1097 is called a "Seyfert galaxy" because its center emits lots of radio waves, but not enough to be considered a full-fledged "quasar". As you can see, the center also emits visible light — and stuff is swirling into it!



3) "Spiral galaxy NGC 1097", European Southern Observatory, `http://www.eso.org/outreach/press-rel/pr-2004/pr-28-04.html#phot-35d-04`

Recently the VLT has gotten a really close look at the center of NGC 1097. VLT stands for "Very Large Telescope". It's actually four 8-meter telescopes and three smaller auxiliary ones, all of which can function as a single unit, making it the biggest telescope in the world. It's run by Europeans as part of the "European Southern Observatory" — but it's based on a mountain called Cerro Paranal in the driest part of the Atacama Desert in northern Chile, which makes for wonderful viewing conditions. They had to carry the enormous mirrors across rugged desert roads to build this observatory:

4) "Mirror transport", European Southern Observatory, `http://www.eso.org/outreach/press-rel/pr-1997/phot-35-97.html`

but the view of the night sky up there makes it all worthwhile:

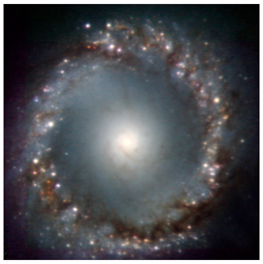
5) "The southern sky above Paranal", European Southern Observatory, `http://www.eso.org/outreach/press-rel/pr-2005/images/phot-40b-05-normal.jpg`

and now they have everything they need to take advantage of that location:

6) "The VLT array on Paranal Mountain", European Southern Observatory, `http://www.eso.org/outreach/press-rel/pr-2000/phot-14a-00-normal.jpg`

So, what did they see when they looked at NGC 1097?

In the middle there seems to be a black hole, emitting radiation as filaments of gas and dust spiral in. This has caused a ring of new stars to form, which are ionizing the hydrogen in their vicinity:



7) "Feeding the monster", European Southern Observatory `http://www.eso.org/outreach/press-rel/pr-2005/phot-33-05.html`

This ring is about 5,500 light-years across! That sounds big, but the galaxy is 45 million light-years away, so this is a stunningly detailed photo. So, we can examine in great detail the process by which black holes eat galaxies.

Anyway. . . .

Talking to the logicians and computer scientists here, I'm hearing lots of gossip that I don't usually get. For example, I just learned that in 2004 there was a successful collision attack on MD5.

Huh? Well, it sounds very technical, but it boils down to this: it means somebody took a function f that is known not to be one-to-one, and found x and y such that $f(x) = f(y)$.

You wouldn't think this would make news! But such functions, called "cryptographic hash functions", are used throughout the computer security business. The idea is that you can take any file and apply the hash function to compute a string of, say, 128 bits. It's supposed to be hard to find two files that give the same bit string. This lets you use the bit string as a kind of summary or "digest" of the file. It's also supposed to be hard to guess the contents of a file from its digest. This lets you show someone the digest of a file without giving away any secrets.

MD5 is a popular hash function invented by Ron Rivest in 1991. People use it for checking the integrity of files: first you compute the digest of a file, and then, when you send the file to someone, you send the digest via some separate channel. If they're worried that the file has been corrupted or tampered with, they compute its digest and compare it to what you sent them.

People also use MD5 and other hash functions for things like keeping passwords safe, digital fingerprinting and copy protection. To illustrate this I'll give a silly example that's easy to understand.

Suppose that this February, Alice proves that $\mathrm{P} = \mathrm{NP}$. She wants to take her time writing a nice paper about it... but she wants to be able to show she was the first to solve this problem, in case anyone else solves it while she's writing her paper.

To do this, she types up a quick note explaining her solution, feeds this file into a cryptographic hash function, and posts the resulting 128-bit string to the newsgroup `sci.math` in an article entitled "I PROVED $\mathrm{P} = \mathrm{NP}$!" A dated copy appears on Google for everyone to see.

Now, if Bob solves the problem in July while Alice is still writing up her solution, Alice can reveal her note. If anyone doubts she wrote her note back in February, they can apply the cryptographic hash function and check that yes, the result matches the bit string she posted on Google!

For this to work, it had better be hard for a nasty version of Alice to take Bob's solution and cook up some note summarizing it whose hash function equals some bit string she already posted.

So, it's very good if the hash function is resistant to a "preimage attack". A preimage attack is where for a given x you have a trick for finding y such that $f(y) = f(x)$.

Nobody has carried out a successful preimage attack on MD5. But, people have carried out a successful "collision attack". This is where you can cook up pairs x, y such that $f(x) = f(y)$. This isn't as useful, since you don't have control over *either* x or y. But, there do exist fiendish schemes for conning people using collision attacks:

8) Magnus Daum and Stefan Lucks, 'Attacking hash functions by poisoned messages: "The Story of Alice and Her Boss" ', `http://www.cits.rub.de/MD5Collisions/`

On this webpage you can see a letter of recommendation for Alice and a letter granting her a security clearance which both have the same MD5 digest! It also explains how Alice could use these to do evil deeds.

For more on cryptographic hash functions and their woes, try these:

9) "Cryptographic hash function", Wikipedia, `http://en.wikipedia.org/wiki/Cryptographic_hash`

Steve Friedl, "An illustrated guide to cryptographic hashes", `http://www.unixwiz.net/techtips/iguide-crypto-hashes.html`

Now, if you're a mathematician, the whole idea of a cryptographic hash function may seem counterintuitive. Just for a change of pace, take another important cryptographic hash function, called SHA-1. This is a function that takes any string of up to 2^{64} bits and gives a digest that's 160 bits long. So, it's just some function

$$f\colon S \to T$$

from a set S of size $2^{2^{64}+1}$ to a set T of size 2^{160}.

The first set is vastly bigger. So, the function f must be far from one-to-one! So why in the world is anyone surprised, much less dismayed, when people find a way to generate two elements in the first set that map to the same element in the second?

One reason is that while the first set is much bigger than the second, the second is mighty big too!

Suppose you're trying to do a preimage attack. Someone hands you an element of T and asks you to find an element of S that maps to it. The brute-force approach, where you keep choosing elements of S, applying the function, and seeing if you get the desired element, will on average take about 2^{160} tries. That's infeasible.

Note that the huge size of S is irrelevant here; what matters most is the size of T.

Or, suppose you're trying a brute-force collision attack. You start looking through elements of S, trying to find two that map to the same thing. On average it will take 2^{80} tries — the square root of the size of T. That's a lot less, but still infeasible.

(Why so much less? This is called the "birthday paradox": it's a lot easier to find two people at a big party who share the same birthday than to find someone with the same birthday as the host.)

Of course, a smarter approach is to use your knowledge about the function f to help you find pairs of elements that map to the same thing. This is what Xiaoyun Wang, Yiqun Lisa Yin, and Hongbo Yu actually did in February 2005. Based on trials with watered-down versions of SHA-1, they argued that they could do a collision attack that would take only 2^{69} tries instead of the expected 2^{80}.

They didn't actually carry out this attack — with the computer power they had, it would have taken them 5 million years. But their theoretical argument was already enough to make people nervous!

And indeed, in August 2005, an improved version of their strategy reduced the necessary number of tries to 2^{63} — in other words, reducing the time to just 78 thousand years, after only a few months of work.

So, people are getting wary of SHA-1. This is serious, because it's a widely used government standard. You can see the algorithm for this function here:

10) "SHA hash functions", Wikipedia, `http://en.wikipedia.org/wiki/SHA_hash_functions`

People still hope that good hash functions exist. They hope that if a function f is cleverly chosen, $f(x)$ will depend in a "seemingly random" way on x, so that given $f(x)$, it's hard to compute some y with $f(y) = f(x)$. People call this a "one-way function".

Now we're finally getting to the really interesting math. It takes work to make the concept of "one-way function" precise, but it can be done. For example, Chapter 2 of this book starts out by defining "strong" and "weak" one-way functions:

11) Oded Goldreich, *Foundations of Cryptography*, Cambridge U. Press, Cambridge, 2004. Older edition available at http://www.wisdom.weizmann.ac.il/~oded/frag.html

Since you can look them up and they're a bit gnarly, I won't give these definitions here. But *roughly*, a "strong one-way function" is a function f that satisfies two conditions:

1. You can compute f in "polynomial time": in other words, there's an algorithm that computes it in a number of steps bounded by some polynomial in the length of the input.
2. Given some input x, any polynomial-time probabilistic algorithm has a very low chance of finding y with $f(y) = f(x)$. Here a "probabilistic algorithm" is just an algorithm equipped with access to a random number generator.

Condition 2 is related to the idea of a "preimage attack". We allow the attacker to use a probabilistic algorithm because it seems this can help them, and a strong one-way function should be able to resist even really nasty attacks.

Unfortunately, nobody has proved that a one-way function exists!

In fact, the existence of a one-way function would imply that "P does not equal NP". But, proving or disproving this claim is one of the most profound unsolved math problems around. If you settle it, you'll get a million dollars from the Clay Mathematics Institute:

11) Clay Mathematics Institute, "P vs NP problem", http://www.claymath.org/millennium/P_vs_NP/

But if you prove that P *does* equal NP, you might make more money by breaking cryptographic hash codes and setting yourself up as the Napoleon of crime.

The existence of one-way-functions is also closely related to the existence of "pseudorandom" sequences. These are sequences of numbers that "seem" random, but can be computed using deterministic algorithms — so they're not *really* random.

To see the relation, recall that a good cryptographic hash code shouldn't let you guess a message from its digest. So, for example, if my message is the EMPTY STRING — no message at all! — its digest should be a pseudorandom sequence.

For example, if we apply SHA-1 to the empty string we get the following 160 bits, written as 40 hexadecimal digits: > `SHA1("") = "da39a3ee5e6b4b0d3255bfef95601890afd80709"` Seems mighty random to me! But of course it comes out the same every time.

Indeed, if you've ever seen a "random number generator" while messing with computers, it probably was a pseudorandom number generator. There are a few people who generate random numbers from physical phenomena like atmospheric noise. In fact, there's a website called `random.org` where you can get such numbers for free:

12) "Random.org: random integer generator", http://random.org/

There's also a website that offers *nonrandom* numbers:

13) "NoEntropy.net: your online source for truly deterministic numbers", http://www.noentropy.net/

But joking aside, it's a really tantalizing and famous problem to figure out what "random" means, and what it means for something to "seem" random even if it's the result of a deterministic process. These are huge and wonderful philosophico-physico-mathematical questions with serious practical implications. Much ink has been spilt regarding these, and I don't have the energy to discuss them carefully, so I'll just say some stuff to pique your curiosity, and then give you some references.

We can define a "random sequence" to be one that no algorithm can generate with a success rate better than chance would dictate. By virtue of this definition, no algorithm can generate truly random sequences. It's easy to prove that most sequences are random — but it's also easy to prove that you can never exhibit any one *particular* sequence and prove it's random! Chaitin has given a wonderful definition of a particular random sequence of bits called Ω using the fact that no algorithm can decide which Turing machines halt... but this random sequence is uncomputable, so you can't really "exhibit" it:

14) Gregory Chaitin, "Paradoxes of randomness", *Complexity* **7** (2002), 14–21. Also available as `http://www.umcs.maine.edu/~chaitin/summer.html`

So, true randomness is somewhat elusive. It seems hard to come by except in quantum mechanics. For example, the time at which a radioactive atom decays is believed to be *really* random. I'll be pissed off if it turns out that God (or his henchman Satan) is fooling us by simulating quantum mechanics with some cheap pseudorandom number generator!

Similarly, we could define a "pseudorandom sequence" to be one that no *efficient* algorithm can generate with a success rate higher than chance would dictate.

Efficiency is a somewhat vague concept. It's popular to define it by saying an algorithm is "efficient" if it runs in "polynomial time": the time it takes to run is bounded by some polynomial function of the size of the input data. If the polynomial is

$$p(n) = 1000000000000000n^{1000000000000000} + 1000000000000000$$

most people wouldn't consider the algorithm efficient, but this definition is good for proving theorems, and usually in practice the polynomial turns out to be more reasonable.

A "pseudorandom number generator" needs to be defined carefully if we want to find efficient algorithms that do this job. After all, no efficient algorithm can produce a sequence that no efficient algorithm can guess: *it* can always guess what it's going to do!

So, the basic idea is that a pseudorandom number generator should be an efficient algorithm that maps short truly random strings to long pseudorandom strings: we give it a short random "seed" and it cranks out a lot of digits that no efficient algorithm can guess with a success rate higher than chance would dictate.

If you want a more precise definition and a bunch of theorems, try these:

15) Oded Goldreich, papers and lecture notes on pseudorandomness, available at `http://www.wisdom.weizmann.ac.il/~oded/pp_pseudo.html`

    Luby, M. *Pseudorandomness and Cryptographic Applications*. Princeton, NJ: Princeton University Press, 1996.

Unfortunately, nobody has proved that pseudorandom number generators exist! So, this whole subject is a bit like axiomatic quantum field theory, or the legendary Ph.D thesis where the student couldn't produce any examples of the mathematical gadgets he was studying. It's a risky business proving results about things that might not exist. But in the case of pseudorandom number generators, the subject is too important not to take the chance.

One of the most interesting things about pseudorandom number generators is that they let us mimic probabilistic algorithms with deterministic ones. In fact there are some nice theorems about this. Let me sketch one of them for you.

As I already mentioned, a "probabilistic algorithm" is just a deterministic algorithm that's been equipped with access to a (true) random number generator. Just imagine a computer with the ability to reach out and flip a coin when it wants to. A problem is said to be in "BPP" — "bounded-error probabilistic polynomial time" — if you can find polynomial-time probabilistic algorithms that solve this problem with arbitrarily high chance of success.

It's a fascinating question whether randomness actually helps you compute stuff. I guess most computer scientists think it does. But, it's tricky. For example, consider the problem of deciding whether an integer is prime. Nobody knew how to do this in polynomial time... but then in 1977, Solovay and Strassen showed this problem was in BPP. This is one of the results that got people really excited about probabilistic algorithms.

However, in 2002, Maninda Agrawal, Neeraj Kayal and Nitin Saxena showed that deciding whether numbers are prime is in P! In other words, it can be solved in polynomial time by a plain old deterministic algorithm:

16) Anton Stiglic, "Primes is in P little FAQ", `http://crypto.cs.mcgill.ca/~stiglic/PRIMES_P_FAQ.html`

Is $\mathrm{BPP} = \mathrm{P}$? Nobody knows! But if good enough pseudorandom number generators could be shown to exist, we would have $\mathrm{BPP} = \mathrm{P}$, since then we could use these pseudorandom numbers as a substitute for truly random ones. This is not obvious, but it was proved by Nisan and Wigderson in 1994.

Here's a great review article that discusses their result:

17) Luca Trevisian, "Pseudorandomness and combinatorial constructions", available as `cs.CC/0601100`.

I recommend reading it along with this:

18) Scott Aaronson, "Is P versus NP formally independent?", available at `http://www.scottaaronson.com/papers/pnp.pdf` and `http://www.scottaaronson.com/papers/pnp.ps`

This is a delightfully funny and mindblowing crash course on logic. It starts with a review of first-order logic and Gdel's theorems, featuring a dialog between a mathematician and the axiom system $\mathrm{ZF}+\mathrm{not}(\mathrm{Con}(\mathrm{ZF}))$. Here ZF is the popular Zermelo-Fraenkel axioms for set theory and $\mathrm{not}(\mathrm{Con}(\mathrm{ZF}))$ is a statement asserting that ZF is not consistent. Thanks to Gdel's second incompleteness theorem, $\mathrm{ZF}+\mathrm{not}(\mathrm{Con}(\mathrm{ZF}))$ is consistent if ZF is! The mathematician is naturally puzzled by this state of affairs, but in this dialog, the axiom system explains how it works.

Then Aaronson zips on through Loeb's theorem, which is even weirder than Gdel's first incompleteness theorem. Gdel's first incompleteness theorem says that the statement

This statement is unprovable in ZF.

is not provable in ZF, as long as ZF is consistent. Loeb's theorem says that the statement

This statement is provable in ZF.

is provable in ZF.

Then Aaronson gets to the heart of the subject: a history of the P vs. NP question. This leads up to an amazing paper of Razborov and Rudich, which I'll now summarize:

19) Alexander A. Razborov and Steven Rudich, "Natural proofs", in *Journal of Computer and System Sciences*, Vol. **55**, No 1, 1997, pages 24–35. Also available at `http://www-2.cs.cmu.edu/~rudich/papers/natural.ps` and `http://genesis.mi.ras.ru/~razborov/int.ps`

The basic point of this paper is that if P is not equal to NP, as most mathematicians expect, then this fact is hard to prove! Or, as Aaron more dramatically puts it, this conjecture "all but asserts the titanic difficulty of finding its own proof".

Zounds! But let's be a bit more precise. A Boolean circuit is a gizmo built of "and", "or" and "not" gates, without any loops. We can think of this as computing a Boolean function, meaning a function of the form:

$$f\colon \{0,1\}^n \to \{0,1\}$$

Razborov and Rudich start by studying a common technique for proving lower bounds on the size of a Boolean circuit that computes a given function. The technique goes like this:

1. Invent some way of measuring the complexity of a Boolean function.
2. Show that any Boolean circuit of a certain size can compute only functions of complexity less than some amount.
3. Show that the function f has high complexity.

They call this style of proof a "natural" proof.

The P versus NP question can be formulated as a question about the size of Boolean circuits — but Razborov and Rudich show that, under certain assumptions, there is no "natural" proof that P is not equal to NP. What are these assumptions? They concern the existence of good pseudorandom number generators. However, the existence of these pseudorandom number generators would follow from the fact that P is not equal to NP. So, if "P is not equal to NP" is true, it has no natural proof.

Aaronson says this is the deepest insight into the P versus NP question so far. I would like to understand it better — I explained it very sketchily because I don't really understand it yet. Aaronson recommends us to these papers for more details:

20) Alexander A. Razborov, "Lower bounds for propositional proofs and independence results in bounded arithmetic", in *Proceedings of ICALP 1996*, 1996, pp. 48–62. Also available at `http://genesis.mi.ras.ru/~razborov/icalp.ps`

R. Raz, "$\mathrm{P} \neq \mathrm{NP}$, propositional proof complexity, and resolution lower bounds for the weak pigeonhole principle", in *Proceedings of ICM 2002*, Vol. **III**, 2002, pp. 685–693. Also available at `http://www.wisdom/weizmann.ac.il/~ranraz/publication/Pchina.ps`

S. Buss, "Bounded arithmetic and propositional proof complexity", in *Logic of Computation*, ed. H. Schwictenberg, Springer-Verlag, 1997, pp. 67–122. Also available at `http://www.wisdom/weizmann.ac.il/\~ranraz/publication/Pchina.ps`

(Why don't these guys use the arXiv??)
Also, here are some lecture notes on Boolean circuits that might help:

21) Uri Zwick, "Boolean circuit complexity", `http://www.math.tau.ac.il/~zwick/scribe-boolean.html`

Aaronson wraps up by musing on the possibility that the P versus NP question is independent from strong axiom systems like Zermelo-Fraenkel set theory. It's possible... and it's possible that this is true and unprovable!

So, there is a fascinating relationship between one-way functions, pseudorandom numbers, and incompleteness — but it's a relationship shrouded in mystery... and perhaps inevitably so. Perhaps it will always remain unknown whether this mystery is inevitable... and perhaps this is inevitable too! And so on.

Here's a question for you experts out there: have people studied Gdel-like self-referential sentences of this form?

> *The shortest proof of this statement has n lines.*

I hear that people *have* studied ones like

> *The shortest proof that $0 = 1$ has n lines.*

and they've proved that the shortest proof of *this* has to keep getting longer with larger n, as long as one is working in a sufficiently powerful and consistent axiom system. This is fairly obvious if you know how Gdel's second incompleteness theorem is proved... but it's possible that some interesting, nonobvious lower bounds have been proved. If so, I'd like to know!

Addenda: Here is some discussion with Wolfgang Brand, Allan Erskine, Scott Aaronson, Aaron Bergman and an entity named tessel. I had written:

> *Chaitin has given a marvelous definition of a particular random sequence of bits called Ω using the fact that no algorithm can decide which Turing machines halt... but this random sequence is uncomputable, so you can't really "exhibit" it.*

However, Wolfgang Brand points out this paper:

22) Cristian Calude, Michael J. Dinneen, and Chi-Kou Shu, "Computing a glimpse of randomness", *Experimental Mathematics* **11** (2002), 361–370. Also available at `http://www.cs.auckland.ac.nz/~cristian/Calude361_370.pdf`

where the first 64 bits of Ω have been computed. There's no contradiction, as the paper explains — but it's fairly surprising!

Next, here is an email from Allan Erskine. As usual I will take the liberty of numbering references and putting them into my own favorite format.

Allan Erskine wrote:

> *I enjoyed week 226! Algorithmic complexity was the area I studied in... Your readers might find this an enjoyable read:*
>
> *24) Leonid Levin, "The tale of one-way functions",* Problems of Information Transmission *(= Problemy Peredachi Informatsii)* ***39*** *(2003), 92–103. Also available as* `cs.CR/0012023`.
>
> *As for your "shortest proof of this statement has n lines" question, you may have noticed that Chaitin asks a very similar question about the shortest proofs that a LISP program is "elegant" (most short) and proves a strong incompleteness result with an actual $410 + n$ character LISP program! Crazy...*
>
> *25) Gregory Chaitin, "Elegant LISP programs", in* People and Ideas in Theoretical Computer Science*, ed. C. Calude, Springer, Singapore, 1999, pp. 32–52. Also available at* `http://www.cs.auckland.ac.nz/CDMTCS/chaitin/lisp.html`

I replied by pointing Allan to the following old article, which is related to the idea of Chaitin's paper:

> ***From: John Baez***
> ***Subject: Re: compression, complexity, and the universe***
> ***Date: 1997/11/20***
> ***Newsgroups:*** `sci.physics.research`, `comp.compression.research`
>
> *Aaron Bergman wrote:*
>
> > *The smallest number not expressible in under ten words*
>
> *Hah! This, by the way, is the key to that puzzle I laid out: prove that there's a constant K such that no bitstring can be proved to have algorithmic entropy greater than K.*
>
> *Let me take this as an excuse to say a bit more about this. I won't give away the answer to the puzzle; anyone who gets stuck can find the answer in this nice survey:*

> *26) Peter Gacs, Lecture notes on descriptional complexity and randomness, available at* `http://www.cs.bu.edu/faculty/gacs/`
>
> *In my more rhapsodic moments, I like to think of K as the "complexity barrier". The world* seems *to be full of incredibly complicated structures — but the constant K sets a limit on our ability to* prove *this. Given any string of bits, we can't rule out the possibility that there's some clever way of printing it out using a computer program less than K bits long. The Encyclopedia Brittanica, the human genome, the detailed atom-by-atom recipe for constructing a blue whale, or for that matter the entire solar system — we are unable to prove that a computer program less than K bits long couldn't print these out. So we can't firmly rule out the reductionistic dream that the whole universe evolved mechanistically starting from a small "seed", a bitstring less than K bits long. (Maybe it did!)*
>
> *So this raises the question, how big is K?*
>
> *It depends on ones axioms for mathematics.*
>
> *Recall that the algorithmic entropy of a bitstring is defined as the length of the shortest program that prints it out. For any finite consistent first-order axiom system A exending the usual axioms of arithmetic, let $K(A)$ be the constant such that no bitstring can be proved, using A, to have algorithmic entropy greater than $K(A)$. We can't compute $K(A)$ exactly, but there's a simple upper bound for it. As Gacs explains, for some constant c we have:*
>
> $$K(A) < L(A) + 2\log_2 L(A) + c$$
>
> *where $L(A)$ denotes the length of the axiom system A, encoded as bits as efficiently as possible. I believe the constant c is computable, though of course it depends on details like what universal Turing machine you're using as your computer.*
>
> *What I want to know is, how big in practice is this upper bound on $K(A)$? I think it's not very big! The main problem is to work out a bound on c.*

Next, some email from the man himself: Scott Aaronson!

> *Hi John,*
>
> *I just read the latest issue of TWF. I'd been waiting a long time for you to tackle computational complexity — thanks! :) Here are some comments and responses that you might find helpful.*
>
> *It's not known whether $\mathrm{P} \neq \mathrm{NP}$ implies the existence of one-way functions (OWF's) — indeed, there's evidence that proving such an implication will be hard:*
>
> *27) Adi Akavia, Oded Goldreich, Shafi Goldwasser and Dana Moshkovitz, "On basing one-way functions on* NP*-hardness", November 22, 2005, available at* `http://theory.csail.mit.edu/~akavia/AGGM.pdf`

What's known is that one-way functions exist iff pseudorandom generators (PRG's) do — that's a long, gnarly result of Hastad et al.:

28) *John Hastad, Russell Impagliazzo, Leonid A. Levin and Michael Luby, "A pseudorandom generator from any one-way function", available at `http://citeseer.ifi.unizh.ch/hastad99pseudorandom.html`*

So when I wrote that Razborov-Rudich illustrates the "self-referentiality" of proving $\mathrm{P} \neq \mathrm{NP}$, really I meant that it illustrates the self-referentiality of proving that OWF's and PRG's exist! But most of us believe not only that $\mathrm{P} \neq \mathrm{NP}$, but also that OWF's exist, and that all of these questions are "morally similar" anyway. . . :)

You write: "It's a fascinating question whether randomness actually helps you compute stuff. I guess most computer scientists think it does."

Well, today most of us think that good pseudorandom generators exist, which (as you pointed out) implies that $\mathrm{P} = \mathrm{BPP}$. That's not to say that randomness never helps; it just doesn't provide more than a polynomial advantage.

Incidentally, it followed already from Andy Yao's work in 1982 that if REALLY good pseudorandom generators exist, then $\mathrm{P} = \mathrm{BPP}$. What Nisan and Wigderson showed in 1994 was that even if only kinda sorta good PRG's exist, then still $\mathrm{P} = \mathrm{BPP}$!

Since you wrote at length about hash functions like MD5, you might be interested in the following abstraction of the collision-finding problem. Suppose we have a two-to-one function

$$f\colon \{0,1\}^n \to \{0,1\}^{n-1},$$

which we can only access in a "black-box" fashion — i.e., we can feed it an input and get an output. The goal is to find distinct x and y such that $f(x) = f(y)$. Classically, it's easy to see that $\sim 2^{\frac{n}{2}}$ accesses to f are necessary and sufficient — that's just the birthday paradox you described. But what if we can access f in quantum superposition, to create states like

$$\sum_x |x\rangle|f(x)\rangle \text{ ?}$$

In that case, it turns out that one can combine the birthday paradox with Grover's quantum search algorithm, to find a collision using $\sim 2^{\frac{n}{3}}$ accesses to f instead of $\sim 2^{\frac{n}{2}}$. But could we do it with $n^{\mathcal{O}(1)}$ accesses, or even a constant *number independent of n? You might be surprised that this was open for years!*

Probably my best-known result was to show that $\sim 2^{\frac{n}{5}}$ accesses are needed:

29) *Scott Aaronson, "Quantum lower bound for the collision problem", available as `quant-ph/0111102`.*

Subsequently Yaoyun Shi improved that to show that $\sim 2^{\frac{n}{3}}$ is indeed the right answer:

30) Yauyon Shi, "Quantum lower bounds for the collision and the element distinctness problems", available as `quant-ph/0112086`.

You ask why theoretical computer scientists don't use the arXiv. The answer is historical and cultural. Quantum computer scientists like me do *post to `quant-ph`, as a result of hanging around physicists for so long. But in classical complexity, if you want your paper to be read, then you (1) put in on your homepage, (2) submit it to STOC or FOCS, which are the top conferences in the field (and are MUCH more important than journals), and (3) post it here:*

31) Electronic Colloquium on Computational Complexity, `http://www.eccc.uni-trier.de/eccc/`

which is our own homegrown arXiv.

I like your finite Gdel statement! Let me restate it as follows:

$$G(n) = \text{"This statement has no ZF proof of at most } n \text{ symbols."}$$

What can we say? Obviously, if ZF *is consistent, then* $G(n)$ *has no proof of at most* n *symbols. Hence* $G(n)$ *is true. Furthermore,* $G(n)$ *has a proof of* $\sim 2^n$ *symbols. This proof simply enumerates all proofs of* $\leqslant n$ *symbols, and shows that none of them work.*

You might wonder: does $G(n)$ *have a proof of* $n+1$ *symbols, or for that matter* $n^{\mathcal{O}(1)}$ *symbols? This turns out to be related to the main open questions of complexity theory! In particular, if* $\mathrm{NP} = \mathrm{coNP}$*, then* $G(n)$ *has a proof of* $n^{\mathcal{O}(1)}$ *symbols.* $\mathrm{NP} = \mathrm{coNP}$ *just means that, whenever a Boolean formula of size* n *is unsatisfiable, there's a proof of that fact of length polynomial in* n*. Clearly if* $\mathrm{P} = \mathrm{NP}$ *then* $\mathrm{NP} = \mathrm{coNP}$*, but the converse isn't known. The prevailing belief is not only that* $\mathrm{P} \neq \mathrm{NP}$*, but also that* $\mathrm{NP} \neq \mathrm{coNP}$*.*

Actually I can say more than that: if $\mathrm{NE} = \mathrm{coNE}$*, then* $G(n)$ *has a proof of* $n^{\mathcal{O}(1)}$ *symbols. Here* NE *is Nondeterministic Exponential-Time, or the exponential-time analogue of* NP*. (See my Complexity Zoo:*

32) Scott Aaronson and Greg Kuperberg, "Complexity Zoo", `http://qwiki.caltech.edu/wiki/Complexity_Zoo`

for lots more about this and other classes.) coNE *is just the complement of* NE*. If* $\mathrm{NP} = \mathrm{coNP}$ *then* $\mathrm{NE} = \mathrm{coNE}$*, but again the converse isn't known.*

What do these weird exponential classes have to do with $G(n)$*? The point is that to describe an integer* n *takes only* $\log(n)$ *symbols, not* n *symbols. So when we ask for a proof of* $G(n)$ *with at most* n *symbols, from "*$\log(n)$*'s standpoint" we're actually asking for an exponentially long proof.*

At this point I'll conjecture that I have an "if and only if" characterization: if $G(n)$ *has a proof of* $n^{\mathcal{O}(1)}$ *symbols, then* $\mathrm{NE} = \mathrm{coNE}$*.*

Incidentally, your "finite Gdel statement" is closely related to the Hartmanis-Stearns time hierarchy theorem, which launched computational complexity in

> *the mid-1960's. The time hierarchy theorem is a finite version of Turing's undecidability theorem, where instead of asking whether a given Turing machine ever halts, you ask whether it halts after at most T steps. Solving this problem turns out to take slightly more than T steps. So in particular, you can solve more problems in n^2 time than in n time, more in n^3 time than in n^2 time, and so on.*
>
> *Let me know if you have other questions.*
>
> *Best,*
> *Scott*

Next, here is my reply to a post on `sci.physics.research` from a mysterious entity named tessel... again, I'll edit it a bit:

> *tessel@um.bot wrote:*
>
>> *On Sat, 11 Feb 2006, John Baez mentioned that md5sum was "broken" about a year ago. I just wanted to add:*
>>
>> 1. *If I am not mistaken, sha-1 and md5sum are different algorithms (IIRC, both are known to be insecure).*
>
> *Yeah, I said SHA-1 and MD5 are different, and I said they were both vulnerable to collision attacks. MD5 is very vulnerable in practice, while the vulnerability of SHA-1 is still theoretical: you'd have to have big computers or lots of time or another clever idea to exploit it. (Guess who's likely to have all three!)*
>
> *For more on this subject, try this:*
>
> 33) *Arjen K. Lenstra, "Further progress in hashing cryptanalysis", February 26, 2005, available at* `http://cm.bell-labs.com/who/akl/hash.pdf`
>
>> 2. *The latest versions of the open source utility gpg supports a more secure algorithm, SHA-512, which AFAIK has not been broken; see*
>>
>> 34) *Tony Stieber, "GnuPG Hacks", Linux Journal, March 2006.*
>>
>> 3. *Even insecure checksum utilities are probably better than none at all. Indeed, checking the given example:*
>>
>> ```
>> gpg --print-md md5 letter_of_rec.ps order.ps
>> A2 5F 7F 0B 29 EE 0B 39  68 C8 60 73 85 33 A4 B9
>> A2 5F 7F 0B 29 EE 0B 39  68 C8 60 73 85 33 A4 B9
>> ```
>>
>> *Oh NOOOOOO!!! But wait, there's more:*
>>
>> ```
>> gpg --print-md sha1 letter_of_rec.ps order.ps
>> 0783 5FDD 04C9 AFD2 8304  6BD3 0A36 2A65 16B7 E216
>> 3548 DB4D 0AF8 FD2F 1DBE  0228 8575 E8F9 F539 BFA6
>>
>> gpg --print-md RIPEMD160 letter_of_rec.ps order.ps
>> 9069 8ACC 6D67 6608 657B 9C26 F047 59A1 DC0E 6CA1
>> C1BB DE12 B312 EAAD DD3D D3B8 4CA1 CB1B BA47 DD13
>> ```

> > *Ah HAAAAAA!!! Gotcha, Alice!*
> >
> > *JB wrote:*
> >
> > > *These are huge and wonderful philosophico-physico-mathematical questions with serious practical implications.*
> >
> > *You mean the Weyl curvature hypothesis? :-/*
>
> *Heh, no — I mean stuff like whether there's such a thing as a provably good cryptographic hash code function, or cipher.*
>
> > *But while we should never neglect incompleteness entirely, I was fascinated to discover from my readings a few years back that even first order logic has its fascinations!*
> >
> > *35) Joel Spencer,* The Strange Logic of Random Graphs*, Springer, Berlin, 2001.*
> >
> > *Here's a thought: "Everyone knows" that if on day D, mathematician M is studying an example of size S in class C, he is more likely to be studying a "secretly special" representative R than a generic representative G of size S. Why? Because the secretly special reps show up in disguise in other areas, and M was probably hacking through the jungle from one of those places when he got lost and ate a poisoned cache.*
>
> *Interesting.*

Finally, here's my reply to an article on `sci.math.research` by Aaron Bergman:

> *Aaron Bergman wrote:*
>
> > *John Baez wrote:*
> >
> > > *In fact, the existence of a one-way function would imply that "*P *does not equal* NP*". But, proving or disproving this claim is one of the most profound unsolved math problems around.*
> >
> > *This brings to mind a question I was wondering about. Given an* NP *(-complete?) problem, is it ever possible to engineer a (partial) differential equation, the solution of which, if known, would solve the* NP *problem?*
> >
> > *I realize this is vague. The general thought was whether a continuous dynamical system can somehow be more computationally powerful than something discrete.*
>
> *I was very interested in this question back in college, and I wrote a paper about it:*
>
> *36) John Baez, "Recursivity in quantum mechanics",* Trans. Amer. Math. Soc. ***280*** *(1983), 339–350.*

> *where I showed that the time evolution for bunch of PDE's — including the wave equation, the Klein-Gordon equation, and Schroedinger's equation for finitely many electrically charged point particles — is "computable" in a precise sense.*
>
> *This seemed interesting at the time because Pour-El and Richards had claimed that the behavior of the wave equation in 3+1 dimensions was uncomputable. Their result is correct, but it's an artifact of using a funny space of solutions! If you use a sensible Hilbert space of solutions for which time evolution is unitary, then time evolution is also computable.*
>
> *Hmm, I see that someone has rediscovered this a few years ago:*
>
> *37) Klaus Weihrauch and Ning Zhong, "Is wave propagation computable or can wave computers beat the Turing machine?",* Proc. Lond. Math. Soc. ***85** (2002) 312–332. Abstract available at `http://www.lms.ac.uk/publications/proceedings/abstracts/p1364a.html`*
>
> *I wonder if they cite my paper.*
>
> *Anyway, I came away feeling that time evolution for any PDE people actually care about is computable once you define your concepts correctly. This is simply a way of saying that you can numerically compute the solutions to any desired accuracy: it's sort of obvious, modulo the crucial technical details.*
>
> *But a more interesting question would be the one you ask, where mere "computability" is replaced by something more refined, like "computable up to a specified accuracy in polynomial time".*

Another addendum: here's a new article on the problems with finding a cryptographic hash function that's hard to crack:

38) Susan Landau, "Find me a hash", *AMS Notices* **53** (March 2006), 330–332. Also available at `http://www.ams.org/notices/200603/fea-landau.pdf` and `http://www.ams.org/notices/200603/fea-landau.ps`

Landau points out that there's no good mathematical theory backing up the most popular hash functions. "There are hash functions based on hard mathematical problems, making them likely to be secure, but these hash function are inefficient and not used in practice."

Next, here's yet another email from Scott Aaronson, dated July 22, 2006:

> *Hi John,*
>
> *A while ago you asked whether anyone had studied sentences of the form*
>
> $P(n) =$ *"This statement has no proof at most n symbols long."*
>
> *Clearly $P(n)$ has a proof with $\sim 2^n$ symbols. I emailed you conjecturing that $P(n)$ has no proof much shorter than that.*
>
> *Well, I just came back from a complexity meeting in Prague, where Pavel Pudlak gave a talk about exactly this sort of question. It turns out that Harvey Friedman made the same conjecture I did in the 1970's, but his conjecture was soon shown*

to be FALSE! Indeed, the number of symbols needed to prove $P(n)$ grows only a little bit faster than n itself.

Let me sketch why, since I’m guessing this will be as surprising to you as it was to me.

To prove $P(n)$, it clearly suffices to show that, if our proof system is inconsistent, then any proof of $0 = 1$ must contain statements with more than n symbols. We can do that using a technique invented by Tarski in the 1930’s for “defining truth” — i.e., for recursively building up a one-to-one correspondence between syntax and semantics. More concretely, let $T_k(A)$ be a function that takes as input a string A with £k symbols. We want T_k to return 1 if A encodes a true statement and 0 otherwise. So, for all k, $T_{k+1}(\text{”not”} + A) := 1$ if $T_k(A) = 0$, or 0 otherwise.

$T_{k+1}(A + \text{”and”} + B) := 1$ *if* $T_k(A) = 1$ *and* $T_k(B) = 1$*, or* 0 *otherwise.*

$T_{k+1}(”There exists an x such that” + A) := 1$ *if there exists an x^* such that $T_k(A[x = x^*]) = 1$, or 0 otherwise.*

and so on. Of course I’m skipping lots of technicalities, like what I mean by $A[x = x^*]$.

Anyway, one can prove that, if $T_n(A) = T_n(B) = 1$, and if C is derivable from A and B via a first-order inference rule, then $T_n(C) = 1$ as well. Also, if A is an axiom of our proof system, then (reasoning within *the system) A is true, and hence $T_n(A) = 1$.*

It follows that, as long as we restrict ourselves to statements of length £n, no contradiction can be proved. Furthermore, we can formalize the whole argument using only $\mathcal{O}(n)$ symbols (or maybe $\mathcal{O}(n \log(n))$, depending on the proof system).

So what’s the catch? Why can’t we generalize this argument to statements of all *lengths, thereby contradicting the Incompleteness Theorem? Because, when we examine the proof that the syntax correctly models the semantics, we’ll always find that it involves statements with more than n quantifiers. I.e. to prove consistency for statements with n quantifiers, we need at least $n+1$ quantifiers.*

So, is there any way to salvage a finitary Gdel’s Theorem? There probably is! In particular, consider the following statement:

$P(n, k) =$*“This statement has no proof at most n symbols long,*
involving statements with at most k quantifiers.”

The new conjecture would be that, for every fixed k, $P(n, k)$ has no proof with less than $2^{\mathcal{O}(n)}$ symbols involving statements with at most k quantifiers. According to Pavel, this conjecture is still open. Indeed, it’s not even known whether $P(n)$ itself has a proof with less than $2^{\mathcal{O}(n)}$ symbols, involving statements with at most k quantifiers for some fixed k.

Best,
Scott

Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.

— *John von Neumann*

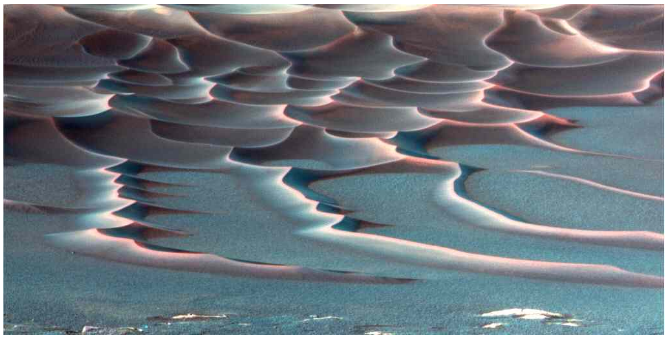
He is the Napoleon of crime, Watson. He is the organizer of half that is evil and of nearly all that is undetected in this great city. He is a genius, a philosopher, an abstract thinker. . . .

— *Sherlock Holmes*

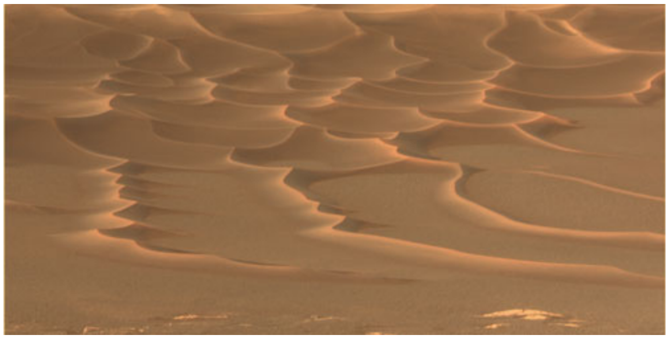
Week 227

March 12, 2006

Today I want to say a bit about physics, and then a bit about about logic, since that's what I was studying for the last month in Marseille. But first, the astronomy pictures of the week:

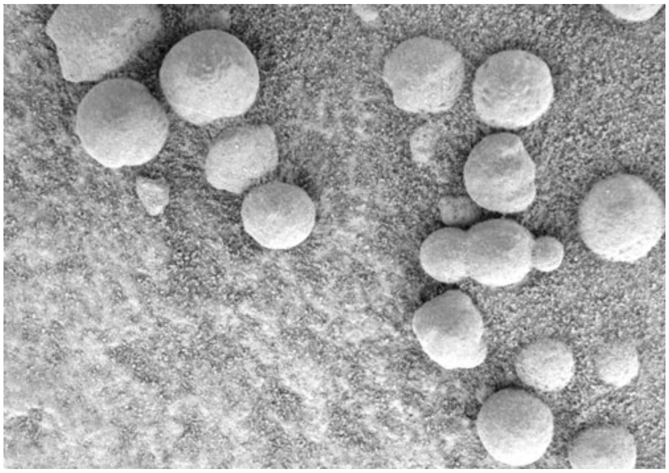


1) "Endurance crater's dazzling dunes", NASA/JPL, available at `http://marsrovers.jpl.nasa.gov/gallery/press/opportunity/20040806a.html`



On August 4, 2004, the Mars rover called Opportunity took these pictures of dunes as it entered Endurance Crater. The red one is what we'd actually see; the blue one is a false-color image designed to bring out certain details.

Both images show show tendrils of sand less than 1 meter high stretching from the big dunes toward the rover, and some rocks in the foreground. The false-color image emphasizes accumulations of millimeter-sized spheres called "blueberries" on the flat parts of the dunes. Here's what they look like close up:



2) 'Mineral in Mars "berries" adds to water story', NASA/JPL, available at `http://marsrover.nasa.gov/newsroom/pressreleases/20040318a.html`

Thanks to a Mssbauer spectroscope aboard the rover, which studies rocks by firing γ rays at them, we know these blueberries contain a lot of hematite.

Hematite is made of ferric oxide, Fe_2O_3, otherwise known as "rust". It's usually formed in the presence of water. For this and other reasons, it's believed that the blueberries in Endurance Crater were formed when Mars was wetter than today. An interesting puzzle is whether they were formed by groundwater leaching ferric oxide from rocks, or deposited in standing water — for example, a lake.

It's amazing how much we can learn from unmanned space probes. And it's amazing to me how some people want to spend billions on manned missions. To read more about what SF writers Larry Niven, Joe Haldeman, Greg Bear and other folks including me think about the merits of manned versus unmanned space missions, try this:

3) "Meme Therapy: Life from a science fiction point of view", `http://memetherapy.blogspot.com/2006/03/brain-parade-feature-where-we-pester.html`

But enough space stuff... now for some physics!

A strange thing happened around the 1980s. Before that, theorists had been making rapid and revolutionary progress in understanding the fundamental laws of physics for almost a century. They kept predicting shocking new effects that were soon found in actual experiments: radio waves, nuclear chain reactions, black holes, lasers, antimatter, neutrinos, quarks,... up to and including the W and Z bosons. The power of human thought never seemed greater.

Since the 1980s, most of the new discoveries in fundamental physics have come from unexpected observations in astronomy. These observations were mostly *not* predicted by theorists. The key examples are: dark matter, dark energy, and neutrino oscillations. The only serious counterexamples that come to mind are the top quark, discovered in 1995, and Alan Guth's inflationary cosmology, dreamt up around 1979 and *partially* confirmed by recent data — though the jury is still out.

Theorists are publishing more than ever, but most of their theories are either not yet testable (string theory, loop quantum gravity) or seem to have been disproved by experiment (grand unified theory predictions of proton decay).

It's interesting to meditate on why this change has happened, what it means, and what will happen next. I spoke about this in Marseille shortly before coming home to Riverside. You can see my slides here:

4) John Baez, "Fundamental physics: where we stand today", lecture at at the Faculty of Sciences, Luminy, February 27th, 2006, available at `http://math.ucr.edu/home/baez/where_we_stand/`

This talk was for nonphysicists, so it's not very technical. The first part is a gentle introduction to the laws of physics as we know them.

For more info on dark matter and dark energy, try this:

5) Varun Sahri, "Dark matter and dark energy", available as `astro-ph/0403324`.

If you read this, you'll learn about the "cuspy core problem" — existing cold dark matter models produce galaxies with a sharp spikes of high density near their cores, sharper than observed. You'll learn about "quintessence", a kind of hypothetical field that some people use to model dark energy, thus "explaining" the accelerating expansion of the universe. You'll learn about the "Chaplygin gas", a hypothetical substance whose properties interpolate between those of cold dark matter and dark energy. And, you'll learn about "phantom energy" models of dark energy, which fit the accelerating expansion of the universe quite nicely now but predict a "Big Rip", in which the expansion rate eventually becomes *infinite*.

In short, you'll see how people are flailing around trying to understand dark matter and dark energy.

For more on neutrino oscillations, try this:

6) K. M. Heeger, "Evidence for neutrino mass: a decade of discovery", available as `hep-ex/0412032`.

I had a great time in Marseille. The area around there is great for mathematicians. Algebraists can visit the beautiful nearby city of Aix — pronounced "x". Logicians will enjoy the dry, dusty island of If — pronounced "eef", just like a French logician would say it. And everyone will enjoy the medieval hill town of Les Baux, which looks like something out of Escher. Here's a picture of it:

Actually the Chateau D'If, on the island of the same name, is where Edmond Dantes was imprisoned in Alexander Dumas' novel "The Count of Monte Cristo". It's in this formidable fortress that the wise old Abbe Faria tells Dantes the location of the treasure that later made him rich.

I guess everyone except me read this story as a kid — I'm just reading it now. But how many of you remember that Faria spent his time in prison studying the works of Aristotle? There's a great scene where Dantes asks Faria where he learned so much about logic, and Faria replies: "If — and only If!"

That Dumas guy sure was a joker.

Luckily I didn't need to be locked up on a deserted island to learn some logic in Marseille. There were lots of great talks on this topic at the conference I attended:

7) *Geometry of Computation 2006* (Geocal06), `http://iml.univ-mrs.fr/geocal06/`

For example, Yves Lafont gave a category-theoretic approach to Boolean logic gates which explains their relation to Feynman diagrams:

8) Yves Lafont, "Towards an algebraic theory of Boolean circuits", *Journal of Pure and Applied Algebra* **184** (2003), 257–310. Also available at `http://iml.univ-mrs.fr/~lafont/publications.html`

and together with Yves Guiraud, Francois Metayer and Albert Burroni, he gave a detailed introduction to the homology of n-categories and its application to rewrite rules. The idea is to study any sort of algebraic gadget (like a group) by creating an n-category where the objects are "expressions" for elements in the gadget, the morphisms are "ways of rewriting expressions" by applying the rules at hand, the 2-morphisms are "ways of passing from one way of rewriting expressions to another" by applying certain "meta-rules", and so on. Then one can use ideas from algebraic topology to study this n-category and prove stuff about the original gadget!

To understand how this actually works, it's best to start with Craig Squier's work on the word problem for monoids. I explained this pretty carefully back in "Week 70" when I first heard Lafont lecture on this topic — it made a big impression on me! You can read more here:

9) Yves Lafont and A. Proute, "Church-Rosser property and homology of monoids", in *Mathematical Structures in Computer Science*, Cambridge U. Press, 1991, pp. 297–326. Also available at `http://iml.univ-mrs.fr/~lafont/publications.html`

10) Yves Lafont, "A new finiteness condition for monoids presented by complete rewriting systems (after Craig C. Squier)", *Journal of Pure and Applied Algebra* **98** (1995), 229–244. Also available at `http://iml.univ-mrs.fr/~lafont/publications.html`

Then you can go on to the higher-dimensional stuff:

11) Albert Burroni, "Higher dimensional word problem with application to equational logic", *Theor. Comput. Sci.* **115** (1993), 43–62. Also available at `http://www.math.jussieu.fr/~burroni/`

12) Yves Guiraud, "The three dimensions of proofs", *Annals of Pure and Applied Logic* (in press). Also available at `http://iml.univ-mrs.fr/%7Eguiraud/recherche/cos1.pdf`

13) Francois Metayer, "Resolutions by polygraphs", *Theory and Applications of Categories* **11** (2003), 148–184. Available online at `http://www.tac.mta.ca/tac/volumes/11/7/11-07abs.html`

I was also lucky to get some personal tutoring from folks including Laurent Regnier, Peter Selinger and especially Phil Scott. Ever since "Week 40", I've been trying to understand something called "linear logic", which was invented by Jean-Yves Girard, who teaches in Marseille. Thanks to all this tutoring, I think I finally get it!

To get a taste of what Phil Scott told me, you should read this:

14) Philip J. Scott, "Some aspects of categories in computer science", *Handbook of Algebra*, Vol. **2**, ed. M. Hazewinkel, Elsevier, New York, 2000. Available as `http://www.site.uottawa.ca/~phil/papers/`

Right now, I'm only up to explaining a microscopic portion of this stuff. But since the typical reader of This Week's Finds may know more about physics than logic, maybe that's good. In fact, I'll use this as an excuse to simplify everything tremendously, leaving out all sorts of details that a real logician would want.

Logic can be divided into two parts: SYNTAX and SEMANTICS. Roughly speaking, syntax concerns the symbols you scribble on the page, while semantics concerns what these symbols mean.

A bit more precisely, imagine some kind of logical system where you write down some theory — like the axioms for a group, say — and use it to prove theorems.

In the realm of syntax, we focus on the form our theory is allowed to have, and how we can deduce new sentences from old ones. So, one of the key concepts is that of a PROOF. The details will vary depending on the kind of logical system we're studying.

In the realm of semantics, we are interested in gadgets that actually satisfy the axioms in our theory — for example, actual groups, if we're thinking about the theory of groups. Such a gadget is called a MODEL of the theory. Again, the details vary immensely.

In the realm of syntax, we say a list of axioms X "implies" a sentence P if we can prove P from X using some deduction rules, and we write this as

$$X \vdash P$$

In the realm of semantics, we say a list of axioms X "entails" a sentence P if every model of X is also a model of P, and we write this as

$$X \models P$$

Syntax and semantics are "dual" in a certain sense — a sense that can be made very precise if one fixes a specific class of logical systems. This duality is akin to the usual relation between vector spaces and their duals, or more generally groups and their categories of representations. The idea is that given a theory T you can figure out its models, which form a category $\mathsf{Mod}(T)$ — and conversely, given the category of models $\mathsf{Mod}(T)$, perhaps with a little extra information, you can reconstruct T.

A little extra information? Well, in some cases a model of T will be a *set* with some extra structure — for example, if T is the theory of groups, a model of T will be a group, which is a set equipped with some operations. So, in these cases there's a functor

$$U\colon \mathsf{Mod}(T) \to \mathsf{Set}$$

assigning each model its underlying set. And, you can easily reconstruct T from $\mathsf{Mod}(T)$ *together* with this functor.

This idea was worked out by Lawvere for a class of logical systems called algebraic theories, which I discussed in "Week 200". But, the same idea goes by the name of "Tannaka-Krein duality" in a different context: a Hopf algebra H has a category of comodules $\mathsf{Rep}(H)$, which comes equipped with a functor

$$U\colon \mathsf{Rep}(H) \to \mathsf{Vect}$$

assigning each comodule its underlying vector space. And, you can reconstruct H from $\mathsf{Rep}(H)$ together with this functor. The proof is even very similar to Lawvere's proof for algebraic theories!

I gave a bunch of talks in Marseille about algebraic theories, some related logical systems called PROPs and PROs, and their relation to quantum theory, especially Feynman diagrams:

14) John Baez, "Universal algebra and diagrammatic reasoning", available as `http://math.ucr.edu/home/baez/universal/`

I came mighty close to explaining how to compute the cohomology of an algebraic theory. . . and you can read more about that here:

15) Mauka Jibladze and Teimuraz Pirashvili, "Cohomology of algebraic theories", *J. Algebra* **137** (1991) 253–296.

Mauka Jibladze and Teimuraz Pirashvili, "Quillen cohomology and Baues-Wirsching cohomology of algebraic theories", *Max-Planck-Institut fr Mathematik*, preprint series **86** (2005).

But alas, I didn't get around to talking about the duality between syntax and semantics. For that Lawvere's original thesis is a good place to go:

16) F. William Lawvere, *Functorial Semantics of Algebraic Theories*, Ph.D. thesis, Columbia University, 1963. Also available at `http://www.tac.mta.ca/tac/reprints/articles/5/tr5abs.html`

Anyway, the stuff Phil Scott told me about was mainly over on the syntax side. Here categories show up in another way. Oversimplifying as usual, the idea is to create a category where an object P is a *sentence* — or maybe a list of sentences — and a morphism

$$f\colon P \to Q$$

is a *proof* of Q from P — or maybe an equivalence class of proofs.

We can compose proofs in a more or less obvious way, so with any luck this gives a category! And, different kinds of logical system give us different kinds of categories.

Quite famously, the multiplicative fragment of intuitionistic logic gives cartesian closed categories. (The "multiplicative fragment" is the portion that deals with "and" and "implies" but leaves out "or" and "not". I'm calling it this because "and" acts like multiplication, while "or" acts like addition.) Similarly, the multiplicative fragment of linear logic gives $*$-autonomous categories. Full-fledged intuitionistic logic gives cartesian closed categories with finite coproducts, and full-fledged linear logic gives us even fancier kinds of categories! If you want to learn about these examples, read the handbook article by Phil Scott mentioned above.

One thing that intrigues me is the equivalence relation we need to get a category whose morphisms are equivalence classes of proofs. In Gentzen's "natural deduction" approach to logic, there are various deduction rules. Here's one:

$$\frac{P \vdash Q \qquad P \vdash Q'}{P \vdash Q\&Q'}$$

This says that if P implies Q and it also implies Q', then it implies $Q\&Q'$.

Here's another:

$$\frac{P \vdash Q \implies R}{P \,\text{and}\, Q \vdash R}$$

And here's a very important one, called the "cut rule":

$$\frac{P \vdash Q \qquad Q \vdash R}{P \vdash R}$$

If P implies Q and Q implies R, then P implies R!

There are a bunch more... and to get the game rolling we need to start with this:

$$P \vdash P$$

In this setup, a proof $f\colon P \to Q$ looks vaguely like this:

$$\begin{array}{c} f\text{-crud} \\ f\text{-crud} \\ f\text{-crud} \\ f\text{-crud} \\ \hline P \vdash Q \end{array}$$

The stuff I'm calling "f-crud" is a bunch of steps which use the deduction rules to get to $P \vdash Q$.

Suppose we also we also have a proof

$$g \colon Q \to R$$

There's a way to stick f and g together to get a proof

$$fg \colon P \to R$$

This proof consists of setting the proofs f and g side by side and then using the cut rule to finish the job. So, fg looks like this:

$$\dfrac{\dfrac{\begin{array}{c} f\text{-crud}\\ f\text{-crud}\\ f\text{-crud}\\ f\text{-crud}\end{array}}{P \vdash Q} \qquad \dfrac{\begin{array}{c} g\text{-crud}\\ g\text{-crud}\\ g\text{-crud}\\ g\text{-crud}\end{array}}{Q \vdash R}}{P \vdash R}$$

Now let's see if composition is associative. Suppose we also have a proof

$$h \colon R \to S$$

We can form proofs

$$(fg)h \colon P \to S$$

and

$$f(gh) \colon P \to S$$

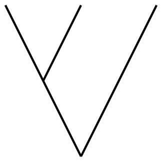
Are they equal? No! The first one looks like this:

$$\dfrac{\dfrac{\dfrac{\begin{array}{c} f\text{-crud}\\ f\text{-crud}\\ f\text{-crud}\\ f\text{-crud}\end{array}}{P \vdash Q} \qquad \dfrac{\begin{array}{c} g\text{-crud}\\ g\text{-crud}\\ g\text{-crud}\\ g\text{-crud}\end{array}}{Q \vdash R}}{P \vdash R} \qquad \dfrac{\begin{array}{c} h\text{-crud}\\ h\text{-crud}\\ h\text{-crud}\\ h\text{-crud}\end{array}}{R \vdash S}}{P \vdash S}$$

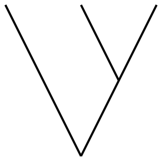
while the second one looks like this:

$$\dfrac{\dfrac{\begin{array}{c} f\text{-crud}\\ f\text{-crud}\\ f\text{-crud}\\ f\text{-crud}\end{array}}{P \vdash Q} \qquad \dfrac{\dfrac{\begin{array}{c} g\text{-crud}\\ g\text{-crud}\\ g\text{-crud}\\ g\text{-crud}\end{array}}{Q \vdash R} \qquad \dfrac{\begin{array}{c} h\text{-crud}\\ h\text{-crud}\\ h\text{-crud}\\ h\text{-crud}\end{array}}{R \vdash S}}{Q \vdash S}}{P \vdash S}$$

So, they're not quite equal! This is one reason we need an equivalence relation on proofs to get a category. Both proofs resemble trees, but the first looks more like this:

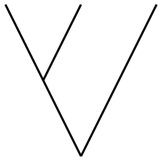


while the second looks more like this:

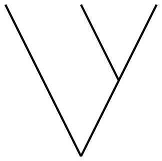


So, we need an equivalence relation that identifies these proofs if we want composition to be associative!

This sort of idea, including this "tree" business, is very familiar from homotopy theory, where we need a similar equivalence relation if we want composition of paths to be associative. But in homotopy theory, people have learned that it's often better NOT to impose an equivalence relation on paths! Instead, it's better to form a *weak* 2*-category* of paths, where there's a 2-morphism going from this sort of composite:



to this one:



This is called the "associator". In our logic context, we can think of the associator as a way to transform one proof into another.

The associator should satisfy an equation called the "pentagon identity", which I explained back in "Week 144". However, it will only do this if we let 2-morphisms be *equivalence classes* of proof transformations.

So, there's a kind of infinite regress here. To deal with this, it would be best to work with a "weak ω-category" with

- sentences (or sequences of sentences) as objects,
- proofs as morphisms,

- proof transformations as 2-morphisms,
- transformations of proof transformations as 3-morphisms, ...

and so on. With this, we would never need any equivalence relations: we keep track of all transformations explicitly. This is almost beyond what mathematicians are capable of at present, but it's clearly a good thing to strive toward.

So far, it seems Seely has gone the furthest in this direction. In his thesis, way back in 1977, he studied what one might call "weak cartesian closed 2-categories" arising from proof theory. You can read an account of this work here:

17) R.A.G. Seely, "Weak adjointness in proof theory", in *Proc. Durham Conf. on Applications of Sheaves*, Springer Lecture Notes in Mathematics **753**, Springer, Berlin, 1979, pp. 697–701. Also available at `http://www.math.mcgill.ca/rags/WkAdj/adj.pdf`

    R.A.G. Seely, "Modeling computations: a 2-categorical framework", in *Proc. Symposium on Logic in Computer Science 1987*, Computer Society of the IEEE, pp. 65–71. Also available at `http://www.math.mcgill.ca/rags/WkAdj/LICS.pdf`

Can we go all the way and cook up some sort of ω-category of proofs? Interestingly, while the logicians at *Geocal06* were talking about n-categories and the geometry of proofs, the mathematician Vladimir Voevodsky was giving some talks at Stanford about something that sounds pretty similar:

17) Vladimir Voevodsky, lectures on homotopy λ calculus, notice at `http://math.stanford.edu/distinguished_voevodsky.htm`

Voevodsky has thought hard about n-categories, and he won the Fields medal for his applications of homotopy theory to algebraic geometry.

The typed λ calculus is another way of thinking about intuitionistic logic — or in other words, cartesian closed categories of proofs. The "homotopy λ calculus" should thus be something similar, but where we keep track of transformations between proofs, transformations between transformations between proofs... and so on ad infinitum.

But that's just my guess! Is this what Voevodsky is talking about??? I haven't managed to get anyone to tell me. Maybe I'll email him and ask.

There were a lot of other cool talks at *Geocal06*, like Girard's talk on applications of von Neumann algebras (especially the hyperfinite type II_1 factor!) in logic, and Peter Selinger's talk on the category of completely positive maps, diagrammatic methods for dealing with these maps, and their applications to quantum logic:

18) Peter Selinger, "Dagger compact closed categories and completely positive maps", available at `http://www.mscs.dal.ca/~selinger/papers.html`

But, I want to finish writing this and go out and have some waffles for my Sunday brunch. So, I'll stop here!

Addendum: I thank Aaron Lauda, Paul Levy and Peter McBurney for corrections. Jeffrey Winkler points out that "hematite" got its name from the Greek word for "blood" because the ancient Greeks thought these rocks marked the locations of battles where the blood of warriors had soaked into the rocks. This is appropriate, since Mars is the god of war!

An anonymous correspondent had this to say about the "homotopy λ calculus":

> *Several years ago, Kontsevich explained to me an idea he had about "homotopy proof theory" (or model theory, or logic, . . .). As soon as I saw Voevodsky's abstract it reminded me of what Kontsevich said; perhaps it's a well-known idea in the Russian-Fields-medallist club. Somewhere I have notes from what Kontsevich said, but as far as I remember it went roughly like this.*
>
> *In certain set-ups (such as Martin-Lf type theory) every statement carries a proof of itself. Of course, a statement may have many proofs. If we imagine that all the statements are of the form "$A = B$", then what we're saying is that every equals sign carries with it a* reason *for equality, or proof of equality. If I remember rightly, Kontsevich's idea was to do a topological analogue, so that every term (like A and B) is assigned a point in some fixed space, and equalities of terms induce paths between points. There was more, pushing the idea further, but I forget what.*

Week 228

March 18, 2006

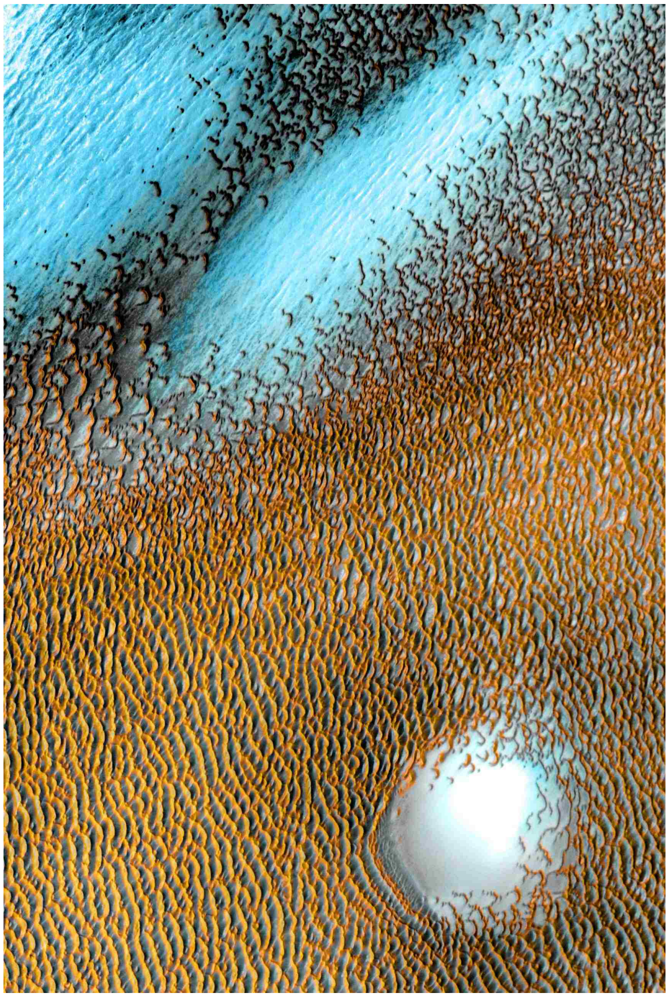
Last week I showed you some pretty pictures of dunes on Mars. This week I'll talk about dunes called "barchans" and their relation to self-organized criticality. Then I'll say a bit about Lauscher and Reuter's work on quantum gravity… and then I'll beg for help on a problem involving so-called "rational tangles".

But first, a demonstration of my psychic powers.

Take any book off the shelf and look at its 10-digit ISBN number. Multiply the first digit by 1, the second digit by 2, the third digit by 3 and so on… up to the *next to last* digit. Add them up.

Then take this sum and see what it equals $\bmod 11$. At the end of this article, I'll say what you got.

Okay. Here's a photo of the icy dunes of northern Mars. I love it because it shows that Mars is a lively place with wind and water:

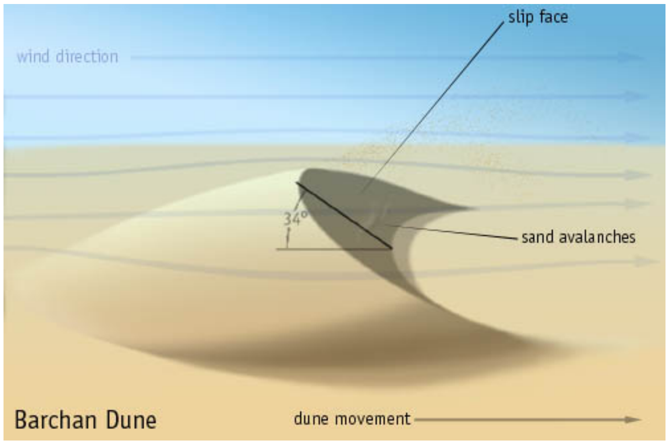


1) "North polar sand sea", Mars Odyssey Mission, THEMIS (Thermal emission imaging system), `http://themis.mars.asu.edu/features/polardunes`

These dunes, occupying a region the size of Texas, have been sculpted by wind into long lines with crests 500 meters apart. Their hollows are covered with frost, which appears bluish-white in this infrared photograph. The big white spot near the bottom is a hill 100 meters high.

Where the dunes become sparser — for example, near that icy hill — they break apart into "barchans". These are crescent-shaped formations whose horns point downwind. Barchans are also found on the deserts of Earth, and surely on many other planets across the Universe. They are one of several basic dune patterns, an inevitable consequence of the laws of nature under fairly common conditions.

The upwind slope of a barchan is gentle, while the downwind slope is between 32 and 34 degrees. This is the "angle of repose" for sand - the maximum angle it can tolerate before it starts slipping down:

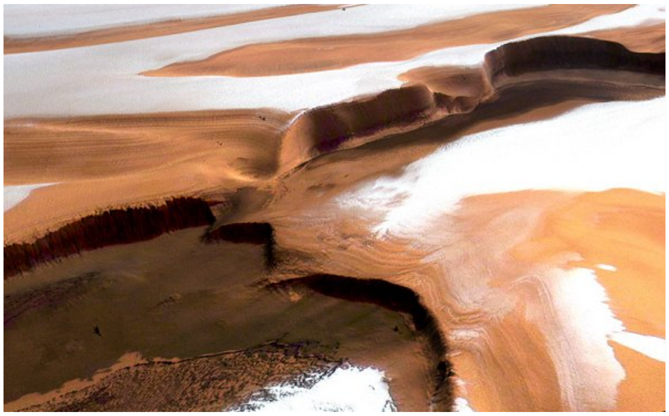


2) Wikipedia, "Barchan", `http://en.wikipedia.org/wiki/Barchan`

Wind-blown sand accumulates on the front of the barchan, and then slides down the "slip face" on the back.

Barchans gradually migrate in the direction of the wind at speeds of about 1-20 meters per year, with small barchans moving faster than big ones. In fact, when they collide, the smaller barchans pass right through the big ones! So, they act like solitons in some ways.

It would be great to see one of these frosty Martian barchans close up. We almost can do it now! The European Space Agency's orbiter called Mars Express took this wonderful closeup, already shown in "Week 211":



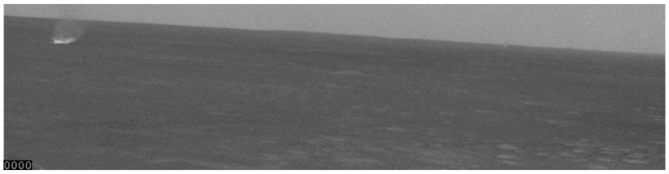
3) ESA/DLR/FU Berlin (G. Neukum), "Glacial, volcanic and fluvial activity on Mars: latest images", `http://www.esa.int/SPECIALS/Mars_Express/SEMLF6D3M5E_1.html`

However, this is not a barchan — it's a lot bigger. On top of the picture we see dunes, but then there's a cliff almost 2 kilometers high leading down into what may be a volcanic caldera. The white stuff is ice, while the dark stuff could be volcanic ash.

It's actually a bit surprising that there's enough wind on Mars to create dunes. After all, the air pressure there is about 1% what it is here on Earth! But in fact the wind speed on Mars often exceeds 200 kilometers per hour, with gusts up to 600 kilometers per hour. There are dust storms on Mars so big they were first seen from telescopes on Earth long ago. So, wind is a big factor in Martian geology:

4) NASA, "Mars exploration program: dust storms", `http://mars.jpl.nasa.gov/gallery/duststorms/`

The Mars rover Spirit even got its solar panels cleaned by some dust devils, and it took some movies of them:



5) NASA, "Exploration rover mission: dust devils at Gusev, Sol 525", `http://marsrovers.nasa.gov/gallery/press/spirit/20050819a.html`

Turning to mathematical physics per se, I can't resist pointing out that sand piles became very fashionable in this subject a while back.

Why? Well, for this I need to explain "self-organized criticality".

First, note that when a pile of sand is exactly at its angle of repose, it will suffer lots of little landslides — and a few of these will become big.

The theory of "critical phenomena" suggests that in this situation, the probability that a landslide grows to size L should satisfy a power law. In other, it should be proportional to

$$\frac{1}{L^c}$$

for some number c called the "critical exponent". At least, this type of behavior is seen in many other situations where a physical system is on the brink of some drastic change — or more precisely, a "critical point".

When a system is not at a critical ponit, we typically see exponential laws, where the probability of a disturbance of size L is proportional to

$$\exp\left(-\frac{L}{L_0}\right)$$

where L_0 is a fixed length scale. This means that our system will look qualitatively different depending how much we zoom in with our microscope. At length scales shorter than L_c, disturbances are really common, while at larger length scales they're incredibly rare.

When a system *is* at a critical point, it's self-similar: you can zoom in or zoom out, and it looks qualitatively the same! It has no specific length scale. This is what the power law says.

Here's a good place to learn the basics of power laws and self-similarity:

6) Manfred R. Schroeder, *Fractals, Chaos, Power Laws*, W. H. Freeman, New York, 1992.

What makes sand dunes interesting is that as they seem to *enjoy* living on the brink of danger. As the wind blows, they heap up until their slip face is right at the angle of repose... ready for landslides!

This is the idea of "self-organized criticality": some physical systems seem to spontaneously bring themselves towards critical points, without any need for us to tune their parameters to special values.

The paper that introduced this idea came out in 1987:

7) Per Bak, Chao Tang and Kurt Wiesenfeld, "Self-organized criticality: an explanation of $1/f$ noise", *Phys. Rev. Lett.* **59** (1987) 381–384.

They came up with a simple model of a sand pile that exhibits self-organized criticality. In the words of Jos Thijssen:

> *Bak and co-workers modelled the sand pile as a regular array of columns consisting of cubic sand grains. Addition of new grains is simply performed by selecting a column at random and increasing its height by one. If the column then exceeds its neighbours in height more than some threshold, it will "collapse": it will lose some grains which are distributed evenly over its nearest neighbours. As this collapse alters the height differences involving those neighbours, there is the possibility that they collapse in turn. A cascade process sets in until all height differences are below the threshold. The size of such an avalanche is defined as the number of sand grains sliding as a result of a single grain of sand being added to the pile.*
>
> *What is so interesting about the sand pile model? It turns out that the sides of the sand pile acquire a specific slope, which is such that the distribution of avalanches as function of size scales as a power law. Power laws indicate the absence of scale and indeed avalanches on all scales are sustained for the equilibrium slope. If the slope is changed artificially from its equilibrium value, the distribution is no longer a power law, but it will have an intrinsic scale (e.g. exponential). Power laws and absence of scale are the signature of a system being critical. Because the sand pile tends to adjust the slope of its sides until the power law scaling sets in, the criticality is called "self-organised".*

If your computer runs Java applets, you can play with Thijssen's simulation sand pile and see the avalanches yourself:

8) Jos Thijssen, "The sand pile model and self organised criticality", `http://www.tn.tudelft.nl/tn/People/Staff/Thijssen/sandexpl.html`

And here's a cellular automaton sand pile you can play with:

9) Albert Schueller, "Cellular automaton sand pile model", `http://schuelaw.whitman.edu/JavaApplets/SandPileApplet/`

This one is only 2-dimensional, so the avalanches are less dramatic, but you can have some fun using the mouse to build structures that impede the motion of sand.

Like a speck of sand landing at the right place at the right time, the original paper of Bak *et al* started a huge landslide of work on self-organized criticality, some of which has been popularized here:

10) Per Bak, *How Nature Works: The Science of Self-Organized Criticality*, Copernicus, New York, 1996.

As you can guess from the title "How Nature Works", some people got a little carried away with the importance of self-organized criticality. Then there was a kind of backlash, just as happened with fractals, chaos, and catastrophe theory. These are all perfectly respectable and interesting topics in mathematical physics that suffered from being oversold. People are always eager to find the secret key that will unlock all the mysteries of the universe. So, if some new idea seems very general, people will run around trying to unlock all the mysteries of the universe with it — and become sorely disappointed when it only unlocks *some*.

I'd be interested to see how well mathematical physicists can model actual sand dunes. These display an interesting complexity of behavior, as the pictures here show:

11) US Army Corps of Engineers, "Dunes", `http://www.tec.army.mil/research/products/desert_guide/lsmsheet/lsdune.htm`

I've only looked at a few papers on the subject, all dealing with barchans:

12) V. Schwaemmle and H. J. Herrmann, "Solitary wave behaviour of sand dunes", *Nature* **426** (Dec. 11, 2003), 619–620.

13) Klaus Kroy, Gerd Sauermann, and Hans J. Hermann, "Minimal model for sand dunes", *Phys. Rev. Lett.* **88** (2002), 054301. Also available at `cond-mat/0101380`.

14) H. Elbelrhiti, P. Claudin, and B. Andreotti, "Field evidence for surface-wave-induced instability of sand dunes", *Nature* **437** (Sep. 29, 2005), 720–723.

The first paper describes how barchans pass through each other like solitons, simulating them by an equation that's described in the second one. (By the way, the term "minimal model" in the title of the second paper is not being used in the sense familiar in conformal field theory!)

The third paper reports the results of a 3-year field study: in reality, barchans are not stable, and big ones (called "megabarchans") can break apart into smaller "elementary barchans".

If you're more interested in Mars than the mathematical physics of sand dunes, you'll be happy to hear that Google has just moved to drastically expand its customer base by introducing "Google Mars":

15) "Google Mars", `http://www.google.com/mars/`

Using this you can explore many features of Mars, including its dunes.

I'm getting a little tired out, but there's one thing I've been meaning to mention for a while. It's actually related to renormalization, which is secretly the same subject as this "critical point" business I just mentioned. But, it's not about sand piles — it's about quantum gravity!

In "Week 222" I spoke about the work of Lauscher and Reuter, who claim to have found evidence for an ultraviolet fixed point in quantum gravity without matter. In other words, as you zoom in closer and closer, they claim quantum gravity without matter acts more and more like some fixed theory. This would be big news: it would suggest that gravity without matter is a sensible theory, contrary to what everyone in string theory says!

Not surprisingly, the string theorist Jacques Distler examined Lauscher and Reuter's work with a critical eye. And, he wrote up a nice explanation of the problems with their work:

16) Jacques Distler, "Unpleasantness", `http://golem.ph.utexas.edu/~distler/blog/archives/000648.html`

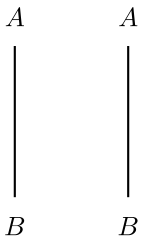
Briefly, the problem is that Lauscher and Reuter make a drastic approximation. They start with the "exact renormalization group equation", which is a beautiful thing: it says how a Lagrangian for a field theory at one length scale gives rise to an effective Lagrangian for the same theory at a larger length scale. However, then they truncate the incredibly complicated formula for a fully general Lagrangian, restricting to Lagrangians with only an Einstein-Hilbert term and a cosmological constant. Like Distler, I see no reason to think this approximation is valid. So, their claimed ultraviolet fixed point could be an artifact of their method.

Whether it's worth going further and checking this by considering a slightly less brutal approximation, using Lagrangians with a few more terms, is a matter of taste. Distler doesn't think so. I hope Lauscher and Reuter do. If they don't, we may never know for sure what happens. I think it's actually rather amazing that they get an fixed point with their brutal approximation, instead of coupling constants that run to infinity or zero, which is what I would have naively expected. But who knows? Maybe this is easily understood if you think hard enough.

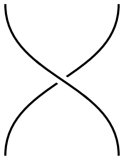
Today I was also going to talk about the 3-strand braid group, the group $\mathrm{PSL}(2,\mathbb{Z})$, and rational tangles, but now I don't have the energy. So instead, I'll just put out a request for help!

There's a wonderful game invented by John Conway called "rational tangles". Here's how it works. It involves two players and a referee.

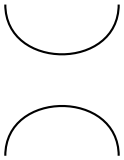
The players, call them A and B, start by facing each other and holding ropes in each hand connecting them together like this:



This is called "position 0". The referee then cries out either *add one!* or *take the negative reciprocal!*. If the referee yells *add one!*, player B has to switch which hand he's using to hold which rope, making sure to pass the right one over the left, like this:



This is called "position 1", since we started with "position 0" and then did *add one!* But if the referee says *take the inverse reciprocal!*, both players must cooperate to move all four ends of the ropes a quarter-turn clockwise, like this:



This is called "position $-1/0$", since we started with 0 and then did *take the negative reciprocal!*

The referee keeps crying *add one!* or *take the negative reciprocal!* in whatever order she feels like, and players A and B keep doing the same sort of thing: either player B switches the ropes right over left, or both players rotate the whole tangle a quarter-turn clockwise. It's actually best if the referee doesn't start with *take the negative reciprocal!*, since some people refuse to divide by zero, for religious reasons. But, it's perfectly legal in this game.

Anyway, after a while the ropes will be in "position p/q" for some complicated rational number p/q. The'll be all tangled up — but in a special way, called a "rational tangle".

Then the players have to *undo* the tangling and get back to "position 0". They may not remember the exact sequence of moves that got them into the mess they are in. In fact the game is much more fun if they *don't* remember. It's best to do it at a party, possibly after a few drinks.

Luckily, any sequence of *add one!* and *take the negative reciprocal!* moves the players make that carry their number back to 0, will carry their tangle back to "position 0". So they just need to figure out how to get their number back to 0, and the tangle will automatically untangle itself. That's the cool part! It's a highly nonobvious theorem due to Conway.

I'm vaguely aware of a few proofs of this fact. As far as I know, Conway's original proof uses the Alexander-Conway polynomial:

16) John Horton Conway, "An enumeration of knots and links and some of their algebraic properties", in *Computational Problems of Abstract Algebra*, ed. John Leech, Pergamon Press, Oxford, 1970, 329–358.

There's also a proof by Goldman and Kauffman using the Jones polynomial:

17) Jay R. Goldman and Louis H. Kauffman, "Rational tangles", *Advances in Applied Mathematics* **18** (1997), 300–332. Also available at `http://www.math.uic.edu/~kauffman/RTang.pdf`

There are also two proofs in here:

18) Louis H. Kauffman and Sofia Lambropoulou, "On the classification of rational tangles", available as `math.GT/0311499`.

But here's what I want to know: is there a proof that makes extensive use of the group $\mathrm{PSL}(2,\mathbb{Z})$ and its relation to topology?

After all, the basic operations on rational tangles are "adding one" and "negative reciprocal", and these generate all the fractional linear transformations

$$z \mapsto \frac{az+b}{cz+d}$$

with a, b, c, d integer and $ad - bc = 1$. The group of these transformations is $\mathrm{PSL}(2,\mathbb{Z})$. It acts on rational tangles, and Conway's theorem says this action is isomorphic to the obvious action of $\mathrm{PSL}(2,\mathbb{Z})$ as fractional linear transformations of the "rational projective line", meaning the rational numbers together with a point at infinity. Since $\mathrm{PSL}(2,\mathbb{Z})$ has lots of relations to topology, there should be some proof of Conway's theorem that *uses* these relations to get the job done.

Does anybody know one?

Finally, the answer to the psychic powers puzzle: if you did the calculation right, you got the last digit of the book's ISBN number — unless your answer was 10, in which case the ISBN number should end in the letter X.

This trick is called a "check sum" or "check digit": it's a way to spot errors. The Universal Product Code, used in those bar codes you see everywhere, also has a check digit. So do credit cards.

Addendum: Aaron Lauda and James Given had comments. Lauda wrote:

> *Usually people describe the scheme in a different way, which is actually equivalent to what you said. Denote the 10 digit ISBN number as N_i, for $1 \leqslant i \leqslant 10$. Compute*
>
> $$\sum_{i=1}^{10} (11-i)N_i \mod 11 \qquad (\star)$$
>
> *which should give you zero. That is take the first digit, multiply it by 10, the second by 9, etc. Add them up and compute the sum* $\mod 11$. *You will always get zero.*
>
> *Some fun things you might like to add:*
>
> 1. *If you make a mistake writing down a single digit in the ISBN then the equation (*) will not equal zero.*

> 2. *The equation (*) may fail to give you zero* $\mod 11$ *if you make a mistake with two of the digits, but it will never fail if you interchange two adjacent digits.*
>
> *Regards,*
> *Aaron*

Given wrote:

> *Self-organized criticality (SOC) does in fact involve special settings of the parameters in a model.*
>
> *SOC occurs in sandpile models because one adds the sand extremely slowly, i.e., one grain at a time. Otherwise a critical state is not obtained. This makes SOC be a special example of dynamical critical phenomena in the case that the flux variable (here the rate of sand addition) is set to* ε_+*, i.e., an infinitesimal value greater than zero. This formulation allows SOC to be studied using quantum field theory.*
>
> *Of course the model is built around an underlying instability, namely the fact that sand piles which are too steep will fall down. Also, one must remove by idealization most sorts of friction between sand grains which will otherwise blur out the transition. So SOC is no magic prescription for generating scale invariant phenomena. SOC systems are "special" in the way that equilibrium critical points are "special". As you note, theories of this kind are easily oversold among those eager to believe in magic formulas.*
>
> *Also, you may have been confusing SOC with a favorite concept of the chaos/dynamical systems people, namely the "edge of chaos". It went through several incarnations. Each one tried to formally specify the domain, intermediate between order and chaos, in which complex systems were most "interesting".*
>
> *Wikipedia summarizes this pretty well. I append the listing for convenience.*
>
> *All My Best,*
> *Jim Given*

The Wikipedia article is:

19) "Edge of chaos", Wikipedia, `http://en.wikipedia.org/wiki/Edge_of_chaos`

Week 229

April 13, 2006

I'm visiting Chicago now. I came just in time for a conference in honor of Saunders Mac Lane, one of the founders of category theory, who taught at the University of Chicago for many years and died last year at the age of 95:

1) Category Theory and its Applications: A Conference in Memory of Saunders Mac Lane, `http://www.math.uchicago.edu/~may/MACLANE/`

On Friday there was a memorial service where the friends and family of Mac Lane spoke about him, and a kind of reminiscence session where everyone could tell their favorite stories involving him. Then there were a bunch of math talks, both by people with strong connections to Mac Lane — Peter Johnstone, Bill Lawvere, Peter Freyd, Ieke Moerdijk, Peter May and Steve Awodey — and by people working on higher categories and their applications.

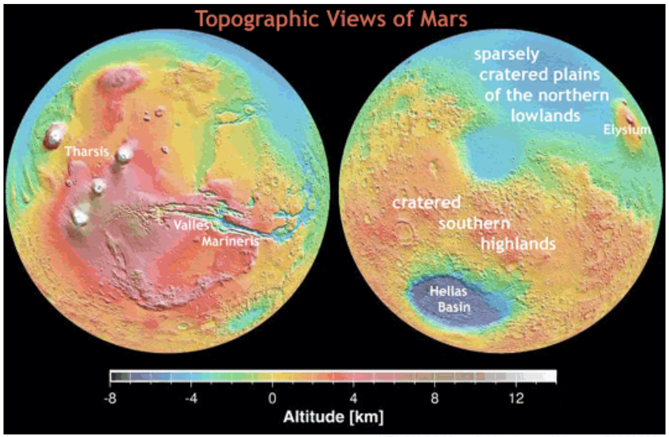
My own connection to Mac Lane is tiny. Everything I do uses his work, but that's true of lots of mathematicians: he discovered so much. Apart from watching him celebrate his 90th birthday at a category theory conference in Portugal back in 1999, the best moment happened when I came to Chicago and gave a talk. He invited me up to his office and we talked a bit. He told me I should write a book explaining n-categories. I promised I would. . . I was too shy to say much.

Now I'm trying to write that book, and I just happen to be staying in Mac Lane's old office, which makes me feel especially obliged to do it. This office is on the third floor of the Ryerson Physical Laboratory. It has a very high ceiling, and one wall is lined with two stacks of metal bookshelves. You'd need a ladder to reach the top! When I spoke to Mac Lane in his office, they were all full of books. Alas, they're empty now.

Next time I'll say a bit about Julie Bergner's talk at the Maclane memorial conference — she spoke about derived categories of quiver representations and quantum groups. But the conference was so intense and exhausting that first I need to recover by thinking about something completely different. So, I'll concentrate on last week's puzzle about rational tangles.

But first: the astronomy picture of the week!

It'll be more fun after a little background. The northern part of Mars is very different from the rest. It's much smoother, and the altitude is much less:


Topographic Views of Mars
sparsely cratered plains of the northern lowlands
Elysium
Tharsis
Valles Marineris
cratered southern highlands
Hellas Basin
-8
-4
0
4
8
12
Altitude [km]

(JPL/NASA Mars Orbiter Laser Altimeter, PIA02820)

2) Linda M. V. Martel, "Ancient floodwaters and seas on Mars", `http://www.psrd.hawaii.edu/July03/MartianSea.html`

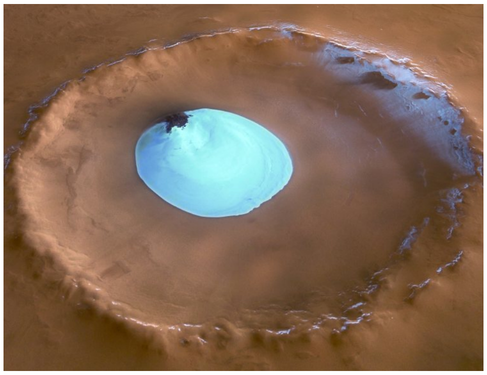
Why is this?

Many scientists believe the north was an ocean during the Hesperian Epoch, a period of Martian history that stretches from about 3.5 to about 1.8 billion years ago. In particular, the beautifully named "Vastitas Borealis", an enormous plain that covers most of northern Mars, has textures that may have been formed by an ocean that froze and then slowly sublimated. (Sublimation is what happens when ice turns directly into water vapor without actually melting.) Mike Carr and James Head wrote a paper suggesting that around the end of the Hesperian, about 30% of the water on Mars evaporated and left the atmosphere, drifting off into outer space... part of the danger of life on a planet without much gravity:

3) M. H. Carr and J. W. Head, III, "Oceans of Mars: An assessment of the observational evidence and possible fate", *Journal of Geophysical Research* **108** (2003), 5042.

The rest of the water is now frozen at the poles or lurking underground.

And that brings us to our picture. Here's some ice in a crater in Vastitas Borealis:



4) “Water ice in crater at Martian north pole”, European Space Agency (ESA), http://www.esa.int/esaMI/Mars_Express/SEMGKA808BE_3.html

Perhaps this is a remnant of a once mighty ocean!

The picture is close to natural color, but the vertical relief is exaggerated by a factor of 3. The crater is 35 kilometers wide and 2 kilometers deep. It’s incredible how they can get this kind of picture from satellite photos and lots of clever image processing. I hope they didn’t do *too* much stuff just to make it look pretty.

Next: rational tangles.

In “Week 228”, I asked for help understanding the connection between rational tangles and the group $\mathrm{PSL}(2,\mathbb{Z})$. I got a great reply from Michael Hutchings, which winds up relating these ideas to the branched double cover of the sphere by the torus. And, this gives me an excuse to tell you some stuff I learned from James Dolan about elliptic functions and a map of the world called “Peirce’s quincuncial”.

So, let’s dive in!

Did you ever try to wrap a sphere around itself twice? Mentally, I mean? Slit it open, grab it, pull it, stretch it, wrap it around itself twice, and glue the seams back together?

It’s not hard. You just take the Riemann sphere — the complex numbers together with a point at infinity — and map it to itself by the function

$$f(z) = z^2$$

If you think of the sphere as the surface of the Earth, with zero at the south pole and infinity as the north pole, this function doubles the longitude. So, it wraps the sphere around itself twice!

I hope you’re visualizing this.

This function is not quite a “double cover”, because it’s not quite two-to-one everywhere. Only one point gets mapped to $z = 0$, namely itself, and only one point gets mapped to $z = \infty$, namely itself. Elsewhere f is two-to-one.

If you walk once around the north pole or south pole, and then apply the function f to your path, you get a path that goes around these points *twice*. Summarizing these properties, we call the function a “branched double cover” of the sphere by itself, with zero and infinity as branch points.

Now, how about wrapping a torus twice around a sphere?

This too can be done. It turns out there’s a nice branched double cover of the sphere by the torus, which has four branch points.

To visualize this, first take the surface of the Earth and mold it into a regular octahedron. There will be six corners: the north pole, the south pole, the east pole, the west pole, the front pole and the back pole. Now take the octahedron and unfold it like this:

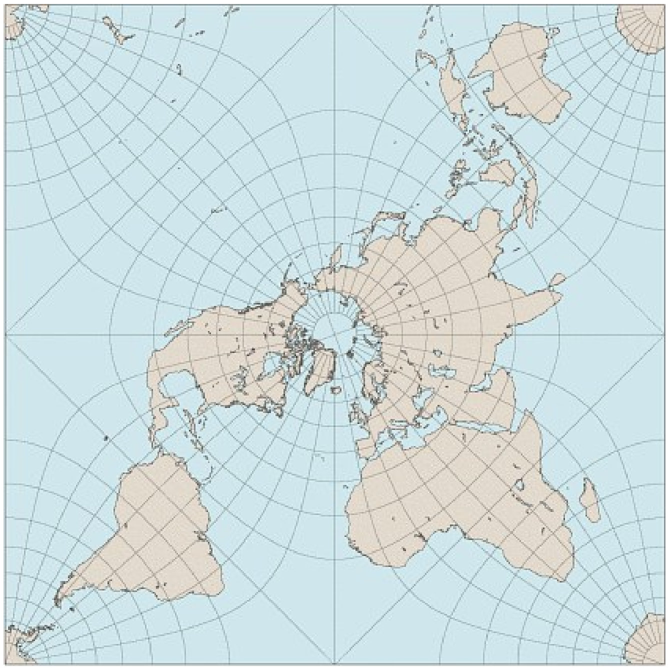
```
S-----B-----S
|    /|\    |
|   / | \   |
|  /  |  \  |
| /   |   \ |
|/    |    |
W-----N-----E
|\    |    /|
| \   |   / |
|  \  |  /  |
|   \ | /   |
|    |/    |
S-----F-----S
```

We get an interesting map of the world, which was invented in 1876 by the American mathematician and philosopher C. S. Peirce while he was working at the U. S. Coast and Geodetic Survey. This map is called “Peirce’s quincuncial”, since when you arrange five dots this way:

```
o   o
  o
o   o
```

it’s called a “quincunx”. (Somehow this word goes back to the name of an ancient Roman coin. I don’t understand how this pattern is related to the coin.)

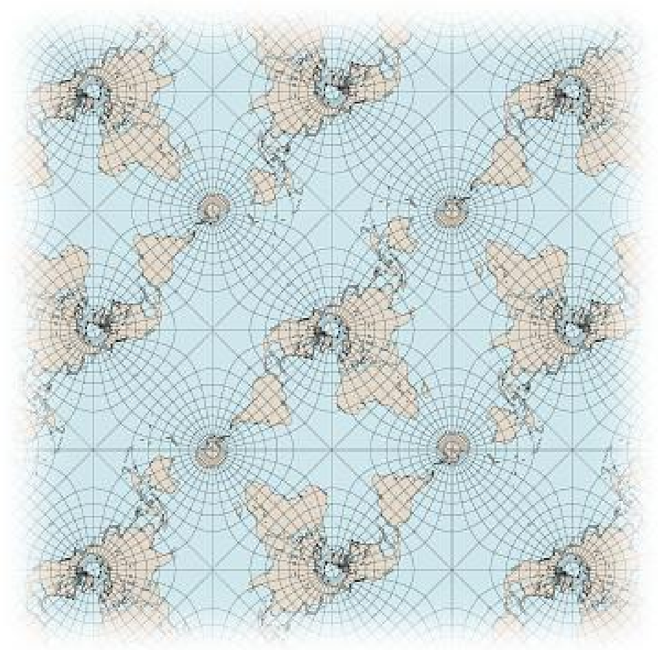
This is how Peirce’s quincuncial looks as an actual map:



5) Carlos A. Furuti, “Conformal projections”, `http://www.progonos.com/furuti/MapProj/Normal/ProjConf/projConf.html`

The cool part is that you can tile the plane indefinitely with this map:

```
S-----B-----S-----F-----S-----B-----S-----F-----S
|    /|\    |    /|\    |    /|\    |    /|\    |
|   / | \   |   / | \   |   / | \   |   / | \   |
|  /  |  \  |  /  |  \  |  /  |  \  |  /  |  \  |
| /   |   \ | /   |   \ | /   |   \ | /   |   \ |
|/    |    |/    |    |/    |    |/    |    |
W-----N-----E-----N-----W-----N-----E-----N-----W
|\    |    /|\    |    /|\    |    /|\    |    /|
| \   |   / | \   |   / | \   |   / | \   |   / |
|  \  |  /  |  \  |  /  |  \  |  /  |  \  |  /  |
|   \ | /   |   \ | /   |   \ | /   |   \ | /   |
|    |/    |    |/    |    |/    |    |/    |
S-----F-----S-----B-----S-----F-----S-----B-----S
|    /|\    |    /|\    |    /|\    |    /|\    |
|   / | \   |   / | \   |   / | \   |   / | \   |
|  /  |  \  |  /  |  \  |  /  |  \  |  /  |  \  |
| /   |   \ | /   |   \ | /   |   \ | /   |   \ |
|/    |    |/    |    |/    |    |/    |    |
E-----N-----W-----N-----E-----N-----W-----N-----E
```



This gives a branched cover of the sphere by the plane! It has branch points at the east, west, front and back poles, since walking once around a point like that on the above map corresponds to walking around it twice on the actual Earth. This is pretty weird, but Peirce cleverly located two of these branch points in the Pacific Ocean, one in the Atlantic, and one in the Indian Ocean.

We can be less extravagant and get a branched cover of the sphere by the torus if we take the smallest parallelogram whose opposite edges match up:

```
            B
           /|\
          / | \
         /  |  \
        /   |   \
       /    |    \
      E-----N-----W
     /|\    |    /|\
    / | \   |   / | \
   /  |  \  |  /  |  \
  /   |   \ | /   |   \
 /    |    |/    |    \
B-----S-----F-----S-----B
 \    |    /|\    |    /
  \   |   / | \   |   /
   \  |  /  |  \  |  /
    \ | /   |   \ | /
     |/    |    |/
      W-----N-----E
       \    |    /
        \   |   /
         \  |  /
          \ | /
           |/
            B
```

This would actually be a square if I could draw it right in ASCII. We can curl this into a torus by gluing together the opposite edges. There's then an obvious function from this torus to the sphere sending both points labelled "N" to the north pole, both points labelled "S" to the south pole, and so on.

This function is mostly two-to-one, but it's one-to-one at the points labelled E, F, W, and B. After all, there's just *one* point of each of these sorts in the above picture after we glue together the opposite edges. There are *two* copies of any other sort of point.

So, our function is a branched double cover of the sphere by the torus, which has four branch points. In fact, this function is quite famous. It's an example of an "elliptic function"!

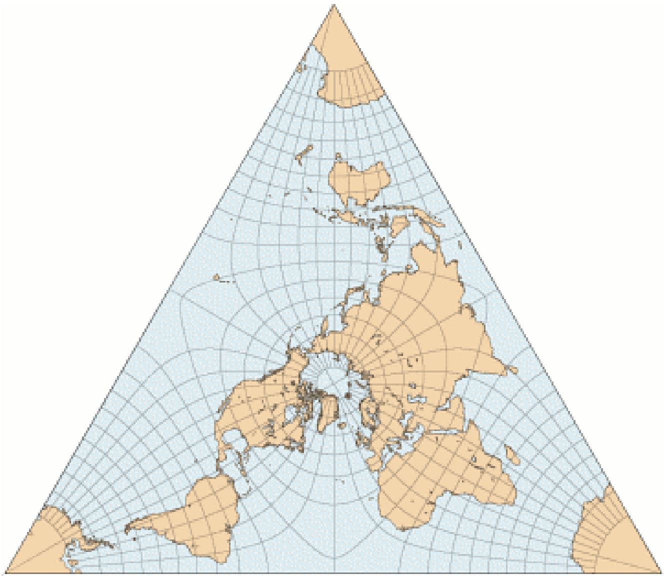
I explained elliptic functions way back in "Week 13". Briefly, what we just did starting with this parallelogram:

```
            B
           /|\
          / | \
         /  |  \
        /   |   \
       /    |    \
      E-----N-----W
     /|\    |    /|\
    / | \   |   / | \
   /  |  \  |  /  |  \
  /   |   \ | /   |   \
 /    |    |/    |    \
B-----S-----F-----S-----B
 \    |    /|\    |    /
  \   |   / | \   |   /
   \  |  /  |  \  |  /
    \ | /   |   \ | /
     |/    |    |/
      W-----N-----E
       \    |    /
        \   |   /
         \  |  /
          \ | /
           |/
            B
```

actually works for a parallelogram of any shape. The parallelogram curls up to give a torus, and we get a map from this torus to the Riemann sphere, called an "elliptic function".

As before, this is a branched double cover with four branch points. However, where the branch points sit on the sphere depends on the shape of the parallelogram. By picking the parallelogram carefully, you can put the branch points wherever you want! Peirce's neat idea was to put them evenly spaced along the equator — at the east, front, west and back poles. This is nice and symmetrical.

It's also especially nice to put the branch points at the vertices of a regular tetrahedron. I'm not sure, but this may give a map developed by the cartographer Laurence P. Lee in 1965. There's also a picture of this on Furuti's webpage:



In fact, these two specially nice locations for branch points correspond to the two most symmetrical lattices in the plane: the square one and the hexagonal one. I talked about these in "Week 125" — they're really important in the theory of elliptic functions, and even in string theory.

Anyway: for any parallelogram we can make a map of the Earth that tiles the plane, with tiles shaped like this parallelogram. A cool thing about these maps is that they're all "conformal" — they preserve angles except at the branch points. If you want to show off, you express this by saying "elliptic functions are complex analytic".

But now I'm digressing a little. Let's get back on track. What does all this have to do with rational tangles??

Recall my puzzle from last time. We build rational tangles by starting with the trivial one, which we call "zero"

```
 |   |
 |   |
 |   |
 |   |
```

and repeatedly doing two operations. The first is a twisting operation that we call "adding one":

```
 |   |                |   |
 |   |                |   |
 |   |                |   |
-------              -------
|  T  |   |---->     |  T  |        =  "T + 1"
-------              -------
 |   |                 \ /
 |   |                  /
 |   |                 / \
```

where the box labelled "T" stands for any tangle we've built so far. The second is a rotation that we call "negative reciprocal":

```
 |   |             |     |
 |   |             |     |    ____
 |   |             |     |   /    \
-------            |    -------    |
|  T  |   |---->   |    |  T  |    |    =  "-1/T"
-------            |    -------    |
 |   |              \___/   |      |
 |   |                      |      |
 |   |                      |      |
```

Using these tricks we can try to assign a rational number to any rational tangle. The shocking theorem is that this number is indeed well-defined, and in fact a complete invariant of rational tangles.

Every operation built from "adding one" and "negative reciprocal" looks like this:

$$z \mapsto \frac{az+b}{cz+d}$$

```
         az + b
z \mapsto   -------
         cz + d
```

with a, b, c, d integer and $ad - bc = 1$. The group of these transformations is called $\mathrm{PSL}(2, \mathbb{Z})$. This group acts on the rational numbers together with a point at infinity (the "rational projective line") by the formula above. It also acts on rational tangles. The puzzle is to see why these actions are isomorphic. The proofs I listed in "Week 228" show it's true; the problem is to understand what's really going on!

Here's the answer given by Michael Hutching on `sci.math.research`:

> *There's a simple topological interpretation of the element of the rational projective line associated to a rational tangle. I don't know how to use this to prove the theorem, and I don't know a reference for it (maybe it is in one of the references you cited). Anyway, regard a rational tangle as a two-component curve C in the 3-ball B^3 whose four boundary points are on the 2-sphere S^2. Consider the double branched cover of B^3 along C. This is a 3-manifold Y whose boundary can be identified with the 2-torus T^2. (In fact Y is a solid torus.) The inclusion of T^2 into Y induces a map from $H_1(T^2)$ to $H_1(Y)$, and the kernel of this map is a one-dimensional subspace of $H_1(T^2) = \mathbb{Z}^2$. If I am not mistaken, this is the element in question of the rational projective line.*

In other words, we take a 3-dimensional ball and draw a picture of a rational tangle in it:

```
    .......
  .   | |   .
 .   -----   .
 .  |  T  |  .
 .   -----   .
  .   | |   .
    .......
```

The boundary of this ball is a sphere with 4 points marked. If we take a branched double cover of the sphere with these as the branch points, we get a torus T^2. If we take a branched double cover of the whole ball with everything along the vertical lines as branched points, we get a solid doughnut Y having T^2 as its boundary.

This gets the torus into the game, and also the branched cover I was talking about. And this gets the group $\mathrm{PSL}(2, \mathbb{Z})$ into the game! $\mathrm{SL}(2, \mathbb{Z})$ is the group of 2×2 matrices with determinant 1. When you mod out by the matrices ± 1, you get $\mathrm{PSL}(2, \mathbb{Z})$. But, topologists know that $\mathrm{SL}(2, \mathbb{Z})$ is the "mapping class group" of the torus — the group of orientation-preserving diffeomorphisms modulo those that can be smoothly deformed to the identity.

So, something nice is happening.

Even better, the rational first homology group of the torus is $\mathbb{Q}^2$ (pairs of rational numbers), and $\mathrm{SL}(2, \mathbb{Z})$ acts in the obvious way, by matrix multiplication:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix}$$

It therefore acts on the set of 1-dimensional subspaces of $\mathbb{Q}^2$. Any such subspace consists of vectors like this:

$$\left(\begin{array}{c} kx \\ ky \end{array}\right)$$

The subspace is determined by the ratio x/y, which however could be infinite — so it's just a point in the rational projective line. So, we get an action of $\mathrm{SL}(2,\mathbb{Z})$ on the rational projective line. Indeed we get an action of $\mathrm{PSL}(2,\mathbb{Z})$ since ± 1 act trivially. And, you can easily check that it's the action we've already seen:

$$\left(\begin{array}{cc} a & b \\ c & d \end{array}\right) : z \mapsto \frac{az+b}{cz+d}$$

In short: "projectivizing" the action of mapping class group of the torus on its first homology gives the usual action of $\mathrm{PSL}(2,\mathbb{Z})$ on the rational projective line.

What we need next is a natural way to assign to any rational tangle a 1-dimensional subspace of the homology of the torus. And this is what Hutchings describes: a rational tangle gives a way of mapping the torus T^2 into the solid torus Y, and this gives a map on rational homology

$$H_1(T^2) \to H_1(Y)$$

whose kernel is a 1-dimensional subspace of $H_1(T^2)$.

There's more stuff to check. . . .

Personally I've been trying to think of the mapping class group of the 4-punctured sphere as acting on pictures like this:

```
    .......
  .   | |   .
 .   -----   .
 .  |  T  |  .
 .   -----   .
  .   | |   .
    .......
```

and show that the resulting action on rational tangles factors through a homomorphism from this mapping class group to $\mathrm{PSL}(2,\mathbb{Z})$. The mapping class group should be generated by the twist

```
  |   |                |   |
  |   |                |   |
  |   |                |   |
 -------              -------
 |  T  |   |---->     |  T  |
 -------              -------
  |   |                 \ /
  |   |                  /
  |   |                 / \
```

and the 90 degree rotation

```
 |   |             |     |
 |   |             |     |    ----
 |   |             |     |   /    \
-------            |    -------    |
|  T  |   |---->   |    |  T  |    |
-------            |    -------    |
 |   |              \___/   |      |
 |   |                      |      |
 |   |                      |      |
```

and our homomorphism should map these to the famous matrices

$$T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \text{“shear”}$$

and

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \text{“90 degree rotation”}$$

respectively. If this works, and I could figure out the kernel of this homomorphism and show it acts trivially on rational tangles, I think I'd be almost done. But, I haven't had time!

By the way, if this works, there's a beautiful little sideshow where we use as generators of $\mathrm{SL}(2,\mathbb{Z})$ not the above matrices but S and

$$ST = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}$$

I explained why these are so great in “Week 125”. S is a symmetry of the square lattice, while ST is a symmetry of the hexagonal lattice. The square lattice gives Peirce's quincuncial map, while the hexagonal one presumably gives Laurence Lee's triangular map!

So, there's some intriguing story about elliptic functions and rational tangles taking shape before our eyes. . . . and if I weren't so darn busy, I'd figure out all the details and write a little paper about it.

Before quitting, there's one more thing I can't resist mentioning. Any ordered 4-tuple of points (a, b, c, d) in the Riemann sphere gives a number called its “cross-ratio”:

$$\frac{(a-b)(c-d)}{(a-d)(c-b)}$$

It's a famous fact that you can find a conformal transformation of the Riemann sphere mapping one ordered 4-tuple to another if and only if their cross-ratios are equal!

So, we can play a little trick. Given a lattice we can get a branched double cover of the Riemann sphere as I sketched earlier. Then we can use the location of the branch points to calculate a cross ratio.

But actually, I'm being a bit sloppy here. To compute a cross ratio from a lattice, we need some extra information to *order* the 4-tuple of branch points. In other words, if one of the points S is the origin here:

```
S-----B-----S-----F-----S-----B-----S-----F-----S
|    /|\    |    /|\    |    /|\    |    /|\    |
|   / | \   |   / | \   |   / | \   |   / | \   |
|  /  |  \  |  /  |  \  |  /  |  \  |  /  |  \  |
| /   |   \ | /   |   \ | /   |   \ | /   |   \ |
|/    |    |/    |    |/    |    |/    |    |
W-----N-----E-----N-----W-----N-----E-----N-----W
|\    |    /|\    |    /|\    |    /|\    |    /|
| \   |   / | \   |   / | \   |   / | \   |   / |
|  \  |  /  |  \  |  /  |  \  |  /  |  \  |  /  |
|   \ | /   |   \ | /   |   \ | /   |   \ | /   |
|    |/    |    |/    |    |/    |    |/    |
S-----F-----S-----B-----S-----F-----S-----B-----S
|    /|\    |    /|\    |    /|\    |    /|\    |
|   / | \   |   / | \   |   / | \   |   / | \   |
|  /  |  \  |  /  |  \  |  /  |  \  |  /  |  \  |
| /   |   \ | /   |   \ | /   |   \ | /   |   \ |
|/    |    |/    |    |/    |    |/    |    |
E-----N-----W-----N-----E-----N-----W-----N-----E
```

and the lattice is taken just big enough so the pattern repeats, we need enough information to *label* the points E, F, W and B. This extra information amounts to "choosing a basis for the 2-torsion subgroup of the plane modulo the lattice". So, the cross ratio gives a "modular function of level 2".

Hmm, this is getting pretty jargonesque! I don't want to explain the jargon now, but you can read all about this trick and its consquences in Lecture 9 here:

6) Igor V. Dolgachev, Lectures on modular forms, Fall 1997/8, available at `http://www.math.lsa.umich.edu/~idolga/modular.pdf`

Addenda: Andrei Sobolevskii points out that the etymology of the word "quincunx" is explained here:

7) "Quincunx", World Wide Words, `http://www.worldwidewords.org/weirdwords/ww-qui2.htm`

Very briefly, "quincunx" was a Latin word for "five twelfths", from *quinque* and *uncia*. The latter word is also the root of the word "ounce". They had a copper coin called the *as* weighing twelve ounces (!), and the quincunx was apparently not a coin a symbol for 5/12 of an *as* — or in other words, 5 ounces of copper.

After reading the above, Peter Dickof clarified and corrected the story:

> *Love "This Week's Finds" (though I seldom follow it all) and can't pass up the opportunity to say something.*
>
> *The* as *was indeed a unit of currency and also a specific bronze coin. Early asses (*aes *= bronzes — hence the AE ligature. . .) were full Roman pounds (*librae

> *— hence the British pound “” sign) of 288 scruples with twelve unciae to the pound. Each uncia was ˜27 grams, a modern ounce near enough. Debasement set in around the time of the first Punic war and accelerated through the second (Hannibal’s), by the end of which an as, still of bronze, weighed only ˜30 grams.*
>
> *Multiples of the as were minted (misnomer, this was a cast currency): the* decussis *(X asses),* quincussis *(V asses),* tressis *(III asses), and the* dupondius *(two-pounder). The asses were marked with the Roman numeral I. Common fractions were the* semis *(half, marked with an S),* triens *(third, sometimes called a quatrunx, marked with four “pellets” or dots),* quadrans *(quarter, also* teruncius*, 3 pellets),* sextans *(sixth, also biunx, 2 pellets),* uncia *and* semuncia *(usually unmarked).*
>
> *There* were *quincunx coins (5 pellets), and also a* dextans *(S + 4 pellets), mostly produced by non-Roman Italians. I have appended photos of three quincunxes produced in Luceria (see Thurlow and Vecchi numbers 274, 281). Note that the “pellets” are sometimes (not always) arranged in a quincunx. Luceria (modern Lucera) is 2/3 of the way across the boot from Neapolis (Naples).*
>
> *During the second Punic war, after they captured Syracuse and its treasure (and killed Archimedes), the Romans introduced the silver* denarius*,* quinarius *and* sesterius*; they were worth 10, 5, 2 1/2 (IIS) asses. The denarius is the origin of the “d” for the British shilling, and was about the size of a dime. Later, circa 141 BC, the value of silver was re-tariffed so that denarii, quinarii and sestertii became worth 16, 8, 4 asses; the names did not change but the quinarius and sestertius became rare.*
>
> *Julius Caesar doubled the pay of a legionary to 300 sestertii per installment (stipendium), 3 installments per year.*
>
> *Later yet, during Imperial Rome, the largest bronze/brass coin minted (no longer a misnomer) was a sestertius; its weight was less than 30 grams by the time of Claudius and falling, always falling. . . .*
>
> *Appropriate references are:*
>
> *— Michael H. Crawford,* Roman Republican Coinage*, Cambridge University Press, Cambridge, 1974. (See picture of Quincunx on plate XVIII.) — Bradbury K. Thurlow,* Italian Cast Coinage: Italian Aes Grave. *Italo G. Vecchi,* Italian Aes Rude, Signatum and the Aes Grave of Sicily. *Printed together by Veechi, London, 1979. — Herbert A. Grueber,* A Catalogue of the Coins of the Roman Republic in the British Museum*, three volumes, reprinted 1970.*

My old pal Squark noted some sloppy language about branched covers. Here’s my reply to what he wrote. I’ve changed what he wrote a tiny bit, for cosmetic reasons.

> *Squark wrote:*
>
> > *Hello John and everyone!*
>
> *Hello! Long time no see! How are you doing?*
>
> *I had written:*

> *There's a simple topological interpretation of the element of the rational projective line associated to a rational tangle. I don't know how to use this to prove the theorem, and I don't know a reference for it (maybe it is in one of the references you cited). Anyway, regard a rational tangle as a two-component curve C in the 3-ball B^3 whose four boundary points are on the 2-sphere S^2. Consider the double branched cover of B^3 along C.*

Squark wrote:

> *What is "*the *double branched cover"? Is there a way to choose a canonical one, or is there only one in this case, for some reason?*

Good point. I hope there's a specially nice one.

To pick a branched cover of B^3 along C, it's necessary and sufficient to pick a homomorphism from the fundamental group of $B^3 \backslash C$ to $\mathbb{Z}/2$. This says whether or not the two sheets switch places as we walk around C following some loop in $B^3 \setminus C$.

> *In the case of a sphere with 4 points removed it should be easy to check.*

Yes.

> *The fundamental group has 4 generators — a, b, c, d (loops around each of the points) and one relation $abc = d$ (since we're on a sphere). Hence, it is freely generated by a, b, c (say).*

Right, the fundamental group of the four-punctured sphere is the free group on 3 generators, F_3. I believe the "specially nice" homomorphism

$$f\colon F_3 \to \mathbb{Z}/2$$

is the one that sends each generator to -1, where I'm thinking multiplicatively:

$$\mathbb{Z}/2 = \{1, -1\}$$

One reason this homomorphism is especially nice is that it also sends $d = abc$ to -1.

So, if you walk around ANY of the four punctures, the two sheets switch!

This is just what you want for the Riemann surface of an elliptic integral, as someone else pointed out in another post: there are four branch points each like the branch point of $\sqrt{z}$. It's also the most symmetrical, beautiful thing one can imagine.

Now let's see if and how this branched cover extends to a branched cover of the ball B^3 with C (two arcs) removed. The fundamental group of $B^3 \setminus C$ is the free group on two generators, say X and Y.

The inclusion of the 4*-punctured sphere in* $B^3 \setminus C$ *gives a homomorphism*

$$g\colon F_3 \to F_2$$

as follows:

$$\begin{aligned} a &\mapsto X \\ b &\mapsto X^{-1} \\ c &\mapsto Y \\ d &\mapsto Y^{-1} \end{aligned}$$

So, to extend our branched cover, we need to write our homomorphism

$$f\colon F_3 \to \mathbb{Z}/2$$

as

$$f = hg$$

for some homomorphism

$$h\colon F_2 \to \mathbb{Z}/2$$

The obvious nice thing to try for h *is*

$$\begin{aligned} X &\mapsto -1 \\ Y &\mapsto -1 \end{aligned}$$

It works, and it's unique!

Lee Rudolph adds:

> *Squark wrote:*
>
> > *What is "*the *double branched cover"? Is there a way to choose a canonical one, or is there only one in this case, for some reason?*
>
> *John Baez wrote:*
>
> > *Good point. I hope there's a specially nice one.*
>
> *In this kind of context, there's always exactly one "double branched cover" that actually* does *branch doubly over every component of the proposed branch locus. In particular, in the context of a rational tangle, of course the pair* (B^3, C) *is homeomorphic to* $(B^2, X) \times I$*, where* X *is a* 2*-point set in* $\mathrm{Int}(B^2)$ *and the homeomorphism isn't required to preserve the tangle structure; so the double branched cover of* B^3 *branched over* C *is the product of the double branched cover of* B^2 *branched over* X *with the interval* I*. Now, because the branching is* double *at each point of* X*, and there are* two *points of* X*, it follows that the monodromy around the boundary of* B^2 *must be trivial, so that we can sew another* B^2 *to that boundary and extend the branched double covering over the resulting* 2*-sphere. But of course the branched double cover of a* 2*-sphere over 2 points is another* 2*-sphere, the model for the situation being* $z \mapsto z^2$ *as a map of*

the Riemann sphere to itself. Now remove the interior of the sewed-on second B^2 from the downstairs S^2, and correspondingly the interiors of its two *preimage B^2s from the upstairs S^2; you see that the double cover of B^2 branched over X is an annulus. (Once you know that, you can see it directly: take an annulus embedded in $\mathbb{R}^3$ as the cylinder where $x^2 + y^2 = 1$ and $-1 \leqslant z \leqslant 1$; rotate it by 180 degrees around the x-axis, and convince yourself that the quotient space is a 2-disk by considering the fundamental domain consisting of those points of the annulus with non-negative y-coordinate.) Then the double cover of B^3 branched over C must be a solid torus. (Again, now that you know this, you can see it directly: take the solid torus to be a tubular neighborhood in $\mathbb{R}^3$ of the circle where $x^2 + y^2 = 1$ and $z = 0$, and again rotate by 180 degrees around the x-axis to give yourself the "deck involution".)*

Lee Rudolph

Well, what if we consider our lives to be formed of a series of interlocking practices, including the very important ones of maintaining a thriving family and community? Then we might learn from a practice with the pedigree of mathematics — *mankind's longest conversation* — about the necessity of certain intellectual and moral virtues.

— *David Corfield*

Week 230

May 4, 2006

As we've seen in previous weeks, Mars is a beautiful world, but a world in a minor key, a world whose glory days — the Hesperian Epoch — are long gone, whose once grand oceans are now reduced to windy canyons and icy dunes. Let's say goodbye to it for now. . . leaving off with this Martian sunset, photographed by the rover Spirit in Gusev Crater on May 19th, 2005:

1) "A moment frozen in time", NASA Mars Exploration Rover Mission, `http://marsrovers.nasa.gov/gallery/press/spirit/20050610a.html`

This week I'll talk about Dynkin diagrams, quivers and Hall algebras. But first, some cool identities!

My student Mike Stay did computer science before he came to UCR. When he was applying, he mentioned a result he helped prove, which relates Goedel's theorem to the Heisenberg uncertainty principle:

2) C. S. Calude and M. A. Stay, "From Heisenberg to Goedel via Chaitin", *International Journal of Theoretical Physics*, **44** (2005), 1053–1065. Also available at `http://math.ucr.edu/~mike/`

Now, this particular combination of topics is classic crackpot fodder. People think "Gee, uncertainty sounds like incompleteness, they're both limitations on knowledge — they must be related!" and go off the deep end. So I got pretty suspicious until I read his paper and saw it was CORRECT. . . at which point I *definitely* wanted him around! The connection they establish is not as precise as I'd like, but it's solid math.

So, now Mike is here at UCR working with me on quantum logic and quantum computation using ideas from category theory. In his spare time, he sometimes fools around with math identities and tries to categorify them — see "Week 184" and "Week 202" if you don't know what that means. Anyway, maybe that's how he stumbled on this:

3) Jonathan Sondow, "A faster product for π and a new integral for $\ln(\pi/2)$", *Amer. Math. Monthly* **112** (2005), 729–734. Also available as math.NT/0401406.

In this paper, Sondow gives eerily similar formulas for some of our favorite math constants. First, one for e:

$$e = \left(\frac{2}{1}\right)^{\frac{1}{1}} \left(\frac{2^2}{1\times 3}\right)^{\frac{1}{2}} \left(\frac{2^3\times 4}{1\times 3^3}\right)^{\frac{1}{3}} \left(\frac{2^4\times 4^4}{1\times 3^6\times 5}\right)^{\frac{1}{4}} \cdots$$

Then, one for $\pi/2$:

$$\frac{\pi}{2} = \left(\frac{2}{1}\right)^{\frac{1}{2}} \left(\frac{2^2}{1\times 3}\right)^{\frac{1}{4}} \left(\frac{2^3\times 4}{1\times 3^3}\right)^{\frac{1}{8}} \left(\frac{2^4\times 4^4}{1\times 3^6\times 5}\right)^{\frac{1}{16}} \cdots$$

Then one for e^γ, where γ is Euler's constant:

$$e^\gamma = \left(\frac{2}{1}\right)^{\frac{1}{2}} \left(\frac{2^2}{1\times 3}\right)^{\frac{1}{3}} \left(\frac{2^3\times 4}{1\times 3^3}\right)^{\frac{1}{4}} \left(\frac{2^4\times 4^4}{1\times 3^6\times 5}\right)^{\frac{1}{5}} \cdots$$

He also points out Wallis' product for $\pi/2$ and Pippenger's for e:

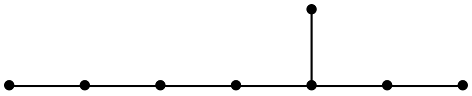
$$\begin{aligned} \frac{\pi}{2} &= \left(\frac{2}{1}\right)^{\frac{1}{1}} \left(\frac{2\times 4}{3\times 3}\right)^{\frac{1}{1}} \left(\frac{4\times 6\times 6\times 8}{5\times 5\times 7\times 7}\right)^{\frac{1}{1}} \cdots \\ e &= \left(\frac{2}{1}\right)^{\frac{1}{2}} \left(\frac{2\times 4}{3\times 3}\right)^{\frac{1}{4}} \left(\frac{4\times 6\times 6\times 8}{5\times 5\times 7\times 7}\right)^{\frac{1}{8}} \cdots \end{aligned}$$

What does it all mean? I haven't a clue! Another mystery thrown down to us by the math gods, like a bone from on high... we can merely choose to chew on it or not, as we wish.

Julie Bergner gave a great talk on "derived Hall algebras" at the Mac Lane memorial conference. I just want to explain the very first result she mentioned, due to Ringel — a surprising trick for constructing certain quantum groups from simply-laced Dynkin diagrams. It's very different from the *usual* method for getting quantum groups from Dynkin diagrams, and it's a miracle that it works.

But, I guess I should start near the beginning!

Way back in 1995, in "Week 62", "Week 63", "Week 64" and "Week 65", I explained how "Dynkin diagrams" — little gizmos like this:



show up all over mathematics. They have a strange way of tying together subjects that superficially seem completely unrelated. In one sense people understand how they work, but in another sense they're very puzzling — their power keeps growing in unexpected ways.

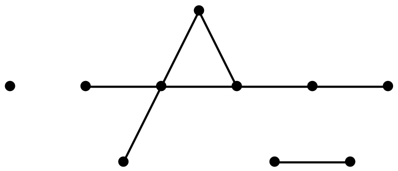
I love mysterious connections, so as soon I understood enough about Dynkin diagrams to appreciate them, I became fascinated by them, and I've been studying them ever since. I explained their relation to geometry in "Week 178", "Week 179", "Week 180", "Week 181" and "Week 182", and their relation to quantum deformation and combinatorics in "Week 186" and "Week 187".

You might think that would be enough — but you'd be wrong, way wrong!

I haven't really talked about the most mysterious aspects of Dynkin diagrams, like their relation to singularity theory and representations of quivers. That's because these aspects were too mysterious! I didn't understand them *at all*. But lately, James Dolan and Todd Trimble and I have been making some progress understanding these aspects.

First, I should remind you how Dynkin diagrams infest so much of mathematics. Let's start with a little puzzle mentioned in "Week 182".

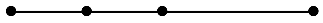
Draw n dots and connect some of them with edges — at most one edge between any pair of dots, please:



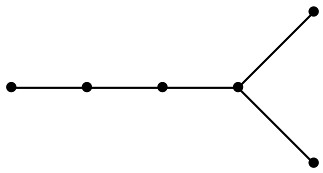
Now, try to find a basis of $\mathbb{R}^n$ consisting of one unit vector per dot, subject to these rules: if two dots are connected by an edge, the angle between their vectors must be 120 degrees, but otherwise their vectors must be at right angles.

This sounds like a silly puzzle that only a mathematician could give a hoot about. It takes a while to see its magnificent depth. But anyway, it turns out you can solve this problem only for certain special diagrams called "simply-laced Dynkin diagrams". The basic kinds are called A_n, D_n, E_6, E_7, and E_8.

The A_n Dynkin diagram is a line of n dots connected by edges like this:

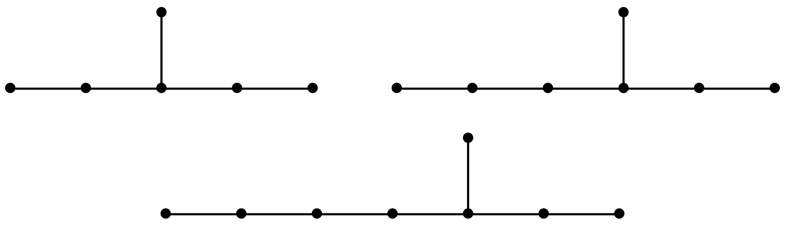


The D_n diagram has n dots arranged like this:



A line of them but then a little fishtail at the end! We should take n to be at least 4, to make the diagram connected and different from A_n.

The E_6, E_7, and E_8 diagrams look like this:



You're also allowed to take disjoint unions of the above diagrams.

So, a weird problem with a weird answer! Its depth is revealed only when we see that many *different* puzzles lead us to the *same* diagrams. For example:

A) the classification of integral lattices in $\mathbb{R}^n$ having a basis of vectors whose length squared equals 2

B) the classification of simply laced semisimple Lie groups

C) the classification of finite subgroups of the 3d rotation group

D) the classification of simple singularities

E) the classification of tame quivers

Let me run through these problems and say a bit about how they're connected:

A) An "integral lattice" in $\mathbb{R}^n$ is a lattice where the dot product of any two vectors in the lattice is an integer. There are zillions of these — but if we demand that they have a basis of vectors whose length squared is 2, we can only get them from simply-laced Dynkin diagrams.

It's not very hard to see that finding a lattice like this is equivalent to the puzzle I mentioned earlier. For example, given a solution of that puzzle, you can just multiply all your vectors by $\sqrt{2}$ and form the lattice of their integer linear combinations.

Here are the lattices you get:

- The diagram A_n gives the n-dimensional lattice of all $(n+1)$-tuples of integers $(x_1, \ldots, x_{n+1})$ with
$$x_1 + \ldots + x_{n+1} = 0.$$
For example, A_2 is a 2-dimensional hexagonal lattice, the sort you use to pack pennies as densely as possible. Similarly, A_3 gives a standard way of packing grapefruit.
- The diagram D_n gives the n-dimensional lattice of all n-tuples of integers $(x_1, \ldots, x_n)$ where
$$x_1 + \ldots + x_n$$
is even. To visualize this, just take an n-dimensional checkerboard, color the cubes alternately red and black, and take the center of each red cube.
- The diagram E_8 gives the 8-dimensional lattice of 8-tuples $(x_1, \ldots, x_8)$ such that the x_i are either all integers or all half-integers — a half-integer being an integer plus $1/2$ — and such that
$$x_1 + \ldots + x_8$$
is even.
- The diagram E_7 gives the 7-dimensional lattice consisting of all vectors in E_8 that are orthogonal to some vector that's closest to the origin (and thus has length $\sqrt{2}$).

- The diagram E_6 gives the 6-dimensional lattice consisting of all vectors in E_7 that are orthogonal to some vector that's closest to the origin (and thus has length $\sqrt{2}$).

For more on these lattices, see "Week 65". They show up in the theory of Lie groups. . . .

B) Lie groups are fundamental throughout math and physics: they're groups of continuous symmetries, like rotations. The nicest of the lot are the semisimple Lie groups. Some familiar examples are the group of rotations in n-dimensional space, which is called $\mathrm{SO}(n)$, and the group of unitary matrices with determinant 1, which is called $\mathrm{SU}(n)$. There are more, but people know what they all are. They're classified by Dynkin diagrams!

Why? The key point is that any semisimple Lie group has a "root lattice". This is an integral lattice spanned by special vectors called "roots". I won't give the details, since I explained this stuff in "Week 63" and "Week 64", but it turns out that root lattices, and thus semisimple Lie groups, are classified by Dynkin diagrams.

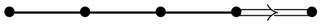
Not all these Dynkin diagrams look like the A, D and E diagrams listed above. But, it turns out that the length squared of any root must be either 1 or 2. If all the roots have length squared equal to 2, we say our semisimple Lie group is "simply laced". In this case, we're back to problem B), which we already solved! So then our Lie group corresponds to a diagram of type A, D, or E — or a disjoint union of such diagrams.

Here's how it goes:

- The diagram A_n gives the compact Lie group $\mathrm{SU}(n+1)$, consisting of $(n+1) \times (n+1)$ unitary matrices with determinant 1. It's the isometry group of complex projective n-space.
- The diagram D_n gives the compact Lie group $\mathrm{SO}(2n)$, consisting of $2n \times 2n$ orthogonal matrices with determinant 1. It's the isometry group of real projective $(2n)$-space.
- The diagram E_6 gives a 78-dimensional compact Lie group that people call E_6. It's the isometry group of the bioctonionic projective plane.
- The diagram E_7 gives a 133-dimensional compact Lie group that people call E_7. It's the isometry group of the quateroctonionic projective plane.
- The diagram E_8 gives a 248-dimensional compact Lie group that people call E_8. It's the isometry group of the octooctonionic projective plane.

In short, two regular series and three exotic weirdos.

You may ask where the rotation groups $\mathrm{SO}(n)$ with n odd went! Well, these correspond to fancier Dynkin diagrams that aren't simply laced, like this:



The funny arrow here indicates that the last two vectors aren't at a 120-degree angle; they're at a 135-degree angle, and the last vector is shorter than the rest: it has length one instead of $\sqrt{2}$.

Why are semisimple Lie groups "better" when they're simply laced? What's the big deal? I don't really understand this, but for one, when all the roots have the same length, they're all alike — a certain symmetry group called the Weyl group acts transitively on them.

Anyway, so far our A, D, E Dynkin diagrams have been classifying things that are clearly related to lattices. But now things get downright spooky. . . .

C) Take a Platonic solid and look at its group of rotational symmetries. You get a finite subgroup of the 3d rotation group $\mathrm{SO}(3)$. But in general, finite subgroups of $\mathrm{SO}(3)$ are classified by ADE Dynkin diagrams!

So, Platonic solids turn out to fit into the game we're playing here!

First I'll say which diagram corresponds to which subgroup of $\mathrm{SO}(3)$. Then I'll explain how the correspondence works:

- The diagram A_n corresponds to the group of obvious rotational symmetries of the regular n-gon. This group is called the "cyclic group" $\mathbb{Z}/n$.
- The diagram D_n corresponds to the group of rotational symmetries of the regular n-gon where you can turn it and also flip it over. By sheer coincidence, this group is called the "dihedral group" D_n. A cosmic stroke of good luck!
- The diagram E_6 corresponds to the group of rotational symmetries of the tetrahedron: the "tetrahedral group". This is also the group of even permutations of 4 elements, the "alternating group" A_4 — not to be confused with the A_n's we were just talking about. A cosmic stroke of bad luck!
- The diagram E_7 corresponds to the group of rotational symmetries of the octahedron or cube: the "octahedral group". This is also the group of all permutations of 4 elements, the "symmetric group" S_4.
- The diagram E_8 corresponds to the group of rotational symmetries of the icosahedron or dodecahedron: the "icosahedral group". This is also the group of even permutations of 5 elements, called A_5. Darn!

So, the exceptional Lie groups E_6, E_7 and E_8 correspond to Platonic solids in a sneaky way.

To understand what's going on here, first we need to switch from $\mathrm{SO}(3)$ to $\mathrm{SU}(2)$. The group $\mathrm{SU}(2)$ is used to describe rotations in quantum mechanics: it's the double cover of the rotation group $\mathrm{SO}(3)$.

It's really finite subgroups of $\mathrm{SU}(2)$ that are classified by ADE Dynkin diagrams! It just so happens that these correspond, in a slightly slippery way, to finite subgroups of $\mathrm{SO}(3)$.

You'll see how if I list the finite subgroups of $\mathrm{SU}(2)$:

- The diagram A_n corresponds to the cyclic subgroup $\mathbb{Z}/n$ of $\mathrm{SU}(2)$. This double covers a cyclic subgroup of $\mathrm{SO}(3)$ when n is even.
- The diagram D_n corresponds to a subgroup of $\mathrm{SU}(2)$ that double covers the dihedral group D_n.

- The diagram E_6 corresponds to a subgroup of $\mathrm{SU}(2)$ that double covers the rotational symmetries of the tetrahedron. This subgroup has 24 elements and it's called the "binary tetrahedral group".
- The diagram E_7 corresponds to a subgroup of $\mathrm{SU}(2)$ that double covers the rotational symmetries of the octahedron. This subgroup has 48 elements and it's called the "binary octahedral group".
- The diagram E_8 corresponds to a subgroup of $\mathrm{SU}(2)$ that double covers the rotational symmetries of the icosahedron. This subgroup has 120 elements and it's called the "binary icosahedral group".

Now, how does the correspondence work? For this, I'm afraid I have to raise the sophistication level a bit — I've been trying to keep things simple, but it's getting tough.

In his book on the icosahedron, Felix Klein noticed it was interesting to let the icosahedral group act on the Riemann sphere, and look for rational functions invariant under this group.

It turned out that every such function depends on a single one: Klein's icosahedral function! The explict formula for it is pretty disgusting, but it's a beautiful thing: you can pick it so that it equals 0 at all the vertices of the icosahedron, 1 at the midpoints of the edges, and infinity at the midpoints of the faces. Even better, if you write the function like this:

$$w = f(z)$$

then Klein showed that knowing how to solve for z as a function of w lets you solve every quintic equation! The reason is that the Galois group of the general quintic is a close relative of the icosahedral group: the former is S_5, the latter is A_5.

Anyway, when I said that "every such function depends on a single one", what I really meant was this. Let $\mathbb{C}(z)$ be the field of rational functions of one variable; then the icosahedral group acts on this, and the invariant functions form a subfield $\mathbb{C}(w)$ where w is Klein's icosahedral function. The Galois group of the little field in the big one is the icosahedral group.

The same kind of thing works for the other finite subgroups of $\mathrm{SO}(3)$, except of course for the connection to the quintic equation.

But, it's actually even better to think about finite subgroups of $\mathrm{SU}(2)$, since $\mathrm{SU}(2)$ acts on $\mathbb{C}^2$, and when we *projectivize* $\mathbb{C}^2$ we get $\mathrm{SO}(3)$ acting on the Riemann sphere. This viewpoint fits more squarely into the worldview of algebraic geometry.

If we take the quotient of $\mathbb{C}^2$ by a finite subgroup G of $\mathrm{SU}(2)$, we don't get a smooth manifold: the quotient has a singularity at 0. But we can "resolve" the singularity, finding a smooth complex manifold with a holomorphic map

$$p\colon M \to \mathbb{C}^2/G$$

that has a holomorphic inverse on a dense open set. There may be lots of ways to do this, but in the present case there's just one "minimal" resolution, meaning a resolution that every other resolution factors through.

Then — and here's the magic part! — the inverse image of 0 in M turns out to be the union of a bunch of Riemann spheres. And if we draw a dot for each sphere, and an edge between these dots whenever their spheres intersect, we get a simply laced Dynkin diagram on the above list!!!

Well, almost. We get this diagram with an extra dot thrown in, connected by some extra edges in a specific way. This is called the "extended" Dynkin diagram. It also shows up naturally from the Lie group viewpoint, when we consider central extensions of loop groups.

That's *one* way the correspondence works. Another way, discovered by McKay, is to draw a dot for each irrep of G. There's always a 2-dimensional representation of G coming from the action of $\mathrm{SU}(2)$ on $\mathbb{C}^2$. Let's just call this irrep $\mathbb{C}^2$. Then, draw an edge from the dot R to the dot S whenever the irrep S shows up in the rep $R \otimes \mathbb{C}^2$. You get the same extended Dynkin diagram as before! The special extra dot in the Dynkin diagram corresponds to the trivial rep of G.

This second way is called the "McKay correspondence". The first way is sometimes called the "geometric McKay correspondence", though I think it was discovered earlier.

Now we're well on the road to the next item. . .

D) Simply-laced Dynkin diagrams also classify the simple critical points of holomorphic functions

$$f\colon \mathbb{C}^3 \to \mathbb{C}$$

A "critical point" is just a place where the gradient of f vanishes. We can try to classify critical points up to a holomorphic change of variables. It's better to classify their "germs", meaning we only look at what's going on *right near* the critical point. But, even this is hopelessly complicated unless we somehow limit our quest.

To do this, we can restrict attention to "stable" critical points, which are those that don't change type under small perturbations. But we can do better: we can classify "simple" critical points, namely those that change into only finitely many other types under small perturbations.

These correspond to simply-laced Dynkin diagrams!

First I'll say which diagram corresponds to which type of critical point. To do this, I'll give a polynomial $f(x, y, z)$ that has a certain type of critical point at $x = y = z = 0$. Then I'll explain how the correspondence works:

- The diagram A_n corresponds to the critical point of $x^{n+1} + y^2 + z^2$.
- The diagram D_n corresponds to the critical point of $x^{n-1} + xy^2 + z^2$.
- The diagram E_6 corresponds to the critical point of $x^4 + y^3 + z^2$.
- The diagram E_7 corresponds to the critical point of $x^3y + y^3 + z^2$.
- The diagram E_8 corresponds to the critical point of $x^5 + y^3 + z^2$.

Here's how the correspondence works. For each of our Dynkin diagrams we have a finite subgroup of $\mathrm{SU}(2)$, thanks to item C). This subgroup acts on the ring of polynomials on $\mathbb{C}^2$, so we can form the subring of invariant polynomials. This turns out to be generated by three polynomials that we will arbitrarily call x, y, and z. But, they satisfy one relation, given by the polynomial above!

Conversely, we can start with the polynomial

$$f\colon \mathbb{C}^3 \to \mathbb{C}$$

The zero set

$$\{f = 0\}$$

has an isolated singularity at the origin. But, we can resolve this singularity, finding a smooth complex manifold N with a holomorphic map

$$q\colon N \to \{f = 0\}$$

that has a holomorphic inverse on a dense open set. There may be lots of ways to do this, but in the present case there's just one "minimal" resolution, meaning one that every other resolution factors through this one.

Then — and here's the magic part! — the inverse image of 0 in N turns out to be the union of a bunch of Riemann spheres. And if we draw a dot for each sphere, and an edge between these dots whenever their spheres intersect, we get back our simply laced Dynkin diagram!!!

This whole section should have given you a feeling of deja vu. It's a lot like section D). If I were smarter, I'd probably see how it's *exactly* the same stuff, repackaged slightly.

The last item on our list seems different. . . .

E) A quiver is just a category freely generated by some set of morphisms. To specify a quiver we just write down some dots and arrows. The dots are the objects of our category; the arrows are the generating morphisms.

A representation of a quiver Q is just a functor

$$F\colon Q \to \mathsf{Vect}$$

So, we get a vector space for each dot and a linear map for each arrow, with no extra restrictions. There's an obvious category of representations $\mathsf{Rep}(Q)$ of a quiver Q.

A guy named Gabriel proved a divine result about these categories $\mathsf{Rep}(Q)$. We say a quiver Q has "finite representation type" if $\mathsf{Rep}(Q)$ has finitely many indecomposable objects — objects that aren't direct sums of others. And, it turns out the quivers of finite representation type are just those coming from simply-laced Dynkin diagrams!!

Actually, for this to make sense, you need to take your Dynkin diagram and turn it into a quiver by putting arrows along the edges. If you have an ADE Dynkin diagram, you get a quiver of finite representation type no matter which way you let the arrows point.

There's clearly a lot of mysterious stuff going on here. In particular, this last item sounds completely unrelated to the rest. But it's not! There are cool relationships between quivers and quantum groups, which tie this item to the rest.

I'll just mention one — the one Julie Bergner started her talk with.

For this, you need to know a bit about abelian categories.

Abelian categories are categories like the category of abelian groups, or more generally the category of modules of any ring, where you can talk about chain complexes, exact sequences and stuff like that. You can see the precise definition here:

4) "Abelian categories", Wikipedia, `http://en.wikipedia.org/wiki/Abelian_category`

and learn more here:

5) Peter Freyd, *Abelian Categories*, Harper and Row, New York, 1964. Also available at `http://www.tac.mta.ca/tac/reprints/articles/3/tr3abs.html`

It's really interesting to study the "Grothendieck group" $K(A)$ of an abelian category A. As a set, this consists of formal linear combinations of isomorphism classes of objects of A, where we impose the relations

$$[a] + [b] = [x]$$

whenever we have a short exact sequence

$$0 \to a \to x \to b \to 0$$

It becomes an abelian group in an obvious way.

For example, if A is the category of representations of some group G, it's an abelian category and $K(A)$ is called the "representation ring" of A — it's a ring because we tensor representations. Or, if A is the category of vector bundles over a space X, it's again abelian, and $K(A)$ is called the "K-theory of X".

The Hall ring $H(A)$ of an abelian category is a vaguely similar idea. As a set, this consists of formal linear combinations of isomorphism classes of objects of A. No extra relations! It's an abelian group with the obvious addition. But the cool part is, with a little luck, we can make it into a *ring* by letting the product $[a][b]$ be the sum of all isomorphism classes of objects $[x]$ weighted by the number of isomorphism classes of short exact sequences

$$0 \to a \to x \to b \to 0$$

This only works if the number is always finite.

So far when speaking of "formal linear combinations" I've been implicitly using integer coefficients, but people seem to prefer complex coefficients in the Hall case, and they get something called the "Hall algebra" instead of the "Hall ring".

The fun starts when we take the Hall algebra of $\mathsf{Rep}(Q)$, where Q is a quiver. We could look at representations in vector spaces over any field, but let's use a finite field — necessarily a field with q elements, where q is a prime power.

Then, Ringel proved an amazing theorem about the Hall algebra $H(\mathsf{Rep}(Q))$ when Q comes from a Dynkin diagram of type A, D, or E:

5) C. M. Ringel, "Hall algebras and quantum groups", *Invent. Math.* **101** (1990), 583–592.

He showed this Hall algebra is a quantum group! More precisely, it's isomorphic to the q-deformed universal enveloping algebra of a maximal nilpotent subalgebra of the Lie algebra associated to the given Dynkin diagram.

That's a mouthful, but it's cool. For example, the Lie algebra associated to A_n is $\mathfrak{sl}(n+1)$, and the maximal nilpotent subalgebra consists of strictly upper triangular matrices. We're q-deforming the universal enveloping algebra of this. One cool thing is that the "q" of q-deformation gets interpreted as a prime power — something we've already seen in "Week 185" and subsequent weeks.

So, it seems that all the ways simply-laced Dynkin diagrams show up in math are related. But, I don't think anyone understands what's really going on! It's like black magic.

And, I've just described *some* of the black magic!

For example, you'll notice I portrayed the Hall algebra $H(A)$ as a kind of evil twin of the more familiar Grothendieck group $K(A)$. They have some funny relations. For example, if you take the minimal resolution of $\mathbb{C}^2/G$ where G is a finite subgroup of $\mathrm{SU}(2)$, you get a variety whose K-theory (as defined above) is isomorphic to the representation ring of G! This was shown here:

6) G. Gonzalez-Springberg and J. L. Verdier, "Construction geometrique de la correspondance de McKay", *Ann. ENS* **16** (1983), 409–449.

For further developments, try this paper, which studies the derived category of coherent sheaves on this minimal resolution of $\mathbb{C}^2/G$:

7) Mikhail Kapranov and Eric Vasserot, "Kleinian singularities, derived categories and Hall algebras", available as `math.AG/9812016`.

Now let me give a bunch of references for further study. For a really quick overview of the whole ADE business, try these:

8) Andrei Gabrielov, "Coxeter-Dynkin diagrams and singularities", in *Selected Papers of E. B. Dynkin with Commentary*, eds. A. A. Yushkevich, G. M. Seitz and A. I. Onishchik, AMS, 1999. Also available at `http://www.math.purdue.edu/~agabriel/dynkin.pdf`

9) John McKay, "A rapid introduction to ADE theory", `http://math.ucr.edu/home/baez/ADE.html`

Here's a more detailed but still highly readable introduction:

10) Joris van Hoboken, *Platonic solids, binary polyhedral groups, Kleinian singularities and Lie algebras of type* A,D,E, Master's Thesis, University of Amsterdam, 2002, available at `http://math.ucr.edu/home/baez/joris_van_hoboken_platonic.pdf`

This classic has recently become available online:

11) M. Hazewinkel, W. Hesselink, D. Siermsa, and F. D. Veldkamp, "The ubiquity of Coxeter-Dynkin diagrams (an introduction to the ADE problem)", *Niew. Arch. Wisk.* **25** (1977), 257–307. Also available at `http://repos.project.cwi.nl:8888/cwi_repository/docs/I/10/10039A.pdf` or `http://math.ucr.edu/home/baez/hazewinkel_et_al.pdf`

Here's a really nice, elementary introduction to Klein's work on the icosahedron and the quintic:

12) Jerry Shurman, *Geometry of the Quintic*, Wiley, New York, 1997. Also available at `http://people.reed.edu/~jerry/Quintic/quintic.html`

I haven't seen this book, but I hear it's good:

13) P. Slodowy, *Simple Singularities and Algebraic Groups*, Lecture Notes in Mathematics **815**, Springer, Berlin, 1980.

Here's a bibliography with links to online references:

14) Miles Reid, Links to papers on McKay correspondence, `http://www.maths.warwick.ac.uk/~miles/McKay/`

Of those references, I especially like this:

15) Miles Reid, "La Correspondence de McKay" (in English), *Seminaire Bourbaki, 52eme annee, November 1999*, no. **867**, to appear in *Asterisque* 2000. Also available as `math.AG/9911165`.

Here you'll also see some material about *generalizations* of the McKay correspondence. For example, if we take a finite subgroup G of $\mathrm{SU}(3)$, we get a quotient $\mathbb{C}^3/G$, which has singularities. If we take a "crepant" resolution of

$$p\colon M \to \mathbb{C}^3/G,$$

which is the right generalization of a minimal resolution, then M is a Calabi-Yau manifold. This gets string theory into the act! Around 1985, Dixon, Harvey, Vafa and Witten used this to guess that the Euler characteristic of M equals the number of irreps of G. A lot of work has been done on this since then, and Reid's article summarizes a bunch.

Apparently a "crepant" resolution is one that induces an isomorphism of canonical bundles; when this fails to happen folks say there's a discrepancy, so a crepant resolution is one with no dis-crepancy. Get it? Since a Calabi-Yau manifold is one whose canonical bundle is trivial, it shouldn't be completely shocking that crepant resolutions yield Calabi-Yaus. This all works in the original 2d McKay correspondence, too — the minimal resolutions we saw there are also crepant.

In fact, string theory also sheds light on the original McKay correspondence. The reason is that the minimal resolution of $\mathbb{C}^2/G$ is a very nice Riemannian 4-manifold (when viewed as a *real* manifold). It's an "asymptotically locally Euclidean" manifold, or ALE manifold for short. Doing string theory on this gives a way of seeing how the extended Dynkin diagrams sneak into the McKay correspondence: they're the Dynkin diagrams for central extensions of loop groups, which show up as gauge groups in string theory! I don't really understand this, but it makes a kind of sense.

I guess this is a famous paper about this stuff:

16) Michael R. Douglas and Gregory Moore, "D-branes, quivers and ALE instantons", available as `hep-th/9603167`.

Addenda: Thanks go to Jeff Barnes for showing how to get ahold of Hazewinkel *et al*'s paper online. I got some nice feedback from Graham Leuschke, David Rusin, and Leslie Coghlan, and I used Leuschke's to fix a mistake.

Graham Leuschke wrote:

> *Hi -*
>
> *A quick correction to your TWF this week. The ADE diagrams are actually the underlying graphs of quivers of* finite *representation type, not tame. You gave the right definition, but the wrong name for it. Tame representation type usually means that there are infinitely many indecomposable representations, but they come in nice one-dimensional families. (The third option is wild representation type, which usually means that classifying the representations would be at least as hard as classifying all modules over the non-commutative polynomial ring k. It's a theorem of Drozd that one of these must hold.)*
>
> *This actually points toward more black magic: the quivers of finite type have ADE diagrams for underlying graphs, while the quivers of tame type have "extended ADE" diagrams underneath them. These extended graphs are the result of adding one (particular) vertex to each of the ADE graphs, and they often arise as the answer to questions that are just slightly weaker than the questions answered by the ADE graphs. For example, the ADE graphs are those for which the Tits form is positive definite, while the extended ADEs are those for which it's positive semi-definite. They correspond to Kac-Moody affine Lie algebras rather than simple Lie algebras, and so on and so on.*
>
> *Harm Derksen and Jerzy Weyman had a nice overview of quiver representations and the theorem of Gabriel in the Notices last year:*
>
> *17) Harm Derksen and Jerzy Weyman, "Quiver representations",* AMS Notices ***52*** *(2005), 200–206. Also available as* `http://www.ams.org/notices/200502/fea-weyman.pdf`
>
> *Idun Reiten had a similar one in the Notices back in 1997, but I can't find it online anywhere:*
>
> *18) Idun Reiten, "Dynkin diagrams and the representation theory of algebras",* AMS Notices ***44*** *(1997), 546–556.*

> *She did a really nice job of explaining the connections with quadratic forms and (sub)additive functions.*
>
> *Cheers,*
> *Graham*

David Rusin wrote:

> *John Baez wrote:*
>
> > *The diagram E_8 corresponds to the critical point of $x^5 + y^3 + z^2$.*
>
> *Milnor has a lovely little book:*
>
> 19) *John Milnor,* Singular points of complex hypersurfaces*, Ann. Math. Studies **61**, Princeton U. Press, Princeton, 1968.*
>
> *which takes the point of view that the RIGHT thing to do at an isolated critical point of a complex-analytic projective variety is to intersect the variety with a small sphere centered at the critical point.*
>
> *Brieskorn did this with the varieties*
>
> $$x^{4k+1} + y^3 + z^2 + w^2 + u^2 = 0$$
>
> *which are 4-dimensional complex varieties with isolated critical points at the origin. So apart from the origin the equation describes an 8-dimensional manifold and the intersection with the sphere in $\mathbb{C}^5 = \mathbb{R}^{10}$ is a 7-dimensional manifold. Brieskorn showed:*
>
> 1. *For every k, these manifolds M_k are homeomorphic to the sphere*
> 2. *For every k, M_k is a smooth manifold.*
> 3. *M_j and M_k are diffeomorphic iff $j = k \mod 28$.*
>
> *Thus in particular, the whole group of diffeomorphism classes of manifolds which are homeomorphic to the sphere has order 28. Milnor had earlier proved that there are 28 diffeomorphism classes of 7-spheres. But here they are very explicit!*
>
> *dave*

Leslie Coghlan wrote:

> *Please add to Week 230 links to copies of these two papers:*
>
> 20) *H. S. M. Coxeter: "The evolution of Coxeter-Dynkin diagrams", in:* Polytopes: Abstract, Convex and Computational*, eds. T. Bisztriczky, P. McMullen, R. Schneider and A. Ivic Weiss, NATO ASI Series C, Vol. **440**, Kluwer, Dordrecht, 1994, pp. 21–42.*
>
> 21) *E. Witt, "Spiegelungsgruppen und Aufzahlung halbeinfacher Liescher Ringe",* Abhandl. Math. Sem. Univ. Hamburg. ***14** (1941), 289–337.*

Yours,

Leslie Coghlan

Here are a couple more online introductions:

22) William Crawley-Boevey, Notes on quiver representations, available at `http://www.amsta.leeds.ac.uk/~pmtwc/quivlecs.pdf`.

23) Alistair Savage, "Finite-dimensional algebras and quivers", available as `arXiv:math/0505082`.

This thesis is an attempt to show an astonishing relation between basic objects from different fields in mathematics. Most peculiarly it turns out that their classification is "the same": the ADE classification. Altogether these objects and the connections between them form a coherent web.

The connections are accomplished by direct constructions leading to bijections between these classes of objects. These constructions however do not always explain or give satisfactory intuition why these classifications [exist], or to say it better, why they should be related in this way. Therefore the deeper reason remains mysterious and when discovered will have to be of great depth. This gives a high motivation to look for new concepts and it shows that simple and since long understood mathematics can still raise very interesting questions, show paths for new research and give a glance at the mystery of mathematics. In my opinion to be aware of a certain truth without having its reason is fundamental to the practise of mathematics.

— *Joris van Hoboken*

Week 231

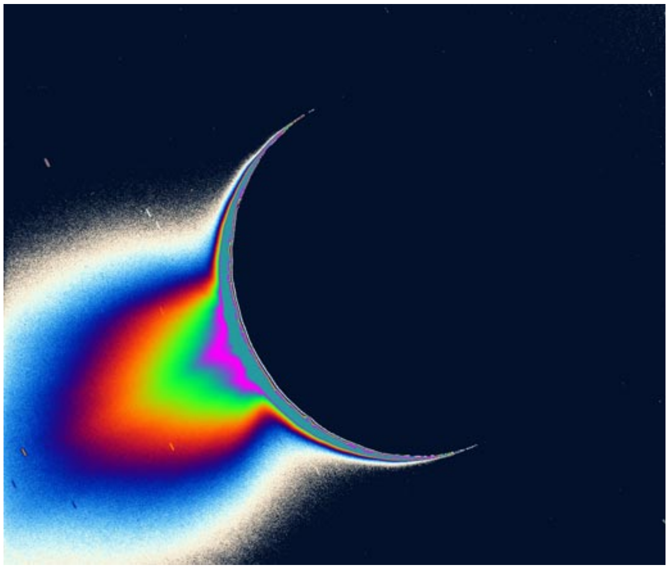
May 9, 2006

Enceladus is a moon of Saturn with a cracked icy surface, twisted and buckled by tidal forces, hinting at mysteries beneath:

1) NASA, "Enceladus the storyteller", `http://www.nasa.gov/mission_pages/cassini/multimedia/pia07800.html`

Recently the NASA space probe Cassini has been getting a good look at Enceladus. In March, Cassini discovered that it has geysers among the cracks near its south pole —

geysers that spray water right out into space!

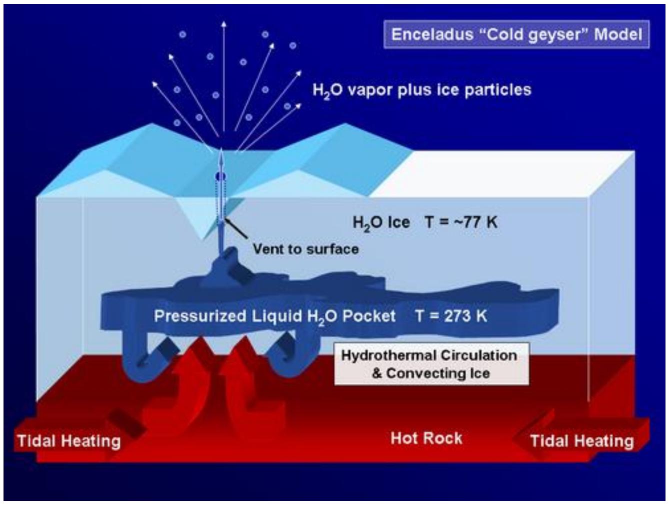


2) "NASA's Cassini discovers potential liquid water on Enceladus", `http://saturn.jpl.nasa.gov/news/press-release-details.cfm?newsID=639`

3) Special issue on Enceladus, *Science* **311** (March 10th 2006).

The water freezes in microscopic crystals, which replenish Saturn's E ring — a diffuse bluish ring that was previously a mystery.

The currently popular theory for the geysers looks like this:


Enceladus "Cold geyser" Model
H_2O vapor plus ice particles
H_2O Ice T = ~77 K
Vent to surface
Pressurized Liquid H_2O Pocket T = 273 K
Hydrothermal Circulation & Convecting Ice
Tidal Heating
Hot Rock
Tidal Heating

4) NASA, 'Enceladus "cold geyser" model', `http://www.nasa.gov/mission_pages/cassini/multimedia/pia07799.html`

Enceladus is now the the only place besides Earth where liquid water has been seen — though people believe Jupiter's moon Europa has oceans under a layer of ice, and maybe Ganymede and Callisto do too.

While we tend to take it for granted, water is a very strange chemical:

5) Martin Chaplin, "Forty-one anomalies of water", `http://www.lsbu.ac.uk/water/anmlies.html`

As you probably know, the specific heat of water is unusually high, which stabilizes the Earth's temperature. And no other simple compound exhibits so many different forms. There at least 18 forms of ice! You can tour them here:

6) Martin Chaplin, "The phase diagram of water", `http://www.lsbu.ac.uk/water/phase.html`

The hexagonal form of ice we find here on earth is called ice Ih. There's also a slightly denser cubic phase, ice Ic, which forms when water vapor is condensed on a cold substrate. Below -130 Celsius, a low-density amorphous solid form called LDA is possible. By compressing ordinary ice Ih to high pressures, you get a different higher-density amorphous form, called HDA. And there's an even denser amorphous form called VHDA.

(It's unusual for a crystal to become amorphous when you compress it or cool it, but ordinary ice is unusually light: it floats on liquid water! That's because the powerful hydrogen bonds of water allow it to maintain a very sparse hexagonal crystal structure — so sparse you could even fit extra water molecules in the gaps. When you crush this, it becomes amorphous.)

There are also crystal forms called ice II through ice XIV, in order of discovery. It would take a few weeks to discuss all these, but luckily Chaplin's website has a separate page on each kind, with nice explanations and pictures of the crystal structures.

Kurt Vonnegut wrote a novel called "Cat's Cradle" starring a substance called ice IX, which was supposedly more stable than liquid water at ordinary temperatures and pressures. When it got loose, it destroyed the world. Luckily the actual ice IX isn't like that, and it couldn't be: the most stable form of water already prevails.

But enough about ice IX. I want to talk about ice X!

This is one of the most extreme forms of ice known. It's only stable at pressures of about 50 gigapascals — in other words, roughly 500,000 atmospheres.

Hmm. Do those quantities mean as little to you as they do to me? A "pascal" is a unit of pressure, or force per area, equal to one newton per square meter. An "atmosphere" is another unit of pressure, basically the average air pressure at sea level here on Earth. This has the annoying value of 101,325 pascals. Personally I have some trouble getting a feel for how much pressure this is, since a newton per square meter isn't much, but 101,325 of them sounds like a lot. So for me, being an American, it's helpful to know that an atmosphere equals 2116 pounds per square foot. If you're a metric sort of person, that's about the weight of 1 kilogram pushing down on each square centimeter. That's a lot of pressure we're under! No wonder we feel stressed sometimes.

(Yes, I know a kilogram is not a unit of weight. I mean the weight corresponding to a mass of a kilogram in the Earth's gravitational field at sea level. Sheesh!)

But I digress. I was saying that ice X only forms at a pressure of about 50 gigapascals. But I've actually read figures ranging from 44 to 80 gigapascals. This raises the question: how do people know these things? Do they actually know, or just guess?

Well, some overgrown kids get paid to study these issues by actually squashing water to enormous pressures using "diamond anvil cells". Not many substances can withstand such huge pressures, but diamonds can: as you know, they're really hard! They're also transparent, so you can see what's going on while you're squashing something. You basically just stick something between two carefully carved diamonds, surrounded by a metal foil gasket, and squash the heck out of it:

7) "Diamond anvil cell", Wikipedia, `http://en.wikipedia.org/wiki/Diamond_Anvil_Cell`

Apparently they can get pressures of up to 360 gigapascals this way, which is the pressure at the center of the Earth.

Another method, which sounds even more fun, is to use a "light gas gun". Here you explode a few kilograms of gunpowder to shoot a piston down a tube. As it shoots forwards, the piston pushes some gas down the tube. The tube narrows to a tiny tip at the end, so the gas is going really fast by the time it shoots out. It shoots out into a much narrower tube, where it pushes a projectile. You can then fire the projectile into something, to generate very high pressures for a very short time.

8) "Light gas gun", Wikipedia, `http://en.wikipedia.org/wiki/Light_Gas_Gun`

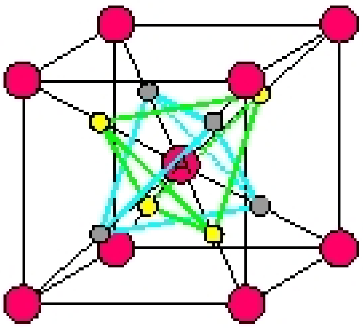
It's not called a "light" gas gun because it's wimpy — in fact they're huge, and everyone evacuates the lab when they run the one at NASA! It's called that because the speed of the projectile is limited only by the speed of sound in the gas, which is higher for a light gas like helium — or even better, hydrogen. Even better, that is, you don't mind exploding gunpowder near highly flammable hydrogen! But, as you can imagine, people who do this stuff are precisely the sort who don't mind. You may enjoy reading how folks at Lawrence Livermore National Laboratory used a light gas gun to compress hydrogen to pressures of up to 200 gigapascals, enough to convert it into a metal:

9) Robert C. Cauble, "Putting more pressure on hydrogen", `http://www.llnl.gov/str/Cauble.html`

This supports the theory that the hydrogen at Jupiter's core is in metallic form, which would explain its enormous magnetic field. They know their hydrogen became a metal because they fired a laser at it and saw it was shiny! In fact, they fired three lasers at it simultaneously, just for kicks.

(By the way, this article erroneously says a "megabar" is 100 pascals. It's a million atmospheres, or 100 gigapascals.)

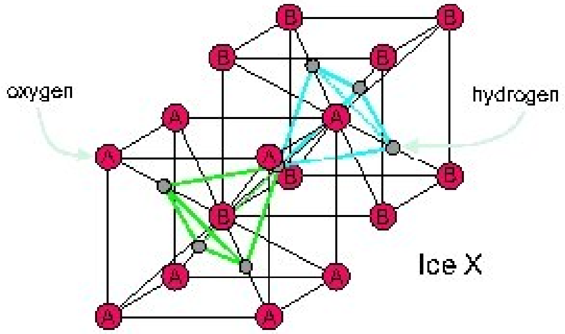
But I'm digressing again. I was saying ice X forms at a pressure of around 50 gigapascals. It's pretty far-out stuff. It's a cubic crystal with density 2.5 times that of ordinary liquid water. It's so compressed that separate water molecules no longer exist! Instead, the oxygen atoms form a body-centered cubic. This means they lie at the corners of a lattice of cubes, but with one at the center of each cube too, like the red dots in this picture by Cavazzoni:



10) Carlo Cavazzoni, *Large scale first-principles simulations of water and ammonia at high pressure and temperature*, Ph.D. thesis, Scuola Internazionale Superiore di Studi Avanzati, October 1998. Figure 4.10: symmetric and super-ionic ice X structures, p. 57. Available at `http://sirio.cineca.it/˜acv0/thesis.html`

Hydrogen ions — in other words, protons — sit at the midpoints of half the edges connecting cube corners to cube centers. There are two ways they can do this, illustrated by the yellow and gray dots shown above. They can form a right-side-up tetrahedron, or an upside-down tetrahedron.

A body-centered cubic can also be visualized as two interpenetrating cubic lattices, labelled A and B here:



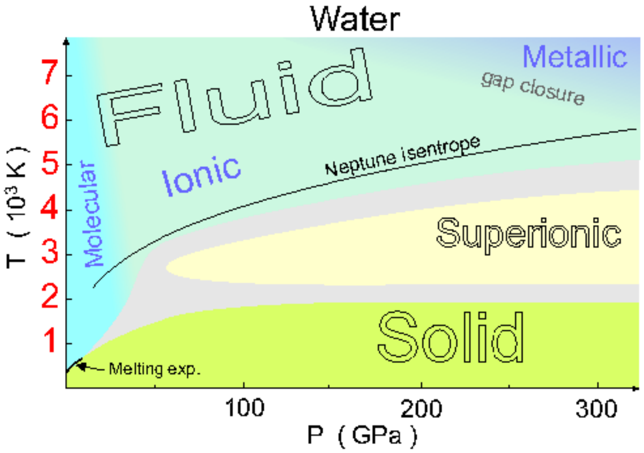
Each oxygen has 4 hydrogens next to it. If you compress water a bit less than enough to make ice X, you get ice VII. This is almost the same, but two of those hydrogens are closer to the oxygen than the other two, so there are still separate water molecules! It's completely random which two hydrogens are closer than the other two. But if you cool down ice VII, you get ice VIII, where it's *not* random.

So, Nature explores all the options.

Recently people have gotten interested in ice at even higher pressures — and also higher temperatures, to understand the interiors of planets like Neptune and Uranus. Here pressures range from 20 to 800 gigapascals, and temperatures from 2000 to 8000 kelvin. In "Week 160" I mentioned that on Neptune it may rain diamonds, formed by methane in the atmosphere. But what happens to the water, and the ammonia? If they became good electrical conductors, that might explain the magnetic fields of these planets.

People have done computer simulations to study this:

12) C. Cavazzoni, G. L. Chiarotti, S. Scandolo, E. Tosatti, M. Bernasconi and M. Parrinello, "Superionic and metallic states of water and ammonia at giant planet conditions", *Science* **283** (January 1999), 44–46. Also available at `http://www.sciencemag.org/cgi/content/full/283/5398/44`



It seems that when you heat up ice X, it goes into a "superionic" state where the little tetrahedra of hydrogen ions in each cube are constantly randomizing themselves, instead of remaining fixed. It's a curious hybrid of a solid and a liquid, since the hydrogens are moving around, while the oxygens stay in their body-centered cubic crystal.

But if you heat it even more, the oxygen melts too! As you can see from the phase diagram above, it then becomes an ionic fluid.

As you heat it even more, you enter the region labelled "gap closure", where the water starts to act like a metallic plasma. Then it's a really good conductor of electricity.

The curve labelled "Neptune isentrope" describes the pressures and temperatures you'd experience if you unwisely jumped into Neptune!

As you fell in, it would keep getting hotter and the pressure would keep rising until you entered this chart, at a temperature of about 2000 kelvin. At this point you'd see molecular fluid water — I say this because at temperatures above 650 kelvin (the critical point for water), there's no sharp difference between liquid and gas. Then the fluid would become ionic. . . and then you'd start drifting towards gap closure and the metallic plasma phase. Down deep, metallic plasmas of water and ammonia might explain the magnetic field of this planet.

Recently people have done some experiments with water at extremely high pressures, checking what theorists like Cavazzoni and company predict. For example, this paper says that using "extremely large lasers", people have studied water at pressures near a terapascal — 1000 gigapascals:

13) P. M. Celliers et al, "Electronic conduction in shock-compressed water", *Plasmas* **11** (2004), L41–L48.

They also mention that "a single datum at 1.4 terapascals from an underground nuclear experiment has never been repeated." Some people just don't know when to stop in the quest for higher pressures.

While I'm at it, I should mention a few more interesting articles on weird forms of ice. There's a lot of research on this subject! Here's a quick overview:

14) Nancy McGuire, "The many phases of water", American Chemical Society, `https://web.archive.org/web/20051201104533/http://www.chemistry.org/portal/a/c/s/1/feature_pro.html?id=c373e9fbed0a01c78f6a4fd8fe800100`

Here's a webpage with some nice pictures and an interesting story:

15) J. L. Finney, "The phase diagram of water and a new metastable form of ice", `http://www.cmmp.ucl.ac.uk/people/finney/soi.html`

And finally, there's a paper that talks about how ordinary ice Ih but also silica and ice XI become amorphous when you squeeze them enough:

16) Koichiro Umemoto, Renata M. Wentzcovitch, Stefano Baroni and Stefano de Cironcoli, "Anomalous pressure-induced transition(s) in ice XI", *Physical Review Letters* **92** (2004), 105502-1. Also available at `http://www.cems.umn.edu/research/wentzcovitch/papers/Phys._Rev._Lett._92_105502_(2004).pdf`

There's some interesting math in here, because they do computer simulations of the transition from a crystal to an amorphous substance, which is interesting to study using Fourier analysis. The idea is that certain vibrational modes of the crystal "go soft", so they get easily excited. When a bunch of modes go soft that have wavelengths not equal to the crystal lattice spacing, the crystal structure becomes unstable, and there can be a transition to an amorphous state.

There's also interesting math lurking in Cavazzoni et al's models of ice X! If you think particle physics is hard, just wait until you try understanding something complicated, like water.

I've been sort of obsessed with ice lately. If you like it too, I recommend this book for general information:

16) Mariana Gosnell, *Ice: The Nature, the History, and the Uses of an Astonishing Substance*, Alfred A. Knopf, New York, 2005.

but I bought this one, because it tells an interesting history of the science of climate change as seen from icy peaks:

17) Mark Bowen, *Thin Ice: Unlocking the Secrets of Climate in the World's Highest Mountains*, Henry Holt & Co., 2005.

Now for some math. Last week I said a bit about quivers, the McKay correspondence, and string theory. I want to dig deeper into the relation between these subjects, because Urs Schreiber has some interesting ideas about them, which he's mentioned here:

18) Urs Schreiber, "A note on RCFT and quiver reps", `http://golem.ph.utexas.edu/string/archives/000794.html`

But, I'm not feeling sufficiently energetic to explain these ideas right now, especially since he already has! For some more clues, try this:

19) Paul Aspinwall, "D-branes on Calabi-Yau manifolds", section 7.3.1, The McKay correspondence, p. 101 and following. Available as `hep-th/0403166`

For more on the the representation theory of quivers, see the references in the "Addenda" to "Week 230", and also this excellent book:

20) David J. Benson, *Representations and Cohomology I*, Cambridge U. Press, Cambridge 1991.

You'll see how the non-simply-laced Dynkin diagrams get into the act! A more thorough treatment, fascinating but somewhat quirky, can be found here:

21) P. Gabriel and A. V. Roiter, *Representations of Finite-Dimensional Algebras*, Enc. of Math. Sci. **73**, Algebra VIII, Springer, Berlin 1992.

If you like category theory you may enjoy this book, because it's all about representations of categories, i.e. functors

$$F\colon \mathcal{C} \to \mathsf{Vect}$$

where $\mathcal{C}$ is a category. It's full of nontrivial theorems about these, starting with Gabriel's classification of quivers into those of finite representation type (see "Week 230"), the tame quivers (which have an infinite but still manageable set of indecomposable representations), and the wild ones. But, you may be puzzled when you read about "svelte" categories, or functors that "preserve heteromorphisms".

I might as well say what those are. A category is "svelte" if its isomorphism classes of objects form a mere set instead of a proper class, like the category of finite-dimensional vector spaces. Most people would say such a category is "essentially small".

And, a functor "preserves heteromorphisms" if it maps heteromorphisms to heteromorphisms. Well, duh! But what's a "heteromorphism"? It's their term for a morphism that's not an isomorphism. Most people would say such a functor "reflects isomorphisms".

You may also be interested in what a "locular" category is, or a "spectroid"... but I won't tell you! Read the book.

Speaking of category theory, this is my last week in Chicago, which is really sad, because Steve Lack is just starting to give us a crash course on "Australian category theory". Australia, you see, is the center of macho category theory, where they're heavy on the calculus of mates, doctrinal adjunctions are a dime a dozen, and everything should be $\mathcal{V}$-enriched if not $\mathcal{W}$-enriched. But Chicago is starting to get macho too: tomorrow Nick Gurski defends his Ph.D. thesis on "Algebraic Tricategories"! So, the Chicago gang wants to learn some tricks from the Australians. But next Monday I'm off to the Perimeter Institute, to indulge the physics side of my personality....

Addenda: I thank Colin Rust for correcting a serious typo. Uncle Al points out that a newton is, quite appropriately, about the weight of an average apple. Aaron Bergman had this to say:

> *John Baez wrote:*
>
> > *Now for some math. Last week I said a bit about quivers, the McKay correspondence, and string theory. I want to dig deeper into the relation between these subjects,*
>
> *You want*
>
> *22) Tom Bridgeland, Alaistair King and Miles Reid, "Mukai implies McKay: the McKay correspondence as an equivalence of derived categories", available as math.AG/9908027.*
>
> *Much coolness there.*
>
> > *For more on the representation theory of quivers, see the references in the "Addenda" to "Week 230", and also this excellent book:*
> >
> > *20) David J. Benson,* Representations and Cohomology I, *Cambridge U. Press, Cambridge 1991.*
>
> *There's a whole lot of cool things to say about non-Dynkin quivers. If you put a D-brane at an ADE singularity, you get the affine Dynkin quivers, but there are plenty of other singularities out there. Another nice set is the canonical bundle over a del Pezzo with the zero section collapsed. Quivers can be associated to this via a work of Bridgeland:*
>
> *23) Tom Bridgeland, "T-structures on some local Calabi-Yau varieties", available as math.AG/0502050.*
>
> *with many of the same results in some physics papers by Herzog, Aspinwall and Wijnholt.*
>
> *My modest contribution to this story is where myself and Nick Proudfoot show that the moduli space of representations of these quivers (for a subclass of examples) has as a component the original cone:*
>
> *24) Aaron Bergman, "Undoing orbifold quivers", available as hep-th/0502105.*
>
> *Aaron Bergman, "Moduli spaces for Bondal quivers", available as math.AG/0512166.*
>
> *Aaron*

That the glass would melt in heat,
That the water would freeze in cold,
Shows that this object is merely a state,
One of many, between two poles.

— *Wallace Stevens*

Week 232

May 18, 2006

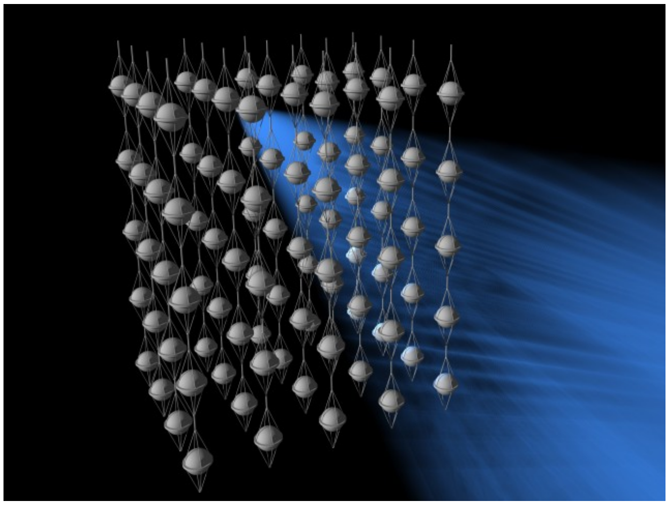
I'm at the Perimeter Institute now. It's great to see how it's developed since I first saw their new building back in 2004 (see "Week 208" for the story).

There's now a busy schedule of seminars and weekly colloquia, with string theorists and loop quantum gravity people coexisting happily. Their program of Superstring Quartets features some really hot bands, like the Julliard and Emerson — unfortunately not playing while I'm here. The Black Hole Bistro serves elegant lunches and dinners, there are at least two espresso machines on each floor, and my friend Eugenia Cheng will be happy to hear that they still have a piano available (after 6 pm).

But don't get the impression that it's overly sophisticated: there are also a couple of guys constantly playing foosball in the Feynman Lounge.

Since I'm here, I should talk about quantum gravity — so I will. But first, let's have the astronomy picture of the week.

This week it comes, not from outer space, but beneath the surface of the South Pole:

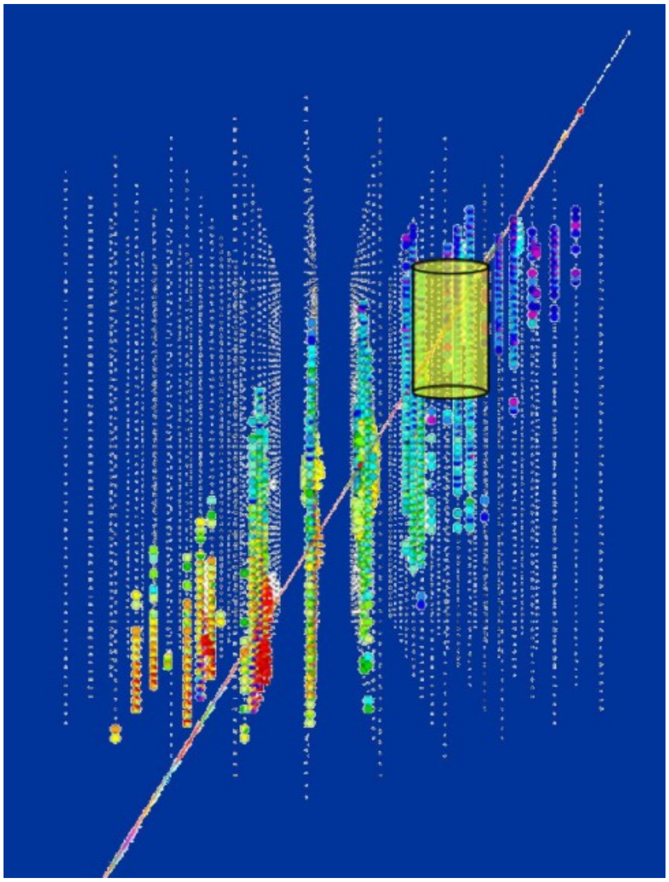


1) Steve Yunck / NSF, "Cerenkov light passing through the IceCube neutrino detector", `http://icecube.wisc.edu/gallery/detector_concepts/ceren_hires`

This is an artist's impression of a huge neutrino observatory called "IceCube". (Maybe they left out the space here so the rap star of that name doesn't sue them for trademark infringement, or go down there and shoot them.)

IceCube is being built in the beautifully clear 18,000-year old ice deep beneath the Amundsen-Scott South Pole Station. When a high-energy neutrino hits a water molecule, sometimes the collision produces a muon zipping faster than the speed of light in ice. This in turn produces something like a sonic boom, but with light instead of sound. It's called "Cerenkov radiation", and it's the blue light in the picture. This will be detected by an array of 5000 photomultiplier tubes — those gadgets hanging on electrical cables.

One thing the artist's impression doesn't show is that IceCube is amazingly large. The whole array is a cubic kilometer in size! It will encompass the already existing AMANDA detector, itself 10,000 meters tall, shown as a yellow cylinder here with a neutrino zipping through:



2) Darwin Rianto / NSF, "Comparison of AMANDA and IceCube", `http://icecube.wisc.edu/gallery/detector_concepts/icecubeencomp_300`

Even the very top of IceCube is 1.4 kilometers beneath the snowy Antarctic surface, to minimize the effect of stray cosmic rays. The station on top looks like this — not very cozy, I'd say:

3) Robert G. Stokstad / NSF, "South Pole Station", `http://icecube.wisc.edu/gallery/antarctica/PC140287_300`

I heard about IceCube from Adrian Burd, one of the old-timers who used to post a lot on `sci.physics`, a former cosmologist turned oceanographer who recently visited Antarctica as part of an NSF-run field course. He ran into some people working on IceCube. It sounds like an interesting community down there! You can read about it in their newspaper, the Antarctic Sun. For example:

4) "Ice Cube turns up the heat", *The Antarctic Sun*, January 29, 2006, `http://antarcticsun.usap.gov/2005-2006/contentHandler.cfm?id=959`

For more on IceCube and Amanda, these are fun to read:

5) Francis Halzen, "Ice fishing for neutrinos", `http://icecube.berkeley.edu/amanda/ice-fishing.html`

6) Katie Yurkiewicz, "Extreme neutrinos", *Symmetry*, volume **1** issue 1, November 2004, `http://symmetrymagazine.org/cms/?pid=1000014`

For some of AMANDA's results, including a map of the sky as seen in neutrinos, try this:

7) M. Ackermann et al, "Search for extraterrestrial point sources of high energy neutrinos with AMANDA-II using data collected in 2000–2002", available as `astro-ph/0412347`.

For much more, try these:

8) AMANDA II Project, `http://amanda.uci.edu/`

9) Welcome to IceCube, `http://icecube.wisc.edu/`

And now, on to gravity.

You may have heard of the gravitational 3-body problem. Well, Richard Montgomery (famous from "Week 181") recently pointed out this movie of the 60-body problem:

10) Davide L. Ferrario, "Periodic orbits for the 60-body problem", `http://www.matapp.unimib.it/~ferrario/mov/index.html`

60 equal masses do a complicated dance while always preserving icosahedral symmetry! First 12 groups of 5 swing past each other, then 20 groups of 3. If you want to know how he found these solutions, read this:

11) Davide L. Ferrario and S. Terracini, "On the existence of collisionless equivariant minimizers for the classical n-body problem". *Invent. Math.* **155** (2004), 305–362.

It's quite math-intensive — though just what you'd expect if you know this sort of thing: they use the G-equivariant topology of loop spaces, where G is the symmetry group in question (here the icosahedral group), to prove the existence of action-minimizing loops with given symmetry properties.

Next, I'd like to say a little about point particles in 3d quantum gravity, and some recent work with Alissa Crans, Derek Wise and Alejandro Perez on string-like defects in 4d topological gravity:

12) John Baez, Derek Wise and Alissa Crans, "Exotic statistics for strings in 4d BF theory", available as gr-qc/0603085.

13) John Baez and Alejandro Perez, "Quantization of strings and branes coupled to BF theory", available as gr-qc/0605087.

(Jeffrey Morton is also involved in this project, a bit more on the n-category side of things, but that aspect is top secret for now.)

In "Week 222" I listed a bunch of cool papers on 3d quantum gravity, but I didn't really explain them. What we're trying to do now is generalize this work to higher dimensions. But first, let me start by explaining the wonders of 3d quantum gravity.

The main wonder is that we actually understand it! The classical version of general relativity is exactly solvable when spacetime has dimension 3, and so is the quantum version. Most of the wonders I want to discuss are already visible in the classical theory, where they are easier to understand, so I'll focus on the classical case.

A nice formulation of general relativity in 3 dimensions uses a "Lorentz connection" A and a "triad field" e. This is a gauge theory where the gauge group is $\mathrm{SO}(2,1)$, the Lorentz group for 3d spacetime. If we're feeling lowbrow we can think of both A and e as $\mathfrak{so}(2,1)$-valued 1-forms on the 3-manifold M that describes spacetime. The action for this theory is:

$$\int_M \mathrm{tr}(e \wedge F)$$

where F is the curvature of A. If you work out the equations of motion one of them says that $F = 0$, so our connection A is flat. The other, $d_A e = 0$, says A is basically just the Levi-Civita connection.

This is exactly what we want, because in the absence of matter, general relativity in 3 dimensions says spacetime is *flat*.

A fellow named Phillipp de Sousa Gerbert came up with an interesting way to couple point particles to this formulation of quantum gravity:

14) Phillipp de Sousa Gerbert, "On spin and (quantum) gravity in 2+1 dimensions", *Nuclear Physics* **B346** (1990), 440–472.

He actually did it for particles with spin, but I'll just do the spin-zero case.

The idea is to fix a 1-dimensional submanifold W in our 3-manifold M and think of it as the worldlines of some particles. Put $\mathfrak{so}(2,1)$-valued functions p and q on these worldlines — think of these as giving the particles' momentum and position as a function of time.

Huh? Well, normally we think of position and momentum as vectors. In special relativity, "position" means "position in spacetime", and "momentum" means "energy-momentum". We can think of both of these as vectors in Minkowksi spacetime. But in 3 dimensions, Minkowski spacetime is naturally identified with the Lorentz Lie algebra $\mathfrak{so}(2,1)$. So, it makes sense to think of q and p as elements of $\mathfrak{so}(2,1)$ which vary from point to point along the particle's worldline.

To couple our point particles to gravity, we then add a term to the action like this:

$$S = \int_M \mathrm{tr}(e \wedge F) - \int_W \mathrm{tr}((e + d_A q) \wedge p)$$

Now if you vary the e field you get a field equation saying that

$$F = p\delta_W$$

Here δ_W is like the Dirac delta function of the worldline W; it's a distributional 2-form defined by requiring that

$$\int_W X = \int_M (X \wedge \delta_W)$$

for any smooth 1-form X on W. This sort of "distributional differential form" is also called a "current", and you can read about them in the classic tome by Choquet-Bruhat et al. But the main point is that the field equation

$$F = p\delta_W$$

says our connection on spacetime is flat except along the worldlines of our particles, where the curvature is a kind of "δ function". This is nice, because that's what we expect in 3d gravity: if you have a particle, spacetime will be flat everywhere except right at the particle, where it will have a singularity like the tip of a cone.

A cone, you see, is intrinsically flat except at its tip: that's why you can curl paper into a cone without crinkling it!

So, our spacetime is flat except along the particles' worldlines, and there it's like a cone. The "deficit angle" of this cone — the angle of the slice you'd need to cut out to curl some paper into this cone - is specified by the particle's momentum p.

Since delta functions are a bit scary, it's actually better to work with an "integrated" form of the equation

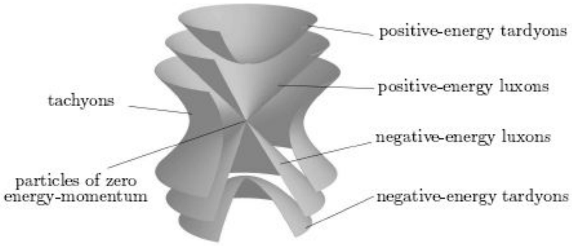
$$F = -p\delta_W$$

The integrated form says that if we parallel transport a little tangent vector around a little loop circling our particle's worldline, it gets rotated and/or Lorentz transformed by the element

$$\exp(p)$$

in $\mathrm{SO}(2,1)$. This will be a rotation if the particle's momentum p is timelike, as it is for normal particles. Again, that's just as it should be: if you parallel transport a little arrow around a massive particle in 3d gravity, it gets rotated!

If p is timelike, our particle is a tachyon and $\exp(p)$ is a Lorentz boost. And so on. . . we get the usual classification of particles corresponding to various choices of p:



There are other equations of motion, obtained by varying other fields, but all I want to note is the one you get by varying q:

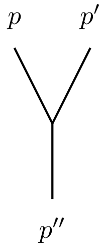
$$d_A p = 0$$

This says that the momentum p is covariantly constant along the particles' worldlines. So, momentum is conserved!

The really cool part is the relation between the Lie algebra element p and the group element $\exp(p)$. Originally we thought of p as momentum — but there's a sense in which $\exp(p)$ is the momentum that really counts!

First, $\exp(p)$ is what we actually detect by parallel transporting a little arrow around our particle.

Second, suppose we let two particles collide and form a new one:



Now our worldlines don't form a submanifold anymore, but if we keep our wits about us, we can see that everything still makes sense, and we get momentum conservation in this form:

$$\exp(p'') = \exp(p)\exp(p')$$

since little loops going around the two incoming particles can fuse to form a loop going around the outgoing particle. Note that we're getting conservation of the *group-valued* momentum, not the Lie-algebra-valued momentum — we don't have

$$p'' = p + p'$$

So, conservation of energy-momentum is getting modified by gravitational effects! This goes by the name of "doubly special relativity":

15) Laurent Freidel, Jerzy Kowalski-Glikman and Lee Smolin, "2+1 gravity and doubly special relativity", *Phys. Rev.* **D69** (2004) 044001. Also available as hep-th/0307085.

This effect is a bit less shocking if we put the units back in. I've secretly been setting $4\pi G = 1$, where G is Newton's gravitational constant. If we put that constant back in — let's call it k instead of $4\pi G$ — we get

$$\exp(kp'') = \exp(kp)\exp(kp')$$

or if you expand things out:

$$p'' = p + p' + \frac{k}{2}[p, p'] + \text{terms of order } k^2 \text{ and higher}$$

So, as long as the momenta are small compared to the Planck mass, the usual law of conservation of momentum

$$p'' = p + p'$$

almost holds! But, for large momenta this law breaks down — we must think of momentum as group-valued if we want it to be conserved!

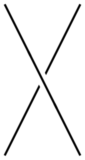
I think this is incredibly cool: as we turn on gravity, the usual "flat" momentum space curls up into a group, and we need to *multiply* momenta in this group, instead of *add* them in the Lie algebra. We can think of this group has having a "radius" of $1/k$, so it's really big and almost flat when the strength of gravity is small. In this limit, multiplication in the group reduces to addition in the Lie algebra.

I should point out that this effect is purely classical! It's still there when we quantize the theory, but it only depends on the gravitational constant, not Planck's constant. Indeed, in 3d quantum gravity, we can build a unit of mass using just G and c: we don't need $\hbar$. This unit is the mass that curls space into an infinitely skinny cone! It would be a bit misleading to call it "Planck mass", but it's the maximum possible mass. Any mass bigger than this acts like a *negative* mass. That's because the corresponding group-valued momenta "wrap around" in the group $\mathrm{SO}(2,1)$.

We also get another cool effect — exotic statistics. In the absence of gravitational or quantum effects, when you switch two particles, you just switch their momenta:

$$(p, p') \mapsto (p', p)$$

But in 3d gravity, you can think of this process of switching particles as a braid:



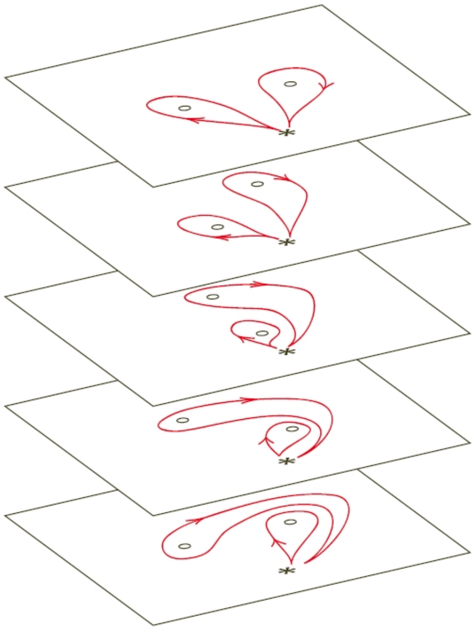
and if you work out what happens to their group-valued momenta, say

$$g = \exp(kp)$$
$$g' = \exp(kp')$$

it turns out that one momentum gets conjugated by the other:

$$(g, g') \mapsto (gg'g^{-1}, g)$$

To see this, remember that we get these group elements by doing parallel transport around loops that circle our particles. When we move our particles, the loops get dragged along, like this:



Note that the left-hand red loop moves until it looks just like the right one did initially, but the right-hand one gets wrapped around the left one. If you ponder this carefully, and you know some math, you can see it yields this:

$$(g, g') \mapsto (gg'g^{-1}, g)$$

So, the process of braiding two particles around each other has a nontrivial effect on their momenta. In particular, if you braid two particles around other twice they don't wind up in their original state!

Thus, our particles are neither bosons nor fermions, but "nonabelian anyons" — the process of switching them is governed not by the permutation group, but by the braid group. But again, if you expand things out in powers of k you'll see this effect is only noticeable for large momenta:

$$(p, p') \mapsto (p' + k[p, p'] + \text{higher order terms}, p)$$

Summarizing, we see quantum gravity is lots of fun in 3 dimensions: it's easy to introduce point particles, and they have group-valued momentum, which gives rise to doubly special relativity and braid group statistics.

Now, what happens when we go from 3 dimensions to 4 dimensions?

Well, we can write down the same sort of theory:

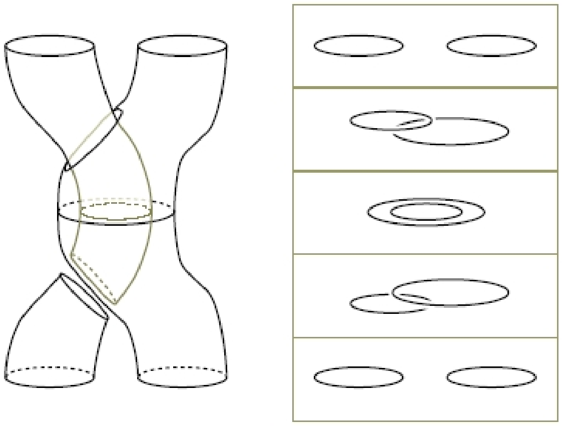
$$S = \int_M \mathrm{tr}(B \wedge F) - \int_W \mathrm{tr}((B + d_A q) \wedge p)$$

The only visible difference is that what I'd been calling "e" is now called "B", so you can see why folks call this "BF theory".

But more importantly, now M is an 4-dimensional spacetime and W is an 2-dimensional "worldsheet". A is again a Lorentz connection, which we can think of as an $\mathfrak{so}(3,1)$-valued 1-form. B is an $\mathfrak{so}(3,1)$-valued 2-form. p is an $\mathfrak{so}(3,1)$-valued function on the worldsheet W. q is an $\mathfrak{so}(3,1)$-valued 1-form on W.

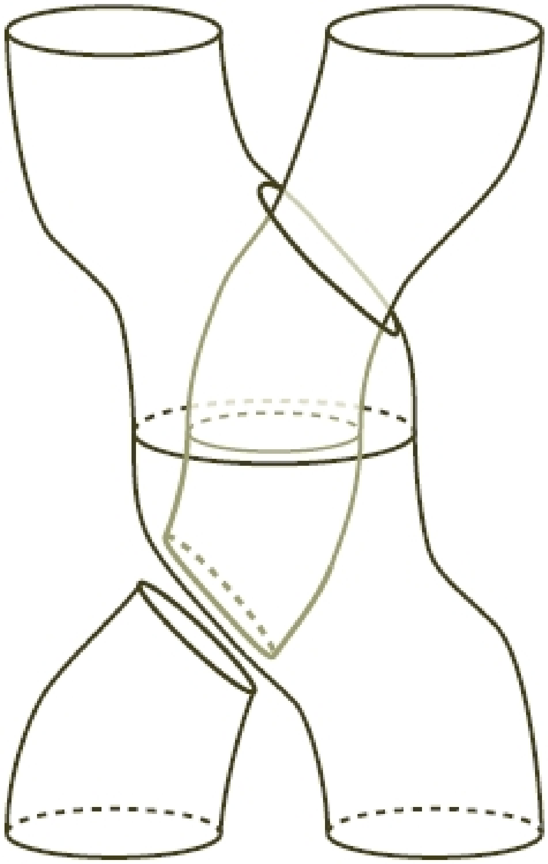
So, only a few numbers have changed... so everything works very similarly! The big difference is that instead of spacetime having a conical singularity along the worldline of a *particle*, now it's singular along the worldsheet of a *string*. When I call it a "string", I'm not trying to say it behaves like the ones they think about in string theory — at least superficially, it's a different sort of theory, a purely topological theory. But, we've got these closed loops that move around, split and join, and trace out surfaces in spacetime.

They can also braid around each other in topologically nontrivial ways, as shown in this "movie":



(By the way, all the math pictures this week were drawn by Derek for our paper.)

So, we get exotic statistics as before, but now they are governed not by the braid group but by the ”loop braid group’’, which keeps track of all the ways we can move a bunch of circles around in 3d space. Let’s take our spacetime M to be $\mathbb{R}^4$, to keep things simple. Then our circles can move around in $\mathbb{R}^3\ldots$ and there are two basic ways we can switch two of them: move them around each other, or pass one *through* the other, like this:



If we just move them around each other, they might as well have been point particles: we get a copy of the permutation group, and all we see are ordinary statistics. But when we consider all the ways of passing them through each other, we get a copy of the braid group!

When we allow ourselves both motions, we get a group called the "loop braid group'' or"braid permutation group'' — and one thing Alissa Derek and I did was to get a presentation of this group. This is an example of a "motion group'': just as the motion group of point particles in the plane is the braid group, and motion group of point particles in $\mathbb{R}^3$ is the permutation group, the motion group of strings in $\mathbb{R}^3$ is the loop braid group.

As before, our strings have group-valued momenta: we can get an element of the Lorentz group $\mathrm{SO}(3,1)$ by parallel transporting a little tangent vector around a string. And, we can see how different ways of switching our strings affect the momenta. When we move two strings around each other, their momenta switch in the usual way:

$$(g, g') \mapsto (g', g)$$

but when we move one through the other, one momentum gets conjugated by the other:

$$(g, g') \mapsto (gg'g^{-1}, g)$$

So, we have exotic statistics, but you can only notice them if you can pass one string through another!

In the paper with Alejandro, we go further and begin the project of quantizing these funny strings, using ideas from loop quantum gravity. Loop quantum gravity has its share of problems, but it works perfectly well for 3d quantum gravity, and matches the spin foam picture of this theory. People have sort of believed this for a long time, but Alejandro demonstrated this quite carefully in a recent paper with Karim Noui:

15) Karim Noui and Alejandro Perez, "Dynamics of loop quantum gravity and spin foam models in three dimensions", to appear in the proceedings of the *Third International Symposium on Quantum Theory and Symmetries (QTS3)*, available as gr-qc/0402112.

The reason everything works so nicely is that the equations of motion say the connection is flat. Since the same is true in BF theory in higher dimensions, we expect that the loop quantization and spin foam quantization of the theory I'm talking about now should also work well.

We find that we get a Hilbert space with a basis of "string spin networks", meaning spin networks that can have loose ends on the stringy defects.

So, there's some weird blend of loop quantum gravity and strings going on here — but I don't really understand the relation to ordinary string theory, if any. It's possible that I can get a topological string theory (some sort of well-defined mathematical gadget) which describes these stringy defects, and that would be quite interesting.

But, I spoke about this today at the Perimeter Institute, and Malcolm Perry said that instead of "strings" I should call these guys $(n-2)$-branes, because the connection has conical singularities on them, "which is what one would expect for any respectable $(n-2)$-brane".

I will talk to him more about this and try to pick his, umm, branes. In fact I took my very first GR course from him, back when he was a postdoc at Princeton and I was a measly undergraduate. I was too scared to ask him many questions then. I'm a bit less scared now, but I've still got a lot to learn. Tomorrow he's giving a talk about this:

17) David S. Berman, Malcolm J. Perry, "M-theory and the string genus expansion", *Phys. Lett.* **B635** (2006) 131–135. Also available as hep-th/0601141.

Addenda: Here's an email from Greg Egan, and my reply:

> *John Baez wrote:*
>
> > *The really cool part is the relation between the Lie algebra element p and the group element $\exp(p)$. Originally we thought of p as momentum — but there's a sense in which $\exp(p)$ is the momentum that really counts!*
>
> *Would it be correct to assume that the ordinary tangent vector p still transforms in the usual way? In other words, suppose I'm living in a 2+1 dimensional universe, and there's a point particle with rest mass m and hence energy-momentum vector in its rest frame of $p = me_0$. If I cross its world line with a certain relative velocity, there's an element g of $\mathrm{SO}(2,1)$ which tells me how to map the particle's tangent space to my own. Would I measure the particle's energy-momentum to be $p' = gp$? (e.g. if I used the particle to do work in my own rest frame) Would there still be no upper bound on the total energy, i.e. by making our relative velocity close enough to c, I could measure the particle's kinetic energy to be as high as I wished?*
>
> *I guess I'm trying to clarify whether the usual Lorentz transformation of the tangent space has somehow been completely invalidated for extreme boosts, or whether it's just a matter of there being a second definition of "momentum" (defined in terms of the Hamiltonian) which transforms differently and is the appropriate thing to consider in gravitational contexts.*
>
> *In other words, does the cut-off mass apply only to the deficit angle, and do boosts still allow me to measure (by non-gravitational means) arbitrarily large energies (at least in the classical theory)?*

I replied:

> *Greg Egan wrote:*
>
> > *John Baez wrote:*
> >
> > > *The really cool part is the relation between the Lie algebra element p and the group element $\exp(p)$. Originally we thought of p as momentum — but there's a sense in which $\exp(p)$ is the momentum that really counts!*
> >
> > *Would it be correct to assume that the ordinary tangent vector p still transforms in the usual way?*
>
> *Hi! Yes, it would.*

> *In other words, suppose I'm living in a 2+1 dimensional universe, and there's a point particle with rest mass m and hence energy-momentum vector in its rest frame of $p = me_0$. If I cross its world line with a certain relative velocity, there's an element g of $\mathrm{SO}(2,1)$ which tells me how to map the particle's tangent space to my own. Would I measure the particle's energy-momentum to be $p' = gp$? (e.g. if I used the particle to do work in my own rest frame) Would there still be no upper bound on the total energy, i.e. by making our relative velocity close enough to c, I could measure the particle's kinetic energy to be as high as I wished?*

To understand this, it's good to think of the momenta as elements of the Lie algebra $\mathfrak{so}(2,1)$ — it's crucial to the game.

Then, if you have momentum p, and I zip past you, so you appear transformed by some element g of the Lorentz group $\mathrm{SO}(2,1)$, I'll see your momentum as

$$p' = gpg^{-1}$$

This is just another way of writing the usual formula for Lorentz transforms in 3d Minkowski space. No new physics so far, just a clever mathematical formalism.

But when we turn on gravity, letting Newton's constant k be nonzero, we should instead think of momentum as group-valued, via

$$h = \exp(kp)$$

and similarly

$$h' = \exp(kp')$$

Different choices of p now map to the same choice of h. In particular, a particle of a certain large mass — the Planck mass — will turn out to act just like a particle of zero mass!

So, if we agree to work with h instead of p, we are now doing new physics. This is even more obvious when we decide to multiply momenta instead of adding them, since multiplication in $\mathrm{SO}(2,1)$ is noncommutative!

But, if we transform our group-valued momentum in the correct way:

$$h' = ghg^{-1}$$

this will be completely compatible with our previous transformation law for vector-valued momentum!

> *I guess I'm trying to clarify whether the usual Lorentz transformation of the tangent space has somehow been completely invalidated for extreme boosts, or whether it's just a matter of there being a second definition of "momentum" (defined in terms of the Hamiltonian) which transforms differently and is the appropriate thing to consider in gravitational contexts.*

Good question! Amazingly, the usual Lorentz transformations still work EXACTLY — even though the rule for adding momentum is new (now it's multiplication in the group). We're just taking $\exp(kp)$ *instead of* p *as the "physical" aspect of momentum.*

This effectively puts an upper limit on mass, since as we keep increasing the mass of a particle, eventually it "loops around" $\mathrm{SO}(2,1)$ *and act exactly like a particle of zero mass.*

But, it doesn't exactly put an upper bound on energy-momentum, since $\mathrm{SO}(2,1)$ *is noncompact. Of course energy and momentum don't take real values anymore, so one must be a bit careful with this "upper bound" talk.*

> *In other words, does the cut-off mass apply only to the deficit angle, and do boosts still allow me to measure (by non-gravitational means) arbitrarily large energies (at least in the classical theory)?*

There's some sense in which energy-momenta can be arbitrarily large. That's because the space of energy-momenta, namely $\mathrm{SO}(2,1)$*, is noncompact. Maybe you can figure out some more intuitive way to express this.*

I was sitting in a chair in the patent office in Bern when all of a sudden a thought occurred to me. If a person falls freely, he will not feel his own weight.

— *Albert Einstein*

Week 233

May 20, 2006

On Tuesday I'm supposed to talk with Lee Smolin about an idea he's been working on with Fotini Markopoulou and Sundance Bilson-Thompson. This idea relates the elementary particles in one generation of the Standard Model to certain 3-strand framed braids:

1) Sundance O. Bilson-Thompson, "A topological model of composite preons", available as hep-ph/0503213.

2) Sundance O. Bilson-Thompson, Fotini Markopoulou, and Lee Smolin, "Quantum gravity and the Standard Model", hep-th/0603022.

It's a very speculative idea: they've found some interesting relations, but nobody knows if these are coincidental or not.

Luckily, one of my hobbies is collecting mysterious relationships between basic mathematical objects and trying to figure out what's going on. So, I already happen to know a bunch of weird facts about 3-strand braids. I figure I'll tell Smolin about this stuff. But if you don't mind, I'll practice on you!

So, today I'll try to tell a story connecting the 3-strand braid group, the trefoil knot, rational tangles, the groups $\mathrm{SL}(2,\mathbb{Z})$ and $\mathrm{PSL}(2,\mathbb{Z})$, conformal field theory, and Monstrous Moonshine.

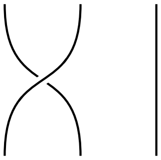
I've talked about some of these things before, but now I'll introduce some new puzzle pieces, which come from two places:

3) Imre Tuba and Hans Wenzl, "Representations of the braid group B_3 and of $\mathrm{SL}(2,\mathbb{Z})$", available as math.RT/9912013.

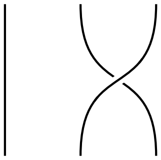
4) Terry Gannon, "The algebraic meaning of genus-zero", available as math.NT/0512248.

You could call it "a tale of two groups".

On the one hand, the 3-strand braid group has generators

$A =$ 

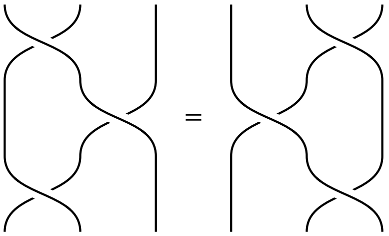
and

$B =$ 

and the only relation is

$$ABA = BAB$$

otherwise known as the “third Reidemeister move”:



On other hand, the group $\mathrm{SL}(2,\mathbb{Z})$ consists of 2×2 integer matrices with determinant 1. It’s important in number theory, complex analysis, string theory and other branches of pure mathematics. I’ve described some of its charms in “Week 125”, “Week 229” and elsewhere.

These groups look pretty different at first. But, there’s a homomorphism from B_3 onto $\mathrm{SL}(2,\mathbb{Z})$! It goes like this:

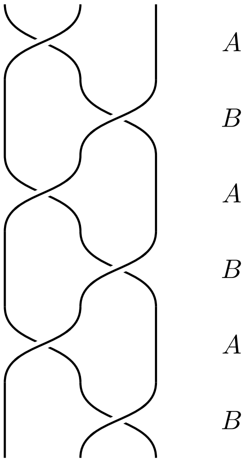
$$A \longmapsto \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$B \longmapsto \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$$

Both these matrices describe “shears” in the plane. You may enjoy drawing these shears and visualizing the equation $ABA = BAB$ in these terms. I did.

I would like to understand this better. . . and here are some clues.

The center of B_3 is generated by the element $(AB)^3$. This element corresponds to a “full twist”. In other words, it’s the braid you get by hanging 3 strings from the ceiling, grabbing them all with one hand at the bottom, and giving them a full 360-degree twist:



This full twist gets sent to -1 in $\mathrm{SL}(2,\mathbb{Z})$:

$$(AB)^3 \longmapsto \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$

So, the double twist gets sent to the identity:

$$(AB)^6 \longmapsto \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

In fact, Tuba and Wenzl say the double twist *generates* the kernel of our homomorphism from B_3 to $\mathrm{SL}(2,\mathbb{Z})$. So, $\mathrm{SL}(2,\mathbb{Z})$ is isomorphic to the group of 3-strand braids modulo double twists!

This reminds me of spinors... since you have to twist an electron around *twice* to get its wavefunction back to where it started. And indeed, $\mathrm{SL}(2,\mathbb{Z})$ is a subgroup of $\mathrm{SL}(2,\mathbb{C})$, which is the double cover of the Lorentz group. So, 3-strand braids indeed act on the state space of a spin-$1/2$ particle, with double twists acting trivially!

(For more on this, check out Trautman's work on "Pythagorean spinors" in "Week 196". There's also a version where we use integers $\mod 7$, described in "Week 219".)

If instead we take 3-strand braids modulo full twists, we get the so-called "modular group":

$$\mathrm{PSL}(2,\mathbb{Z}) = \mathrm{SL}(2,\mathbb{Z})/\{\pm 1\}$$

Now, $\mathrm{SL}(2,\mathbb{Z})$ is famous for being the "mapping class group" of the torus — that is, the group of orientation-preserving diffeomorphisms, modulo diffeomorphisms connected to the identity. Similary, $\mathrm{PSL}(2,\mathbb{Z})$ is famous for acting on the rational numbers together with a point at infinity by means of fractional linear transformations:

$$z \mapsto \frac{az+b}{cz+d}$$

where a, b, c, d are integers and $ad - bc = 1$. The group $\mathrm{PSL}(2,\mathbb{Z})$ also acts on certain 2-strand tangles called "rational tangles". In "Week 229", I told a nice story I heard from Michael Hutchings, explaining how these three facts fit together in a neat package.

But now let's combine those facts with the stuff I just said! Since $\mathrm{PSL}(2,\mathbb{Z})$ acts on rational tangles, and there's a homomorphism from B_3 to $\mathrm{PSL}(2,\mathbb{Z})$, 3-strand braids must act on rational tangles. How does that go?

There's an obvious guess, or two, or three, or four, but let's just work it out.

I just said that the 3-strand braid A gets mapped to this shear:

$$A \longmapsto \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

In "Week 229" I said what this shear does to a rational tangle. It gives it a 180 degree twist at the bottom, like this:

```
 |   |                 |   |
 |   |                 |   |
 |   |                 |   |
-------               -------
|  T  |   |---->      |  T  |
-------               -------
 |   |                  \ /
 |   |                   /
 |   |                  / \
```

Next, Tuba and Wenzl point out that

$$ABA = BAB \longmapsto \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

which is a quarter turn. From "Week 229" you can see how this quarter turn acts on a rational tangle:

```
  |   |                       |     |
  |   |              ----     |     |
  |   |             /     \   |     |
 -------           |     -------    |
 |  T  |   |---->  |     |  T  |    |
 -------           |     -------    |
  |   |            |      |   \____/
  |   |            |      |
  |   |            |      |
```

It gives it a quarter turn!

From these facts, we can figure out what the braid B does to a rational tangle. So, let me do the calculation.

Scribble, scribble, curse and scribble.... Eureka!

Since we know what A does, and what ABA does, we can figure out what B must do. But, to make the answer look cute, I needed a sneaky fact about rational tangles, which is that A *also* acts like this:

```
  |   |                 \ /
  |   |                  /
  |   |                 / \
 -------              -------
 |  T  |   |---->     |  T  |
 -------              -------
  |   |                |   |
  |   |                |   |
  |   |                |   |
```

This is proved in Goldman and Kauffman's paper cited in "Week 228". With the help of this, I can show B acts like this:

```
  |   |              |          |
  |   |              |   ___    |
  |   |              |  /   \   |
 -------             | /    -------
 |  T  |   |---->     \     |  T  |
 -------             / \    -------
  |   |             |   \___/   |
  |   |             |           |
  |   |             |           |
```

And this is *great!* It means our action of 3-strand braids on rational tangles is really easy to describe. Just take your tangle and let the upper left strand dangle down:

```
           |
   ----    |
  /    \   |
 |    -------
 |    |  T  |
 |    -------
 |     |   |
 |     |   |
 |     |   |
```

To let a 3-strand braid act on this, just attach it to the bottom of the picture!

(That's why there were *four* obvious guesses about this would work: one can easily imagine four variations on this trick, depending on which strand is the "odd man out" — here it's the upper right. It's just a matter of convention which we use, but my conventions give this.)

In fact, even the group of 4-strand braids acts on rational tangles in an obvious way, but the 3-strand braid group is enough for now.

Let me summarize. The 3-strand braid group B_3 acts on rational tangles in an obvious way. The subgroup that acts trivially is precisely the center of B_3, generated by the full twist. Using stuff from "Week 229", it follows that the quotient of B_3 by its center acts on the projectivized rational homology of the torus. We thus get a topological explanation of why B_3 mod its center is $\mathrm{PSL}(2,\mathbb{Z})$.

But there's more.

For starters, the 3-strand braid group is also the fundamental group of S^3 minus the trefoil knot!

And, S^3 minus the trefoil knot is secretly the same as $\mathrm{SL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{Z})$!

In fact, Terry Gannon writes that the 3-strand braid group can be regarded as "the universal central extension of the modular group, and the universal symmetry of its modular forms". I'm not completely sure what that means, but here's *part* of what it means.

Just as $\mathrm{PSL}(2,\mathbb{C})$ is the Lorentz group in 4d spacetime, $\mathrm{PSL}(2,\mathbb{R})$ is the Lorentz group in 3d spacetime. This group has a double cover $\mathrm{SL}(2,\mathbb{R})$, which shows up when you study spinors. But, it also has a universal cover, which shows up when you study anyons. The universal cover has infinitely many sheets. And up in this universal cover, sitting over the subgroup $\mathrm{SL}(2,\mathbb{Z})$, we get... the 3-strand braid group!

In math jargon, we have this commutative diagram where the rows are short exact sequences:

$$\begin{tikzcd}
1 \arrow[r] & \mathbb{Z} \arrow[r] \arrow[d] & B_3 \arrow[r] \arrow[d] & \mathrm{SL}(2,\mathbb{Z}) \arrow[r] \arrow[d] & 1 \\
1 \arrow[r] & \mathbb{Z} \arrow[r] & \mathrm{SL}(2,\mathbb{R})^\sim \arrow[r] & \mathrm{SL}(2,\mathbb{R}) \arrow[r] & 1
\end{tikzcd}$$

Here $\mathrm{SL}(2,\mathbb{R})^\sim$ is the universal cover of $\mathrm{SL}(2,\mathbb{R})$. Since $\pi_1(\mathrm{SL}(2,\mathbb{R})) = \mathbb{Z}$, this is a $\mathbb{Z}$-fold cover. You can describe this cover explicitly using the Maslov index, which is a formula that actually computes an integer for any loop in $\mathrm{SL}(2,\mathbb{R})$, or indeed any symplectic group.

But anyway, fiddling around with this diagram and the long exact sequence of homotopy groups for a fibration, you can show that indeed:

$$\pi_1(\mathrm{SL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{Z})) = B_3.$$

This also follows from the fact that $\mathrm{SL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{Z})$ looks like S^3 minus a trefoil.

Gannon believes that number theorists should think about all this stuff, since he thinks it's lurking behind that weird network of ideas called Monstrous Moonshine (see "Week 66").

And here's the basic reason why. I'll try to get this right....

Any rational conformal field theory has a "chiral algebra" A which acts on the left-moving states. Mathematicians call this sort of thing a "vertex operator algebra". A representation of this on some vector space V is a space of states for the circle in some "sector" of our theory. Let's pick some state v in V. Then we can define a "one-point function" where we take a Riemann surface with little disk cut out and insert this state on the boundary. This is a number, essentially the amplitude for a string in the give state to evolve like this Riemann surface says.

In fact, instead of chopping out a little disk it's nice to just remove a point — a "puncture", they call it. But, we get an ambiguous answer unless we pick coordinates at this point, or in the lingo of complex analysis, a choice of "uniforming parameter". Then our one-point function becomes a function on the moduli space of Riemann surfaces equipped with a puncture and a choice of uniformizing parameter.

If we didn't have this uniformizing parameter to worry about, we'd just have the moduli space of tori equipped with a marked point, which is nothing but the usual moduli space of elliptic curves,

$$H/\mathrm{PSL}(2,\mathbb{Z})$$

where H is the complex upper halfplane. Then our one-point function would have nice transformation properties under $\mathrm{PSL}(2,\mathbb{Z})$.

But, with this uniformizing parameter to worry about, our one-point function only has nice transformation properties under B_3. This is somehow supposed to be related to how B_3 is the "universal central extension" of $\mathrm{PSL}(2,\mathbb{Z})$: in conformal field theory, all sorts of naive symmetries hold only up to a phase, so you have to replace various groups by central extensions thereof... and here that's what's happening to $\mathrm{PSL}(2,\mathbb{Z})$!

That last paragraph was pretty vague. If I'm going to understand this better, either someone has to help me or I've got to read something like this:

5) Yongchang Zhu, "Modular invariance of characters of vertex operator algebras", *J. Amer. Math. Soc* **9** (1996), 237–302. Also available at `http://www.ams.org/jams/1996-9-01/S0894-0347-96-00182-8/home.html`

But I shouldn't need any conformal field theory to see how the moduli space of punctured tori with uniformizing parameter is related to the 3-strand braid group! I bet this moduli space is X/B_3 for some space X, or something like that. There's something simple at the bottom of all this, I'm sure.

Addenda: Another relation between the trefoil and the punctured torus: the trefoil has genus 1, meaning that it bounds a torus minus a disc embedded in $\mathbb{R}^3$. You can see this in the lecture "Genus and knot sum" in this course on knot theory:

6) Brian Sanderson, "The knot theory MA3F2 page", `http://www.maths.warwick.ac.uk/\~bjs/MA3F2-page.html`

This course also has material on rational tangles.

The fact that B_3 is a central extension of $\mathrm{PSL}(2,\mathbb{Z})$ by $\mathbb{Z}$, and the quantum-mechanical interpretation of a central extension in terms of phases, plays an important role here:

7) R. Voituriez, "Random walks on the braid group B_3 and magnetic translations in hyperbolic geometry", *Nucl. Phys.* **B621** (2002), 675–688. Also available as `http://arxiv.org/abs/math-ph/0103008`.

Among other things, he points out that the homomorphism $B_3 \to \mathrm{SL}(2,\mathbb{Z})$ described above is the "Burau representation" of B_3 evaluated at $t = 1$. In general, the Burau representation of B_3 is given by:

$$A \longmapsto \begin{pmatrix} t & 1 \\ 0 & 1 \end{pmatrix}$$

$$B \longmapsto \begin{pmatrix} 1 & 0 \\ -t & t \end{pmatrix}$$

(Conventions differ, and this may not be the best, but it's the one he uses.) The Burau representation can also be used to define a knot invariant called the Alexander polynomial. I believe that with some work, one can use this to explain why Conway could calculate the rational number associated to a rational tangle in terms of the ratio of Alexander polynomials of two links associated to it, called its "numerator" and "denominator". In fact he computed this ratio of polynomials and then evaluate it at a special value of t — presumably the same special value we're seeing here (modulo differences in convention).

Another issue: I wrote

> *For starters, the* 3*-strand braid group is also the fundamental group of* S^3 *minus the trefoil knot!*
>
> *And,* S^3 *minus the trefoil knot is secretly the same as* $\mathrm{SL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{Z})$*!*

The first one is pretty easy to see; you start with the "Wirtinger presentation" of the fundamental group of S^3 minus a trefoil, and show by a fun little calculation that this isomorphic to the braid group on 3 strands. A more conceptual proof would be very nice, though. (See "Week 261" for such a proof — and much more on all this stuff.)

What about the second one? Why is S^3 minus the trefoil knot diffeomorphic to $\mathrm{SL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{Z})$? Terry Gannon says so in his paper above, but doesn't say why. Some people asked about this, and eventually some people found some explanations. First of all, there's a proof on page 84 of this book:

8) John Milnor, *Introduction to Algebraic K-theory*, Annals of Math. Studies **72**, Princeton U. Press, Princeton, New Jersey, 1971.

Milnor credits it to Quillen. Joe Christy summarizes it below. I can't tell if this proof is essentially the same as another sketched below by Swiatoslaw Gal, which exhibits a diffeomorphism using functions called the Eisenstein series g_2 and g_3. They are probably quite similar arguments.

Joe Christy writes:

> *I wouldn't be surprised if this was known to Seifert in the 30's, though I can't lay my hands on Seifert & Threfall at the moment to check. Likewise for Hirzebruch, Brieskorn, Pham & Milnor in the 60's in relation to singularities of complex hypersurfaces and exotic spheres. When I was learning topology in the 80's it was considered a warm up case of Thurston's Geometrization Program — the trefoil knot complement has* $\mathrm{PSL}(2,\mathbb{R})$ *geometric structure.*
>
> *In any case, peruse Milnor's Annals of Math Studies for concrete references. There is a (typically) elegant proof on p.84 of "Introduction to Algebraic K-theory" [study 72], which Milnor credits to Quillen. It contains the missing piece of John's argument: introducing the Weierstrass* $\wp$*-function and remarking that the differential equation that it satisfies gives the diffeomorphism to* S^3*-trefoil as the boundary of the pair (discriminant of diff-eq,* $\mathbb{C}^2 = (\wp, \wp')$*-space).*
>
> *This point of view grows out of some observations of Zariski, fleshed out in "Singular Points of Complex Hypersurfaces" [study 61]. The geometric viewpoint is made explicit in the paper "On the Brieskorn Manifolds* $M(p,q,r)$*" in "Knots, Groups, and 3-manifolds" [study 84].*
>
> *It is also related to the intermediate case between the classical Platonic solids and John's favorite Platonic surface — the Klein quartic. By way of a hint, look to relate the trefoil, qua torus knot, the seven-vertex triangulation of the torus, and the dual hexagonal tiling of a (flat) Clifford torus in* S^3.
>
> *Joe*

Swiatoslaw Gal writes:

> *In fact the isomorphism is a part of the modular theory:*
>
> *Looking for*
>
> $$f\colon \mathrm{GL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{Z}) \to \mathbb{C}^2 \setminus \{x^2 = y^3\}$$
>
> *(there is an obvious action of* $\mathbb{R}^+$ *on both sides:*
>
> $$\begin{aligned} M &\mapsto tM \textit{for } M \textit{ in } \mathrm{GL}(2,\mathbb{R}), \\ x &\mapsto t^6 x, \\ y &\mapsto t^4 y, \end{aligned}$$
>
> *and the quotient is what we want).*
>
> $\mathrm{GL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{Z})$ *is a space of lattices in* $\mathbb{C}$*. Such a lattice* L *has classical invariants*
>
> $$g_2(L) = 60 \sum_{z \in L'} z^{-4},$$

and

$$g_3(L) = 140 \sum_{z \in L'} z^{-6},$$

where $L' = L \setminus \{0\}$.

The modular theory asserts that:

1. *For every pair* (g_2, g_3) *there exist a lattice* L *such that* $g_2(L) = g_2$ *and* $g_3(L) = g_3$ *provided that* g_2^3 *is not equal to* $27g_3^2$.
2. *Such a lattice is unique.*

Best,
S. R. Gal

The quantity $g_2^3 - 27g_3^2$ is called the "discriminant" of the lattice L, and vanishes as the lattice squashes down to being degenerate, i.e. a discrete subgroup of $\mathbb{C}$ with one rather than two generators.

Week 234

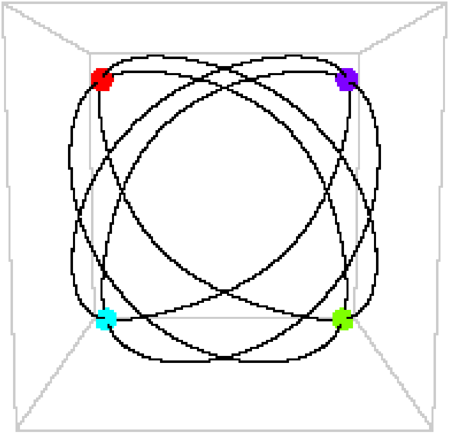
June 12, 2006

Today I'd like to talk about the math of music — including torsors, orbifolds, and maybe even Mathieu groups. But first, some movies of the n-body problem:

1) Cris Moore, "The 3-body (and n-body) problem", `http://www.santafe.edu/~moore/gallery.html`

In 1993 Cris Moore discovered solutions of the gravitational n-body problem where the particles' paths lie in a plane and trace out braids in spacetime! I spoke about these in "Week 181".

More recently, Moore and Michael Nauenberg have found solutions with cubic symmetry and vanishing angular momentum, and made movies of these:



For the mathematical details, try this:

2) Cristopher Moore and Michael Nauenberg, "New periodic orbits for the n-body problem", available at `math.DS/0511219`.

Next, math and music.

Some of you have been in this situation. A stranger at a party asks what you do. You reluctantly admit you're a mathematician, expecting one of the standard responses: "Oh! I hate math!" or "Oh! I was pretty good at math until. . . ."

But instead, after a strained moment they say: "Oh! Do you play an instrument too? Isn't music really mathematical?"

I guess it's like meeting a Martian and asking them if they like Arizona: an attempt to humanize something alien and threatening. You may not have much in common, but at least you can chat about red rocks.

Of course there *is* something mathematical about music, and lots of mathematicians play music. I rarely think about music in a mathematical way. But I know they have something in common: the transcendent beauty of pure form.

Indeed, in the Middle Ages, music was part of a "quadrivium" of mathematical arts: arithmetic, geometry, music, and astronomy. These were studied after the "trivium" of grammar, rhetoric and logic. This is why mathematicians scorn a result as "trivial" when it's easy to see using straightforward logic. So when a result seems more profound, they should call it "quadrivial".

Try saying it sometime: "Cool! That's quadrivial!" It might catch on.

There are also modern applications of math to music theory. I had never heard of "neo-Riemannian theory" until Tom Fiore explained it to me while I was visiting Chicago. Tom is a postdoc who works on categorified algebraic theories, double categories and the like — but he's also into music theory:

3) Thomas M. Fiore, "Music and mathematics", available at `http://www.math.uchicago.edu/~fiore/1/music.html`

4) Thomas M. Fiore and Ramon Satyendra, "Generalized contextual groups", *Music Theory Online* **11** (2005), available at http://www.math.uchicago.edu/~fiore/1/music.html

The first of these is a very nice gentle introduction, suitable both for musicians who don't know group theory and mathematicians who don't know a triad from a tritone!

When Tom first mentioned "neo-Riemannian theory", I thought this was some bizarre application of differential geometry to music. But no — we're not talking about the 19th-century mathematician Bernhard Riemann, we're talking about the 19th-century music theorist Hugo Riemann!

Based on the work on Euler — yes, *the* Euler — Hugo Riemann introduced diagrams called "tone nets" to study the network of relations between similar chords. You can see his original setup here:

5) Joe Monzo, "Tonnetz: the tonal lattice invented by Riemann", "Tonalsoft: the Encyclopedia of Microtonal Music Theory", `http://www.tonalsoft.com/enc/t/tonnetz.aspx`

6) Paul Dysart, "Tonnetz: musics, harmony and donuts", `http://members2.boo.net/~knuth/`

Apparently Riemann's ideas have caught on in a big way. Monzo says that "use of lattices is endemic on internet tuning lists", as if they were some sort of infectious disease.

Dysart seems more gung-ho about it all. The "donuts" he mentions arise when you curl up tone nets by identifying notes that differ by an octave. He has some nice pictures of them!

In neo-Riemannian theory, people like Lewin and Hyer started extending Riemann's ideas by using *group theory* to systematize operations on chords. The best easy introduction to this is Fiore's paper "Music and mathematics". Here you can read about math lurking in the music of Elvis and the Beatles! Or, if you're more of a highbrow sort, see what he has to say about Hindemith and Liszt's "Transcendental Etudes". And if you like doughnuts and music, you'll love the section where he explains how Beethoven's Ninth traces out a systematic path in a torus-shaped tone net! This amazing fact was discovered by Cohn, Douthett, and Steinbach.

(If I weren't so darn honest, I'd add that Liszt wrote the "Transcendental Etudes" as a sequel to his popular "Algebraic Etudes", and explain how Mozart's "eine kleine Nachtmusik" tours a tone net shaped like a Klein bottle. But alas. . . .)

Let me explain a bit about group theory and music — just enough to reach something really cool Tom told me.

If you're a musician, you'll know the notes in an octave go like this, climbing up:

C, C#, D, D#, E, F, F#, G, G#, A, A#, B

until you're back to C. If you're a mathematician, you might be happier to call these notes

$$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11$$

and say that we're working in the group of integers $\mod 12$, otherwise known as $\mathbb{Z}/12$. Let's be mathematicians today.

The group $\mathbb{Z}/12$ has been an intrinsic feature of Western music ever since pianos were built to have "equal temperament" tuning, which makes all the notes equally spaced in a certain logarithmic sense: each note vibrates at a frequency of $2^{\frac{1}{12}}$ times the note directly below it.

Only 7 of the 12 notes are used in any major or minor key — for example, C,D,E,F,G,A,B is C major and A,B,C,D,E,F,G is A minor. So, as long as Western composers stuck to writing pieces in a single fixed key, the $\mathbb{Z}/12$ symmetry was "spontaneously broken" by their choice of key, only visible in the freedom to change keys.

But, as composers gradually started changing keys ever more frequently within a given piece, the inherent $\mathbb{Z}/12$ symmetry became more visible. In the late 1800s this manifested itself in trend called "chromaticism". Roughly speaking, music is "chromatic" when it freely uses all 12 notes, but still within the context of an — often changing — key. I guess Wagner and Richard Strauss are often mentioned as pinnacles of chromaticism.

Chromaticism then led to full-fledged "twelve-tone music" starting with Schoenberg in the early 1900s. This is music that fully exploits the $\mathbb{Z}/12$ symmetry and doesn't seek to privilege a certain 7-element subset of notes defining a key. People found Schoenberg's music disturbing and dissonant at the time, but I find it very beautiful.

Now comes the really exciting thing Tom told me: two other symmetry groups lurking in music, and a relationship between them.

First, the transposition-inversion group. This acts as permutations of the set $\mathbb{Z}/12$. It's generated by two especially nice permutations. The first is "transposition". This raises each note a step:

$$x \mapsto x + 1$$

Musicians would call this a half-step, just like physicists measure spin in multiples of $1/2$, but we're being mathematicians! The second is "inversion". This turns notes upside down:

$$x \mapsto -x$$

The relevance of this to music is a bit less obvious: composers like Bach and Schoenberg used it explicitly, but we'll see it playing a subtler role, relating major and minor chords.

The transposition-inversion group has 24 elements. Mathematicians call it the 24-element "dihedral group", since it consists of the symmetries of a regular 12-sided polygon where you're allowed to rotate the polygon (transposition) and also flip it over (inversion). I hope you see that this geometrical picture is just a way of visualizing the 12 notes.

So, the transposition-inversion group obviously on the 12-element set of notes. But, it also acts on the 24-element set of "triads"!

Triads are among the most basic chords in music. Mathematically they are certain 3-element subsets of $\mathbb{Z}/12$. They come in two kinds, major and minor. There are 12 major triads, namely

$$\{0,4,7\} = \{\mathrm{C},\mathrm{E},\mathrm{G}\} = \text{C major triad}$$

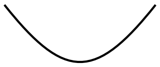
and everything you can get from this by transposition. If you invert these, you get the 12 minor triads, namely

$$\{0,-4,-7\} = \{5,8,0\} = \{\mathrm{F},\mathrm{A}^\flat,\mathrm{C}\} = \text{F minor triad}$$

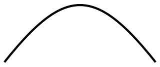
and everything you can get from *this* by transposition.

(Note that $\{0,-4,-7\} = \{5,8,0\}$ because we're working $\bmod 12$ and the order doesn't matter. I've also included the way musicians talk about these triads, in case you care.)

Major triads sound happy; when you invert them they sound sad, just like an upside-down smile looks sad. There could be some profound truth lurking here. A smile has a positive second derivative:



which says that things are "looking up", while a frown has negative second derivative:



which says that things are "looking down". An upside-down smile is a frown.

(On the other hand, a backwards smile is still a smile, and a backwards frown is still a frown. So, if you're a company and the second derivative of your profits is positive, you can say business is looking up — and you could still say this if time were reversed!)

But never mind. We had this transposition-inversion group acting on our set of notes, namely $\mathbb{Z}/12$. Since tranposition and inversion act on notes, they also act on triads. For example, transposition does this:

$$\{0,4,7\} \mapsto \{1,5,8\}$$
$$\text{C major triad} \mapsto \text{C}^\sharp \text{ major triad}$$

while inversion does this:

$$\{0,4,7\} \mapsto \{5,8,0\}$$
$$\text{C major triad} \mapsto \text{F minor triad}$$

So, we've got this 24-element transposition-inversion group acting on the 24-element set of triads!

But here's really cool part: there's *another* important 24-element group acting on the same set! It's easy to define mathematically, but it also has a musical meaning.

Mathematically, it's just the "centralizer" of the transposition- inversion group. In other words, it consists of all ways of permuting triads that *commute* with transposition and inversion!

Musically, it's called the "PLR" group, because it's generated by 3 famous transformations.

To describe these transformations, I'll need to talk about the "bottom", "middle" and "top" note of a triad. If you know a wee bit of music theory this should be obvious as long as you know I'm talking about triads in root position. If you're a mathematician who has never studied music theory and you think of triads as 3-element subsets of $\mathbb{Z}/12$, it might be less obvious, since $\mathbb{Z}/12$ doesn't have a nice ordering — it only has a *cyclic* ordering. But this is enough. The point is that major triads are sets of the form

$$\{n, n+4, n+7\},$$

while minor triads are of the form

$$n, n+3, n+7.$$

So, we can call the note n the "bottom", the note $n+3$ or $n+4$ the "middle", and $n+7$ the "top". Musicians call them the "root", "third" and "fifth", but let's be simple-minded mathematicians.

Okay, what are the transformations P, L, and R? They stand for "parallel", "leading tone change", and "relative" — but what *are* they?

Each of these transformations keeps exactly 2 of the notes in our triad the same. Also, each changes major triads into minor triads and vice versa. These features make these transformations musically interesting.

The transformation "P" keeps the top and bottom notes the same. I've now said enough for you to figure out what it does... at least in principle. For example:

$$P\colon \{0,4,7\} \mapsto \{0,3,7\}$$

$$\text{C major triad} \mapsto \text{C minor triad}$$

$$P\colon \{0,3,7\} \mapsto \{0,4,7\}$$

$$\text{C minor triad} \mapsto \text{C major triad}$$

The tranformation "L" turns the middle and top note into the bottom and middle note when you start with a MAJOR triad. It turns the bottom and middle note into the middle and top note when you start with a MINOR triad. For example:

$$L\colon \{0,4,7\} \mapsto \{4,7,11\}$$

$$\text{C major triad} \mapsto \text{E minor triad}$$

$$L\colon \{0,3,7\} \mapsto \{8,0,3\}$$

$$\text{C minor triad} \mapsto \text{G major triad}$$

The transformation "R" works the other way around. It turns the middle and top note into the bottom and middle note when you start with a MINOR triad. And it turns the bottom and middle note into the middle and top note when you start with a MAJOR triad:

$$R\colon \{0,4,7\} \mapsto \{9,0,4\}$$

C major triad ↦ A minor triad

$$R\colon \{0,3,7\} \mapsto \{3,7,10\}$$

C minor triad ↦ D$^\sharp$ major triad

Can you see why the transformations P, L, and R commute with transposition and inversion? It should be easy to see that they commute with transposition. Commuting with inversion means that if I switch the words "top" and "bottom" and also the words "major" and "minor" in my descriptions above, these transformations don't change!

You should be left wondering why P, L, and R generate the group of *all* transformations of triads that commute with transposition and inversion — and why this group, like the transposition-inversion group itself, has exactly 24 elements!

It turns out some of this has a simple explanation, which has very little to do with the details of triads or even the 12-note scale.

Imagine a scale with n equally spaced notes. Transpositions and inversions will generate a group with $2n$ elements. Let's call this group G. If you take any "sufficiently generic" chord in our scale, G will act on it to give a set S consisting of $2n$ different chords. Then it's a mathematical fact that the group of permutations of S that commute with all transformations in G will be isomorphic to G! So, it too will have $2n$ elements.

To explain *why* this is true, I need a bit more math.

First of all, I need to define my terms. I'm defining a chord to be "sufficiently generic" if no element of G maps it to itself. We then say G acts *freely* on S. By the way we've set things up, G also acts *transitively* on S. A nonempty set on which G acts both freely and transitively is called a "G-torsor". You can read about torsors here:

7) John Baez, "Torsors made easy", `http://math.ucr.edu/home/baez/torsors.html`

They're philosophically very interesting, since they're related to gauge symmetries in physics... but right now the only fact we need is that any G-torsor is isomorphic to G. So, we can identify S with G, with G acting by left multiplication.

Then, it's a well-known fact that any permutation of G that commutes with left multiplication by all elements of G must be given by *right* multiplication by some element of G. And these right multiplications form a group of transformations that is isomorphic to G... just as we were trying to show!

In other words: the group of permutations of G has a subgroup isomorphic to G, namely the left translations. It also has another subgroup isomorphic to G, namely the right translations. Each of these subgroups is the "centralizer" of the other. That is, each one consists of all permutations that commute with every permutation in the other one! Fiore and Satyendra call them "dual groups".

In our application to music, the first copy of G is our good old transposition-inversion group, while the second copy is a generalization of the PLR group. Fiore and Satyendra call it the "generalized contextual group".

All this is indeed very general. I don't know a similarly general explanation of why the operations P, L, and R succeed in generating all transformations that commute with transposition and inversion.

I asked Tom Fiore if he and Ramon Satyendra were the first to show that the PRL group was the centralizer of the transposition- inversion group. His reply was packed with information, so I'll quote it:

> *The initial insight about the duality between the T/I group and the PLR group was at least 20 year ago. Dual groups in the musical sense were introduced in David Lewin's seminal 1987 book "Generalized Musical Intervals and Transformation Theory." This book stimulated interest in neo-Riemannian theory, since Lewin recalled the transformations P, L, and R as objects of study.*
>
> *Major-minor duality was a concern of Hugo Riemann, a theorist from the second half of the 19th century. Given his interest in duality, Riemann may have had some intuition about a duality between T/I and PLR, though it wasn't until after his death that this duality was formulated in algebraic terms. An algebraic proof of the duality of T/I and PLR was in the thesis of Julian Hook in 2002.*
>
> *Ramon and I were the first to prove that the "generalized contextual group" is dual to the T/I group acting on a set generated by an arbitrary pitch-class segment satisfying the tritone condition. (The tritone condition says that the inital pitch-class segment contains an interval other than a tritone and unison.) Our theorem has the PLR group and major/minor triads as a special case, since the generalized contextual group becomes the PLR group when one takes the generating pitch class segment to be the three pitches of a major chord. The advantage of our generalization is that one can now apply the PLR insight to passages that are not triadic. There was a general move toward this in practice for the past decade (Childs and Gollin considered seventh chords rather than triads, Lewin analyzed instances of a non-diatonic phrase in a piano work of Schoenberg, we analyzed Hindemith, and so on). Most music does not consist entirely of triads (e.g. late 19th century chromatic music), so the restriction of PLR to triads was not conclusive.*
>
> *We did a literature review of recent neo-Riemannian theory in Part 5 of our article "Generalized Contextual Groups", since there have been a lot of insights in the past 10 years. One of the main thinkers is Rick Cohn, who came up with (among other things) a nice tiling of the plane which one navigates using P, L, and R (Richard Cohn, "Neo-Riemannian operations, parsimonious trichords, and their Tonnetz representations",* Journal of Music Theory*, 1997). It is quite geometric.*

You read more about these matters here... I'll list these references in the order Tom mentions them:

8) David Lewin, *Generalized Musical Intervals and Transformations*, Yale University Press, New Haven, Connecticut, 1987.

9) Julian Hook, *Uniform Triadic Transformations*, Ph.D. thesis, Indiana University, 2002.

10) Adrian P. Childs, "Moving beyond neo-Riemannian triads: exploring a transformational model for seventh chords", *Journal of Music Theory* **42/2** (1998), 191–193.

11) Edward Gollin, "Some aspects of three-dimensional Tonnetze", *Journal of Music Theory* **42/2** (1998), 195–206.

12) Richard Cohn, 'Neo-Riemannian operations, parsimonious trichords, and their "Tonnetz" representations', *Journal of Music Theory* **41/1** (1997), 1–66.

13) David Lewin, "Transformational considerations in Schoenberg's Opus 23, Number 3", preprint.

In fact, the notion of "torsor" pervades the work of David Lewin, but not under this name — Lewin calls it a "general interval system". Stephen Lavelle noticed the connection to torsors in 2005:

14) Stephen Lavelle, "Some formalizations in musical set theory", June 3, 2005, available at `http://www.maths.tcd.ie/~icecube/lewin.pdf` and `http://www.maths.tcd.ie/~icecube/lewin.ps`

Unfortunately the music theorists seem not to have set up an "arXiv", so some of their work is a bit hard to find. For example, all of Volume 42 Issue 2 of the Journal of Music Theory is dedicated to neo-Riemannian theory, but I don't think it's available online. Luckily, the music theorists have set up some free online journals, like this:

15) *Music Theory Online*, `http://mto.societymusictheory.org/`

and this one has links to others. The Society for Music Theory also has online resources including a nice bibliography on the basics of music theory:

16) Society for Music Theory, "Fundamentals of music theory, selected bibliography", `http://societymusictheory.org/index.php?pid=37`

Now let me turn up the math level a notch. . . .

If you're the right sort of mathematician, you'll have noticed by now that we're doing some fun stuff starting with the abelian group $A = \mathbb{Z}/12$. First we're forming the group G consisting of all "affine transformations" of A. These are the transformations that preserve all these operations:

$$(x, y) \mapsto cx + (1 - c)y$$

where c is an integer. For $A = \mathbb{Z}/n$, the group of affine transformations has the transposition-inversion group as a subgroup. The whole affine group has 48 elements, but for now we only keep this subgroup with 24 elements. Call it G.

Then, we're saying that we can take any "sufficiently generic" subset of A, hit it with all elements of G, and get a G-torsor, say S. G is then seen as a subgroup of the group of permutations of S, and the centralizer of this subgroup is again isomorphic to G.

You may be more familiar with affine transformations on a vector space, where we get to use any real number for c. Then

$$cx + (1 - c)y$$

describes the line through x and y, so you can say that affine transformations are those that preserve lines. Vector spaces are R-modules for R the reals, while abelian groups are R-modules for R the integers. The concept of "affine transformations" of an R-module works pretty much the same way whenever R is any commutative ring. And, indeed, everything I just said in the last paragraph works if we let A be an R-module for any commutative ring R.

So, there's some very simple nice abstract stuff going on here: we're taking an abelian group A, looking at a subgroup G of its affine transformations, and seeing that sufficiently generic subsets of A give rise to G-torsors!

These are nice examples of G-torsors, since nobody is likely to accidentally confuse them with the group G. If you read my webpage on torsors, you'll see it's often easy to mix up a G-torsor with the group G itself.

In fact, I just committed this sin myself! The set of notes is not naturally an abelian group until we pick an origin — a place for the chromatic scale to start. It's really just an A-torsor, where A is the abelian group generated by transposition.

So, there lots of torsors lurking in music. . . .

The pretty math I've just described only captures a microscopic portion of what makes music interesting. It doesn't, for example, have anything to say about what makes some intervals more dissonant than others. As Pythagoras noticed, simple frequency ratios like $3/2$ or $4/3$ make for less dissonant chords than gnarly fractions like $1259/723$. The equal tempered tuning system, where the basic frequency ratio is $2^{\frac{1}{12}}$, would have made Pythagoras roll in his grave! Advocates of other tuning systems say these irrational frequency ratios are driving us crazy, making wars break out and plants wilt — but there's an unavoidable conflict between the desire for simple ratios and the desire for evenly spaced notes, built into the fabric of mathematics and music. Every tuning system is thus a compromise. I would like to understand this better; there's bound to be a lot of nice number theory here.

To study different tuning systems in a unified way, one first step is replace the group $\mathbb{Z}/12$ by a continuous circle. Points on this circle are "frequencies modulo octaves", since for many — though certainly not all — purposes it's good to consider two notes "the same" if they differ by an octave. Mathematically this circle is $\mathbb{R}^+/2$, namely the multiplicative group of positive real numbers modulo doubling. As a group, it's isomorphic to the usual circle group, $\mathrm{U}(1)$.

This "pitch class circle" plays a major role in the work of Dmitri Tymoczko, a composer and music theorist from Princeton, who emailed me after I left a grumpy comment on the discussion page for this fascinating but slightly obscure article:

17) Wikipedia, "Musical set theory", `http://en.wikipedia.org/wiki/Musical_set_theory`

He's recently been working on voice leading and orbifolds. They're related topics, because if you have a choir of n indistinguishable angels, each singing a note, the set of possibilities is:

$$T^n/S_n$$

where T^n is the n-torus — the product of n copies of the pitch class circle — and S_n is the permutation group, acting on n-tuples of notes in the obvious way. This quotient is not usually a manifold, because it has singularities at certain points where more than one voice sings the same note. But, it's an *orbifold*. This kind of slightly singular quotient space is precisely what orbifolds were invented to deal with.

Tymoczko is coming out with an article about this in Science magazine. For now, you can learn more about the geometry of music by playing with his "ChordGeometries" software:

18) Dmitri Tymoczko, ChordGeometries, `http://music.princeton.edu/~dmitri/ChordGeometries.html`

As for "voice leading", let me just quote his explanation, suitable for mathematicians, of this musical concept:

> *BTW, if you're writing on neo-Riemannian theory in music, it might be helpful to keep the following basic distinction in mind. There are chord progressions, which are essentially functions from unordered chords to unordered chords (e.g. the chord progression (function) that takes C major to E minor).*
>
> *Then there are voice leadings, which are mappings from the notes of one chord to the notes of the other E.g. "take the C in a C major triad and move it down by semitone to the B." This voice leading can be written:*
>
> $$(C, E, G) \mapsto (B, E, G).$$
>
> *This distinction is constantly getting blurred by neo-Riemannian music theorists. But to really understand "neo-Riemannian chord progressions" you have to be quite clear about it.*
>
> *To form a generalized neo-Riemannian chord progression, start with an ordered pair of chords, say (C major, E minor). Then apply all the transpositions and inversions to this pairs, producing (D major, F# minor), (C minor, Ab major), etc. The result is a function that commutes with the isometries of the pitch class circle. As a result, it identifies pairs of chords that can be linked by exactly similar collections of voice leading motions.*
>
> *For example, I can transform C major to E minor by moving C down by semitone to B.*
>
> *Similarly, I can transform D major to F# minor by moving D down by semitone to C#.*
>
> *Similarly, I can transform C minor to Ab major by moving G up to Ab.*
>
> *This last voice leading,*
>
> $$(C, E^\flat, G) \mapsto (C, E^\flat, A^\flat)$$
>
> *is just an inversion (reflection) of the voice leading*
>
> $$(C, E, G)| \to (B, E, G).$$
>
> *As a result it moves one note up by semitone, rather than moving one note down by semitone.*
>
> *More generally: if you give me* any *voice leading between C major and E minor, I can give you an exactly analogous voice leading between D major and F# minor,*

> *or C minor and Ab major, etc. So "neo-Riemannian" progressions identify a class of* harmonic *progressions (functions between unordered collections of points on the circle) that are interesting from a* voice leading *perspective. (They identify pairs of chord progressions that can be linked by the same voice leadings, to within rotation and reflection.)*

You can learn more about this here:

19) Dmitri Tymoczko, "Scale theory, serial theory, and voice leading", available at `http://music.princeton.edu/~dmitri/scalesarrays.pdf`

I'd like to conclude tonight's performance with a "chromatic fantasy" — some wild ideas that you shouldn't take too seriously, at least as far as music theory goes. In this rousing finale, I'll list some famous subgroups of the permutations of a 12-element set. They may not be relevant to music, but I can't resist mentioning them and hoping somebody dreams up an application.

So far I've only mentioned two: the cyclic or "transposition" group, $\mathbb{Z}/12$, and the dihedral or "transposition/inversion" group with 24 elements. These are motivated by thinking of $\mathbb{Z}/12$ as a discrete analogue of a circle and considering either just its rotations, or rotations together with reflections. But, mathematically, it's nice to loosen up this rigid geometry and consider *projective* transformations of a circle, now viewed as a line together with a point at infinity — a "projective line".

Indeed, the group $\mathbb{Z}/11$ becomes a field with 11 elements if we multiply as well as add $\mod 11$. If we throw in a point at infinity, we get a projective line with 12 elements. It looks just like our circle of 12 notes. But now we see that the group $\mathrm{PGL}(2, \mathbb{Z}/11)$ acts on this projective line in a natural way. This group consists of invertible 2×2 matrices with entries in $\mathbb{Z}/11$, mod scalars. People call it $\mathrm{PGL}(2, 11)$ for short.

So, $\mathrm{PGL}(2, 11)$ acts on our 12-element set of notes. And, it's a general fact for any field $\mathbb{F}$ that $\mathrm{PGL}(2, \mathbb{F})$ acts on the corresponding projective line in a "triply transitive" way. In other words, given any ordered triple of distinct points on the projective line, we can find a group element that maps it to any *other* ordered triple of distinct points.

Even better, the action is "sharply" triply transitive, meaning there's *exactly one* group element that does the job!

This lets us count the elements in $\mathrm{PGL}(2, 11)$. Since we can find exactly one group element that maps our favorite ordered triple of distinct elements to any other, we just need to count such triples, and there are

$$12 \times 11 \times 10 = 1320$$

of them — so this is the size of $\mathrm{PGL}(2, 11)$.

This may be too much symmetry for music, since this group carries *any* three-note chord to any other, not just in the sense of chord progressions but in the sense of voice leadings. Still, it's cute.

We might go further and look for a quadruply transitive group of permutations of our 12-element set of notes — in other words, one that maps any ordered 4-tuple of distinct notes to any other.

But if we do, we'll run smack dab into MATHIEU GROUPS!

Here's an utterly staggering fact about reality. Apart from the group of *all* permutations of an n-element set and the group of *even* permutations of an n-element set, there are only FOUR groups of permutations that are k-tuply transitive for $k > 3$. Here they are:

- The Mathieu group M_{11}. This is a quadruply transitive group of permuations of an 11-element set — and sharply so! It has
$$11 \times 10 \times 9 \times 8 = 7920$$
elements.
- The Mathieu group M_{12}. This is a quintuply transitive group of permutations of a 12-element set — and sharply so! It has
$$12 \times 11 \times 10 \times 9 \times 8 = 95,040$$
elements.
- The Mathieu group M_{23}. This is a quadruply transitive group of permutations of a 23-element set — but not sharply so. It has
$$23 \times 22 \times 21 \times 20 \times 48 = 10,200,960$$
elements. As you can see, 48 group elements carry any distinct ordered 4-tuple to any other.
- The Mathieu group M_{24}. This is a quintuply transitive group of permutations of a 24-element set — but not sharply so. It has
$$24 \times 23 \times 22 \times 21 \times 20 \times 48 = 244,823,040$$
elements. As you can see, 48 group elements carry any distinct ordered 4-tuple to any other.

These groups all arise as symmetries of certain discrete geometries called Steiner systems. An "$S(L, M, N)$ Steiner system" is a set of N "points" together with a collection of "lines", such that each line contains M points, and *any* set of L points lies on a unique line. The symmetry group of a Steiner system consists of all permutations of the set of points that map lines to lines. It turns out that:

- There is a unique $S(5, 6, 12)$ Steiner system, and the Mathieu group M_{12} is its symmetry group. The stabilizer group of any point is isomorphic to M_{11}.
- There is a unique $S(5, 8, 24)$ Steiner system, and the Mathieu group M_{24} is its symmetry group. The stabilizer group of any point is isomorphic to M_{23}.

So, the group M_{12} could be related to music if there were a musically interesting way of taking the chromatic scale and choosing 6-note chords such that any 5 notes lie in a unique chord. I can't imagine such a way — most of these chords would need to be wretchedly dissonant. Another way to put the problem is that such a big group of permutations would impose more symmetry on the set of chords than I can imagine my ears hearing. It's like those grand unified theories that posit symmetries interchanging particles that look completely different. They could be true, but they've got their work cut out for them.

Luckily, the Mathieu groups appear naturally in other contexts - wherever the numbers 12 and 24 cast their magic spell over mathematics! For example, M_{24} is related to the 24-dimensional Leech lattice, and M_{12} can be nicely described in terms of 12 equal-sized balls rolling around the surface of another ball of the same size. See "Week 20" for more on this — and the book by Conway and Sloane cited there for even more.

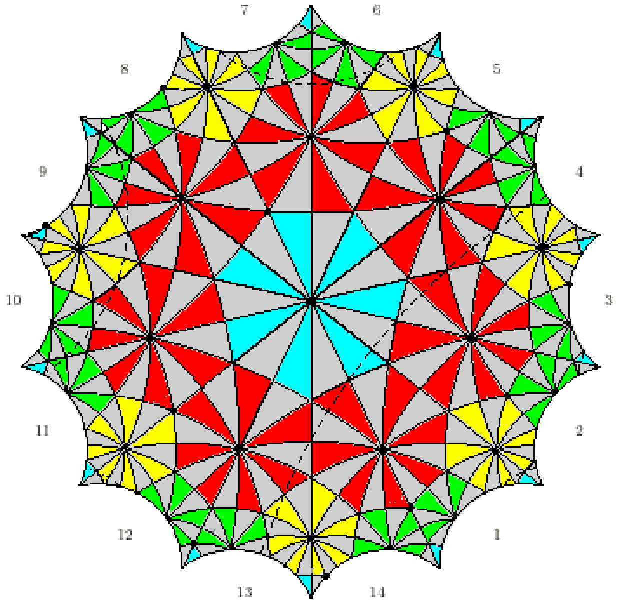
For a pretty explanation of M_{24}, also try this:

20) Steven H. Cullinane, Geometry of the 4×4 square, `http://finitegeometry.org/sc/16/geometry.html`

For explanations of both M_{24} and M_{12}, try this:

21) Peter J. Cameron, *Projective and Polar Spaces*, QMW Math Notes **13**, 1991. Also available at `http://www.maths.qmul.ac.uk/~pjc/pps/` Chapter 9: The geometry of the Mathieu groups, available at `http://www.maths.qmul.ac.uk/~pjc/pps/pps9.pdf`

It would be fun to dream up more relations between incidence geometry and music theory. Could Klein's quartic curve play a role? Remember from "Week 214", "Week 215" and "Week 219" that this 3-holed torus can be nicely tiled by 24 regular heptagons:



Its orientation-preserving symmetries form the group $\mathrm{PSL}(2,7)$, which consists of all $2x2$ matrices with determinant 1 having entries in $\mathbb{Z}/7$, modulo scalars. This group has

$24 \times 7 = 168$ elements. Since there are 7 notes in a major or minor scale, and 24 of these scales, it's hard to resist wanting to think of each heptagon as a scale!

Indeed, after I mentioned this idea to Dmitri Tymoczko, he said that David Lewin and Bob Peck have written about related topics.

Alas, the heptagonal tiling of Klein's quartic has a total of 56 vertices, not a multiple of 12, so there's no great way to think of the vertices as notes. But, it has $84 = 7 \times 12$ edges, so maybe the edges are labelled by notes and each note labels 7 edges.

Unlike some groups I mentioned earlier, $\mathrm{PSL}(2, 7)$ is not a transitive subgroup of the permutations of a 12-element set. And while $\mathrm{PSL}(2, 7)$ has lots of 12-element subgroups, these are not cyclic groups but instead copies of A_4. These facts put some further limitations on any crazy ideas you might try.

On the bright side, mathematically if not musically, there is a fascinating way to embed $\mathrm{PSL}(2, 7)$ into the Mathieu group M_{24}, which can be described by getting M_{24} to act on the set of 24 heptagons in the Klein quartic:

22) David Richter, "How to make the Mathieu group M_{24}", `http://homepages.wmich.edu/~drichter/mathieu.htm`

He works in the Poincar dual picture, where the Klein quartic is tiled by 56 triangles, but that's no big deal.

By the way, in "Week 79" I explained how $\mathrm{PSL}(2, \mathbb{F})$ acts on the projective line over the field $\mathbb{F}$; the same thing works for $\mathrm{PGL}(2, \mathbb{F})$. I also passed on some interesting facts mentioned by Bertram Kostant, which relate $\mathrm{PSL}(2, 5)$, $\mathrm{PSL}(2, 7)$ and $\mathrm{PSL}(2, 11)$ to the symmetry groups of the tetrahedron, cube/octahedron and dodecahedron/icosahedron. Kostant put these together to give a nice description of the buckyball!

Kepler would be pleased. But, he'd be happier if we could find the music of the spheres lurking in here, too.

Addenda: This week's issue provoked more discussion than any in recent history! You can read a lot on `sci.math.research`. Here are some comments from Dave Rusin, David Corfield, Mike Stay, Dmitri Tymoczko, Cris Moore, Robert Israel, Noam Elkies, Stephen Lavelle, and Steve Lubin.

Dave Rusin explained the logic behind having 12 notes in the chromatic scale. David Corfield mentioned a book on topos theory in music, and a paper by Noam Elkies on Mathieu groups. Mike Stay pointed out William Sethares' work on how the timbre of an instrument affects which scales sound good. Dmitri Tymoczko had more comments on this issue. Cris Moore mentioned an interesting microtonal composer named Easley Blackwood. Robert Israel pointed out an unusual fact about Riemann and Einstein. Noam Elkies explained what David Lewin was trying to do with $\mathrm{PSL}(2, 7)$ in music theory. And Stephen Lavelle gave some more references on torsors and topoi in music, and said more about the origin of the 12-note scale.

So, here we go! Dave Rusin wrote:

> *You wrote:*

> > *. . . there's an unavoidable conflict between the desire for simple ratios and the desire for evenly spaced notes, built into the fabric of mathematics and music. Every tuning system is thus a compromise. I would like to understand this better; there's bound to be a lot of nice number theory here.*
>
> *Sure there is. You want to choose a number N of intervals into which to divide the octave, so that there are two tones in the scale that, like C and G, have frequencies very nearly in a $3:2$ ratio. (This also gives a bonus pair like G and the next C up, which are then in a $4:3$ ratio.) But that just means you want $2^{\frac{n}{N}}$ to be nearly $3/2$, i.e. n/N is a good rational approximation to $\log_2(3/2)$. Use continued fractions or Farey sequences as you like. You'll find that a five-note octave is not a bad choice (roughly giving you just the black keys on a piano, and roughly corresponding to ancient Oriental musical sounds) but a 12-note octave is a really good choice. So it's not just happenstance that we have a firmly-entrenched system of 12-notes-per-octave. I'm sure you've seen this "7–12" magic before, e.g. the circle-of-fifths in music takes you through 7 octaves, or the simple arithmetic that $2^{19} \sim 3^{12}$ (i.e. $524288 \sim 531441$). Long ago I programmed an old PC to play a 41-tone scale because the next continued-fractions approximant calls for such a scale.*
>
> *Of course you could argue that music consists of more than just (musical) fourths and fifths and so the REAL number theory comes about by choosing numbers of tones which allow lots of sets of notes to be in (or nearly in) simple Pythagorean harmonies. How, exactly, you balance the conflicting goals is a matter of personal choice.*
>
> *What with the musicians in your family and all that, I'm guessing you probably knew all this already and simply withheld the comments because of space limitations, but just in case, I thought I'd complete your train of thought for you. This stuff is pretty classic and it's all over the web. I get more hits on my web page about this than any of my math pages!*
>
> *dave*

Here's Dave's web page:

23) Dave Rusin, "Mathematics and music", `http://www.math.niu.edu/~rusin/uses-math/music/`

David Corfield wrote:

> *Hi,*
>
> *Next you need to wade through all 1300 pages of The Topos of Music. This is "topos" in the category theoretic sense. Check out the table of contents!*
>
> 24) *Guerino Mazzola,* The Topos of Music: Geometric Logic of Concepts, Theory and Performance*, Birkhauser, Berlin, 2002. Preface and contents available at* `http://www.encyclospace.org/tom/tom_preface_toc.pdf`
>
> *Guerino Mazzola, homepage,* `http://www.ifi.unizh.ch/staff/mazzola/mazzola.html`

> *Colin McLarty reviewed it for MathSciNet. I like the part:*
>
> > *Symmetries within scores, and structural relations between scores, drive the mathematics up to sheaves, and very briefly to toposes and Grothendieck topologies. The author candidly states he is unsure whether this musicological perspective can use topos cohomology (p. 436).*
>
> *Did you ever hear about Conway's M_{13}?*
>
> 25) *John H. Conway, Noam D. Elkies, Jeremy L. Martin, "The Mathieu group M_{12} and its pseudogroup extension M_{13}", available as* math.GR/0508630*.*
>
> *I can't remember whether it was this that Alexander Borovik mentioned to me as a sign that the simple sporadic groups are just islands sticking up above the water.*
>
> *Best, David*

Needless to say, David doesn't write me emails with numbered references; I often polish the emails I get, with the permission of the authors, trying not to violate the spirit of the thing.

My student Mike Stay wrote:

> *Music really does sound better if the piano is tuned to the particular key, i.e. the Pythagorean intervals.*
>
> *Start with a frequency for C. At each step, multiply by 2 (up an octave) or by $2/3$ (down a fourth). Go down a fourth unless it will take you out of the octave; in the latter case, multiply by 2 first.*

```
*2    C'
*2/3  G
*2/3  D
*2    D'
*2/3  A
*2/3  E
*2    E'
*2/3  B
*2/3  F#
*2/3  C#
```

> *etc.*
>
> *Classical music was written for a particular key because the keys sounded different! Using the tuning above induces a "distance" on the keys–how in tune they are. Pieces would use the dissonant tunings of other keys for effect. My friend is an organist and piano tuner; he says that with the logarithmic tuning all keys sound "equally bad."*

> *But the timbre of the instrument — the harmonics and overtones — apparently have a great deal to do with whether a particular chord is consonant or not. This is a really cool paper that illustrates how to choose nearly any collection of frequencies as a scale and then come up with a timbre for which it sounds natural and right:*
>
> 26) *William Sethares, "Relating tuning and timbre",* `http://eceserv0.ece.wisc.edu/~sethares/consemi.html`
>
> *Sethares' home page has a bunch of MP3's on it for people who want to listen:*
>
> 27) *William Sethares, "MP3 Download Central",* `http://eceserv0.ece.wisc.edu/~sethares/otherperson/all_mp3s.html`
>
> *I like "Truth on a Bus", played in a 19-note scale.*
>
> *There are some tracks from the CD mentioned above here:*
>
> 28) *William Sethares,* Tuning, Timbre, Spectrum, Scale*, 2nd edition, Springer Verlag, Berlin, 2004. Author's guide available at* `http://eceserv0.ece.wisc.edu/~sethares/ttss.html`*. Sound examples available at* `http://eceserv0.ece.wisc.edu/~sethares/html/soundexamples.html`
>
> *The first several tracks play a tune on a typical* 12*-tone instrument. Then they change its timbre by adjusting the harmonics. Now if played in 12 divisions of a perfect octave (twice the frequency), it sounds perfectly awful; but if played in 12 divisions of* 2.1*, it sounds "right" again.*
>
> *Mike Stay*
> `http://math.ucr.edu/~mike`

In response to my comment "Every tuning system is thus a compromise. I would like to understand this better...", Dmitri Tymoczko wrote:

> *William Sethares' "Tuning, Timbre, Spectrum, Scale" is the best book about this. He has a convincing demonstration that "pure ratios" are not in themselves important: what's important is that the overtones of two simultaneously-sounding notes match. Since harmonic tones have partials that are integer multiples of the fundamental, you get pure ratios.*
>
> *However, for inharmonic tones, such as bell-sounds, the overtones are not integer multiples of the lowest tone. Hence, to get the partials to match you often need to use non-integer ratios. Sethares' book comes with a CD demonstrating this. It has to be heard to be believed.*

It seems that a bunch of the music on Sethares' CD is available online, as Mike pointed out above. I find most of his music interesting but unpleasant, not because of the tuning systems, but because it lacks soul. I haven't listened to "Truth on a Bus" yet.

Cris Moore wrote:

> *By the way, you should check out the music of Easley Blackwood. He wrote a series of microtonal etudes, where the number of tones per octave ranges from 13 to 24. Some of them (17, I think) are quite beautiful.*
>
> *Cris*

Here's an interesting syllabus and list of references that gives a feel for what mathematically sophisticated music theorists need to know these days:

29) John Rahn, "Music 575: Music and Mathematics", November 2004, syllabus available at `http://faculty.washington.edu/jrahn/5752004.htm`

Rahn makes some interesting comments on David Lewin's book Generalized Musical Intervals and Transformations, which defines a concept of "generalized interval system", or GIS. As far as I can tell without having read the book, a generalized interval system is a G-torsor for some group G, where quite likely we might wish to restrict G to be abelian or even cyclic. Thus, concretely, a generalized interval system is a set S of "pitch classes" on which some group G acts, and such that for any two elements s, s' in S there is a unique g in G with $gs = s'$. In this situation we say g is the "difference in pitch" between s and s'.

A subtle feature of G-torsors is that they are isomorphic to G, but not in a canonical way, because they don't have a god-given "identity element". I explain the importance of this in my webpage Torsors made easy. However, as in physics and mathematics, some people in music theory seem willing to ignore this subtlety and identity any G-torsor with G.

Rahn has the following comments on Lewin's book. I find them interesting because it shows music theorists grappling with ideas like category theory.

> *There are some problems in the formal ideas in this book, and extensions to them:*
>
> *GIS: Oren Kolman has recently shown (Kolman 2003) that every GIS can be rewritten as a group, so that all group theory applies directly ("transfers") to GIS. Among other things, this points up a possible flaw in the definition of GIS; a more intuitive definition would restrict a group of intervals to some cyclic group of one generator (my assertion). (See Kolman 2003.)*
>
> *Definitions in Ch 9: There is a problem here which prevents having more than one arrow-label between any two nodes. Lewin defines an arrow in his node-arrow def (p. 193) as an ordered pair of points, then maps ARROW into SGP, so each ordered pair of nodes has exactly one transformation in the semigroup that labels the arrow (one arrow). This probably originates in Lewin's work with groups of intervals, which are constrained to work this way. Of course in most groups, such as D_{24}, you need multiple arrows. There are various alternatives which would work for networks with multiple arrow(-labels) for a given ordered pair of nodes. Multiple arrows (or labels on an arrow, depending on the definitional system) in digraphs are standard, and it is hard to see what is accomplished by not allowing more than one relationship between any two nodes in the model. You also need multiple arrows for groups applied to graphs, category theory, etc.*

> *With this change, a Lewin network is formally a commutative diagram in some musical category — a directed graph with arrows labeled in a monoid, such that the composition of paths in the underlying category is associative and so on (definition of category and of commutative diagram.) Lewin says the labels are in a semigroup but his definition of node-arrow system makes every graph reflexive, providing the identities that augment a semigroup to a monoid.* So it is possible to use category theory to explore Lewin networks, much as GIS turned out to be groups: group theory transfers into GIS theory, and category theory transfers into Lewin network theory.
>
> *I made this connection in my paper, "The Swerve and the Flow: Music's Relation to Mathematics," delivered at IRCAM in October 2003 and subsequently published in PNM 42/1; I think I was the first to say this. I expanded on this idea in a talk at the ICMC, Miami, Nov 2 2004, called "Musical Acts"; in this talk I expanded into the relation of Lewin nets to the fundamental group of a topological space, and to homotopy classes, and adding category theory as a solution to part of a set of criteria for a general music theory. Later in this seminar I'll give a talk about all this.*

I don't know what a Lewin network is, except from the above. Unfortunately, Lewin's book is out of print. Lewin died in 2003, and Milton Babbitt said that a fair portion of his writing remains unpublished:

30) "Ken Gewertz, Composer, music theorist David Lewin dies at 69", *Harvard University Gazette*, `http://www.news.harvard.edu/gazette/2003/05.15/13-lewinobit.html`

Robert Israel pointed out an interesting connection between Einstein and Riemann:

> *John Baez wrote:*
>
> > *When Tom first mentioned "neo-Riemannian theory", I thought this was some bizarre application of differential geometry to music. But no — we're not talking about the 19th-century mathematician Bernhard Riemann, we're talking about the 19th-century music theorist Hugo Riemann!*
>
> *Interestingly, both Bernhard Riemann and Hugo Riemann had a connection with Einstein. But in Hugo's case, it was not Albert but his cousin Alfred, the musicologist, who edited Hugo Riemann's "Musik Lexicon"*
>
> *Robert Israel*
> *Department of Mathematics*
> *University of British Columbia*

Noam Elkies explained what David Lewin was doing with $\mathrm{PSL}(2,7)$:

> *John Baez wrote:*

> *Today I'd like to talk about the math of music — including torsors, orbifolds, and maybe even Mathieu groups. [. . .]*

Lots of interesting points there (though curiously none about Fourier analysis, which seem to be a natural connection in the context of mathematical physics). For now I'll just address Mathieu groups and the like. First, though, a trivial etymological point:

> *[. . .] in the Middle Ages, music was part of a "quadrivium" of mathematical arts: arithmetic, geometry, music, and astronomy. These were studied after the "trivium" of grammar, rhetoric and logic. This is why mathematicians scorn a result as "trivial" when it's easy to see using straightforward logic. When a result seems more profound, they should call it "quadrivial"!*
>
> *Try saying it sometime: "Cool! That's quadrivial!" It might catch on.*

I hope not; while the usual sense "commonplace, ordinary, unimportant" of "trivial" does come from a Latin word "trivium", it's not the academic trivium but a fork in the road — the meeting of "tres viae", three ways. Apparently it was common for people to make gossip and other trivial small talk at these three-way crossings. See for instance `http://www.m-w.com/dictionary/trivial`, which to its credit includes also the technical meaning

> *2b : relating to or being the mathematically simplest case; specifically : characterized by having all variables equal to zero*

— it would surely be too much to ask a general dictionary to mention identity elements or maps involving an arbitrary group. To its discredit, *the same dictionary glosses "trivium" as "crossroads", which should of course be where at least* four *ways meet, that is, a quadrivium! In fact the `www.m-w.com` entry for "quadrivium" also glosses that Latin word as "crossroads". . .*

[Given the TWFiMP context, perhaps this is the place to digress further and suggest that a Feynman diagram is a network of trivialities, in the original meaning of "3-way meetings" :-)]

Back to groups and music:

> *I'd like to conclude tonight's performance with a "chromatic fantasy" - some wild ideas that you shouldn't take too seriously, at least as far as music theory goes. In this rousing finale, I'll list some famous subgroups of the permutations of a 12-element set. They may not be relevant to music, but I can't resist mentioning them and hoping somebody dreams up an application. [. . .]*
>
> - *The Mathieu group M_{12}. This is a quintuply transitive group of permutations of a 12-element set — and sharply so! [. . .]*
>
> *These groups all arise as symmetries of certain discrete geometries called Steiner systems. An "$S(L, M, N)$ Steiner system" is a set of*

> *N "points" together with a collection of "lines", such that each line contains M points, and any set of L points lies on a unique line. The symmetry group of a Steiner system consists of all permutations of the set of points that map lines to lines. It turns out that:*
>
> - *There is a unique $S(5, 6, 12)$ Steiner system, and the Mathieu group M_{12} is its symmetry group. The stabilizer group of any point is isomorphic to M_{11}. [...]*
>
> *So, the group M_{12} could be related to music if there were a musically interesting way of taking the chromatic scale and choosing 6-note chords such that any 5 notes lie in a unique chord. I can't imagine such a way — most of these chords would need to be wretchedly dissonant.*

This from the man who finds Schoenberg's music so beautiful?... (Actually not all of Schoenberg's music is 12-tone, and some of it is not even atonal; perhaps you had heard only his tonal music and/or early explorations outside tonality such as the Op.19 piano pieces.)

> *Another way to put the problem is that such a big group of permutations would impose more symmetry on the set of chords than I can imagine my ears hearing.*

Curiously the real problem here is that M_{12}, big as it is, is missing *an important kind of symmetry: a 12-cycle. While one wouldn't need all the symmetries of the Steiner system to be evident in the music, it would be nice for any transposition of any Steiner chord to be again a Steiner chord. But, since M_{12} is a simple group, it contains only even permutations, whereas a 12-cycle is odd. It follows that M_{12} does not contain the triply-transitive group $PGL_2(\mathbb{Z}/11\mathbb{Z})$ – though it* does *contain the index-2 subgroup $PSL_2(\mathbb{Z}/11\mathbb{Z})$, in two different ways that are switched by an outer automorphism of M_{12}!*

While M_{12} does not contain any 12-cycles, it does contain double 6-cycles. We can choose one of them (doesn't matter which, because they're all conjugate in M_{12}), and then choose an action of M_{12} on $\mathbb{Z}/12\mathbb{Z}$ that makes our double 6-cycle act by translation by 2. This gives a Steiner $(5, 6, 12)$ system on $\mathbb{Z}/12\mathbb{Z}$ that's invariant under all even translations. Now there are $\mathrm{Binom}(12, 5)/\mathrm{Binom}(6, 5) = 132$ hexads in the system, in 66 complementary pairs, and it turns out that none of the hexads is taken to itself or its complement by a nontrivial element of M_{12}. So, under the standard identification of $\mathbb{Z}/12\mathbb{Z}$ with the chromatic pitch classes, we get 11 six-note chords that, together with their whole-tone transpositions and complements of whole-tone transpositions (all different), form a $(5, 6, 12)$ Steiner system and thus contain every 5-note chord in just one way.

One problem with this is that our musical training does not prepare us to distinguish whole-tone transpositions from non-whole-tone transpositions. So, let's form a double Steiner system by allowing arbitrary transpositions. This ruins most of the M_{12} symmetry (most of which wasn't audible anyway), but gives us *symmetry under the full group of musical transpositions, inversion, and complements. It so happens that this group still acts freely, so we get 11 six-note chords that, together with their transpositions and complements of transpositions, contain every 5-note chord in exactly two ways.*

Back in 1991 I figured out what these 11 chords are. Written as subsets of $\mathbb{Z}/12\mathbb{Z}$ *rather than of* $\{C,C\#,D,\dots\}$*, they are:*

$$\begin{gathered}\{0,1,2,3,4,6\},\quad \{0,1,2,3,5,7\},\quad \{0,1,2,3,6,7\},\\ \{0,1,2,4,5,8\},\quad \{0,1,3,4,6,9\},\quad \{0,1,3,5,7,9\},\\ \{0,1,2,4,5,9\},\quad \{0,1,2,4,7,8\},\quad \{0,1,2,5,6,8\},\\ \{0,1,2,4,7,9\},\quad \{0,1,3,5,6,8\}.\end{gathered}$$

Each five-note chord can be regarded as a link between two of these chords (not necessarily different ones; for instance, the two completions of a symmetrical chord such as $\{0,1,2,3,4\}$ *must be each others inversion, here* $\{0,1,2,3,4,6\}$ *and* $\{-2,0,1,2,3,4\}$*). The resulting graph is here, with some cryptic labeling to indicate things like whether one of the chords linked by an edge must be inverted to get a 5-note overlap:*

`http://www.math.harvard.edu/~elkies/m12.pdf`

> *It would be fun to dream up more relations between incidence geometry and music theory. Could Klein's quartic curve play a role? [. . .] Its orientation-preserving symmetries form the group* $\mathrm{PSL}(2,7)$*, which consists of all* $2x2$ *matrices with determinant* 1 *having entries in* $\mathbb{Z}/7$*, modulo scalars. This group has* $24\times 7 = 168$ *elements. Since there are 7 notes in a major or minor scale, and 24 of these scales, it's hard to resist wanting to think of each heptagon as a scale!*
>
> *Indeed, after I mentioned this idea to Dmitri Tymoczko, he said that David Lewin and Bob Peck have written about related topics.*

Right. In fact this group, call it G_{168}*, also has a doubly-transitive action on 7 objects, which includes a* 7*-cycle and thus gives a natural way for this group to act on a* 7*-note diatonic scale. David Lewin noted this possibility in one of his last papers.*

The key (no pun intended) is that this group G_{168} *is also isomorphic with the invertible* 3×3 *matrices over* $\mathbb{Z}/2\mathbb{Z}$ *— one of those remarkable isomorphisms between small matrix groups. So,* G_{168} *permutes the* $2^3-1=7$ *nonzero vectors in a* 3*-dimensional vector space* V *over* $\mathbb{Z}/2\mathbb{Z}$*. The existence of a* 7*-cycle is automatic because 7 is prime, though it is true in general that for any finite field* k *and any positive integer* n *the group of* $n\times n$ *matrices over* k *contains elements that cyclically permute the nonzero vectors of* k^n*. The reason is that* k *has a degree-*n *extension* k'*; any finite field has cyclic unit group, and multiplication by a generator is a cyclic permutation of the nonzero elements that acts linearly over* k'*, so a fortiori over k!*

But back to G_{168}*. This group also permutes the 7 codimension-*1 *subspaces (planes) of* V*, each of which contains three nonzero vectors. Any two distinct*

> *nonzero vectors are contained in a unique such plane, so we get a $(2, 3, 7)$ Steiner system, a.k.a. projective plane of order 2, whose group of symmetries is G_{168}.*
>
> *If we use a 7-cycle to identify the nonzero vectors of V with a diatonic scale, then the Steiner system gives us a distinguished collection of 7 three-note chords, which form a single orbit under translation (or "diatonic transposition" in music-theory lingo). It is well-known that these can be chosen to be either $\{n, n+1, n+3\}$ or $\{n, n-1, n-3\}$. Using the first one yields the chords*
>
> *{CDF}, {DEG}, {EFA}, {FGB}, {GAC}, {ABD}, {BCE},*
>
> *and either this or its inversions is what David Lewin suggested using.*
>
> > *Unlike some groups I mentioned earlier,* $\mathrm{PSL}(2, 7)$ *is not a subgroup of the permutations of a* 12*-element set.*
>
> *Not a transitive subgroup, you mean. Since the group acts on 7 objects, it can certainly act on 12 while keeping five of them unmoved. That's basically what we did above, with the five untouched objects being the "black keys" C#, D#, F#, G#, A#. Or, use the action on the 8 points of the projective line over $\mathbb{Z}/7\mathbb{Z}$ and leave four of the 12 untouched.*
>
> *–Noam D. Elkies*
>
> *P.S.*
>
> > *(If I weren't so darn honest, I'd add that Liszt wrote the "Transcendental Etudes" as a sequel to his popular "Algebraic Etudes", and explain how Mozart's "eine kleine Nachtmusik" tours a tone net shaped like a Klein bottle. But alas. . . .)*
>
> *Ha. Here's another kind of transcendental etude:*
>
> *http://math.harvard.edu/~elkies/stego.pdf*
>
> *which sounds like this:*
>
> *http://math.harvard.edu/~elkies/stego.mid*
>
> *(an automatic realization that has all the notes but makes no attempt at interpretations). I premiered the piece here last year on March 14; for more hints see*
>
> *http://math.harvard.edu/~elkies/stego.hints*
>
> *:-)*

I used Noam Elkies' remark to correct my claim that $\mathrm{PSL}(2, 7)$ isn't a subgroup of permutations of a 12-element set. Interestingly, the triples $(n, n+1, n+3)$ and $(n, n-1, n-3)$ in $\mathbb{Z}/7$ are also famous as two equivalent ways of defining the octonion multiplication tables! If we call the seven unit imaginary octonions A, B, C, D, E, F, and G, then we can define octonion multiplication using the chords Elkies lists by saying that

$$CD = F, \quad DE = G, \quad EF = A, \quad FG = B,$$
$$GA = C, \quad AB = D, \quad BC = E$$

along with knowing they anticommute and square to -1.

Unfortunately, these chords *don't* seem particularly fundamental to music! Alas, if only they were triads. . . .

Here are some more nice references and comments from Stephen Lavelle:

> *I may have been the first to say torsor, but there's a lot of stuff in existence about group actions as they apply to music theory. In particular, Fripertinger's page here:*
>
> *31) Harald Fripertinger, "Mathematical music theory", `http://www.uni-graz.at/~fripert/index_11.html`*
>
> *has a lot of pretty damned useful articles.*
>
> *There're also some pretty cool, and possibly musical, applications of proper Topos Theory to music by Mazzola's school — see, for instance:*
>
> *32) Thomas Noll,* The topos of triads*, available at `http://www.cs.tu-berlin.de/~noll/ToposOfTriads.pdf`*
>
> *John Baez wrote:*
>
>> *Apparently Riemann's ideas have caught on in a big way. Monzo says that "use of lattices is endemic on internet tuning lists", as if they were some sort of infectious disease.*
>>
>> *Dysart seems more gung-ho about it all. The "donuts" he mentions arise when you curl up tone nets by identifying notes that differ by an octave. He has some nice pictures of them!*
>
> *More general "nets", which look quite like commutative diagrams, are put into use by Lewin. Actually, they are commutative diagrams, and one can conceptually quite reasonably formulate a category of musically meaningful limits — see:*
>
> *33) Guerino Mazzola and Moreno Andreatta, "From a categorical point of view: K-nets as limit denotators", available at `recherche.ircam.fr/equipes/repmus/mamux/documents/mazzola-andreatta.pdf`*
>
>> *In neo-Riemannian theory, people like Lewin and Hyer started extending Riemann's ideas by using group theory to systematize operations on chords. The best easy introduction to this is Fiore's paper "Music and mathematics". Here you can read about math lurking in the music of Elvis and the Beatles! Or, if you're more of a highbrow sort, see what he has to say about Hindemith and Liszt's "Transcendental Etudes". And if you like doughnuts and music, you'll love the section where he explains how Beethoven's Ninth traces out a systematic path in a torus-shaped tone net! This amazing fact was discovered by Cohn, Douthett, and Steinbach.*

Mazzola has a nice argument as well (in his book "The Topos of Music"), where he shows an inconsistency in Riemann's logic of harmonic functions by showing that his construction would allow one to fix an orientation on a Moebius strip :) (if you take the seven notes in the major scale, and connect any three points with a triangle when they form a major, minor, or diminished chord, you get a Moebius strip — this was first come up with by Schoenberg).

> *(If I weren't so darn honest, I'd add that Liszt wrote the "Transcendental Etudes" as a sequel to his popular "Algebraic Etudes", and explain how Mozart's "eine kleine Nachtmusik" tours a tone net shaped like a Klein bottle. But alas. . . .)*

Encore! Encore!

> *C, C#, D, D#, E, F, F#, G, G#, A, A#, B*
> *until you're back to C. If you're a mathematician, you might be happier to call these notes*
>
> $$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11$$
>
> *and say that we're working in the group of integers* $\mod 12$*, otherwise known as* $\mathbb{Z}/12$*. Let's be mathematicians today.*

Here's a question for you, that you might want to know the answer to: Historically, *why are there twelve notes in the scale? And why are seven white and five black?*

The answer is that one that ties in lots of stuff about continued fractions, but goes along these lines: one is looking at the octave, and divides it up by looking at the first n *fifths (in our scale c,g,d,a,e,b,. . .) — this divides up the scale.*

Pythagoras et al. thought that one should try to keep the variety of intervals between consecutive notes as small as possible — in the end, deciding that the fewer different intervals present the better. Scales generated by fifths that have only two intervals present between side-by-side notes are called Pythagorean. None have just one interval, and the first three Pythagorean scales have 5,7, and 12 notes. 12 was thought pretty much enough, I'm guessing, and it can have nicely embedded into it the two smaller scales (as white and black notes).

I should have a reference for the original article where I read this (some Irish maths society bulletin I think), but I've said enough that the material should be findable online. Ah yes, here it is:

34) Maria Jose Garmendia Rodriguez and Juan Antonio Navarro Gonzalez, "Musical scales", IMS Bulletin ***35*** *(Christmas 1995), 24.*

for all the good it'll do you.

Steve Lubin wrote:

Dear John,

I just came across your charming page on the web about donuts and music and related matters, and I'd like to contribute a bit of info. Your statement:

> *And if you like doughnuts and music, you'll love the section where he [Fiore] explains how Beethoven's Ninth traces out a systematic path in a torus-shaped tone net! This amazing fact was discovered by Cohn, Douthett, and Steinbach.*

isn't quite accurate; or at least, doesn't do justice to priority of discovery. In 1974 I wrote a dissertation in musicology at NYU entitled Techniques for the Analysis of Development in Middle-Period Beethoven, in which I laid out four ideas:

> *The 18th-century European composers behaved as though they inhabited a communally shared abstract tonal space that existed independently of individual works and served as a substrate for them.*
>
> *This space evolved in the course of the century, with respect to its internal features and the ways in which it encouraged or discouraged composers' choices in navigating through it.*
>
> *A suitable graphic representation of this space could be used as a map for describing composers' itineraries in particular pieces; and these intineraries in themselves constitute an important contribution to the effect and beauty of their works (in coordination with other factors: rhythmic, textural, and so on).*
>
> *The tonal space inhabited by the late-18th-century composers had the form of a torus, for reasons arising from the internal geometry of the key-relationships they favored.*

Recently I've been gratified to discover that some neo-Riemannian theorists have been acknowledging my contribution in their publications: e.g., Cohn, Lerdahl and Gollin.

I wonder what Bernhard Riemann might have thought of this! My understanding of what his namesake Otto did is as follows: Otto described two-dimensional tone nets that extended out indefinitely in all directions like wallpaper -in this I believe he followed out earlier work of Gottfried Weber. The initial insight of taking the patterns on this flat surface and giving them closure by covering a torus with them was mine–I'm not aware of any prior precedents.

BTW, I'm a physics groupie, always starved for laypersons' accounts of the latest stuff. (The collapse of Scientific American as a classy journal was a catastrophe for me.) I've decided to set aside the rest of August for reading your diaries.
Thanks and best,
Steve
`http://www.stevenlubin.com/`

A guiding principle in modern mathematics is this lesson: Whenever you have to do with a structure-endowed entity S, try to determine its group of automorphisms, the group of those element-wise transformations which leave all structural relations undisturbed. You can expect to gain a deep insight into the constitution of S in this way.

— *Hermann Weyl*

Week 235

July 15, 2006

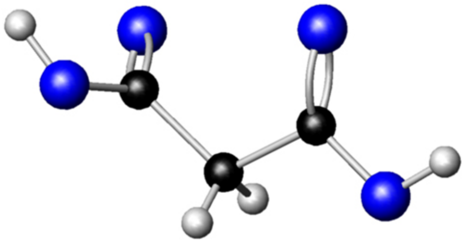
After leaving the Perimeter Institute near the end of June, I went home to Riverside and then took off for a summer in Shanghai. That's where I am now. I'm having a great time — you can read about it in my online diary!

Today I'll talk about classical and quantum computation, then quantum gravity, and finally a bit about higher gauge theory.

My interest in quantum computation was revived when Scott Aaronson invited me to this building near the Perimeter Institute:

1) Institute for Quantum Computing (IQC), `http://www.iqc.ca/`

Raymond Laflamme gave me a fun tour of the labs, especially the setup where he's using nuclear magnetic resonance to control the spins of three carbon-13 nuclei in a substance called malonic acid. Each molecule is its own little quantum computer:



This picture was drawn by Osama Moussa, a grad student working with Laflamme on this project. The black spheres are carbons, the whites are hydrogens and the blues are oxygens.

One of the banes of quantum computation is "decoherence", in which the quantum state of the computer interacts with its environment and becomes correlated with it, or "entangled", in a way that appears to "collapse its wavefunction" and ruin the calculation.

In general it's good to keep things cool if you don't want things to get messed up. Surprisingly, Laflamme said that for liquids, keeping them hot *reduces* the rate of decoherence: the different molecules zip around so fast they don't stay near each other long enough to affect each other much!

But malonic acid is a solid, so the main way to keep the molecules from disturbing each other is to keep them far apart on average. So, they dilute the malonic acid made using carbon-13 nuclei in a lot of ordinary malonic acid made using carbon-12. Since 12 is an even number, such carbon atoms have no unpaired spinning neutron in their nuclei. So, the ordinary malonic acid serves as an inert "shield" that keeps the active molecules well separated most of the time.

So, each active molecule acts like an isolated system, doing its own computation as they zap it with carefully timed pulses of microwaves and the three spinning nuclei interact. About a quadrillion of these molecules are doing their thing in parallel, mixed in with a bunch more made using carbon-12. For more details, see:

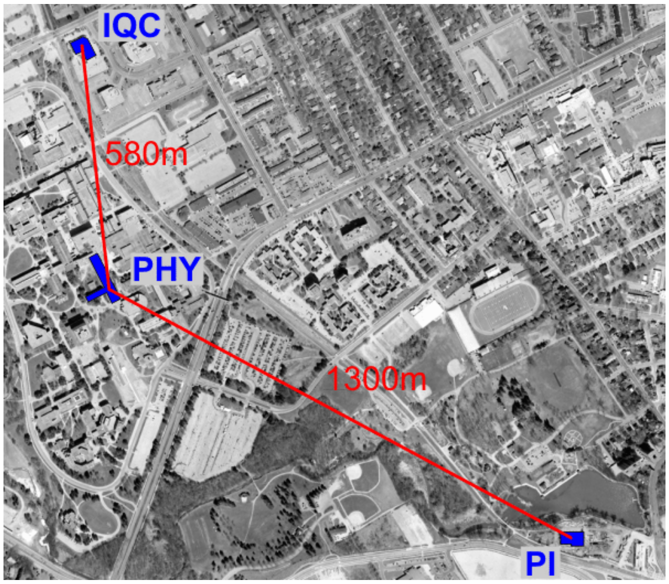
2) Jonathan Baugh, Osama Moussa, Colm A. Ryan, Raymond Laflamme, Chandrasekhar Ramanathan, Timothy F. Havel and David G. Cory, "Solid-state NMR three-qubit homonuclear system for quantum information processing: control and characterization", *Phys. Rev. A* **73** (2006), 022305. Also available as `quant-ph/0510115`.

Laflamme also showed me some beams of spin-entangled photons which they can use as keys for quantum cryptography. Nobody can peek at these photons without affecting them! It's a great scheme. If you don't know it, try this simple explanation:

3) Artur Ekert, "Cracking codes, part II", *Plus Magazine*, `http://pass.maths.org.uk/issue35/features/ekert/index.html`

There are already two companies — idQuantique and MagiQ — selling quantum key distribution systems that send entangled photons down optical fibers. But the folks at the IQC are planning to send them right through the air!

Eventually they want to send them from satellites down to the Earth. But as a warmup, they'll send beams of entangled photons from an intermediate building to the Institute of Quantum Computing and the Perimeter Institute.



4) IQC, "Free-space quantum key distribution", `http://www.iqc.ca/laboratories/peg/free_space.php`

Then they can share secrets with nobody able to spy on them unnoticed. They should do something to dramatize this capability! Unfortunately they don't actually *have* any secrets. So, they might need to make some up.

The really cool part, though, is that Scott helped me see that at least in principle, quantum computers could keep from drifting off course without the computation getting ruined by quantum entanglement with the environment. I had long been worried about this.

You see, to make any physical system keep acting "digital" for a long time, one needs a method to keep its time evolution from drifting off course. It's easiest to think about this issue for an old-fashioned, purely classical digital computer. It's already an interesting problem.

What does it mean for a physical system to act "digital"? Well, we like to idealize our computers as having a finite set of states; with each tick of the clock it jumps from one state to another in a deterministic way. That's how we imagine a digital computer.

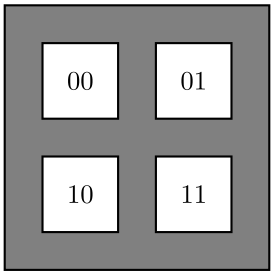
But if our computer is actually a machine following the laws of classical mechanics, its space of states is actually continuous — and time evolution is continuous too! Physicists call the space of states of a classical system its "phase space", and they describe time evolution by a "flow" on this phase space: states move continuously around as time passes, following Hamilton's equations.

So, what we like to idealize as a single state of our classical computer is actually a big bunch of states: a blob in phase space, or "macrostate" in physics jargon.

For example, in our idealized description, we might say a wire represents either a 0 or 1 depending on whether current is flowing through it or not. But in reality, there's a blob of states where only a little current is flowing through, and another blob of states where a lot is flowing through. All the former states count as the "0" macrostate in our idealized description; all the latter count as the "1" macrostate.

Unfortunately, there are also states right on the brink, where a medium amount of current is flowing through! If our machine gets into one of these states, it won't act like the perfect digital computer it's trying to mimic. This is bad!

So, you should imagine the phase space of our computer as having a finite set of blobs in it — macrostates where it's doing something good — separated by a no-man's land of states where it's not doing anything good. For a simple 2-bit computer, you can imagine 4 blobs like this:



though in reality the phase space won't be 2-dimensional, but instead much higher-dimensional.

Now, as time evolves for one tick of our computer's clock, we'd like these nice macrostates to flow into each other. Unfortunately, as they evolve, they sort of spread out. Their volume doesn't change — this was shown by Liouville back in the 1800s:

5) Wikipedia, "Liouville's theorem (Hamiltonian)", `http://en.wikipedia.org/wiki/Liouville's_theorem_(Hamiltonian)`

But, they get stretched in some directions and squashed in others. So, it seems hard for each one to get mapped completely into another, without their edges falling into the dangerous no-man's-land (shaded in gray in the diagram).

We want to keep our macrostates from getting hopelessly smeared out. It's a bit like herding a bunch of sheep that are drifting apart, getting them back into a tightly packed flock. Unfortunately, Liouville's theorem says you can't really "squeeze down" a flock of states! Volume in phase space is conserved. . . .

So, the trick is to squeeze our flock of states in some directions while letting them spread out in other, irrelevant directions.

The relevant directions say whether some bit in memory is a zero or one — or more generally, anything that affects our computation. The irrelevant ones say how the molecules in our computer are wiggling around. . . or the molecules of air *around* the computer — or anything that doesn't affect our computation.

So, for our computer to keep acting digital, it should pump out *heat!*

Here's a simpler example. Take a ball bearing and drop it into a wine glass. Regardless of its initial position and velocity — within reason - the ball winds up motionless at the bottom of the glass. Lots of different states seem to be converging to one state!

But this isn't really true. In fact, information about the ball's position and velocity has been converted into *heat*: irrelevant information about the motion of atoms.

In short: for a fundamentally analogue physical system to keep acting digital, it must dispose of irrelevant information, which amounts to pumping out waste heat.

In fact, Rolf Landauer showed back in 1961 that getting rid of one bit of information requires putting out this much energy in the form of heat:

$$kT\ln(2)$$

where T is the temperature and k is Boltzmann's constant. That's not much — about 3×10^{-21} joules at room temperature! But, it's theoretically important.

What had me worried was how this would work for quantum computation. A bunch of things are different, but some should be the same. When we pump information — i.e., waste heat — from the computer into the environment, we inevitably correlate its state with that of the environment.

In quantum mechanics, correlations often take the form of "entanglement". And this is a dangerous thing. For example, if our quantum computer is in a superposition of lots of states where it's doing interesting things, and we peek at it to see *which*, we get entangled with it, and its state seems to "collapse" down to one specific possibility. We say it "decoheres".

Won't the entanglement caused by pumping out waste heat screw up the coherence needed for quantum computation to work its wonders?

I finally realized the answer was: maybe not. Yes, the quantum state of the computer gets entangled with that of the environment — but maybe if one is clever, only the *irrelevant* aspects of its state will get entangled: aspects that don't affect the computation. After all, it's this irrelevant information that one is trying to pump out, not the relevant information.

So, maybe it can work. I need to catch up on what people have written about this, even though most of it speaks the language of "error correction" rather than thermodynamics. Here are some things, including material Scott Aaronson recommended to me.

Gentle introductions:

6) Michael A. Nielsen and Isaac L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, Cambridge, 2000.

7) John Preskill, Quantum computation — lecture notes, references etc. at `http://www.theory.caltech.edu/people/preskill/ph229/`

8) John Preskill, “Fault-tolerant quantum computation”, to appear in *Introduction to Quantum Computation*, eds. H.-K. Lo, S. Popescu, and T. P. Spiller. Also available as `quant-ph/9712048`.

Chapter 7 of Preskill’s lecture notes is about error correction.

This is a nice early paper on getting quantum computers to work despite some inaccuracy and decoherence:

9) Peter Shor, “Fault-tolerant quantum computation”, *37th Symposium on Foundations of Computing*, IEEE Computer Society Press, 1996, pp. 56–65. Also available as `quant-ph/9605011`.

This more recent paper shows that in a certain model, quantum computation can be made robust against errors that occur at less than some constant rate:

10) Dorit Aharonov and Michael Ben-Or, “Fault-tolerant quantum computation with constant error rate”, available as `quant-ph/9906129`.

Here’s a paper that assumes a more general model:

11) Barbara M. Terhal and Guido Burkard, “Fault-tolerant quantum computation for local non-markovian noise”, *Phys. Rev. A* **71**, 012336 (2005). Also available as `quant-ph/0402104`.

Rolf Landauer was a physicist at IBM, and he discovered the result mentioned above — the “thermodynamic cost of forgetting” — in a study of Maxwell’s demon. This is a fascinating and controversial subject, and you can learn more about it in this book of reprints:

12) H. S. Leff and Andrew F. Rex, editors, *Maxwell’s Demon: Entropy, Information and Computing*, Institute of Physics Publishing, 1990.

I think Landauer’s original paper is in here. He figured out why you can’t get free energy from heat by using a little demon to watch the molecules and open a door to let the hot ones into a little box. The reason is that it takes energy for the demon to forget what it’s seen!

Finally, on a somewhat different note, if you just want a great read on the interface between physics and computation, you’ve got to try this:

13) Scott Aaronson, “NP-complete problems and physical reality”, *ACM SIGACT News*, March 2005. Also available as `quant-ph/0502072`.

Can a soap film efficiently solve the traveling salesman problem by minimizing its area? If quantum mechanics were slightly nonlinear, could quantum computers solve NP problems in polynomial time? And what could quantum *gravity* computers do? Read and learn and the state of the art on puzzles like these.

At the Perimeter Institute I also had some great discussions with Laurent Freidel and his student Aristide Baratin. They have a new spin foam model that reproduces ordinary quantum field theory — in other words, particle physics in flat spacetime. It's not interesting as a model of quantum gravity — it doesn't include gravity! Instead, it serves as a convenient target for spin foam models that *do* include gravity: it should be the limit of any such model as the gravitational constant approaches zero.

14) Aristide Baratin and Laurent Freidel, "Hidden quantum gravity in 4d Feynman diagrams: emergence of spin foams". Available as `hep-th/0611042`.

It's the sequel of this paper for 3d spacetime:

15) Aristide Baratin and Laurent Freidel, "Hidden quantum gravity in 3d Feynman diagrams". Available as `gr-qc/0604016`.

Freidel, Kowalski-Glikman and Starodubtsev have also just come out with a paper carrying out some of the exciting project I mentioned in "Week 208":

16) Laurent Freidel, Jerzy Kowalski-Glikman and Artem Starodubtsev, "Particles as Wilson lines in the gravitational field", available as `gr-qc/0607014`.

Their work is based on the MacDowell-Mansouri formulation of gravity. This is a gauge theory with gauge group $\mathrm{SO}(4,1)$ — the symmetry group of deSitter spacetime. DeSitter spacetime is a lot like Minkowski spacetime, but it has constant curvature instead of being flat. It's really just a hyperboloid in 5 dimensions:

$$\{(w,x,y,z,t) \mid w^2+x^2+y^2+z^2-t^2=k^2\}$$

for some constant k. It describes an exponentially expanding universe, a lot like ours today. It's the most symmetrical solution of Einstein's equation with a positive cosmological constant. The cosmological constant is proportional to $1/k^2$.

When you let the cosmological constant approach zero, which is the same as letting $k \to \infty$, DeSitter spacetime flattens out to Minkowski spacetime, and the group $\mathrm{SO}(4,1)$ contracts to the symmetry group of Minkowski spacetime: the Poincare group.

So, MacDowell-Mansouri gravity is similar to the formulation of gravity as gauge theory with the Poincare group as gauge group. I explained that pretty carefully back in "Week 176".

But, there's one way $\mathrm{SO}(4,1)$ is better than the Poincare group. It's a "simple" Lie group, so it has an inner product on its Lie algebra that's invariant under conjugation. This lets us write down the BF Lagrangian:

$$\mathrm{tr}(B \wedge F)$$

where tr is defined using the inner product, F is the curvature of an $\mathrm{SO}(4,1)$ connection A, and B is an $\mathfrak{so}(4,1)$-valued 2-form. Spin foam models of BF theory work really well:

17) John Baez, "An introduction to spin foam models of BF theory and quantum gravity", in *Geometry and Quantum Physics*, eds. Helmut Gausterer and Harald Grosse, Lecture Notes in Physics, Springer-Verlag, Berlin, 2000, pp. 25–93. Also available as gr-qc/9905087.

So, the MacDowell-Mansouri approach is a natural for spin foam models. It's not that MacDowell-Mansouri gravity *is* a BF theory — but its Lagrangian is the BF Lagrangian plus extra terms. So, we can think of it as a perturbed version of BF theory.

There's also one way $\mathrm{SO}(4,1)$ is worse than the Poincare group. It's a simple Lie group — so it doesn't have a god-given "translation" subgroup the way the Poincare group does. The Poincare gauge theory formulation of general relativity requires that we treat translations differently from boosts and rotations. We can't do this in an $\mathrm{SO}(4,1)$ gauge theory unless we break the symmetry down to a smaller group: the Lorentz group, $\mathrm{SO}(3,1)$.

So, to get MacDowell-Mansouri gravity from $\mathrm{SO}(4,1)$ BF theory, we need to add extra terms to the Lagrangian that break the symmetry group down to $\mathrm{SO}(3,1)$. This isn't bad, just a bit sneaky.

The new paper by Freidel, Kowalski-Glikman and Starodubtsev is mainly about the $\mathrm{SO}(4,1)$ BF theory rather than full-fledged MacDowell-Mansouri gravity. They show that if you cut out curves in spacetime and couple them to the A field in the right way, they act like the worldlines of point particles. In particular, they have a mass and spin, and they trace out geodesics when their spin is zero. Spinning particles do something a bit fancier, but it's the right thing.

This generalizes some results for 3d gravity that I explained in detail back in "Week 232". It's nice to see it working in 4 dimensions too.

Back then I also explained something else about 4d BF theory: if you cut out *surfaces* in spacetime and couple them to the B field, they act like the worldsheets of 1-dimensional extended objects, which one might call *strings*. I don't think they're the wiggling stretchy strings that string theorists like; I think their equation of motion is different. But I should actually check! It's stupid; I should have checked this a long time ago.

Ahem. Anyway, it's really neat how particles couple to the A field and "strings" couple to the B field in BF theory.

This is vaguely reminiscent of how the A and B field form two parts of a "2-connection" — a gadget that lets you define parallel transport along curved and surfaces. You can read about 2-connections here:

18) John Baez and Urs Schreiber, "Higher gauge theory", to appear in the volume honoring Ross Street's 60th birthday, available as math.DG/0511710.

The cool thing is that a pair consisting of an A field and a B field gives well-behaved parallel transport for curves and surfaces only if they satisfy an equation. . . which is *implied* by the basic equation of BF theory!

The above paper is a summary of results without proofs. Before one can talk about 2-connections, one needs to understand 2-bundles, which are a "categorified" sort of bundle where the fiber is not a smooth manifold but a smooth category. My student

Toby Bartels recently finished writing an excellent thesis that defines 2-bundles and relates them to "gerbes" — another popular approach to higher gauge theory, based on categorifying the concept of "sheaf" instead of "bundle":

19) Toby Bartels, "Higher Gauge Theory I: 2-bundles", available as math.CT/0410328.

The detailed study of 2-connections will show up in the next installment — a paper I'm writing with Urs Schreiber.

You can also see transparencies of some talks about this stuff:

20) John Baez, Alissa Crans and Danny Stevenson, "Chicago lectures on higher gauge theory", available at http://math.ucr.edu/home/baez/namboodiri/

21) John Baez, "Higher gauge theory", 2006 Barrett lectures, available at http://math.ucr.edu/home/baez/barrett/

It'll be lots of fun if higher gauge theory and the work relating MacDowell-Mansouri gravity to BF theory fit together and develop in some nontrivial direction. But the funny thing is, I don't how they fit together yet.

Here's why. In gauge theory, there's a famous way to get a number from a connection A and a loop. First you take the "holonomy" of A around the loop, and then you take the trace (in some representation of your gauge group) to get a number. This number is called a "Wilson loop".

This is an obvious way to define an *action* for a particle coupled to a connection A — at least if the particle moves around a loop. For example, it's this action that let us compute knot invariants from BF theory: you use the BF action for your fields, you use the Wilson loop as an action for your particle, and you compute the amplitude for your particle to trace out some knot in spacetime.

One might guess from the title "Particles as Wilson lines in the gravitational field" that this is the action Freidel and company use. But it's not!

Instead, they use a different action, which involves extra fields on the particle's worldline, describing its position and momentum. I explained a close relative of this action back in "Week 232", when I was coupling particles to 3d gravity.

The same funny difference shows up when we couple strings to the B field. In higher gauge theory you can define holonomies and Wilson loops using the A field, but you can also define "2-holonomies" and "Wilson surfaces" using both the A and B fields. The 2-holonomy describes how a string changes as it moves along a surface, just as the holonomy describes how a particle changes as it moves along a curve. If you have a closed surface you can take a "trace" of the 2-holonomy and get a number, which deserves to be called a "Wilson surface".

This is an obvious way to define an action for a string coupled to the A and B fields — at least if it traces out a closed surface. But, it's not the one Perez and I use! Why not? Because we were trying to do something analogous to what people did for particles in 3d gravity.

So, there's some relation between this "particles and strings coupled to 4d BF theory" business and the mathematics of higher gauge theory, but it's not the obvious one you might have guessed at first.

Mysteries breed mysteries. For more musings on these topics, try my talk at the Perimeter Institute:

22) John Baez, "Higher-dimensional algebra: a language for quantum spacetime", colloquium talk at Perimeter Institute, available at `http://math.ucr.edu/home/baez/quantum_spacetime/`

Addenda: I thank Chris Weed for catching some errors, Osama Moussa for letting me display his picture and for catching some more errors, and Ben Rubiak-Gould, Nathan Urban and K. Eric Drexler for some interesting comments. Here's my reply to Ben Rudiak-Gould:

> *John Baez wrote:*
>
> > *In fact, Rolf Landauer showed back in 1961 that getting rid of one bit of information requires putting out this much energy in the form of heat:*
> >
> > $$kT\ln(2)$$
>
> *Ben Rubiak-Gould replied:*
>
> > *It's easy to understand where this formula comes from. Getting rid of a bit means emitting one bit of entropy, which is $k\ln 2$ in conventional units. The associated quantity of heat is $ST = kT\ln 2$.*
>
> *Thanks; I should have said that.*
>
> *Landauer's analysis showing that "forgetting information" costs energy is still interesting, and it was surprising at the time. There had been a number of other analyses of why Maxwell's demon can't get you something for nothing, by Szilard and others, but none (I think) had focussed on the key importance of resetting the demon's memory to its initial state.*
>
> > *But it seems to me that you're conflating two different issues here. One is the cost of forgetting a bit, which only affects irreversible computation, and the other is the cost of keeping the computation on track, which affects reversible computation also. Landauer's formula tells you the former, but I don't think there's any lower bound on the latter.*
>
> *Not in principle: with a perfectly tuned dynamics, an analogue system can act perfectly digital, since each macrostate gets mapped perfectly into another one with each click of the clock. But with imperfect dynamics, dissipation is needed to squeeze each macrostate down enough so it can get mapped into the next — and the dissipation makes the dynamics irreversible, so we have to pay a thermodynamic cost.*
>
> *If I were smarter I could prove an inequality relating the "imperfection of the dynamics" (how to quantify that?) to the thermodynamic cost of computation, piggybacking off Landauer's formula.*

Here's what Nathan Urban wrote:

> *John Baez wrote:*
>
> > *[quantum computation] So, maybe it can work. I need to catch up on what people have written about this, even though most of it speaks the language of "error correction" rather than thermodynamics.*
>
> *A nice recent overview of some of this work can be found in the latest* Physics Today*:*
>
> *23) Sarma, Freedman, and Nayak, "Topological quantum computation",* Physics Today *(July 2006).*
>
> *In this approach, error-free computation is accomplished using topological quantum field theories, as topological theories are robust against local perturbations.*
>
> *The article has some nice discussion of anyons, braidings, non-Abelian topological phases of condensed matter systems, etc. It speculates that the $\nu = 12/5$ state of the fractional quantum Hall effect might support universal topological quantum computation (meaning that its braiding operators could realize any desired unitary transformation).*

Here's my reply:

> *Long time no see, Nathan!*
>
> *Nathan Urban wrote:*
>
> > *John Baez wrote:*
> >
> > > *[quantum computation] So, maybe it can work. I need to catch up on what people have written about this, even though most of it speaks the language of "error correction" rather than thermodynamics.*
> >
> > *A nice recent overview of some of this work can be found in the latest Physics Today (July 2006), in the article "Topological quantum computation" by Sarma, Freedman, and Nayak. (Nayak is at UCLA if you ever get out that way.)*
>
> *Thanks, I'll check that out.*
>
> *I'm usually too lazy to drive into LA, but now that I'm in Shanghai, I thought I'd take the chance to visit Zhenghan Wang in the nearby city of Hangzhou and talk to him about topological quantum computation.*
>
> *Wang and Freedman both work for "Project Q", aptly named after the Star Trek villain — it's Microsoft's project to develop quantum computers using non-abelian anyons:*
>
> *24) Topological quantum computing at Indiana University,* `http://www.tqc.iu.edu/`

> > *The article has some nice discussion of anyons, braidings, non-Abelian topological phases of condensed matter systems, etc. It speculates that the $\nu = 12/5$ state of the fractional quantum Hall effect might support universal topological quantum computation (meaning that its braiding operators could realize any desired unitary transformation).*
>
> *Freedman, Larsen and Wang have already proved that certain versions of Chern-Simons theory support universal quantum computation:*
>
> *25) Michael Freedman, Michael Larsen, and Zhenghan Wang, "A modular functor which is universal for quantum computation", available as quant-ph/0001108.*
>
> *The fractional quantum Hall effect is supposedly described by Chern-Simons theory, so this is relevant. I don't know anything about the "$\nu = 12/5$ state" of the fractional quantum Hall effect, but the folks at Project Q* do *want to use the fractional quantum Hall effect for quantum computation, and some people are looking for nonabelian anyons in the $\nu = 5/2$ state:*
>
> *26) Parsa Bonderson, Alexei Kitaev and Kirill Shtengel, "Detecting non-abelian statistics in the $\nu = 5/2$ fractional quantum Hall state",* Phys. Rev. Lett. ***96** (2006) 016803. Also available as cond-mat/0508616.*
>
> *Apparently there's just one lab in the world that has the capability of producing these fractional quantum Hall states!*

The article in the latest Physics Today isn't free for nonsubscribers, but this is, and it seems to cover similar ground:

27) Charles Day, "Devices based on the fractional quantum Hall effect may fulfill the promise of quantum computing", *Physics Today* (October 2005), also available at `http://www.physicstoday.org/vol-58/iss-10/p21.html`

It discusses both the $\nu = 12/5$ and $\nu = 5/2$ states.

Alas, I never managed to visit Zhenghan Wang in Hangzhou.

K. Eric Drexler writes:

> *Dear John,*
>
> *To continue a thread in Week 235:*
>
> *John Baez wrote:*
>
> > *[. . .] with a perfectly tuned dynamics, an analogue system can act perfectly digital, since each macrostate gets mapped perfectly into another one with each click of the clock. But with imperfect dynamics, dissipation is needed to squeeze each macrostate down enough so it can get mapped into the next — and the dissipation makes the dynamics irreversible, so we have to pay a thermodynamic cost.*

> *Logically reversible computation can, in fact, be kept on track without expending energy and without accurately tuned dynamics. A logically reversible computation can be embodied in a constraint system resembling a puzzle with sliding, interlocking pieces, in which all configurations accessible from a given input state correspond to admissible states of the computation along an oriented path to the output configuration. The computation is kept on track by the contact forces that constrain the motion of the sliding pieces. The computational state is then like a ball rolling along a deep trough; an error would correspond to the ball jumping out of the trough, but the energy barrier can be made high enough to make the error rate negligible. Bounded sideways motion (that is, motion in computationally irrelevant degrees of freedom) is acceptable and inevitable.*
>
> *Keeping a computation of this sort on track clearly requires no energy expenditure, but moving the computational state in a preferred direction (forward!) is another matter. This requires a driving force, and in physically realistic systems, this force will be resisted by a "friction" caused by imperfections in dynamics that couple motion along the progress coordinate to motion in other, computationally irrelevant degrees of freedom. In a broad class of physically realistic systems, this friction scales like viscous drag: the magnitude of the mean force is proportional to speed, hence energy dissipation per distance travelled (equivalently, dissipation per logic operation) approaches zero as the speed approaches zero.*
>
> *Thus, the thermodynamic cost of keeping a classical computation free of errors can be zero, and the thermodynamic cost per operation of a logically reversible computation can approach zero. Only Landauer's* $\ln(2)kT$ *cost of bit erasure is unavoidable, and the number of bits erased is a measure of how far a computation deviates from logical reversibility. These results are well-known from the literature, and are important in understanding what can be done with atomically-precise systems.*
>
> *With best wishes,*
>
> *Eric*

For an introduction to Drexler's plans for atomically-precise reversible computers, see:

28) K. Eric Drexler, *Nanosystems: Molecular Machinery, Manufacturing, and Computation*, John Wiley and Sons, New York, 1992.

The issue of heat dissipation in such devices is also studied here:

29) Ralph C. Merkle, "Two types of mechanical reversible logic", *Nanotechnology* **4** (1993), 114–131. Also available at `http://www.zyvex.com/nanotech/mechano.html`

I need to think about this stuff more!

Week 236

July 26, 2006

This week I'd like to catch you up on some papers about categorification and quantum mechanics.

But first, since it's summer vacation, I'd like to take you on a little road trip — to infinity. And then, for fun, a little detective story about the history of the icosahedron.

Cantor invented two kinds of infinities: cardinals and ordinals. Cardinals are more familiar. They say how big sets are. Two sets can be put into 1-1 correspondence iff they have the same number of elements — where this kind of "number" is a cardinal.

But today I want to talk about ordinals. Ordinals say how big "well-ordered" sets are. A set is well-ordered if it's linearly ordered and every nonempty subset has a smallest element.

For example, the empty set

$$\{\}$$

is well-ordered in a trivial sort of way, and the corresponding ordinal is called

$$0.$$

Similarly, any set with just one element, like this:

$$\{0\}$$

is well-ordered in a trivial sort of way, and the corresponding ordinal is called

$$1.$$

Similarly, any set with two elements, like this:

$$\{0, 1\}$$

becomes well-ordered as soon as we decree which element is bigger; the obvious choice is to say $0 < 1$. The corresponding ordinal is called

$$2.$$

Similarly, any set with three elements, like this:

$$\{0, 1, 2\}$$

becomes well-ordered as soon as we linearly order it; the obvious choice here is to say $0 < 1 < 2$. The corresponding ordinal is called

$$3.$$

Perhaps you're getting the pattern — you've probably seen these particular ordinals before, maybe sometime in grade school. They're called finite ordinals, or "natural numbers".

But there's a cute trick they probably didn't teach you then: we can *define* each ordinal to *be* the set of all ordinals less than it:

- $0 = \{\}$ (since no ordinal is less than 0)
- $1 = \{0\}$ (since only 0 is less than 1)
- $2 = \{0, 1\}$ (since 0 and 1 are less than 2)
- $3 = \{0, 1, 2\}$ (since 0, 1 and 2 are less than 3)

and so on. It's nice because now each ordinal *is* a well-ordered set of the size that ordinal stands for. And, we can define one ordinal to be "less than or equal" to another precisely when its a subset of the other.

Now, what comes after all the finite ordinals? Well, the set of all finite ordinals is itself well-ordered:

$$\{0, 1, 2, 3, \ldots\}$$

So, there's an ordinal corresponding to this — and it's the first *infinite* ordinal. It's usually called omega (ω). Using the cute trick I mentioned, we can actually define

$$\omega = 0, 1, 2, 3, \ldots$$

Now, what comes after this? Well, it turns out there's a well-ordered set

$$\{0, 1, 2, 3, \ldots, \omega\}$$

containing the finite ordinals together with ω, with the obvious notion of "less than": ω is bigger than the rest. Corresponding to this set there's an ordinal called

$$\omega + 1$$

As usual, we can simply define

$$\omega + 1 = \{0, 1, 2, 3, \ldots, \omega\}$$

(At this point you could be confused if you know about cardinals, so let me throw in a word of reassurance. The sets ω and $\omega + 1$ have the same "cardinality", but they're different as ordinals, since you can't find a 1-1 and onto function between them that *preserves the ordering*. This is easy to see, since $\omega + 1$ has a biggest element while ω does not.)

Now, what comes next? Well, not surprisingly, it's

$$\omega + 2 = \{0, 1, 2, 3, \ldots, \omega, \omega + 1\}$$

Then comes

$$\omega + 3, \omega + 4, \omega + 5, \ldots$$

and so on. You get the idea.

What next?

Well, the ordinal after all these is called $\omega + \omega$. People often call it "ω times 2" or "$\omega 2$" for short. So,

$$\omega 2 = \{0, 1, 2, 3, \ldots, \omega, \omega + 1, \omega + 2, \omega + 3, \ldots\}$$

What next? Well, then comes

$$\omega 2+1, \omega 2+2, \ldots$$

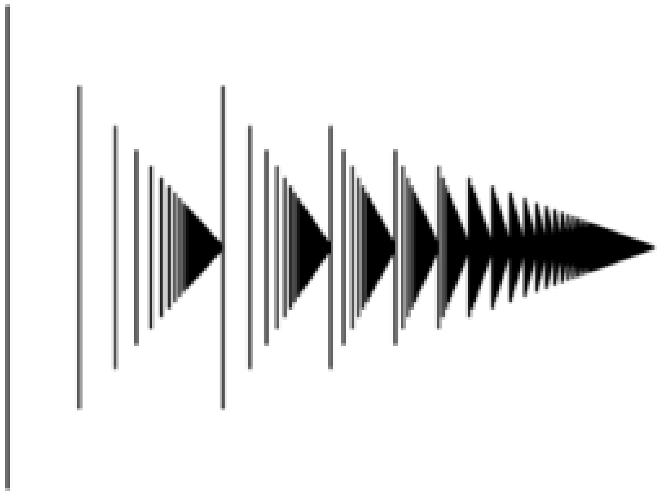
and so on. But you probably have the hang of this already, so we can skip right ahead to $\omega 3$.

In fact, you're probably ready to skip right ahead to $\omega 4$, and $\omega 5$, and so on.

In fact, I bet now you're ready to skip all the way to "ω times ω", or ω^2 for short:

$$\omega^2=\{0,1,2, \ldots \omega, \omega+1, \omega+2, \ldots, \omega 2, \omega 2+1, \omega 2+2, \ldots\}$$

It would be fun to have a book with ω pages, each page half as thick as the previous page. You can tell a nice long story with an ω-sized book. But it would be even more fun to have an encyclopedia with ω volumes, each being an ω-sized book, each half as thick as the previous volume. Then you have ω^2 pages — and it can still fit in one bookshelf!



What comes next? Well, we have

$$\omega^2+1, \omega^2+2, \ldots$$

and so on, and after all these come

$$\omega^2+\omega, \omega^2+\omega+1, \omega^2+\omega+2, \ldots$$

and so on — and eventually

$$\omega^2+\omega^2=\omega^2 2$$

and then a bunch more, and then

$$\omega^2 3$$

and then a bunch more, and then

$$\omega^2 4$$

and then a bunch more, and more, and eventually

$$\omega^2 \omega=\omega^3 .$$

You can probably imagine a bookcase containing ω encyclopedias, each with ω volumes, each with ω pages, for a total of ω^3 pages.

I'm skipping more and more steps to keep you from getting bored. I know you have plenty to do and can't spend an *infinite* amount of time reading This Week's Finds, even if the subject is infinity.

So, if you don't mind me just mentioning some of the high points, there are guys like ω^4 and ω^5 and so on, and after all these comes

$$\omega^\omega.$$

And then what?

Well, then comes $\omega^\omega + 1$, and so on, but I'm sure that's boring by now. And then come ordinals like

$$\omega^\omega 2, \ldots, \omega^\omega 3, \ldots, \omega^\omega 4, \ldots$$

leading up to

$$\omega^\omega \omega = \omega^{\omega+1}$$

Then eventually come ordinals like

$$\omega^\omega \omega^2, \ldots, \omega^\omega \omega^3, \ldots, \omega^\omega \omega^4, \ldots$$

and so on, leading up to:

$$\omega^\omega \omega^\omega = \omega^{\omega+\omega} = \omega^{\omega 2}$$

This actually reminds me of something that happened driving across South Dakota one summer with a friend of mine. We were in college, so we had the summer off, so we drive across the country. We drove across South Dakota all the way from the eastern border to the west on Interstate 90.

This state is huge — about 600 kilometers across, and most of it is really flat, so the drive was really boring. We kept seeing signs for a bunch of tourist attractions on the western edge of the state, like the Badlands and Mt. Rushmore — a mountain that they carved to look like faces of presidents, just to give people some reason to keep driving.

Anyway, I'll tell you the rest of the story later — I see some more ordinals coming up:

$$\omega^{\omega 3}, \ldots, \omega^{\omega 4}, \ldots, \omega^{\omega 5}, \ldots$$

We're really whizzing along now just to keep from getting bored — just like my friend and I did in South Dakota. You might fondly imagine that we had fun trading stories and jokes, like they do in road movies. But we were driving all the way from Princeton to my friend Chip's cabin in California. By the time we got to South Dakota, we were all out of stories and jokes.

Hey, look! It's

$$\omega^{\omega\omega} = \omega^{\omega^2}$$

That was cool. Then comes

$$\omega^{\omega^3}, \ldots \omega^{\omega^4}, \ldots \omega^{\omega^5}, \ldots$$

and so on.

Anyway, back to my story. For the first half of our half of our trip across the state, we kept seeing signs for something called the South Dakota Tractor Museum.

Oh, wait, here's an interesting ordinal — let's slow down and take a look:

$$\omega^{\omega^\omega}$$

I like that! Okay, let's keep driving:

$$\omega^{\omega^\omega} + 1, \omega^{\omega^\omega} + 2, \ldots$$

and then

$$\omega^{\omega^\omega} + \omega, \ldots, \omega^{\omega^\omega} + \omega 2, \ldots, \omega^{\omega^\omega} + 3, \ldots$$

and then

$$\omega^{\omega^\omega} + \omega^2, \ldots, \omega^{\omega^\omega} + \omega^3, \ldots$$

and eventually

$$\omega^{\omega^\omega} + \omega^\omega$$

and eventually

$$\omega^{\omega^\omega} + \omega^{\omega^\omega} = \omega^{\omega^\omega} 2$$

and then

$$\omega^{\omega^\omega} 3, \ldots, \omega^{\omega^\omega} 4, \ldots, \omega^{\omega^\omega} 5, \ldots$$

and eventually

$$\omega^{\omega^\omega} \omega = \omega^{\omega^\omega + 1}$$

and then

$$\omega^{\omega^\omega + 2}, \ldots, \omega^{\omega^\omega + 3}, \ldots, \omega^{\omega^\omega + 4}, \ldots$$

This is pretty boring; we're already going infinitely fast, but we're still just picking up speed, and it'll take a while before we reach something interesting.

Anyway, we started getting really curious about this South Dakota Tractor Museum — it sounded sort of funny. It took 250 kilometers of driving before we passed it. We wouldn't normally care about a tractor museum, but there was really nothing else to think about while we were driving. The only thing to see were fields of grain, and these signs, which kept building up the suspense, saying things like "ONLY 100 MILES TO THE SOUTH DAKOTA TRACTOR MUSEUM!"

We're zipping along really fast now:

$$\omega^{\omega^{\omega^\omega}}, \ldots, \omega^{\omega^{\omega^{\omega^\omega}}}, \ldots, \omega^{\omega^{\omega^{\omega^{\omega^\omega}}}}, \ldots$$

What comes after all these?

At this point we need to stop for gas. Our notation for ordinals runs out at this point!

The ordinals don't stop; it's just our notation that gives out. The set of all ordinals listed up to now — including all the ones we zipped past — is a well-ordered set called

$$\varepsilon_0$$

or "epsilon-nought". This has the amazing property that

$$\varepsilon_0 = \omega^{\varepsilon_0}$$

And, it's the smallest ordinal with this property.

In fact, all the ordinals smaller than ε_0 can be drawn as trees. You write them in "Cantor normal form" like this:

$$\omega^{\omega^{\omega^\omega}+\omega+1} + \omega^{\omega^\omega+\omega^\omega} + \omega^\omega + \omega + \omega + 1 + 1 + 1$$

using just $+$ and exponentials and 1 and ω, and then you turn this notation into a picture of a tree. I'll leave it as a puzzle to figure out how.

So, the set of (finite, rooted) trees becomes a well-ordered set whose ordinal is ε_0. Trees are important in combinatorics and computer science, so ε_0 is not really so weird after all.

Another cool thing is that Gentzen proved the consistency of the usual axioms for arithmetic — "Peano arithmetic" — with the help of ε_0. He did this by drawing proofs as trees, and using this to give an inductive argument that there's no proof in Peano arithmetic that $0 = 1$. But, this inductive argument goes beyond the simple kind you use to prove facts about all natural numbers. It uses induction up to ε_0.

You can't formalize Gentzen's argument in Peano arithmetic: thanks to Gdel, this system can't proof itself consistent unless it's *not*. I used to think this made Gentzen's proof pointless, especially since "induction up to ε_0" sounded like some sort of insane logician's extrapolation of ordinary mathematical induction.

But now I see that induction up to ε_0 can be thought of as induction on trees, and it seems like an obviously correct principle. Of course Peano's axioms also seem obviously correct, so I don't know that Gentzen's proof makes me *more sure* Peano arithmetic is consistent. But, it's interesting.

Induction up to ε_0 also lets you prove other stuff you can't prove with just Peano arithmetic. For example, it lets you prove that every Goodstein sequence eventually reaches zero!

Huh?

To write down a Goodstein sequence, you start with any natural number and write it in "recursive base 2", like this:

$$2^{2^2+1} + 2^1$$

Then you replace all the 2's by 3's:

$$3^{3^3+1} + 3^1$$

Then you subtract 1 and write the answer in "recursive base 3":

$$3^{3^3+1} + 1 + 1$$

Then you replace all the 3's by 4's, subtract 1 and write the answer in recursive base 4. Then you replace all the 4's by 5's, subtract 1 and write the answer in recursive base 5. And so on.

You can try some examples using the applet on this site:

1) National Curve Bank, "Goodstein's theorem", `http://curvebank.calstatela.edu/goodstein/goodstein.htm`

You'll see that for any starting number bigger than 3, the Goodstein sequence seems to keep growing forever. So, it's shocking to learn that they all eventually go to zero!

For example, if you start with the number 4, you get this sequence:

$$4, 26, 41, 60, 83, 109, 139, 173, 211, 253, 299, 348, \ldots$$

Believe it or not, this goes to zero after about 7×10^{121210694} steps! For a proof, see Kevin Buzzard's calculations in the Addendum at the bottom of this article.

In fact, it takes nothing but algebra, careful reasoning and persistence to work out when any given Goodstein sequence reaches zero. But, proving they *all* eventually reach zero uses induction up to ε_0.

How can that be?

The point is that these numbers in "recursive base n" look a lot like ordinals in Cantor normal form. If we translate them into ordinals by replacing n by ω, the ordinals keep getting smaller at each step, even when the numbers get bigger!

For example, when we do the translation

$$2^{2^2+1} + 2^1 \mapsto \omega^{\omega^\omega+1} + \omega^1$$
$$3^{3^3+1} + 1 + 1 \mapsto \omega^{\omega^\omega+1} + 1 + 1$$

we see the ordinal got smaller even though the number got bigger. Since ε_0 is well-ordered, the ordinals must bottom out at zero after a finite number of steps — that's what "induction up to ε_0" tells us. So, the numbers must too!

In short, induction up to ε_0 lets us prove all Goodstein sequences eventually reach zero. But Kirby and Paris showed something much deeper: they showed that you *need* induction up to ε_0 to get the job done.

And, they noted a big consequence of this fact. Suppose Peano arithmetic is consistent. Then you can't do induction up to ε_0 in this system — since if you could, Gentzen's work would let this system prove itself consistent, and Gdel's theorem would kick in and say it's *not*. But since you *need* induction up to ε_0 to prove all Goodstein sequences go to zero, you can't prove this in Peano arithmetic.

So, the fact that Goodstein sequences go to zero is obvious if you think about it the right way, but it's not provable in Peano arithmetic.

I don't know any results in mathematical physics that use induction up to ε_0, but these could be one — after all, trees show up in the theory of Feynman diagrams. That would be pretty interesting.

There's a lot more to say about this, but I hear what you're asking: what comes after ε_0?

Well, duh! It's

$$\varepsilon_0 + 1$$

Then comes

$$\varepsilon_0 + 2$$

and then eventually we get to

$$\varepsilon_0 + \omega$$

and then

$$\varepsilon_0 + \omega^2, \ldots, \varepsilon_0 + \omega^3, \ldots, \varepsilon_0 + \omega^4, \ldots$$

and after a long time

$$\varepsilon_0 + \varepsilon_0 = \varepsilon_0 2$$

and then eventually

$$\varepsilon_0^2$$

and then eventually..

Oh, I see! You want to know the first *really interesting* ordinal after ε_0.

Well, this is a matter of taste, but you might be interested in ε_1. This is the first ordinal after ε_0 that satisfies this equation:

$$x = \omega^x$$

How do we actually reach this ordinal? Well, just as ε_0 was the limit of this sequence:

$$\omega, \omega^\omega, \omega^{\omega^\omega}, \omega^{\omega^{\omega^\omega}}, \ldots$$

ε_1 is the limit of this:

$$\varepsilon_0 + 1, \omega^{\varepsilon_0+1}, \omega^{\omega^{\varepsilon_0+1}}, \omega^{\omega^{\omega^{\varepsilon_0+1}}}, \ldots$$

In other words, it's the *union* of all these well-ordered sets.

In what sense is ε_1 the "first really interesting ordinal" after ε_0? I'm not sure! Maybe it's the first one that can't be built out of 1, ω and ε_0 using finitely many additions, multiplications and exponentiations. Does anyone out there know?

Anyway, the next really interesting ordinal I know after ε_1 is ε_2. It's the next solution of

$$x = \omega^x$$

and it's defined to be the limit of this sequence:

$$\varepsilon_1 + 1, \omega^{\varepsilon_1+1}, \omega^{\omega^{\varepsilon_1+1}}, \omega^{\omega^{\omega^{\varepsilon_1+1}}}, \ldots$$

Maybe now you get the pattern. In general, ε_α is the αth solution of

$$x = \omega^x$$

and we can define this, if we're smart, for any ordinal α.

So, we can keep driving on through fields of ever larger ordinals:

$$\varepsilon_2, \ldots, \varepsilon_3, \ldots, \varepsilon_4, \ldots$$

and eventually

$$\varepsilon_\omega, \ldots, \varepsilon_{\omega+1}, \ldots, \varepsilon_{\omega+2}, \ldots$$

and eventually

$$\varepsilon_{\omega^2}, \ldots, \varepsilon_{\omega^3}, \ldots, \varepsilon_{\omega^4}, \ldots$$

and eventually

$$\varepsilon_{\omega^\omega}, \ldots, \varepsilon_{\omega^{\omega^\omega}}, \ldots$$

As you can see, this gets boring after a while — it's suspiciously similar to the beginning of our trip through the ordinals, with them now showing up as subscripts under this

"ε" notation. But this is misleading: we're moving much faster now. I'm skipping over much bigger gaps, not bothering to mention all sorts of ordinals like

$$\varepsilon_{\omega^\omega} + \varepsilon_{\omega 248} + \omega^{\omega^{\omega+17}} + 1$$

Anyway... so finally we *got* to this South Dakota Tractor Museum, driving pretty darn fast at this point, about 85 miles an hour... and guess what?

Oh — wait a minute — it's sort of interesting here:

$$\varepsilon_{\varepsilon_0}, \ldots, \varepsilon_{\varepsilon_1}, \ldots, \varepsilon_{\varepsilon_2}, \ldots$$

and now we reach

$$\varepsilon_{\varepsilon_\omega}$$

and then

$$\varepsilon_{\varepsilon_{\omega^\omega}}, \ldots, \varepsilon + \varepsilon_{\omega^{\omega^\omega}}, \ldots$$

and then as we keep speeding up, we see:

$$\varepsilon_{\varepsilon_{\varepsilon_0}}, \ldots \varepsilon_{\varepsilon_{\varepsilon_{\varepsilon_0}}}, \ldots \varepsilon_{\varepsilon_{\varepsilon_{\varepsilon_{\varepsilon_0}}}}, \ldots$$

So, anyway: by the time we got that tractor museum, we were driving really fast. And, all we saw as we whizzed by was a bunch of rusty tractors out in a field! It was over in a split second! It was a real anticlimax — just like this little anecdote, in fact.

But that's the way it is when you're driving through these ordinals. Every ordinal, no matter how large, looks pretty pathetic and small compared to the ones ahead — so you keep speeding up, looking for a really big one... and when you find one, you see it's part of a new pattern, and that gets boring to...

Anyway, when we reach the limit of this sequence:

$$\varepsilon_0, \varepsilon_{\varepsilon_0}, \varepsilon_{\varepsilon_{\varepsilon_0}}, \varepsilon_{\varepsilon_{\varepsilon_{\varepsilon_0}}}, \varepsilon_{\varepsilon\varepsilon_{\varepsilon_{\varepsilon_0}}}, \ldots$$

our notation breaks down, since this is the first solution of

$$x = \varepsilon_x$$

We could make up a new name for this ordinal, like η_0.

Then we could play the whole game again, defining η_α to be the αth solution of

$$x = \varepsilon_x$$

sort of like how we defined the epsilons. This kind of equation, where something equals some function of itself, is called a "fixed point" equation.

But since we'll have to play this game infinitely often, we might as well be more systematic about it!

As you can see, we keep running into new, qualitatively different types of ordinals. First we ran into the powers of ω, then we ran into the epsilons, and now these etas. It's gonna keep happening! For each type of ordinal, our notation runs out when we reach the first "fixed point" — when the xth ordinal of this type is actually equal to x.

So, instead of making up infinitely many Greek letters, let's use φ_γ for the γth type of ordinal, and $\varphi_\gamma(\alpha)$ for the αth ordinal of type γ.

We can use the fixed point equation to define $\varphi_{\gamma+1}$ in terms of φ_γ. In other words, we start off by defining

$$\varphi_0(\alpha) = \omega^\alpha$$

and then define

$$\varphi_{\gamma+1}(\alpha)$$

to be the αth solution of

$$x = \varphi_\gamma(x)$$

We can even define this stuff when γ itself is infinite. For a more precise definition see the Wikipedia article cited below... but I hope you get the rough idea.

This defines a lot of really big ordinals, called the "Veblen hierarchy".

There's a souped-up version of Cantor normal form that can handle every ordinal that's a finite sum of guys in the Veblen hierarchy: you can write them *uniquely* as finite sums of the form

$$\varphi_{\gamma_1}(\alpha_1) + \ldots + \varphi_{\gamma_k}(\alpha_k)$$

where each term is less than or equal to the previous one, and each α_i is not a fixed point of φ_{γ_i}

But as you might have suspected, not *all* ordinals can be written in this way. For one thing, every ordinal we've reached so far is *countable*: as a set you can put it in one-to-one correspondence with the integers. There are much bigger *uncountable* ordinals — at least if you believe you can well-order uncountable sets.

But even in the realm of the countable, we're nowhere near done!

As I hope you see, the power of the human mind to see a pattern and formalize it gives the quest for large countable ordinals a strange quality. As soon as we see a systematic way to generate a sequence of larger and larger ordinals, we know this sequence has a limit that's larger then all of those! And this opens the door to even larger ones...

So, this whole journey feels a bit like trying to outrace our car's own shadow — the faster we drive, the faster it chases after us. But, it's interesting to hear what happens next. At this point we reach something a bit like the Badlands on the western edge of South Dakota — something a bit spooky!

It's called the Feferman-Schtte ordinal, Γ_0. This is just the limit, or union if you prefer, of all the ordinals mentioned so far: all the ones you can get from the Veblen hierarchy. You can also define Γ_0 by a fixed point property: it's the smallest ordinal x with

$$\varphi_x(0) = x$$

Now, we've already seen that induction up to different ordinals gives us different amounts of mathematical power: induction up to ω is just ordinary mathematical induction as formalized by Peano arithmetic, but induction up to ε_0 buys us more — it lets us prove the consistency of Peano arithmetic!

Logicians including Feferman and Schtte have carried out a detailed analysis of this subject. They know a lot about how much induction up to different ordinals buys you. And apparently, induction up to Γ_0 lets us prove the consistency of a system called "predicative analysis". I don't understand this, nor do I understand the claim I've seen that Γ_0 is the first ordinal that cannot be defined predicatively — i.e., can't be defined without reference to itself. Sure, saying Γ_0 is the first solution of

$$\varphi_x(0) = x$$

is non-predicative. But what about saying that Γ_0 is the union of all ordinals in the Veblen hierarchy? What's non-predicative about that?

If anyone could explain this in simple terms, I'd be much obliged.

As you can see, I'm getting out my depth here. That's pretty typical in This Week's Finds, but this time — just to shock the world — I'll take it as a cue to shut up. So, I won't try to explain the outrageously large Bachmann-Howard ordinal, or the even more outrageously large Church-Kleene ordinal — the first one that can't be written down using *any* computable system of notation. You'll just have to read the references.

For a fun introduction to the infinite, try

2) Rudy Rucker, *Infinity and the Mind: The Science and Philosophy of the Infinite*, Princeton University Press, Princeton, 2004.

Rucker also wrote a novel, *White Light*, about a mathematician who actually goes to the transfinite realms and climbs some transfinite mountains.

For online self-study, I urge you to start by reading the Wikipedia article on ordinal numbers, then the article on ordinal arithmetic, and then the one on large countable ordinals — they're really well-written:

3) Wikipedia, "Ordinal numbers", `http://en.wikipedia.org/wiki/Ordinal_number`

"Ordinal arithmetic", `http://en.wikipedia.org/wiki/Ordinal_arithmetic`

"Large countable ordinals", `http://en.wikipedia.org/wiki/Large_countable_ordinals`

The last one has a tempting bibliography, but warns us that most books on this subject are hard to read and out of print. Apparently nobody can agree on notation for ordinals beyond the Veblen hierarchy, either.

Gentzen proved the consistency of Peano arithmetic in 1936:

4) Gerhard Gentzen, "Die Widerspruchfreiheit der reinen Zahlentheorie", *Mathematische Annalen* **112** (1936), 493–565. Translated as "The consistency of arithmetic" in M. E. Szabo ed., *The Collected Works of Gerhard Gentzen*, North-Holland, Amsterdam, 1969.

Goodstein's theorem came shortly afterwards:

5) R. Goodstein, "On the restricted ordinal theorem", *Journal of Symbolic Logic* **9** (1944), 33–41.

but Kirby and Paris proved it independent of Peano arithmetic only in 1982:

6) L. Kirby and J. Paris, "Accessible independence results for Peano arithmetic", *Bull. London. Math. Soc.* **14** (1982), 285–93.

That marvelous guy Alan Turing wrote his PhD thesis at Princeton under the logician Alonzo Church. It was about ordinals and their relation to logic:

7) Alan M. Turing, "Systems of logic defined by ordinals", *Proc. London Math. Soc.*, Series 2, **45** (1939), 161–228.

This is regarded as his most difficult paper. The idea is to take a system of logic like Peano arithmetic and throw in an extra axiom saying that system is consistent, and then another axiom saying *that* system is consistent, and so on ad infinitum — getting a new system for each ordinal. These systems are recursively axiomatizable up to (but not including) the Church-Kleene ordinal.

These ideas were later developed much further. . .

But, reading original articles is not so easy, especially if you're in Shanghai without access to a library. So, what about online stuff — especially stuff for the amateur, like me?

Well, this article is great fun if you're looking for a readable overview of the grand early days of proof theory, when Hilbert was battling Brouwer, and then Gdel came and blew everyone away:

8) Jeremy Avigad and Erich H. Reck, '"Clarifying the nature of the infinite": the development of metamathematics and proof theory', Carnegie-Mellon Technical Report CMU-PHIL-120, 2001. Also available as `http://www.andrew.cmu.edu/user/avigad/Papers/infinite.pdf`

But, it doesn't say much about the newer stuff, like the idea that induction up to a given ordinal can prove the consistency of a logical system — the bigger the ordinal, the stronger the system. For work up to 1960, this is a good overview:

9) Solomon Feferman, "Highlights in proof theory", in *Proof Theory*, eds. V. F. Hendricks et al, Kluwer, Dordrecht (2000), pp. 11–31. Also available at `http://math.stanford.edu/~feferman/papers.html`

For newer stuff, try this:

10) Solomon Feferman, "Proof theory since 1960", prepared for the *Encyclopedia of Philosophy Supplement*, Macmillan Publishing Co., New York. Also available at `http://math.stanford.edu/~feferman/papers.html`

Also try the stuff on proof theory, trees and categories mentioned in "Week 227", and this book mentioned in "Week 70" — now you can get it free online:

11) Jean-Yves Girard, Y. Lafont and P. Taylor, *Proofs and Types*, Cambridge Tracts in Theoretical Computer Science **7**, Cambridge U. Press, 1989. Also available at `http://www.cs.man.ac.uk/\~pt/stable/Proofs+Types.html`

Finally, sometime I want to get ahold of this book by someone who always enlivened logic discussions on the internet until his death in April this year:

12) Torkel Franzen, *Inexhaustibility: A Non-Exhaustive Treatment*, Lecture Notes in Logic **16**, A. K. Peters, Ltd., 2004.

The blurb sounds nice: "The inexhaustibility of mathematical knowledge is treated based on the concept of transfinite progressions of theories as conceived by Turing and Feferman."

Okay, now for a bit about the icosahedron — my favorite Platonic solid.

I've been thinking about the "geometric McKay correspondence" lately, and among other things this sets up a nice relationship between the symmetry group of the icosahedron and an amazing entity called E_8. E_8 is the largest of the exceptional Lie groups — it's 248-dimensional. It's related to the octonions (the number "8" is no coincidence) and it shows up in string theory. It's very beautiful how this complicated sounding stuff can be seen in distilled form in the icosahedron.

I have a lot to say about this, but you're probably worn out by our road trip through the land of big ordinals. So for now, try "Week 164" and "Week 230" if you're curious. Let's talk about something less stressful — the early history of the icosahedron.

I spoke about the early history of the dodecahedron in "Week 63". It's conjectured that the Greeks got interested in this shape from looking at crystals of iron pyrite. These aren't regular dodecahedra, since normal crystals can't have 5-fold symmetry — though "quasicrystals" can. Instead, they're "pyritohedra". The Greeks' love of mathematical perfection led them to the regular dodecahedron. . .

. . . and it also led them to invent the icosahedron:

13) Benno Artmann, "About the cover: the mathematical conquest of the third dimension", *Bulletin of the AMS* **43** (2006), 231–235. Also available at `http://www.ams.org/bull/2006-43-02/S0273-0979-06-01111-6/`

According to Artmann, an ancient note written in the margins of a copy of Euclid's Elements says the regular icosahedron and octahedron were discovered by Theaetetus!

If you're a cultured sort, you may know Theaetetus through Plato's dialog of the same name, where he's described as a mathematical genius. He's also mentioned in Plato's "The Sophist". He probably discovered the icosahedron between 380 and 370 BC, and died at an early age in 369. Euclid wrote his construction of the icosahedron that we find in Euclid's Elements:

14) Euclid, *Elements, Book XIII*, Proposition 16, online version due to David Joyce at `http://aleph0.clarku.edu/~djoyce/java/elements/bookXIII/propXIII16.html`

Artmann says this was the first time a geometrical entity appeared in pure thought before it was seen! An interesting thought.

Book XIII also contains a complete classification of the Platonic solids — perhaps the first really interesting classification theorem in mathematics, and certainly the first "ADE classification":

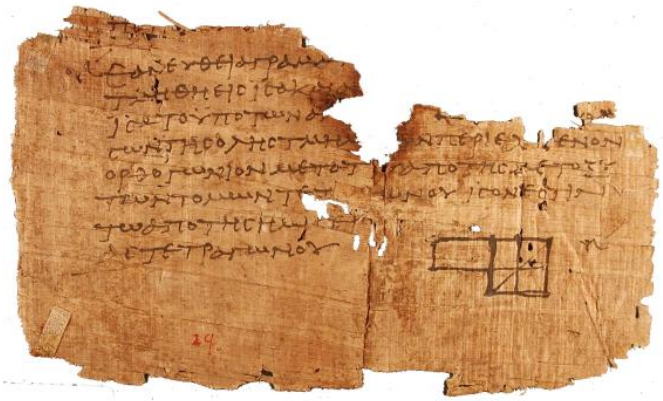
15) Euclid, *Elements, Book XIII*, Proposition 18, online version due to David Joyce at `http://aleph0.clarku.edu/~djoyce/java/elements/bookXIII/propXIII18.html`

If you don't know about ADE classifications, see "Week 62".

I got curious about the "ancient note written in the margins of a copy of Euclid" that Artmann mentions. It seemed too good to be true. Just for fun, I tried to track down the facts about this, using only my web browser here in Shanghai.

First of all, if you're imagining an old book in a library somewhere with marginal notes scribbled by a pal of Theaetetus, dream on. It ain't that simple! Our knowledge of Euclid's original Elements relies on copies of copies of copies. . . and centuries of detective work, with each detective having to root through obscure journals and dim-lit library basements to learn what the previous detectives did.

The oldest traces of Euclid's Elements are pathetic fragments of papyrus. People found some in a library roasted by the eruption of Mount Vesuvius in 79 AD, some more in a garbage dump in the Egyptian town of Oxyrhynchus (see "Week 221"), and a couple more in the Fayum region near the Nile. All these were written centuries after Euclid died. Here's one from Oxyrhynchus, photographed by Bill Casselman, who kindly let me include this picture:



What math is being discussed here? What's that red "29" doing there? For the answer to these and other questions, check out his website!

16) Bill Casselman, "One of the oldest extant diagrams from Euclid", http://www.math.ubc.ca/~cass/Euclid/papyrus/

The oldest nearly complete copy of the Elements lurks in a museum called the Bodleian at Oxford. It dates back to 888 AD, about a millennium after Euclid.

More copies date back to the 10th century; you can find their stories here:

17) Thomas L. Heath, editor, *Euclid's Elements*, chap. V: the text, Cambridge U. Press, Cambridge, 1925. Also available at http://www.perseus.tufts.edu/cgi-bin/ptext?lookup=Euc.+5

18) Menso Folkerts, "Euclid's Elements in Medieval Europe", http://www.math.ubc.ca/~cass/Euclid/folkerts/folkerts.html

All these copies are somewhat different. So, getting at Euclid's original Elements is as hard as sequencing the genome of Neanderthal man, seeing a quark, or peering back to the Big Bang!

A lot of these copies contain "scholia": comments inserted by various usually unnamed copyists. These were collected and classified by a scholar named Heiberg in the late 1800s:

19) Thomas L. Heath, editor, *Euclid's Elements*, chap. VI: the scholia, Cambridge U. Press, Cambridge, 1925. Also available at http://www.perseus.tufts.edu/cgi-bin/ptext?lookup=Euc.+6

One or more copies contains a scholium about Platonic solids in book XIII. Which copies? Ah, for that I'll have to read Heiberg's book when I get back to UC Riverside — our library has it, I'm proud to say.

And, it turns out that another scholar named Hultsch argued that this scholium was written by Geminus of Rhodes.

Geminus of Rhodes was an astronomer and mathematician who may have lived between 130 and 60 BC. He seems like a cool dude. In his Introduction to Astronomy, he broke open the "celestial sphere", writing:

> *... we must not suppose that all the stars lie on one surface, but rather that some of them are higher and some are lower.*

And in his Theory of Mathematics, he proved a classification theorem stating that the helix, the circle and the straight line are the only curves for which any portion is the same shape as any other portion with the same length.

Anyway, the first scholium in book XIII of Euclid's Elements, which Hultsch attributes to Geminus, mentions

> *... the five so-called Platonic figures which, however, do not belong to Plato, three of the five being due to the Pythagoreans, namely the cube, the pyramid, and the dodecahedron, while the octahedron and the icosahedron are due to Theaetetus.*

So, that's what I know about the origin of the icosahedron! Someday I'll read more, so let me make a note to myself:

20) Benno Artmann, "Antike Darstellungen des Ikosaeders", *Mitt. DMV* **13** (2005), 45–50. (Here the drawing of the icosahedron in Euclid's elements is analysed in detail.)

21) A. E. Taylor, *Plato: the Man and His Work*, Dover Books, New York, 2001, page 322. (This discusses traditions concerning Theaetetus and Platonic solids.)

22) Euclid, *Elementa: Libri XI-XIII cum appendicibus*, postscript by Johan Ludvig Heiberg, edited by Euangelos S. Stamatis, Teubner BSB, Leipzig, 1969. (Apparently this contains information on the scholium in book XIII of the Elements.)

Now for something a bit newer: categorification and quantum mechanics. I've said so much about this already that I'm pretty much talked out:

23) John Baez and James Dolan, "From finite sets to Feynman diagrams", in *Mathematics Unlimited — 2001 and Beyond*, vol. 1, eds. Bjrn Engquist and Wilfried Schmid, Springer, Berlin, 2001, pp. 29–50.

24) John Baez and Derek Wise, "Quantization and Categorification", *Quantum Gravity Seminar lecture notes*, available at: `http://math.ucr.edu/home/baez/qg-fall2003/` `http://math.ucr.edu/home/baez/qg-winter2004/` `http://math.ucr.edu/home/baez/qg-spring2004/`

As I explained in "Week 185", many basic facts about harmonic oscillators, Fock space and Feynman diagrams have combinatorial interpretations. For example, the commutation relation between the annihilation operator a and the creation operator a^*:

$$aa^* - a^*a = 1$$

comes from the fact that if you have some balls in a box, there's one more way to put a ball in and then take one out than to take one out and then put one in! This way of thinking amounts to using finite sets as a substitute for the usual eigenstates of the number operator, so we're really "categorifying" the harmonic oscillator: giving it a category of states instead of a set of states.

Working out the detailed consequences takes us through Joyal's theory of "structure types" or "species" — see "Week 202" — and on to more general "stuff types". Some nice category and 2-category theory is needed to make the ideas precise. For a careful treatment, see this thesis by a student of Ross Street:

25) Simon Byrne, *On Groupoids and Stuff*, honors thesis, Macquarie University, 2005, available at `http://www.maths.mq.edu.au/~street/ByrneHons.pdf` and `http://math.ucr.edu/home/baez/qg-spring2004/ByrneHons.pdf`

However, none of this work dealt with the all-important *phases* in quantum mechanics! For that, we'd need a generalization of finite sets whose cardinality can be be complex. And that's what my student Jeffrey Morton introduces here:

26) Jeffrey Morton, "Categorified algebra and quantum mechanics", *Theory and Application of Categories* **16** (2006), 785–854. Available at `http://www.emis.de/journals/TAC/volumes/16/29/16-29abs.html`; also available as math.QA/0601458.

He starts from the beginning, explains how and why one would try to categorify the harmonic oscillator, introduces the "$\mathrm{U}(1)$-sets" and "$\mathrm{U}(1)$-stuff types" needed to do this, and shows how the usual theorem expressing time evolution of a perturbed oscillator as a sum over Feynman diagrams can be categorified. His paper is now *the* place to read about this subject. Take a look!

Addendum: I thank Tim Chow, Phillip Helbig, Rudy Rucker, Jeffrey Winkler and especially Kevin Buzzard for interesting comments.

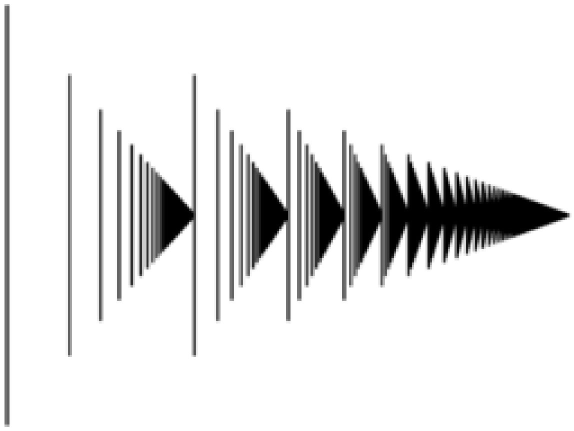
Jeffrey Winkler wrote:

> *Are the ridiculously infinite forms of infinity you discuss in your article ever used for anything?*
>
> *If they only refer to themselves, if the only thing these infinite sets refer to is other infinite sets, then it's pointless self-reflexive recursive circular reasoning. A set that has the same number of elements as the number of apples on your table, or the number of points in a line or a plane, or in infinitely dimensional Hilbert space, could be said to refer to something, but give an example of something, other than one of the sets you're talking about, which would have ε_0 elements.*

> *Where in mathematics or physics would you ever use such a concept? If you never would, then what's the point? In other words, is there anything where you could have ε_0 "many" of something, other than the sets themselves, and if there isn't, then are these actually numbers? If nothing is ever "that many", then is it a number? Of course, we've expanded the concept of "number" to include complex numbers, quaternions, octonions, vectors, tensors, matrices, etc. where they don't literally correspond to "how many" something is, yet in all those examples, they have uses in various fields of mathematics and physics, other than just when talking about themselves, so there is a reason for inventing them.*
>
> *Jeffery*

I replied:

> *Jeffery Winkler wrote:*
>
> > *Are the ridiculously infinite forms of infinity you discuss in your article ever used for anything?*
>
> *Without ε_0 you can't prove that Goodstein sequences converge to zero — an obviously true fact. As I mentioned, the main use of these ordinals is to measure the strength of axiom systems. But, I didn't write about these ordinals because they're useful. I wrote about them because they're fun.*
>
> *They're not "ridiculously infinite", though. The ordinals I mentioned are all countable ordered sets, and you can describe them all* explicitly *as subsets of the rational numbers.*
>
> *More precisely: any one of the ordinals I mentioned, up to and including the Feferman-Schtte ordinal (and quite a ways beyond), is isomorphic as an ordered set to a subset of the rational numbers. Moreover, you can write a computer program that will decide whether or not any given fraction is in this subset. As a consequence, you can also write a computer program that lists the fractions in this set.*
>
> *It's pretty obvious how to do this for ω^2:*



> *But you can do it for any one of the ordinals I mentioned! David Madore has drawn a picture of ε_0, for example.*
>
> *So, for someone to reject these ordinals as "ridiculously infinite", they must have some doubts about the legitimacy of computable subsets of the rational numbers as valid objects of study. That seems like an extreme position.*
>
> *The Church-Kleene ordinal is much larger than any of the ordinals I discussed in detail. It's still countable. Any ordinal below it can be described in a computable way — but it itself can't. So, if you believe that only computable mathematical entities are worth studying, you might want to stop shy of this one. I stopped* far *short of this one.*
>
> > *If they only refer to themselves, if the only thing these infinite sets refer to is other infinite sets, then it's pointless self-reflexive recursive circular reasoning.*
>
> *I hope you see that this is not true for the countable ordinals I was discussing. I deliberately refrained from mentioning the large cardinals that logicians often discuss, precisely because I share your distaste for such stuff.*
>
> *I realized after I posted "Week 236" that some people might think I was talking about mystical entities, when I was actually talking about very concrete things. So, I'm glad you brought this up.*
>
> *Best,*
> *jb*

And now for Kevin Buzzard's wonderful calculation! In the original version of this Week's Finds, I wrote:

> *if you start with the number 4, you get this Goodstein sequence:*
>
> $$4, 26, 41, 60, 83, 109, 139, 173, 211, 253, 299, 348, \ldots$$
>
> *and apparently it takes about 3×10^{60605351} steps to reach zero!*

I got this figure from the "National Curve Bank" website mentioned above, but I got the details wrong: they said the sequence "can increase for approximately 2.6×10^{60605351} steps", whatever that means.

Kevin Buzzard then sent me an email in which he worked out the number himself. After correcting a few small mistakes, we seem to have settled on a different answer.

Kevin writes:

> *You write this as if it were some kind of mystery. I remember working out this number explicitly when I was a graduate student! There is some nice form for it, as I recall. Let's see if I can reconstruct what I did.*
>
> *If I've understood the sequence correctly, it should be (where "n" at the beginning of a line denotes we're working in base n on this line, so strictly speaking it's probably the $n-1$st term in the sequence)*

2) $2^2 = 4$

3) $3^3 - 1 = 2.3^2 + 2.3 + 2 = 26$ *[note: base 3, ends in 2, and 3+2=5]*

4) $2 \cdot 4^2 + 2 \cdot 4 + 1 = 41$ *[note: base 4, ends in 1, and 4+1=5]*

5) $2 \cdot 5^2 + 2 \cdot 5 = 60$ *[we're at a limit ordinal here, note 3+2=4+1=5]*

6) $2 \cdot 6^2 + 2 \cdot 6 - 1 = 2 \cdot 6^2 + 6 + 5 = 83$ *[note: base 6, ends in 5]*

7) $2 \cdot 7^2 + 7 + 4$ *[note: base 7, ends in 4]*

8) $2 \cdot 8^2 + 8 + 3$ *[note: base 8, ends in 3, so we next get a limit ordinal at. . .]*

. . .

11) $2 \cdot 11^2 + 11$

12) $2 \cdot 12^2 + 12 - 1 = 2 \cdot 12^2 + 11$

13) $2 \cdot 13^2 + 10$

. . .

23) $2 \cdot 23^2$ *(as* $23 = 12 + 11 = 13 + 10 = \ldots$*)*

24) $24^2 + 23 \cdot 24 + 23$

. . .

47) $47^2 + 23 \cdot 47$

48) $48^2 + 22 \cdot 48 + 47$

. . .

95) $95^2 + 22 \cdot 95$

96) $96^2 + 21 \cdot 96 + 95$

. . .

and now we spot a pattern: we're just doubling—getting a limit ordinal at bases $24 - 1$*,* $48 - 1$*,* $96 - 1$ *and so on. Let's look again at those limit ordinals:*

47) $47^2 + 23 \cdot 47$

48) $95^2 + 22 \cdot 95$

. . .

$24 \cdot 2^{t-1}$*)* $(24 \cdot 2^{t-1})^2 + (24 - t) \cdot (24 \cdot 2^{t-1})$

. . .

so the last one with a square in it will be the case $t = 24$*, corresponding to*

r) r^2

where

$$r = 24 \cdot 2^{24} - 1 = 402653183.$$

> *All those 24s, but I'm sure you'll not get carried away. Let's define*
>
> $$n = r + 1 = 24 \cdot 2^2 4$$
>
> *and continue on. At the next step, the ordinal decreases sharply:*
>
> *n)* $n^2 - 1 = (n-1)n + (n-1)$ *n + 1)* $(n-1)(n+1) + (n-2)$ *[note: now back to the usual tricks]*
>
> ...
>
> $2n-1$*)* $(n-1)(2n-1)$ *[the next limit, at base* $2n-1$*]* $2n$*)* $(n-2)(2n)+(2n-1)$
>
> ...
>
> $4n-1$*)* $(n-2)(4n-1)$ $4n$*)* $(n-2)(4n)+(4n-1)$
>
> ...
>
> *and the limit ordinals we're running into now (and we're going to run into about n of them, which is a lot), are*
>
> $2n-1$*)* $(n-1)(2n-1)$ $4n-1$*)* $(n-2)(4n-1)$ $8n-1$*)* $(n-3)(8n-1)$
>
> ...
>
> $n2^s-1$*)* $(n-s)(n2^s-1)$
>
> ...
>
> *and finally when $s = n-1$*
>
> *m) m*
>
> *where $m = n2^{n-1} - 1$. The sequence now looks like*
>
> $m+1$*)* $(m+1)-1 = m$ $m+2$*)* $m-1$ $m+3$*)* $m-2$
>
> ...
>
> $2m+1$*)* 0
>
> *So the sequence becomes zero at base $n2^n - 1$, where $n = 24 \cdot 2^2 4$. If 2^2 is the first term in the sequence, I guess this is the $(n2^n-2)$th term. I make this about $6.9 \times 10^{121210694}$ — curses, you got something else! Actually, I have about the square of what you wrote and hence I have most likely made a slip. On the other hand you can see that it's not a mystery at all, it's just an elementary exercise. It really helps you learn about why the countable ordinals are well-ordered too: as you continue working out the numbers, you always have this impending sense of doom telling you that your gut feeling that the sequence tends to infinity might just be wrong...*
>
> *Kevin*

So, in simple terms, the 4th Goodstein sequence starts out by shooting up faster and faster, reaching almost n^2 by the nth term, where

$$n = 24 \cdot 2^2 4^{=} 402653184.$$

By the $(2n)$th term it reaches almost $2n^2$.

Then it grows by steps of $n-2$, and by the $(4n)$th term it reaches almost $4n^2$.

Then it grows by steps of $n-3$, and by the $(8n)$th term it reaches almost $8n^2$.

And so on. . . it's slowing down now.

After about the $(2^{n-2}n)$th term it levels off completely, and by the $(2^{n-1}n)$th term it equals almost $2^{n-1}n$. Note these numbers are the same.

From this point on the sequence decreases by one each time, and it hits zero at about the $(2^n n)$th term.

So, it spends the last half of its life decreasing by one each time. Right before that, it spends a quarter of its life remaining constant. Right before that, it spends an eighth of its life growing by one each time. Right before that it, it spends a sixteenth of its life growning by two each time. And so on, but not indefinitely: this description covers the phase after the nth term.

In particular, it reaches its maximum and then levels off after about quarter of its life is done. It spends about the next quarter of its life being constant, and the next half going down one step at a time.

Let's be a bit more precise. For this let's call the ith term of the 4th Goodstein sequence a_i, where we count things so that

$$a1 = 4, a2 = 26, \ldots$$

Then a_i first reaches its maximum value at exactly

$$i = \frac{1}{4} \cdot 24 \cdot 2^{24} \cdot 2^(24 \cdot 2^{24}) - 2 \sim 1.72 \times 10^{121210694}$$

Its value at this point is exactly

$$a_i = \frac{1}{2} \cdot 24 \cdot 2^{24} \cdot 2^(24 \cdot 2^{24}) - 1 \sim 3.45 \times 10^{121210694}$$

It then stays constant for about twice as many more terms, and then goes to zero one step at a time, hitting zero precisely when

$$i = 24 \cdot 2^{24} \cdot 2^(24 \cdot 2^{24}) - 2 \sim 6.89 \times 10^{121210694}$$

Believe me, it's a lot more fun to figure this stuff out than to read it.

I was worried that Kevin and I had made a mistake until we found this thesis which gets the same answer (though the author starts the sequence at $a_0 = 4$, so it looks one off from ours):

27) Justin T. Miller, *On the Independence of Goodstein's Theorem*, Masters thesis, University of Arizona, 2001. Also available as `http://www.u.arizona.edu/~miller/thesis/thesis.html`

Tim Chow had some comments about why the Feferman-Schtte ordinal is considered the first "impredicative" ordinal. He wrote:

> *John Baez wrote:*
>
> > *Logicians including Feferman and Schtte have carried out a detailed analysis of this subject. They know a lot about how much induction*

> *up to different ordinals buys you. And apparently, induction up to Γ_0 lets us prove the consistency of a system called "predicative analysis". I don't understand this, nor do I understand the claim I've seen that Γ_0 is the first ordinal that cannot be defined predicatively — i.e., can't be defined without reference to itself. Sure, saying Γ_0 is the first solution of*
>
> $$\varphi_x(0) = x$$
>
> *is non-predicative. But what about saying that Γ_0 is the union of all ordinals in the Veblen hierarchy? What's non-predicative about that?*

The situation is somewhat akin to the situation with the Church-Turing thesis, in that one is tentatively equating an informal notion (predicativity or computability) with a precise mathematical notion. Therefore there is no definitive answer to your question, and Feferman himself has articulated potential objections to the "standard view" that Γ_0 marks the boundary of predicativity.

Having said that, I'll also say that one of the reasons for the standard view is that Γ_0 marks the boundary of "autonomous progressions" of arithmetical theories. The book by Torkel Franzen that you cited is probably the most accessible introduction to this subject. Roughly speaking, the idea is that if anyone fully accepts first-order Peano arithmetic PA*, then implicitly he accepts its consistency* $\mathrm{Con(PA)}$*, as well as* $\mathrm{Con(PA + Con(PA))}$*, etc. If one tries to articulate exactly what is "implicitly" involved in accepting* PA *in this sense, then one can make a plausibility argument that Γ_0 is a natural stopping point. I think you have a better shot at grasping the underlying intuition via this approach than by staring at Γ_0 itself and trying to figure out what is non-predicative about its definition.*

I replied asking if "etc." means there's one theory like this per ordinal. I also asked for more clues about this "plausibility argument", and noted:

There's also someone named Nik Weaver who has debated Feferman on this subject:

http://www.cs.nyu.edu/pipermail/fom/2006-April/010472.html
http://www.math.wustl.edu/~nweaver/conceptualism.html

He seems to claim that Γ_0 and even larger ordinals have predicative definitions. However, I'm too ignorant to follow this debate. Usually in physics I have a sense for when people are being reasonable even if I don't follow the details. In this debate I can't even do that.

Tim Chow replied:

Let's look more closely at what the notion of "one theory like this per ordinal" means. There's no difficulty figuring out what "$\mathrm{Con(PA)}$*" means or how to express that statement in the first-order language of arithmetic. Ditto with "*$\mathrm{Con(PA + Con(PA))}$*". However, once you start ascending the ordinal hierarchy, a difficulty appears. The language of arithmetic doesn't let you talk about*

> *"ordinals" directly—that's a set-theoretical concept. In order to express a statement like "*$\mathrm{Con}(T)$*" for some theory T, you need at minimum to be able to give some sort of "recursive description" or "recursive axiomatization" of T (where here I use the word "recursive" in the technical sense of recursive function theory) in the first-order language of arithmetic. This observation already yields the intuition that we're not going to be able to ascend beyond the Church-Kleene ordinal, because we won't even be able to figure out how to* say *"T is consistent" for a theory T that requires that many iterations to reach from* PA.
>
> *There are other problems, though, that potentially get in the way before we reach the Church-Kleene ordinal. Once we realize that what we need is a system of "ordinal notations" to "fake" the relevant set theory, we may (if we are predicativists) worry about issues such as:*
>
> 1. *As we ascend the ordinal hierarchy, isn't it illegitimate to make a jump to an ordinal α unless we've already proved, at the level of some ordinal β that we've already reached, that an ordinal of type α exists?*
> 2. *And isn't it illegitimate to create sets by quantification over things other than the natural numbers themselves and sets that we've already created?*
>
> *Condition 1 goes by the name of "autonomy" and condition 2 goes by the name of "ramification." If one formalizes these notions in a certain plausible manner, then one arrives at Γ_0 as the least upper bound of theories that you can get to, starting with (for example)* PA.
>
> *One can of course wonder whether 1 and 2 above really capture the concept of "predicativity." Some secondary evidence has accumulated of the following form: Some argument that intuitively seems to be predicative but that is not immediately seen to be provable in the Feferman-Schtte framework is shown, after some work, to indeed be provable below Γ_0.*
>
> *It's still possible, of course, for someone—you mentioned Nik Weaver— to come along and argue that our intuitive notion of predicativism, fuzzy though it is, can't possibly be identified with the level Γ_0. The reason you can't seem to decide immediately whether Weaver's position is nonsensical or not is probably because the critical questions are not mathematical but philosophical, and of course it's usually harder to arrive at definitive answers in philosophy than in mathematics.*

Finally, let me record a number of papers that treat notations for ordinals above the Feferman-Schtte ordinal. In the discussion of this Week's Finds on `sci.math.research`, Dave Renfro pointed out a really nice readable paper on large countable ordinals which goes far beyond what I discussed:

28) Hilbert Levitz, "Transfinite ordinals and their notations: For the uninitiated", available at `http://www.cs.fsu.edu/~levitz/research.html`

This paper introduced the "Schtte Klammersymbole", which generalize the Veblen hierarchy:

29) Kurt Schtte, “Kennzeichnung von Orgnungszahlen durch rekursiv erklrte Funktionen”, *Math. Ann* **127** (1954), 15–32.

These papers discuss a general concept of “ordinal notation system”, which includes the Schtte Klammersymbole and also something called the “n-ary Veblen hierarchy”:

30) Anton Setzer, “An introduction to well-ordering proofs in Martin-Lf’s type theory”, in *Twenty-Five Years of Constructive Type Theory*, eds. G. Sambin and J. Smith, Clarendon Press, Oxford, 1998, pp. 245–263. Also available at `http://www.cs.swan.ac.uk/~csetzer/index.html`

    Anton Setzer, “Ordinal systems”, in *Sets and Proofs*, Cambridge U. Press, Cambridge, 2011, pp. 301-331. Also available at `http://www.cs.swan.ac.uk/~csetzer/index.html`

This paper has a nice expository section on generalizations of the Veblen hierarchy:

31) Jean H. Gallier, “What’s so special about Kruskal’s theorem and the ordinal Γ_0? A survey of some results in proof theory”, sec. 7, A glimpse at Veblen hierarchies, *Ann. Pure Appl. Logic* **53** (1991), 199–260. Also available at `http://www.cis.upenn.edu/~jean/gallier-old-pubs.html`

This paper is very useful, since it compares different notations:

32) Larry W. Miller, “Normal functions and constructive ordinal notations”, *J. Symb. Log.* **41** (1976), 439–459.

You can get it through JSTOR if you have access to that.

This webpage gives a nice definition of “ordinal notation system” as a coalgebra of a certain functor — nice if you understand categories, that is:

33) Peter Hancock, “Ordinal notation systems”, `http://homepages.inf.ed.ac.uk/v1phanc1/ordinal-notations.html`

Kevin Watkins pointed out this website, which contains several papers on ordinal notations:

34) Harold Simmons, Abstracts of papers and notes, `http://www.cs.man.ac.uk/~hsimmons/DOCUMENTS/papersandnotes.html`

Finally, the Wikipedia article on “large countable ordinals” has some references to books which are, alas, out of print.

Said Conrad Cornelius O’Donald O’Dell,
My very young friend who is learning to spell,
“The A is for Ape. And the B is for Bear.
The C is for Camel. The H is for Hare.
The M is for Mouse. And the R is for Rat.

I know all the twenty-six letters like that...
... through to Z is for Zebra. I know them all well.
So now I know everything anyone knows
From beginning to end. From the start to the close.
Because Z is as far as the alphabet goes."

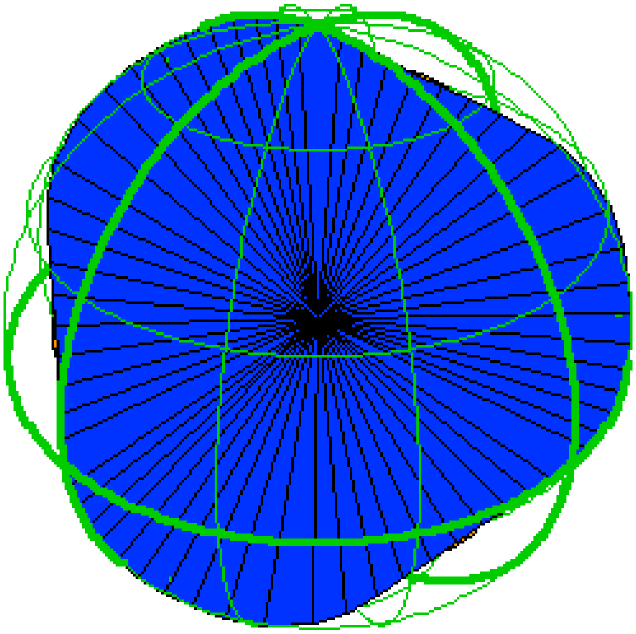
Then he almost fell flat on his face on the floor
When I picked up the chalk and drew one letter more!
A letter he never had dreamed of before!
And I said, "You can stop, if you want, with the Z
Because most people stop with the Z
But not me!
In the places I go there are things that I see
That I never could spell if I stopped with the Z.
I'm telling you this 'cause you're one of my friends.
My alphabet starts where your alphabet ends!"

— *Doctor Seuss*

Week 237

August 10, 2006

This Week I'd like to talk about math books in Shanghai, and Urs Schreiber's blog entry on the gauge 3-group of M-theory. But first:



1) Greg Egan, "Klein's quartic equation", `http://gregegan.customer.netspace.net.au/SCIENCE/KleinQuartic/KleinQuarticEq.html`

I discussed Klein's quartic curve in "Week 214" and "Week 215". The idea is to take the nontrivial complex solutions of

$$u^3v + v^3w + w^3u = 0$$

and "projectivize" them — in other words, count two as the same if one is just a multiple of the other:

$$(u', v', w') = c(u, v, w)$$

The result is a 3-holed Riemann surface with the maximum number of symmetries! Here by a "symmetry" I mean a conformal transformation mapping the surface to itself. Back in 1893 Hurwitz proved something quite bizarre: an n-holed Riemann surface can't have more than $84(n-1)$ symmetries if $n > 1$. So, a 3-holed Riemann surface can't have more than 168 symmetries — and Klein's quartic curve has exactly that many!

These 168 symmetries were constructed by Klein way back in 1879, but Egan gives an elementary proof that uses only algebra and a bit of calculus... and a lot of cleverness.

And, his page has a wonderful spinning picture of the *real* solutions of

$$u^3v + v^3w + w^3u = 0.$$

This is what you see above.

As you can see, it consists of lots of lines through the origin, including the u, v, and w axes. When we "projectivize", we get one point for each of these lines, so we get a curve which is the real version of Klein's quartic curve. This curve has an obvious 3-fold symmetry, from cyclically permuting the coordinate axes. The rest of the 168 symmetries are only easy to visualize when we go to the complex version — as Egan explains.

It's great that Egan can draw this thing in Perth and I can easily see it my apartment here in Shanghai — I feel like I'm living in a futuristic world, and I'm only 45. What it'll be like when I'm 64?

Another pleasant thing about life in Shanghai, at least for a well-off visitor from America, is how cheap everything is. It's clear why the US is running an enormous trade deficit: there's a vast economic differential. Stopping the flow of goods one way and dollars the other would be like damming the Niagara Falls.

For example, last night I saw this excellent hardcover book on sale for 78 yuan, or about $10:

2) Yu. I. Manin and A. A. Panchishkin, *Introduction to Modern Number Theory*, second edition, Science Press, 2005.

It's a great overview of number theory, from the basics through class field theory to L-functions, modular forms and the Langlands program! It's wisely divided into three sections: "problems and tricks", "ideas and theories", and "analogies and visions". Back when I used to hate number theory, I thought it was all problems and tricks. Now I'm beginning to learn some of the ideas and theories, and I hope eventually to grasp the analogies and visions discussed here — for example, the analogy between Arakelov geometry and noncommutative geometry.

$10 is a nice price for a math book. If you buy this one from Springer Verlag, you'll pay ten times that. Illegal knockoffs of Western books are common in China, but I think the one I saw is legal, since Springer has signed an agreement with Science Press, which is run by the Chinese Academy of Sciences. In exchange for letting Science Press publish Springer books in China at affordable prices, Springer gets to publish translations of Chinese journals in the West at unaffordable prices.

By the way — after checking out the bookstore, I went out to the street vendors and bought an excellent dinner of rice, sausage and vegetables for 35 yuan — about 40 cents US. It was cooked by a husband and wife in a wok on a cart.

Just after I bought it, someone yelled the Chinese equivalent of "cops!", and all the street vendors suddenly dashed away with their carts, leaving only the woman, who kindly handed me my dinner in a styrofoam pack before walking off. They clearly had this down to a fine art: it all happened faster than my brain could process. I guess the cops don't allow street vendors there.

I only wish I'd noticed: did the street vendors turn off their gas stoves before running, or run while still cooking?

Anyway, on to some serious math and physics.

You've probably heard of some mysterious thing called "M-theory" that lives in 11 dimensions. Back in "Week 158" and "Week 159" I took a stab at understanding this. Now I'll try again, with a lot of help from Urs Schreiber:

3) Urs Schreiber, "Castellani on free differential algebras in supergravity: gauge 3-group of M-theory", `http://golem.ph.utexas.edu/string/archives/000840.html`

Calling M-theory a "theory" is a bit misleading, because nobody knows what this theory is! There's just a lot of clues pointing to its existence. It seems to be the quantum version of a well-defined classical field theory called "11-dimensional supergravity". And, it seems to involve 2-branes and 5-branes: 2- and 5-dimensional membranes that trace out 3- and 6-dimensional surfaces in spacetime, just like strings trace out 2-dimensional surfaces.

Back in "Week 158" I wrote down a Lagrangian for 11d supergravity. This is a truly monstrous thing involving three fields:

A) a frame field e — the "graviton",

B) a field ψ taking values in the real spin-$3/2$ representation of the 11d Lorentz group — the "gravitino",

C) a 3-form A.

When it was discovered back in 1978, people were interested in 11d supergravity mainly because it was the highest-dimensional theory they could concoct that includes general relativity and supersymmetry — a symmetry that interchanges bosons and fermions, in this case gravitons and gravitinos — without including any particles of spin > 2. So, the fact that it looked like a mess wasn't such a big deal. But now that some people are taking it very seriously, it's worth trying to understand the math behind it more deeply, to see what makes it tick.

For example: what's so great about 11 dimensions? And: what's the reason for that 3-form?

I'm not a huge fan of string theory, but I like puzzles of this sort — finding patterns that make certain things work only in certain dimensions, and stuff like that. So, I got intrigued when I learned that super-Yang-Mills theory and superstring theory are nice in dimension 10 because of special properties of the octonions — see "Week 104". Maybe a little extra stretch could bring us to dimension 11?

I got even more intrigued when I ran across two competing explanations for that 3-form in 11d supergravity. One was that it's a connection on a twice categorified version of a $\mathrm{U}(1)$ bundle. The other was that it's the Chern-Simons form for an E_8 gauge theory.

Let me say a bit about what these means. I talked about categorified $\mathrm{U}(1)$ bundles in "Week 210", so I'll be sort of brief about those. . . .

A connection on a $\mathrm{U}(1)$-bundle looks locally like a 1-form, so we can integrate it along a path and compute how the phase of charged particle changes when we move it along that path:

$$x \xrightarrow{f} y$$

a path f from the point x to the point y;
we write this as $f\colon x \to y$.

Believe it or not, this is the basis of all modern ideas on electromagnetism!

If we categorify this whole idea once, we get a kind of connection that looks locally like a 2-form. Folks call this a "connection on a $\mathrm{U}(1)$ gerbe", but don't let the use of French here intimidate you: they just do that so they can charge more for the wine. It's just a gadget that you can integrate over a surface, to compute how the phase of a charged *string* moves when we slide it along that surface:

$$\begin{tikzcd} x \arrow[r, bend left=50, "f"{name=U}] \arrow[r, bend right=50, "g"'{name=D}] & y \arrow[from=U, to=D, Rightarrow, "F"] \end{tikzcd}$$

a path f from the point x to the point y;
we write this as $f\colon x \to y$.

And, if we categorify once more, we get a "connection on a $\mathrm{U}(1)$ 2-gerbe". This is something that looks locally like a 3-form, which describes what happens when we move 2-*branes* around!

If you're wondering why I'm talking about "categorifying", it's because this:

$$x \xrightarrow{f} y$$

is also a picture of a morphism in a category, while this:

$$\begin{tikzcd} x \arrow[r, bend left=50, "f"{name=U}] \arrow[r, bend right=50, "g"'{name=D}] & y \arrow[from=U, to=D, Rightarrow, "F"] \end{tikzcd}$$

is a picture of a 2-morphism in a 2-category and so on. We're talking about processes between processes between processes. . . so we're climbing up the ladder of n-categories.

Anyway: since 11d supergravity has a 3-form in it, and M-theory apparently has 2-branes in it, maybe we need to categorify the concept of a $\mathrm{U}(1)$ bundle twice to understand what's going on here!

I came up with this crazy idea on my own back in "Week 158", but it's an obvious guess after you learn that the 2-form field called B in 10d superstring theory really *is* a connection on a $\mathrm{U}(1)$ gerbe:

4) Alan L. Carey, Stuart Johnson and Michael K. Murray, "Holonomy on D-branes", available as arXiv:hep-th/0204199.

Unfortunately there are some problems with naively pushing this idea up a dimension. For example, a crucial factor of $1/6$ in the Lagrangian for 11d supergravity is not explained by thinking of A this way.

Another possible explanation was that this 3-form is the Chern-Simons form of an E_8 bundle over spacetime:

5) Emanuel Diaconescu, Gregory Moore and Edward Witten, "E_8 gauge theory, and a derivation of K-theory from M-theory", *Adv. Theor. Math. Phys.* **6** (2003) 1031–1134. Also available as arXiv:hep-th/0005090.

6) Emanuel Diaconescu, Daniel S. Freed and Gregory Moore, "The M-theory 3-form and E_8 gauge theory", available as arXiv:hep-th/0312069.

This idea explains that factor of $1/6$. And, it might move towards an explanation of how the octonions get into the act, because the group E_8 is deeply related to the octonions. But as the authors of the above paper say, "the E_8 gauge field plays a purely topological role and appears, in some sense, to be a 'fake'." In particular, you don't see any E_8 connection staring you in the face in the Lagrangian for 11d supergravity that I wrote down in "Week 158".

Later, it started becoming clear that both ideas — the twice categorified $\mathrm{U}(1)$ connection and the E_8 gauge theory — fit together in some way:

7) Paolo Aschieri and Branislav Jurco, Gerbes, "M5-brane anomalies and E_8 gauge theory", *JHEP* **0410** (2004), 068. Also available as arXiv:hep-th/0409200.

It all became a lot clearer to me when Urs Schreiber read these papers and translated them into a language I like:

8) Leonardo Castellani, "Lie derivatives along antisymmetric tensors, and the M-theory superalgebra", available as arXiv:hep-th/0508213.

9) Pietro Fr and Pietro Antonio Grassi, "Pure spinors, free differential algebras, and the supermembrane", available as arXiv:hep-th/0606171.

The idea is to think of 11d supergravity as a twice categorifed gauge theory — not just the 3-form field in 11d supergravity, but all the fields, in a unified way!

For this, we need to do something much more clever than taking 11d spacetime and slapping a $\mathrm{U}(1)$ 2-gerbe on top of it. We need to combine the graviton, the gravitino and the 2-form field into a connection on a *nonabelian* 2-gerbe.

Here things get a bit technical, but Urs has covered the technical points quite nicely in his blog, so right now I'll just try to give you some hand-wavy intuitions.

Very roughly speaking, an connection on a bundle takes any path in spacetime

$$x \xrightarrow{f} y$$

and gives you an element of some *group*, which says how a particle would transform if you moved it along this path. This group could be $\mathrm{U}(1)$ — the group of phases — or it could be something more fun, like a *nonabelian* group.

If we categorify this concept, we get the concept of a connection on a "2-bundle" (which is more or less the same as a gerbe). Such a connection takes any path and gives you an *object* in some 2-*group*, but it also takes any surface like this:

$$\begin{tikzcd} x \arrow[r, bend left=50, "f", ""{name=U, below}] \arrow[r, bend right=50, "g"', ""{name=D}] & y \arrow[Rightarrow, from=U, to=D, "F"] \end{tikzcd}$$

and gives you a *morphism* in this 2-group. You see, 2-group is a kind of category that acts like a group, and a category has "objects" and "morphisms". The morphisms go between objects. For more on 2-groups, try:

10) "Higher-dimensional algebra V: 2-Groups", with Aaron D. Lauda, *Theory and Applications of Categories* **12** (2004), available at `http://www.tac.mta.ca/tac/volumes/12/14/12-14abs.html`. Also available as `math.QA/0307200`.

If we categorify once more, we get connections on a "3-bundle", which is more or less the same thing as a "2-gerbe" — unfortunately the numbering systems are off by one. This gives us objects, morphisms and 2-morphisms in a 3-*group*, which describe what happens when we move particles, strings and 2-branes.

And so on:

group	point particles
2-group	point particles and strings
3-group	point particles, strings and 2-branes
4-group	point particles, strings, 2-branes and 3-branes

etc.

So, if 11d supergravity is a twice categorified gauge theory, we need to know its symmetry 3-*group*.

But actually, since we're doing geometry, this 3-group should be a "Lie 3-group". In other words, very roughly speaking, a 3-group that has a *manifold* of objects, a manifold of morphisms, and a manifold of 2-morphisms, where all the operations are smooth.

But actually, since we're doing supersymmetric geometry, we need a "Lie 3-supergroup"! In other words, very roughly speaking, a 3-group that has a *supermanifold* of objects, a supermanifold of morphisms, and a supermanifold of 2-morphisms, where all the operations are smooth. (Maybe I should say "supersmooth", just to be consistent.)

If you don't know what a supermanifold is, now is probably not the time to learn. I mean, not right this second. The point is just this: supersymmetry infests everything once you let it in the door, just like n-categories, and just like manifolds — and now we're doing all three.

In fact, nobody has even written down a rigorous definition of a Lie 3-supergroup yet! But, Lie algebras are in some ways simpler than Lie groups, and they're a good start, so we can be glad that people *do* know what a Lie 3-superalgebra is!

And Urs describes, in his blog, the relevant Lie 3-superalgebra for 11d supergravity!

I would like to say more about this, but it's getting a bit tough trying to talk about this stuff in a fun, easily accessible style, and I have the feeling I'm no longer succeeding. In fact, I don't think I can give a "fun, easily accessible" description of this specific Lie 3-superalgebra — at least not yet. So, now I'll completely give up trying to be comprehensible, and simply state some facts.

As shown here:

11) "Higher-dimensional algebra VI: Lie 2-Algebras", with Alissa Crans, *Theory and Applications of Categories* **12** (2004), available at `http://www.tac.mta.ca/tac/volumes/12/15/12-15abs.html` Also available as `math.QA/0307200`.

the category of Lie n-algebras is equivalent to the category of L_∞ algebras which as chain complexes have only n nonvanishing terms, the 0th to the $(n-1)$st. L_∞ algebras are just algebras of Stasheff's L_∞ operad in the category of chain complexes of vector spaces — see "Week 191" and especially these:

12) Martin Markl, Steve Schnider and Jim Stasheff, *Operads in Algebra, Topology and Physics*, AMS, Providence, Rhode Island, 2002.

James Stasheff, Hartford/Luminy talks on operads, available at `http://www.math.unc.edu/Faculty/jds/operadchik.ps`

But, we can replace vector spaces by $\mathbb{Z}/2$-graded vector spaces and everything still works. Physicists call $\mathbb{Z}/2$-graded vector spaces "super vector spaces". So, a "Lie n-superalgebra" is an algebra of the L_∞ operad in the category of chain complexes of super vector spaces.

Given this, to specify a Lie 3-superalgebra we first need to specify the 0-chains, then the 1-chains, then the 2-chains.

For the particular one Urs mentions, we have

- $\{0\text{-chains}\}$ = 11d Poincar Lie superalgebra
- $\{1\text{-chains}\} = \{0\}$
- $\{2\text{-chains}\} = \mathbb{R}$

Here $\mathbb{R}$ is the real numbers, and this 1-dimensional thing is what ultimately gives the 3-form field A in 11d supergravity. As a vector space, the 11d Poincar Lie superalgebra is the direct sum of an even part, which is the usual Poincar Lie algebra $\mathfrak{iso}(11,1)$, and an odd part, which is the 32-dimensional real spinor rep of $\mathfrak{so}(11,1)$. These give the graviton (or more precisely the Levi-Civita connection) and the gravitino in 11d supergravity.

Next we need to make this stuff into a chain complex. That's easy: the differential *has* to be zero.

Next, we need to specify the L_∞ structure on this chain complex. First, we need a binary bracket operation, like in an ordinary Lie superalgebra. The bracket of 0-chains is the usual bracket in the 11d Poincar Lie superalgebra. All the other binary brackets are zero.

Then, we need a ternary bracket operation, which expresses how the Jacobi identity holds only up to chain homotopy. This is zero.

Then, we need a quaternary bracket operation (since that chain homotopy satisfies its own identity only up to chain homotopy). This is nonzero: when we take the quaternary bracket of four 0-chains we get a 2-chain, and there's a nontrivial way to define this! This is the interesting bit, since ultimately it relates the graviton/gravitino to the 3-form field.

How do we get that quaternary bracket? Well, here's where things get funky: D'Auria and Fr dreamt up a formula that gives a number from 2 spinors and 2 vectors:

$$(\psi, \varphi, v, w) \mapsto \overline{\psi}\Gamma^{ab}\varphi v_a w_b.$$

And, magically, in 11 dimensions this gives a 4-cocycle on the Poincar Lie superalgebra! The proof of this uses some Fierz identity in 11 dimensions:

12) R. D'Auria and Pietro Fr, "Geometric supergravity in $D = 11$ and its hidden supergroup", *Nucl. Phys.* **B201** (1982), 101–140. Also available at `http://www.math.uni-hamburg.de/home/schreiber/sdarticle.pdf`

And, from HDA6 we know that the 4-cocycle condition is just what's needed to make the quaternary bracket satisfy the identity we need for a Lie 3-superalgebra. (Alissa and I just did the calculation for Lie n-algebras, but the "super" stuff should work too with a few signs thrown in.)

So, this is all very cool, but I need to understand Fierz identities in different dimensions to see what if anything is special to 11d here — or, alternatively, work out the cohomology of Poincar Lie superalgebras, to see when they can be deformed to Lie n-superalgebras. Sounds like a lot of work — maybe someone already did it. Actually D'Auria and Fre make it look like a matter of understanding tensor products of irreps of $\mathfrak{so}(n, 1)$, which is not bad. A worthwhile project in any event.

I also need to understand what all this has to do with E_8. For that the paper by Diaconescu, Freed and Moore should help.

Well, this is just the beginning, but Urs explains the rest.

Addenda: I thank Noam Elkies for a correction. Aaron Bergman has this to say about E_8 and M-theory:

> *John Baez wrote:*
>
> > *I also need to understand what all this has to do with* E_8.
>
> E_8 *might not have much of anything to do with this. As mentioned in Diaconescu, Freed and Moore,* E_8 *appears to function solely as a convenient stand-in for* $K(\mathbb{Z}, 3)$.
>
> *On the other hand, the split form of the* E*-series (up to* E_{11} *if you're feeling particularly speculative) is known to show up in describing the fields of 11D SUGRA, but the I don't think anyone knows of a connection between the two* E_8*s. Just to add to the fun,* E_8 *gauge fields also show up on the fixed points of M-theory on* $S^1/\mathbb{Z}_2$ *giving the* $\mathrm{E}_8 \times \mathrm{E}_8$ *heterotic string.*
>
> *Aaron*

Urs Schreiber replied:

> *Aaron Bergman wrote:*
>
> > E_8 *gauge fields also show up on the fixed points of M-theory on* $S^1/\mathbb{Z}_2$ *giving the* $\mathrm{E}_8 \times \mathrm{E}_8$ *heterotic string.*
>
> *The topological part of the membrane action involves the integral of the sugra* 3*-form over the worldvolume. By DFW, part of that* 3*-form can be thought of as an* E_8 *CS* 3*-form.*
>
> *So part of the membrane action looks similar to an* E_8 *CS-theory over the worldvolume.*
>
> *Now let the membrane have a boundary. A bulk* E_8 *CS-theory is well known to induce an* E_8 *WZW theory on the boundary.*
>
> *Could this be the connection between the DWF* E_8 *and the Horava-Witten* E_8*?*
>
> *I have asked this question before:* `http://golem.ph.utexas.edu/string/archives/000791.html`*. Jarah then agreed that this must be about right. But it is not completely clear to me yet.*
>
> *One problem is that in CS-theory we vary the connection, while in the topological membrane the* E_8 *connection on the background is fixed and we vary the embedding by which we pull it back to the worldvolume. Under suitable assumptions that might be equivalent to varying an* E_8 *connection on the worldvolume itself?*

Aaron Bergman replied:

> *Urs Schreiber wrote:*
>
> > *Now let the membrane have a boundary. A bulk* E_8 *CS-theory is well known to induce an* E_8 *WZW theory on the boundary.*
> >
> > *Could this be the connection between the DWF* E_8 *and the Horava-Witten* E_8*?*
>
> *A similar observation was made by Horava way back in* `arXiv:hep-th/9712130` *(in the final section).*
>
> *Aaron*

Week 238

August 16, 2006

NASA is trying to built up suspense with this “media advisory”:

1) NASA, “NASA Announces Dark Matter Discovery”, `http://www.nasa.gov/home/hqnews/2006/aug/HQ_M06128_dark_matter.html`

which says simply:

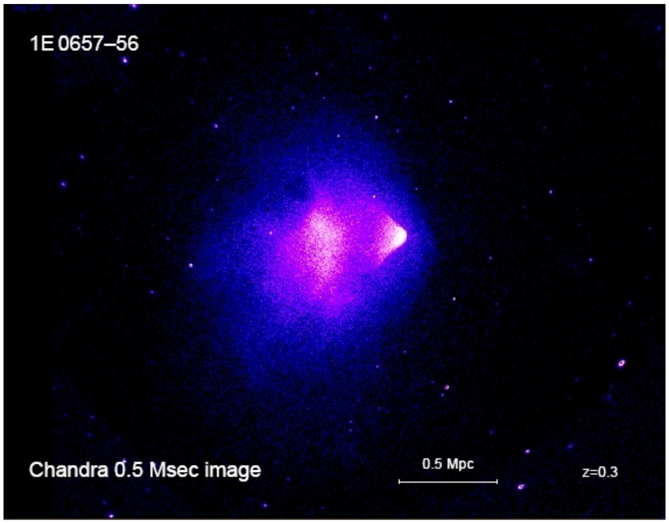
> *Astronomers who used NASA’s Chandra X-ray Observatory will host a media teleconference at 1 p.m. EDT Monday, Aug. 21, to announce how dark and normal matter have been forced apart in an extraordinarily energetic collision.*

Hmm! What’s this about?

Someone nicknamed “riptalon” at Slashdot made a good guess. The media advisory lists the “briefing participants” as Maxim Markevitch, Doug Clowe and Sean Carroll. Markevitch and Clowe work with the Chandra X-ray telescope to study galaxy collisions and dark matter. Last November, Markevitch gave a talk on this work, which you can see here:

2) Maxim Markevitch, Scott Randall, Douglas Clowe, and Anthony H. Gonzalez, “Insights on physics of gas and dark matter from cluster mergers”, available at `http://cxc.harvard.edu/symposium_2005/proceedings/theme_energy.html#abs23`

So, barring any drastic new revelations, we can guess what’s up. Markevitch and company have been studying the “Bullet Cluster”, a bunch of galaxies that has a small bullet-shaped subcluster zipping away from the center at 4,500 kilometers per second. Here’s a picture of it from the above paper:



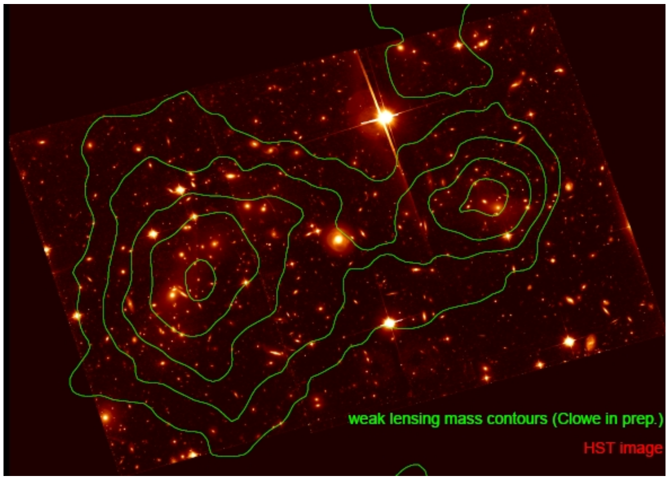
To help you understand this picture a bit: the official name of the Bullet Cluster is 1E0657-56. The "exposure" for this X-ray photograph taken by Chanda was apparently 0.5 million seconds — 140 hours! The distance scale shown, 0.5 megaparsecs, is about 1.6 million light years. The cluster itself has a redshift $z = 0.3$, meaning its light has wavelengths stretched by a factor of 1.3. Under currently popular ideas on cosmology, this means it's roughly 4 billion light years away.

Anyway, what are we seeing here?

You can see rapidly moving galaxy cluster with a shock wave trailing behind it. It seems to have hit another cluster at high speed. When this kind of thing happens, the *gas* in the clusters is what actually collides — the individual galaxies are too sparse to hit very often. And when the gas collides, it gets hot. In this case, it heated up to about 160 million degrees and started emitting X-rays like mad! The picture shows these X-rays. This may be hottest known galactic cluster.

That's fun. But that's not enough reason to call a press conference. The cool part is not the crashing of gas against gas. The cool part is that the dark matter in the clusters was unstopped — it kept right on going!

How do people know this? Simple. Folks can see the *gravity* of the dark matter bending the light from more distant galaxies! It's called "gravitational lensing". Here are the mass density contours, as seen by this effect. I guess Clowe took this photo using the Hubble Space Telescope:



So: X-rays show the gas in one place, but gravity shows most of the mass is somewhere else — two lumps zipping along unstopped. That's good evidence that dark matter is for real.

For more try these:

3) M. Markevitch, S. Randall, D. Clowe, A. Gonzalez, and M. Bradac, "Dark matter and the Bullet Cluster", available at `http://www.cosis.net/abstracts/COSPAR2006/02655/COSPAR2006-A-02655.pdf`

4) M. Markevitch, A. H. Gonzalez, D. Clowe, A. Vikhlinin, L. David, W. Forman, C. Jones, S. Murray, and W. Tucker, "Direct constraints on the dark matter self-

interaction cross-section from the merging galaxy cluster 1E0657-56", available as astro-ph/0309303.

5) Maxim Markevitch, "Chandra observation of the most interesting cluster in the Universe", available as astro-ph/0511345.

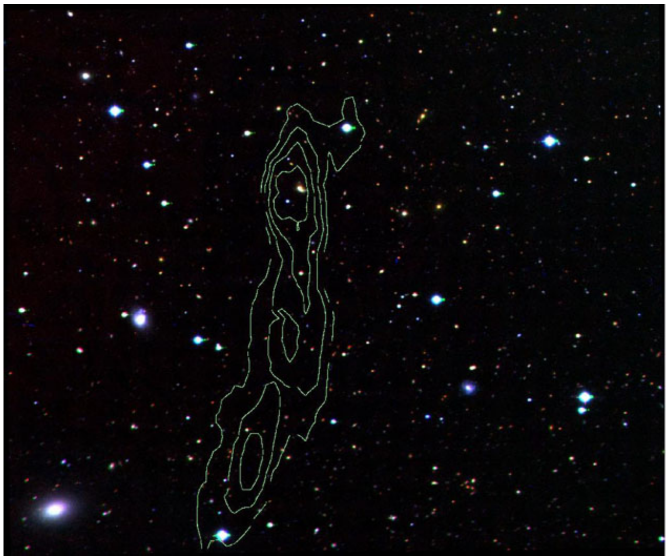
6) M. Markevitch, A. H. Gonzalez, L. David, A. Vikhlinin, S. Murray, W. Forman, C. Jones and W. Tucker, "A textbook example of a bow shock in the merging galaxy cluster 1E0657-56", *Astrophys. J.* **567** (2002), L27. Also available as astro-ph/0110468.

7) Eric Hayashi and Simon D. M. White, "How rare is the Bullet Cluster?", *Mon. Not. Roy. Astron. Soc. Lett.* **370** (2006), L38–L41, available as astro-ph/0604443.

The first of these is, alas, only the abstract of a talk. But it's worth reading, so I'll quote it in its entirety here:

> *1E0657-56, the "Bullet Cluster", is a merger with a uniquely simple geometry. From the long Chandra X-ray observation which revealed a classic bow shock in front of a small subcluster, we can derive the velocity of the subcluster and its direction of motion. Recent accurate weak and strong lensing total mass maps clearly show two merging subclusters, including the host of the gas bullet seen in X-rays. This cluster provided the first direct, model-independent proof of the dark matter existence (as opposed to any modified gravity theory) and a direct constraint on the self-interaction cross-section of the dark matter particles. I will review these and other related results.*

The Bullet Cluster is not the only direct evidence for dark matter. In fact, last year folks claimed to have found a "ghost galaxy" made mainly of dark matter and cold hydrogen, with very few stars:



8) PPARC, "New evidence for a dark matter galaxy", `http://www.interactions.org/cms/?pid=1023641`

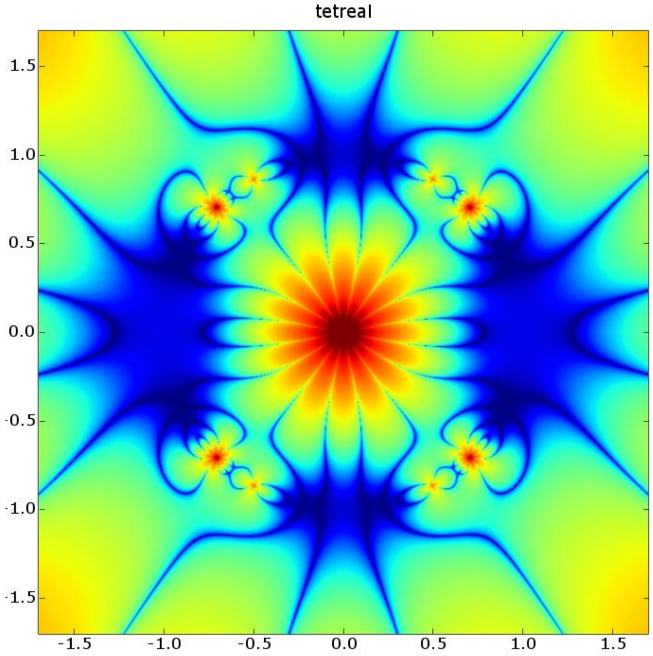
However, Matt Owers informs me that the consensus on this ghost, VIRGOHI 21, is that it's hydrogen stripped off from a galaxy by the "wind" it felt as it fell into the Virgo Cluster. This effect is called "ram pressure stripping" — the gas of a galaxy can be stripped off if the galaxy is moving rapidly through a cluster, due to interaction with the gas in the cluster.

Nonetheless, dark matter is seeming more and more real. It thus becomes ever more interesting to find out what dark matter actually *is*. The lightest neutralino? Axions? Theoretical physicists are good at inventing plausible candidates, but finding them is another thing.

Since I'd like to send this off in time to beat NASA, I won't say a lot more today... just a bit.

Dan Christensen and Igor Khavkine have discovered some fascinating things by plotting the amplitude of the tetrahedral spin network — the basic building block of spacetime in 3d quantum gravity — as a function of the cosmological constant.

They get pictures like this:



9) Dan Christensen and Igor Khavkine, "Plots of q-deformed tets", `http://jdc.math.uwo.ca/spinnet/`

Here the color indicates the real part of the spin network amplitude, and it's plotted as a function of q, which is related to the cosmological constant by a funky formula I won't bother to write down here.

You can get some nice books on category theory for free these days:

10) Jiri Adamek, Horst Herrlich and George E. Strecker, *Abstract and Concrete Categories: the Joy of Cats*, available at `http://katmat.math.uni-bremen.de/acc/acc.pdf`

11) Robert Goldblatt, *Topoi: the Categorial Analysis of Logic*, available at `http://cdl.library.cornell.edu/cgi-bin/cul.math/docviewer?did=Gold010`

12) Michael Barr and Charles Wells, *Toposes, Triples and Theories*, available at `http://www.case.edu/artsci/math/wells/pub/ttt.html`

The first two are quite elementary — don't be scared of the title of Goldblatt's book; the only complaints I've ever heard about it boil down to the claim that it's too easy!

You can also download this classic text on synthetic differential geometry, which is an approach to differential geometry based on infinitesimals, formalized using topos theory:

13) Anders Kock, *Synthetic Differential Geometry*, available at `http://home.imf.au.dk/kock/`

He asks that we not circulate it in printed form — electrons are okay, but not paper.

Next I want to say a *tiny* bit about Koszul duality for Lie algebras, which plays a big role in the work of Castellani on the M-theory Lie 3-algebra, which I discussed in "Week 237".

Let's start with the Maurer-Cartan form. This is a gadget that shows up in the study of Lie groups. It works like this. Suppose you have a Lie group G with Lie algebra $\mathrm{Lie}(G)$. Suppose you have a tangent vector at any point of the group G. Then you can translate it to the identity element of G and get a tangent vector at the identity of G. But, this is nothing but an element of $\mathrm{Lie}(G)$!

So, we have a god-given linear map from tangent vectors on G to the Lie algebra $\mathrm{Lie}(G)$. This is called a "$\mathrm{Lie}(G)$-valued 1-form" on G, since an ordinary 1-form eats tangent vectors and spits out numbers, while this spits out elements of $\mathrm{Lie}(G)$. This particular god-given $\mathrm{Lie}(G)$-valued 1-form on G is called the "Maurer-Cartan form", and denoted ω.

Now, we can define exterior derivatives of $\mathrm{Lie}(G)$-valued differential forms just as we can for ordinary differential forms. So, it's interesting to calculate $d\omega$ and see what it's like.

The answer is very simple. It's called the Maurer-Cartan equation:

$$d\omega = -\omega \wedge \omega$$

On the right here I'm using the wedge product of $\mathrm{Lie}(G)$-valued differential forms. This is defined just like the wedge product of ordinary differential forms, except instead of multiplication of numbers we use the bracket in our Lie algebra.

I won't prove the Maurer-Cartan equation; the proof is so easy you can even find it on the Wikipedia:

14) Wikipedia, "Maurer-Cartan form", `http://en.wikipedia.org/wiki/Maurer-Cartan_form`

An interesting thing about this equation is that it shows everything about the Lie algebra $\mathrm{Lie}(G)$ is packed into the Maurer-Cartan form. The reason is that everything about the bracket operation is packed into the definition of $\omega \wedge \omega$.

If you have trouble seeing this, note that we can feed $\omega \wedge \omega$ a pair of tangent vectors at any point of G, and it will spit out an element of $\mathrm{Lie}(G)$. How will it do this? The two copies of ω will eat the two tangent vectors and spit out elements of $\mathrm{Lie}(G)$. Then we take the bracket of those, and that's the final answer.

Since we can get the bracket of *any* two elements of $\mathrm{Lie}(G)$ using this trick, $\omega \wedge \omega$ knows everything about the bracket in $\mathrm{Lie}(G)$. You could even say it's the bracket viewed as a geometrical entity — a kind of "field" on the group G!

Now, since

$$d\omega = -\omega \wedge \omega$$

and the usual rules for exterior derivatives imply that

$$d^2\omega = 0$$

we must have

$$d(\omega \wedge \omega) = 0$$

If we work this concretely what this says, we must get some identity involving the bracket in our Lie algebra, since $\omega \wedge \omega$ is just the bracket in disguise. What identity could this be?

THE JACOBI IDENTITY!

It has to be, since the Jacobi identity says there's a way to take 3 Lie algebra elements, bracket them in a clever way, and get zero:

$$[u,[v,w]] + [v,[w,u]] + [w,[u,v]] = 0$$

while $d(\omega \wedge \omega)$ is a $\mathrm{Lie}(G)$-valued 3-form that happens to vanish, built using the bracket.

It also has to be since the equation $d^2 = 0$ is just another way of saying the Jacobi identity. For example, if you write out the explicit grungy formula for d of a differential form applied to a list of vector fields, and then use this to compute d^2 of that differential form, you'll see that to get zero you need the Jacobi identity for the Lie bracket of vector fields. Here we're just using a special case of that.

The relationship between the Jacobi identity and $d^2 = 0$ is actually very beautiful and deep. The Jacobi identity says the bracket is a derivation of itself, which is an infinitesimal way of saying that the flow generated by a vector field, acting as an operation on vector fields, preserves the Lie bracket! And this, in turn, follows from the fact that the Lie bracket is *preserved by diffeomorphisms* — in other words, it's a "canonically defined" operation on vector fields.

Similarly, $d^2 = 0$ is related to the fact that d is a natural operation on differential forms — in other words, that it commutes with diffeomorphisms. I'll leave this cryptic; I don't feel like trying to work out the details now.

Instead, let me say how to translate this fact:

$d^2\omega = 0$ *IS SECRETLY THE JACOBI IDENTITY*

into pure algebra. We'll get something called "Kozsul duality". I always found Koszul duality mysterious, until I realized it's just a generalization of the above fact.

How can we state the above fact purely algebraically, only using the Lie algebra $\mathrm{Lie}(G)$, not the group G? To get ourselves in the mood, let's call our Lie algebra simply L.

By the way we constructed it, the Maurer-Cartan form is "left-invariant", meaning it doesn't change when you translate it using maps like this:

$$\begin{aligned} Lg\colon G &\to G \\ x &\mapsto gx \end{aligned}$$

that is, left multiplication by any element g of G. So, how can we describe the left-invariant differential forms on G in a purely algebraic way? Let's do this for *ordinary* differential forms; to get $\mathrm{Lie}(G)$-valued ones we can just tensor with $L = \mathrm{Lie}(G)$.

Well, here's how we do it. The left-invariant vector fields on G are just

$$L$$

so the left-invariant 1-forms are

$$L^*$$

So, the algebra of all left-invariant diferential forms on G is just the exterior algebra on L^*. And, defining the exterior derivative of such a form is precisely the same as giving the bracket in the Lie algebra L! And, the equation $d^2 = 0$ is just the Jacobi identity in disguise.

To be a bit more formal about this, let's think of L as a graded vector space where everything is of degree zero. Then L^* is the same sort of thing, but we should *add one to the degree* to think of guys in here as 1-forms. Let's use S for the operation of "suspending" a graded vector space — that is, adding one to the degree. Then the exterior algebra on L^* is the "free graded-commutative algebra on SL^*".

So far, just new jargon. But this lets us state the observation of the penultimate paragraph in a very sophisticated-sounding way. Take a vector space L and think of it as a graded vector space where everything is of degree zero. Then:

> *Making the free graded-commutative algebra on SL^* into a* differential *graded-commutative algebra is the same as making L into a Lie algebra.*

This is a basic example of "Koszul duality". Why do we call it "duality"? Because it's still true if we switch the words "commutative" and "Lie" in the above sentence!

> *Making the free graded Lie algebra on SL^* into a* differential *graded Lie algebra is the same as making L into a commutative algebra.*

That's sort of mind-blowing. Now the equation $d^2 = 0$ secretly encodes the *commutative law*.

So, we say the concepts "Lie algebra" and "commutative algebra" are Koszul dual. Interestingly, the concept "associative algebra" is its own dual:

> *Making the free graded associative algebra on SL^* into a* differential *graded associative algebra is the same as making L into an associative algebra.*

This is the beginning of a big story, and I'll try to say more later. If you get impatient, try the book on operads mentioned in "Week 191", or else these:

15) Victor Ginzburg and Mikhail Kapranov, "Koszul duality for quadratic operads", *Duke Math. J.* **76** (1994), 203–272. Also "Erratum", *Duke Math. J.* **80** (1995), 293.

16) Benoit Fresse, "Koszul duality of operads and homology of partition posets", *Homotopy theory and its applications (Evanston, 2002)*, Contemp. Math. **346** (2004), 115–215. Also available at `http://math.univ-lille1.fr/~fresse/PartitionHomology.html`

The point is that Lie, commutative and associative algebras are all defined by "quadratic operads", and one can define for any such operad $\mathcal{O}$ a "dual" operad $\mathcal{O}^*$ such that:

> *Making the free graded $\mathcal{O}$-algebra on SL^* into a* differential *graded $\mathcal{O}$-algebra is the same as making L into an $\mathcal{O}^*$-algebra.*

And, we have $\mathcal{O}^{**} = \mathcal{O}$, hence the term "duality".

This has always seemed incredibly cool and mysterious to me. There are other meanings of the term "Koszul duality", and if really understood them I might better understand what's going on here. But, I'm feeling happy now because I see this special case:

> *Making the free graded-commutative algebra on SL^* into a* differential *graded-commutative algebra is the same as making L into a Lie algebra.*

is really just saying that the exterior derivative of left-invariant differential forms on a Lie group encodes the bracket in the Lie algebra. That's something I have a feeling for. And, it's related to the Maurer-Cartan equation... though notice, I never completely spelled out how.

Addenda: Let me say some more about how $d^2 = 0$ is related to the fact that d is a canonically defined operation on differential forms. Being "canonically defined" means that d commutes with the action of diffeomorphisms. Saying that d commutes with "small" diffeomorphisms — those connected by a path to the identity — is the same as saying

$$dL_v = L_v d$$

where v is any vector field and L_v is the corresponding "Lie derivative" operation on differential forms. But, Weil's formula says that

$$L_v = i_v d + d i_v$$

where i_v is the "interior product with v", which sends p-forms to $(p-1)$-forms. If we plug Weil's formula into the equation we're pondering, we get

$$d(i_v d + d i_v) = (i_v d + d i_v)d$$

which simplifies to give

$$d^2 i_v = i_v d^2$$

So, as soon as we know $d^2 = 0$, we know d commutes with small diffeomorphisms. Alas, I don't see how to reverse the argument.

Similarly, as soon as we know the Jacobi identity, we know the Lie bracket operation on vector fields is preserved by small diffeomorphisms, by the argument outlined in the body of this Week. This argument is reversable.

So, maybe it's an exaggeration to say that $d^2 = 0$ and the Jacobi identity say that d and the Lie bracket are preserved by diffeomorphisms — but at least they *imply* these operations are preserved by *small* diffeomorphisms.

Week 239

September 8, 2006

David Corfield, Urs Schreiber and I have started up a new blog!

David is a philosopher, Urs is a physicist, and I'm a mathematician, but one thing we all share is a fondness for n-categories. We also like to sit around and talk shop in a public place where our friends can drop by. Hence the title of our blog:

1) *The n-Category Caf*, http://golem.ph.utexas.edu/category

Technologically speaking, the cool thing about this blog is that it uses itex and MathML to let us (and you) write pretty equations in TeX. For this we thank Jacques Distler, who pioneered the technology on his own blog:

2) Jacques Distler, Musings, `http://golem.ph.utexas.edu/~distler/blog/`

It's great!

Urs began by posting about 11d supergravity and higher gauge theory (see "Week 237"). Now he's discussing Barrett and Connes' new work on the Standard Model. Meanwhile, I've been obsessed with the categorical semantics of quantum computation, and David has been running discussions on categorifying Klein's Erlangen program (see "Week 213"), the differences between mathematicians and historians when it comes to writing histories of math, and so on.

And, it's all free.

Meanwhile, in the bad old world of extortionist math publishers, we see a gleam of hope. The entire editorial board of the journal Topology resigned to protest Reed-Elsevier's high prices!

3) Topology board of editors, letter of resignation, `http://math.ucr.edu/home/baez/topology-letter.pdf`

The board includes some topologists I respect immensely. It takes some guts for full-fledged memmbers of the math establishment to do something like this, and I congratulate them for it. It'll be fun to see what stooges Reed-Elsevier rounds up to form a new board of editors. I can't imagine they'll just declare defeat and let the journal fold.

This is part of trend where journal editors "declare independence" from their publishers and move toward open access:

4) Open Access News, "Journal declarations of independence", `http://www.earlham.edu/%7Epeters/fos/lists.htm#declarations`

Speaking of open access, you can now get the notes from the course Freeman Dyson taught on quantum electrodynamics when he first became a professor of physics at Cornell:

5) Freeman J. Dyson, *1951 Lectures on Advanced Quantum Mechanics*, second edition, available as `quant-ph/0608140`. For historical context and original mimeographs, see `http://hrst.mit.edu/hrs/renormalization/dyson51-intro/`

These notes are from an exciting period in physics, shortly after the 1947 Shelter Island conference where Feynman and Schwinger presented their approaches to quantum electrodynamics to an audience of luminaries including Bohr, Oppenheimer, von Neumann, and Weisskopf. Nobody understood Feynman's diagrams except Schwinger and maybe Feynman's thesis advisor, John Wheeler.

Every true fan of physics loves reading about this heroic era and its figures, especially Feynman. So, if you haven't read these yet, run to the bookstore and buy them now!

6) James Gleick, *Genius: the Life and Science of Richard Feynman*, Vintage Press, 1993.

7) Jagdish Mehra, *The Beat of a Different Drum: the Life and Science of Richard Feynman*, Oxford U. Press, 1996.

8) Silvan S. Schweber, *QED and the Men Who Made It*, Princeton U. Press, Princeton, 1994.

The first book is a barrel of fun but doesn't get into the nitty-gritty details of Feynman's work. The second more scholarly treatment also has lots of Feynman anecdotes — even some new ones! But, it covers his work in enough detail to intimidate any non-physicist. The third offers a broader panorama of the development of quantum electrodynamics. Taken together, they add up to quite a nice story.

Of course, I'm *assuming* you've read these:

9) Richard P. Feynman, *Surely You're Joking, Mr. Feynman! (Adventures of a Curious Character)*, W. W. Norton and Company, New York, 1997.

10) Richard P. Feynman, *What Do* You *Care What Other People Think? (Further Adventures of a Curious Character)*, W. W. Norton and Company, New York, 2001.

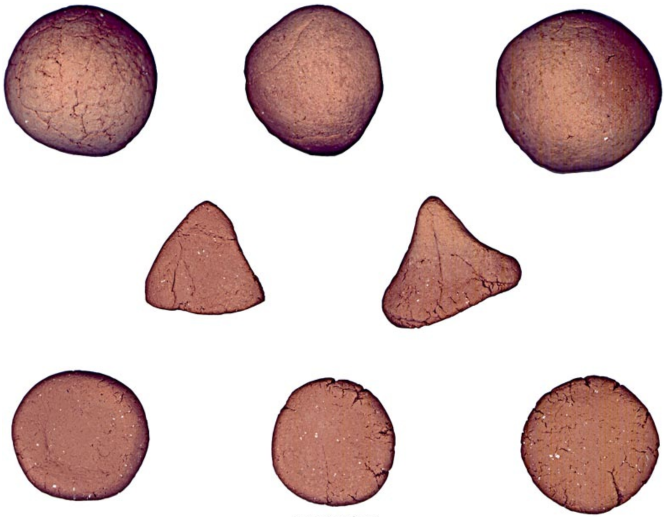
They're more fun than everything else I've ever recommended on This Week's Finds, combined. If you haven't read them, don't just *run* to the nearest bookstore — get in a time machine, go back, and make sure you *did* read them.

Today I'd like to wrap up the discussion of Koszul duality which I began last Week. As we'll see, this gives a really efficient way of categorifying the theory of Lie algebras and defining "Lie n-algebras". And, as Urs Schreiber notes, these seem to be just what we need to understand 11-dimensional supergravity in a nice geometric way.

But before I dive into this heavy stuff, something fun. Thanks to Christine Dantas' blog, I just saw a webpage on the origins of math and writing in Mesopotamia:

11) Duncan J. Melville, "Tokens: the origin of mathematics", from his website Mesopotamian Mathematics, `http://it.stlawu.edu/%7Edmelvill/mesomath/`

Before people in the Near East wrote on clay tablets, there were "tokens":


MS 5067/1-8
Neolithic plain counting tokens. Near East, ca. 8000--3500 BC

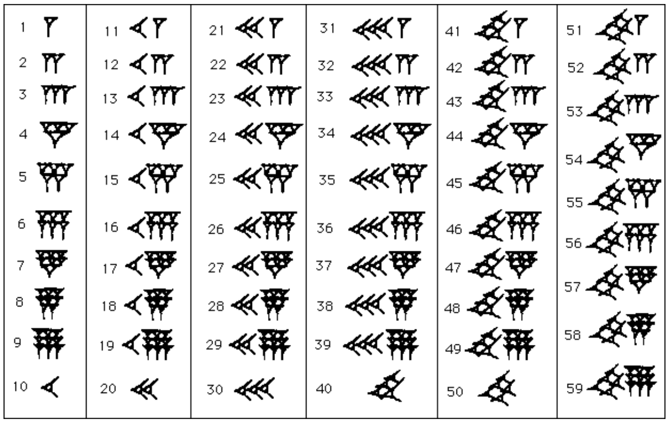
12) The Schyen Collection, MS 5067/1-8, "Neolithic plain counting tokens possibly representing 1 measure of grain, 1 animal and 1 man or 1 day's labour, respectively", `http://www.nb.no/baser/schoyen/5/5.11/index.html`

These are little geometric clay figures that represented things like sheep, jars of oil, and various amounts of grain. They are found throughout the Near East starting with the agricultural revolution in about 8000 BC. Apparently they were used for contracts!

Eventually groups of tokens were sealed in clay envelopes, so any attempt to tamper with them would be visible.

But, it's annoying to have to break a clay envelope just to see what's in it. So, after a while, they started marking the envelopes to say what was inside. At first, they did this simply by pressing the tokens into the soft clay of the envelopes.

Later, these marks were simply drawn on tablets. Eventually they gave up on the tokens — a triumph of convenience over security. The marks on tablets then developed into the Babylonian number system! The transformation was complete by 3000 BC.



So, five millennia of gradual abstraction led to the writing of numbers! From three tokens representing jars of oil, we eventually reach the abstract number "3" applicable to anything.

Of course, all history is detective work. The story I just told is an interpretation of archaeological evidence. It could be wrong. This particular interpretation is due to Denise Schmandt-Besserat. It seems to be fairly well accepted in broad outline, but scholars are still arguing about it.

For more on her ideas, try this:

13) Denise Schmandt-Besserat, "Accounting with tokens in the ancient Near East", `http://www.utexas.edu/cola/centers/lrc/numerals/dsb/dsb.html`

For a bibliography of her many papers, try:

14) Denise Schmandt-Besserat, "Publications", `http://www.utexas.edu/cola/centers/lrc/iedocctr/ie-pubs/dsb-pubs.html`

For more work on this subject — I want to read more! — try:

15) Eleanor Robson, "Bibliography of Mesopotamian mathematics", `http://it.stlawu.edu/˜dmelvill/mesomath/erbiblio.html`

And for a fun intro to writing on clay tablets, try this:

16) John Heise, "Cuneiform writing system", `http://xoomer.alice.it/bxpoma/akkadeng/cuneiform.htm`

Next, from 8000 BC, let's shoot forward ten millennia straight into the 20th century. Last week I gave three examples of Koszul duality:

> *Making the free graded-commutative algebra on SL^* into a differential graded-commutative algebra is the same as making L into a Lie algebra.*

> *Making the free graded Lie algebra on SL^* into a differential graded Lie algebra is the same as making L into a commutative algebra.*
>
> *Making the free graded associative algebra on SL^* into a differential graded associative algebra is the same as making L into an associative algebra.*

Here L is a vector space, which we think of as a graded vector space concentrated in degree zero. L^* is its dual, and SL^* is the "shifted" or "suspended" version of L^*, where we add one to the degree of everything.

Now, what if we replace L by a graded vector space that can have stuff of any degree? We get a fancier version of Koszul duality, which goes like this:

> *Making the free graded-commutative algebra on SL^* into a differential graded-commutative algebra is the same as making L into an L_∞-algebra.*
>
> *Making the free graded Lie algebra on SL^* into a differential graded Lie algebra is the same as making L into a C_∞-algebra.*
>
> *Making the free graded associative algebra on SL^* into a differential graded associative algebra is the same as making L into an A_∞-algebra.*

Here an "L_∞-algebra" is a chain complex that's like a Lie algebra, except the Jacobi identity holds up to a chain homotopy called the "Jacobiator", which in turn satisfies its own identity up to a chain homotopy called the "Jacobiatorator", and so on ad infinitum. Keeping track of all these higher homotopies is quite a chore. Well, it's sort of fun when you get into it, but the great thing about Koszul duality is that you don't need to remember any fancy formulas: all the higher homotopies are packed into the *differential* on SL^*.

Similarly, a "C_∞-algebra" is a chain complex that's like a graded-commutative algebra up to homotopy, ad infinitum.

Similarly, an "A_∞-algebra" is a chain complex that's like an associative algebra up to homotopy, ad infinitum. Here you can read off all the higher homotopies from the Stasheff associahedra, which you know and love from "Week 144" — but again, Koszul duality means you don't have to!

As mentioned last week, all this stuff generalizes to any kind of algebraic gadget in Vect — the category of vector spaces — which is defined by a "quadratic operad" $\mathcal{O}$. Any such operad has a "Koszul dual" operad $\mathcal{O}^*$ such that:

> *Making the free graded $\mathcal{O}^*$-algebra on SL^* into a differential graded $\mathcal{O}^*$-algebra is the same as making L into an $\mathcal{O}_\infty$-algebra.*

Here $\mathcal{O}_\infty$ is an operad in the category of chain complexes defined by "weakening" $\mathcal{O}$ in a systematic way — replacing all the laws by chain homotopies, ad infinitum. We can define $\mathcal{O}_\infty$ using the "bar construction", as nicely described here:

16) Todd Trimble, "Bar constructions", `http://math.ucr.edu/home/baez/trimble/bar.html`

or in the book by Markl, Schnider and Stasheff:

17) Martin Markl, Steve Schnider and Jim Stasheff, *Operads in Algebra, Topology and Physics*, AMS, Providence, Rhode Island, 2002.

See "Week 191" for more on this book, and what the heck an "operad" is. By the way, we have

$$\mathcal{O}^{**} \cong \mathcal{O}$$

so we can also say:

> *Making the free graded $\mathcal{O}$-algebra on SL^* into a differential graded $\mathcal{O}$-algebra is the same as making L into an $\mathcal{O}_\infty^*$-algebra.*

Anyway, I don't have much intuition for how Koszul duality lets us magically sidestep the bar construction of $\mathcal{O}_\infty$. Someday I hope I'll understand this.

But, once we have the concept of "L_∞-algebra", we can restrict ourselves to chain complexes that vanish except for their first n terms — that is, degrees $0, 1, \ldots, n-1$ — and get the concept of "Lie n-algebra".

In fact, a Lie n-algebra is like a hybrid of a Lie algebra and an n-category! The definition I just gave says a Lie n-algebra is an L_∞-algebra which as a chain complex vanishes above degree $n-1$. But, such chain complexes are equivalent to strict n-category objects in Vect! So, we can think of Lie n-algebras as strict n-categories that do their best to act like Lie algebras, but with some of the laws holding up to isomorphism, with these isomorphisms satisfying their own laws up to isomorphism, etcetera.

But, the really cool part is that we can do *gauge theory* using Lie n-algebras instead of Lie algebras, and taking $n = 3$ we get an example that seems to explain the geometry of 11d supergravity... that is, the classical limit of that mysterious thing called M-theory.

For this, you really need to read Urs Schreiber's stuff:

18) Urs Schreiber, "Castellani on free differential algebras in supergravity: gauge 3-group of M-theory", `http://golem.ph.utexas.edu/string/archives/000840.html`

19) Urs Schreiber, "SuGra 3-connection reloaded", `http://golem.ph.utexas.edu/category/2006/08/sugra_3connection_reloaded.html`

and many other things he's been writing on the n-Category Caf lately.

Addenda: You can see discussion of this Week's Finds at the n-Category Caf. In his blog Not Even Wrong, Peter Woit has more to say about the open access movement and a questionable plan broached by CERN to pay for-profit journals to make their papers freely available. Some comments on this blog article dig deeper into the evolution of Babylonian numerals:

19) John Baez and Richard Elwes, "Babylon and the square root of 2", Azimuth, December 2nd, 2011.

In particular, Duncan Melville points out that when number systems first evolved in Babylonia, they had about a dozen *different* systems for different kinds of products! A base-60 system called the S system, was used to count most discrete objects, such as sheep or people. For 'rations' such as cheese or fish, they used a base 120 system, the B system. Another system, the E system, was used to measure quantities of grain, and so on. So, number systems were a bit more like business software today, with different kinds used by different trades.

> I never once doubted that I would eventually succeed in getting to the bottom of things.
>
> — *Alexander Grothendieck*

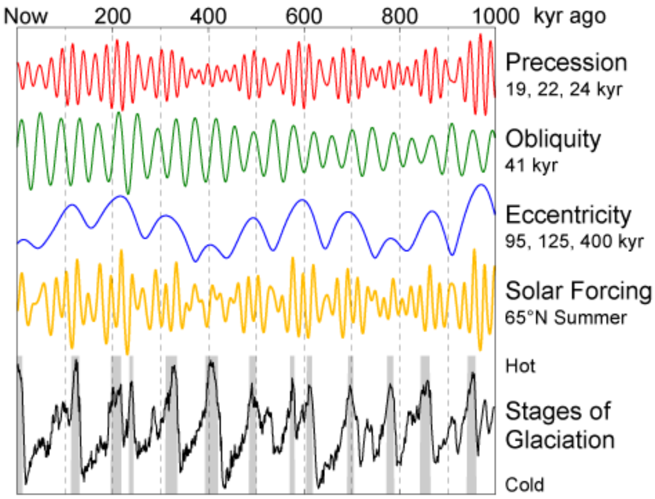
Week 240

October 22, 2006

I'm back from Shanghai, and classes are well underway now. For the last few weeks I'd been frantically preparing a talk for Stewart Brand's "Seminars About Long-Term Thinking", up in San Francisco. I talked about how we need to "zoom out" of our short-term perspective to understand the history of the earth's climate and what we're doing to it now:

1) John Baez, "Zooming out in time", `http://math.ucr.edu/home/baez/zoom/`

There's a lot of tricky physics in this business. Consider, for example, this graph of cycles governing the Earth's precession, the obliquity of its orbit, and the eccentricity of its orbit:



2) Wikipedia, "Milankovitch cycles", `http://en.wikipedia.org/wiki/Milankovitch_cycles`

Here a "kyr" is a thousand years. The yellow curve combines information from all three of these cycles and shows the amount of solar radiation at 65 degrees north latitude. The bottom black curve shows the amount of glaciation. As Milankovitch's theory predicts, you can sort of see a correlation between the yellow and black curves — but it's nothing simple or obvious. One reason is the complex feedback mechanisms within the Earth's climate.

Here's a great place to read about this stuff:

3) Barry Saltzman, *Dynamical Paleoclimatology: Generalized Theory of Global Climate Change*, Academic Press, New York, 2002.

Anyway, now this talk is done, and I can focus more on teaching.

In my seminar this year, we're focusing on two topics: quantization and cohomology, and classical versus quantum computation. I'm trying out something new: not only are the notes available on the web, there's also a blog entry for each class, where you can ask questions, make comments and correct my mistakes!

4) John Baez, Fall 2006 seminars: "Quantization and cohomology", and "Classical versus quantum computation". Notes by Derek Wise, homeworks and blog entries available at `http://math.ucr.edu/home/baez/qg-fall2006/`

I hope more people blend teaching with blogging. It's not too much work if someone with legible handwriting takes notes and the lectures can actually be followed from the notes. You can use blogging to interactively teach people scattered all over the planet!

This week, James Dolan gave a talk on something he's been working on for a long time: games and cartesian closed categories. Lately he's been working with Todd Trimble, and they reproved some important results in a fun new way. Let me sketch the ideas for you. . . .

Let's play a game. I have a set X in my pocket, and I'm not telling you what it is. Can you pick an element of X in a systematic way?

No, of course not: you don't have enough information. X could even be empty, in which case you're clearly doomed! But even if it's nonempty, if you don't know anything about it, you can't pick an element in a systematic way.

So, you lose.

Okay, let's play another game. Can you pick an element of

$$X^X$$

in a systematic way? Here A^B means the set of functions from B to A. So, I'm asking if you can pick a function from X to itself in a systematic way.

Yes! You can pick the identity function! This sends each element of X to itself:

$$x \mapsto x$$

You don't need to know anything about X to describe this function. X can even be empty.

So, you win.

Are there any other ways to win? No.

Now let's play another game. Can you pick an element of

$$X^{X^X}$$

in a systematic way?

An element in here takes functions from X to itself and turns them into elements of X. When X is the set of real numbers, people call this sort of thing a "functional", so let's use that term. A functional eats functions and spits out elements.

You can scratch your head for a while trying to dream up a systematic way to pick a functional for any set X. But, there's no way.

So, you lose.

Let's play another game. Can you pick an element of

$$(X^X)^{X^X}$$

in a systematic way?

An element in here eats functions and spits out functions. When X is the set of real numbers, people often call this sort of thing an "operator", so let's use that term.

Given an unknown set X, can you pick an operator in a systematic way? Sure! You can pick the identity operator. This operator eats any function from X to itself and spits out the same function:

$$f \mapsto f$$

Anyway: you win.

Are there any other ways to win? Yes! There's an operator that takes any function and spits out the identity function:

$$f \mapsto (x \mapsto x)$$

This is a bit funny-looking, but I hope you get what it means: you put in any function f, and out pops the identity function $x \mapsto x$.

This arrow notation is very powerful. It's usually called the "λ calculus", since when Church invented it in the 1930s, he wrote it using the Greek letter λ instead of an arrow: instead of

$$x \mapsto y$$

he wrote

$$\lambda x.y$$

But this just makes things more confusing, so let's not do it.

Are there more ways to win this game? Yes! There's also an operator called "squaring", which takes any function f from X to itself and "squares" it — in other words, does it twice. If we write the result as f^2, this operator is

$$f \mapsto f^2$$

But, we can express this operator without using any special symbol for squaring. The function f is the same as the function

$$x \mapsto f(x)$$

so the function f^2 is the same as

$$x \mapsto f(f(x))$$

and the operator "squaring" is the same as

$$f \mapsto (x \mapsto f(f(x)))$$

This looks pretty complicated. But, it shows that our systematic way of choosing an element of

$$(X^X)^{X^X}$$

can still be expressed using just the λ calculus.

Now that you know "squaring" is a way to win this particular game, you'll immediately guess a bunch of other ways: "cubing", and so on. It turns out all the winning strategies are of this form! We can list them all using the λ calculus:

$$\begin{aligned}
&f \mapsto (x \mapsto x) \\
&f \mapsto (x \mapsto f(x)) \\
&f \mapsto (x \mapsto f(f(x))) \\
&f \mapsto (x \mapsto f(f(f(x))))
\end{aligned}$$

etc. Note that the second one is just a longer name for the identity operator. The longer name makes the pattern clear.

So far, all these methods of picking an element of $(X^X)^{X^X}$ for an unknown set X can be written using the λ calculus. There are other sneakier ways. For example, there's the operator that sends functions with fixed points to the identity function, and sends functions without fixed points to themselves. It's an interesting challenge to figure out all these sneaky ways, but it's way too hard for me. So, from now on, just to keep things simple, let's only consider "systematic ways" that can be expressed using the λ calculus. To win one of my games, you need to use the λ calculus to pick an element of the set I write down.

So, let's play another game. Can you write down an element of

$$X^{X^{X^X}}$$

using the λ calculus?

An element in here eats functionals and spits out elements of X. So, it's called a "functionalal" on X. At least that's what Jim calls it.

If I have an unknown set in my pocket, can you write down functionalal on this set using the λ calculus?

Yes! You need to dream up a recipe that takes functionals on X and turns them into elements of X. Here's one recipe: take any functional and evaluate it on the *identity* function, getting an element of x.

In λ calculus notation, this recipe looks like this:

$$f \mapsto f(x \mapsto x)$$

Can you think of other ways to win this game? I hope so: there are infinitely many! Jim and Todd figured out a systematic way to list them all.

Now let's play another game. Can you write down an element of

$$X^{X^{X^{X^X}}}$$

using the λ calculus? A thing in here eats functionalals and spits out elements of X, so it's called a "functionalalal".

So, can you write down a functionalalal on an unknown set using the λ calculus?

The answer is no: you lose.

How about writing down an element of

$$((X^X)^{X^X})^{((X^X)^{X^X})}$$

using the λ calculus? Such a thing eats operators and spits out operators, so it's called an "operatorator".

The answer is yes: there are lots of ways to win this game. The real challenge is listing all of them! This is the sort of question Dolan and Trimble figured out the answer to — though as we'll see, they weren't the first.

In fact, instead of moving on to functionalators, operatorals, operatoralatorals, and so on, let me just tell you trick for instantly deciding which of all these games you can win.

You just take your game, like this:

$$((X^X)^{X^X})^{((X^X)^{X^X})}$$

and evaluate it by setting $X = 0$. If you get 0, there's no way to win. If you get 1, there's at least one way to win.

To use this trick, you need to know that

$$0^0 = 1$$

This is something they don't teach in school! In analysis, X^Y can approach anything between 0 and 1 when X and Y approach 0 from above. So, teachers like to say 0^0 is undefined. But X^X approaches 1 when $X \to 0$. More importantly, in set theory, A^B stands for the set of functions from B to A, and the number of elements in this set is

$$|A^B| = |A|^{|B|}$$

When A and B are empty, there's just one function from B to A, namely the identity. So, for our purposes we should define $0^0 = 1$.

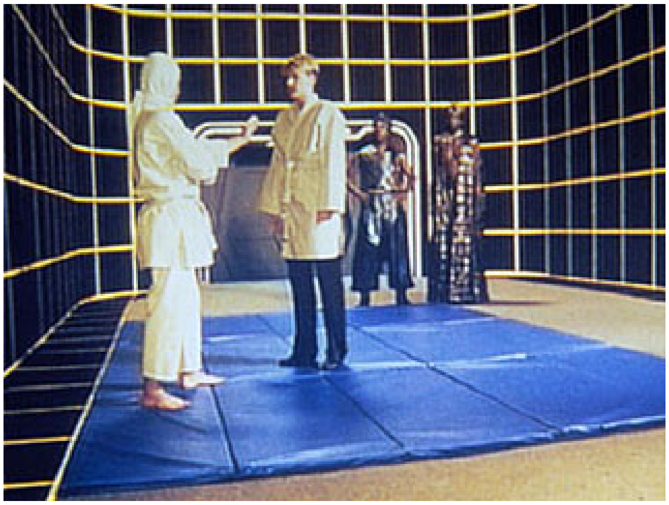
Consider the case of functionals, which are elements of X^{X^X}. If we evaluate this at $X = 0$ we get

$$0^{0^0} = 0^1 = 0$$

So, there are no functionals when X is the empty set. So, you can't pick a functional on a unknown set in *any* systematic way. That's why you lose when your game evaluates to 0. It's more interesting to prove that for games evaluating to 1, there's a way to win, using the λ calculus.

But we'd really like to understand *all* the ways to win using the λ calculus. And for this, Dolan and Trimble used the theory of holodeck games.

In Star Trek, the "holodeck" is a virtual reality environment where you can play various games:



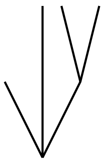
On the holodeck, if you regret a move you made, you can back up to any earlier point in the game and make a new move.

Actually I'm deviating from the technical specifications of the holodeck on Star Trek, as explained here:

6) Wikipedia, "Holodeck", `http://en.wikipedia.org/wiki/Holodeck`

So, if you're a Star Trek purist, it's better to imagine a video game where you can save your game at any state of play, and go back to these saved games whenever you want. And, you have to imagine being so paranoid that you *always* save your game before making a move. This allows games to go on forever, so we only say you have a winning strategy if you can win in a finite number of moves, no matter what the other player does.

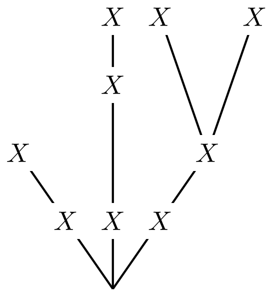
To make this completely precise, we consider two-player games where the players take turns making moves. When a player can't make a move, they lose. Any such game can be written as a "game tree", like this:



In this example, the first player has three choices for her first move. If she picks the middle branch, the second player has one choice for his first move. Then the first player has one choice for her second move. Then the second player has no choice for his second move — so he loses.

So, in this particular example the second player has no winning strategy.

A cool thing about such a game is that we can take its game tree and turn it into an expression built from some variable X using products and exponentials. To do this, just put an X at each vertex of the tree except the root:



Then blow on the tree with a strong westerly wind, so strong that the branches blow away and only the X's are left:

$$X^X \quad X^{X^X} \quad X^{X^{XX}}$$

This is just a way of writing an expression built from X using products and exponentials:

$$X^X X^{X^X} X^{X^{XX}}$$

Conversely, any such expression can be turned back into a tree, at least after we simplify it using these rules:

$$(AB)^C = A^C B^C$$
$$(A^B)^C = A^{BC}$$

For example, consider the set of operators:

$$(X^X)^{X^X}$$

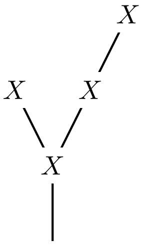
If we simplify this, we get

$$X^{XX^X}$$

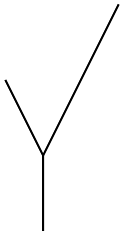
or

$$X^{XX^{X}}$$

giving the tree



or in other words



And here's a cool fact: if you take any expression built from X using products and exponentials, and evaluate it at $X = 0$, you can tell which player has a winning strategy for the game described by the corresponding tree! If you get 1, the second player has a winning strategy; if you get 0, they don't.

It's pretty easy to prove: try it.

But if you've been paying attention, you'll have noticed something weird.

I've told you *two* ways to get a game from any expression built from X using products and exponentials. First, the game of defining an element of the resulting set, using the λ calculus. Second, the game we get by turning this expression into a game tree, like I just did.

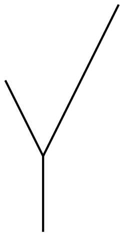
For *both* these games, you can decide if there's a winning strategy by evaluating the expression at $X = 0$.

But are they the same game? No! One is the holodeck version of the other!

Let's look at the familiar example of operators:

$$(X^X)^{X^X} = X^{XX^X}$$

This evaluates to 1 at $X = 0$. So, if we turn it into a tree



we get a game where the second player has a winning strategy.

This game is not very exciting, but it becomes more exciting if you call it "The Lady or the Tiger". In this game, the first player has only one first move: he takes the second player to a room with two doors, corresponding to the two branches of the above tree.

Then it's the second player's turn.

If he opens the left door, a beautiful lady pops out and they instantly get married and live happily ever after. If he opens the right door, the first player opens a tiger cage. Then the tiger jumps out and eats the second player.

In this game, the second player has just *one* winning strategy: on his first move he should choose the left door.

Next look at the game of defining an element of

$$(X^X)^{X^X} = X^{XX^X}$$

using the λ calculus. We've seen there are *infinitely many* strategies for winning this:

$$\begin{aligned} &f \mapsto (x \mapsto x) \\ &f \mapsto (x \mapsto f(x)) \\ &f \mapsto (x \mapsto f(f(x))) \\ &f \mapsto (x \mapsto f(f(f(x)))) \end{aligned}$$

and so on. These correspond to 2nd-player winning strategies for the *holodeck version* of The Lady or the Tiger.

What are these strategies?

One is just to play the game and win by choosing the left door.

Another is to choose the right door — and then, just when the tiger is about to eat you, back up and choose the left door!

Another is to choose the right door — and then, just when the tiger is about to eat you, back up and choose... the right door!

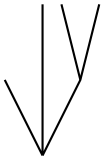
Then, when the tiger is about to devour you again, back up again, and this time choose the left door.

And so on: for each n, there's a strategy where you choose the right door n times before wising up and choosing the left door.

Now, if you want a really nice math project, ponder the pattern relating all these strategies to the corresponding λ calculus expressions:

$$
\begin{aligned}
&f \mapsto (x \mapsto x)\\
&f \mapsto (x \mapsto f(x))\\
&f \mapsto (x \mapsto f(f(x)))\\
&f \mapsto (x \mapsto f(f(f(x))))
\end{aligned}
$$

Then, figure out how to prove that for *any* 2-person game, say:



there's a 1-1 correspondence between winning second-person strategies for the holodeck verson of this game and ways of using the λ calculus to define elements of the corresponding set:

$$X^X X^{X^X} X^{X^{X X}}$$

Apparently this result goes back to work of Hyland and Ong in the early 1990s. Dolan rediscovered the idea, and Trimble and he have recently worked out a new proof.

If you get stuck proving this result yourself, first try these notes from Dolan's talk, for some hints:

7) James Dolan, "Holodeck strategies and cartesian closed categories", lecture at UCR, notes by John Baez, Oct. 19, 2006, available at `http://math.ucr.edu/home/baez/qg-fall2006/f06week03b.pdf`

Then try Trimble's more rigorous, technical treatment, and the original paper by Hyland and Ong:

8) Todd Trimble, "Holodeck games and CCCs", available at `http://math.ucr.edu/home/baez/trimble/holodeck.html`

9) Martin Hyland and C.-H. Luke Ong, "On full abstraction for PCF", *Information and Computation* **163** (2000), 285–408. Also available at `ftp://ftp.comlab.ox.ac.uk/pub/Documents/techpapers/Luke.Ong/pcf.ps.gz`

Dolan's talk also explains some other fun stuff, like how to multiply and exponentiate games. So, if you read these notes, you'll learn how to play

$$\text{chess} \times \text{go}$$

and

$$\text{chess}^{\text{go}}$$

at least after chess and go have been "improved" so games never last forever and the last player able to make a move wins.

But, if you're planning to study this stuff, I'd better admit right now that Dolan and Trimble make heavy use of the relation between the λ calculus and cartesian closed categories.

A category is "cartesian" if it has finite products — or in other words, binary products and a terminal object. It's "cartesian closed" if it also has exponentials. All these terms are carefully defined in the week 2 and week 3 notes of my classical versus quantum computation course, so let me just illustrate them with an example: the category of sets. Here the product $A \times B$ of two sets A and B is their usual Cartesian product. The exponential A^B is the set of functions from B to A. Any 1-element set is a terminal object.

Dolan and Trimble don't really talk about an unknown set X, as I did above. What they really study is the "free cartesian closed category on one object x", which I like to call $\mathrm{CCC}[x]$. Any object in $\mathrm{CCC}[x]$ is built from the object x by means of binary products, exponentials and the terminal object. For example, we have objects like this:

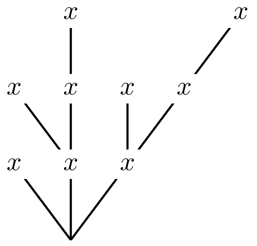
$$x^1 1^{x^x} (xx)^{x1x^{x^x}}$$

where I've omitted the times symbols for products.

However, every object is isomorphic to one in "tree form". For example, the above object is isomorphic to

$$xx^{xx^{x^x}} x^{xx^{x^x}}$$

which we can draw as a tree:



Dolan and Trimble consider the set of elements of any object in $\mathrm{CCC}[x]$, where an "element" is a morphism from the terminal object, e.g.

$$f\colon 1 \to xx^{xx^{x^x}} x^{xx^{x^x}}$$

And, they show these elements are in 1-1 correspondence with second-player winning strategies for the holodeck version of the game whose tree is constructed as above.

If we pick any set X, the universal property of $\mathrm{CCC}[x]$ gives a functor

$$F\colon \mathrm{CCC}[x] \to \mathsf{Set}$$

This maps elements of any object in $\mathrm{CCC}[x]$ to elements of the corresponding object in Set:

$$F(f)\colon 1 \to XX^{XX^{X^X}} X^{XX^{X^X}}$$

So, the element f gives a systematic way of picking elements of any set built from any arbitrary set X using finite products and exponentials.

By the way, in a cartesian closed category, there's a 1-1 correspondence between morphisms

$$f\colon B \to A$$

and elements

$$f\colon 1 \to A^B$$

So, one can use games to describe *all* the objects and morphisms in the free cartesian closed category on one object! One can also describe *composition* of morphisms using games. In short, there's a complete description of $\mathrm{CCC}[x]$ in terms of games.

Now let me give you some references on cartesian closed categories, the λ calculus, categorical semantics, and games. It's an interesting network of subjects.

Categorical semantics was born in Lawvere's celebrated 1963 thesis on algebraic theories:

10) F. William Lawvere, *Functorial Semantics of Algebraic Theories*, Dissertation, Columbia University, 1963. Also available at `http://www.tac.mta.ca/tac/reprints/articles/5/tr5abs.html`

Semantics deals with theories and their models. Dual to the concept of semantics is the concept of "syntax", which deals with proofs. In the case of algebraic theories, the syntax was studied before Lawvere in the subject called "universal algebra":

11) Stanley Burris and H.P. Sankappanavar, "A Course in Universal Algebra", available at `http://www.math.uwaterloo.ca/~snburris/htdocs/ualg.html`

Lawvere modernized universal algebra by realizing that an algebraic theory is just a cartesian category, and a model is a product-preserving functor from this theory into Set or some other cartesian category — hence his thesis title, "Functorial Semantics". I explained this in much more detail back in week200.

The relevance of all this to computer science becomes visible when we note that a proof in universal algebra can be seen as a rudimentary form of computation. The "input" of the computation is a set of assumptions, while the "output" is the equation to be proved.

Treating proofs as computations may seem strained, but it becomes less so when we move to richer formalisms which allow for more complex logical reasoning. One of best-known of these is the λ calculus, invented by Church and Kleene in the 1930s as a model of computation. Any function computable by the λ calculus is also computable by a Turing machine, and according to the Church-Turing thesis these are all the functions computable by any sort of systematic process. Moreover, computations in the λ calculus can actually be seen as proofs.

The usefulness of this way of thinking was brought out in Landin's classic paper:

12) P. Landin, "A correspondence between ALGOL 60 and Church's λ-notation", *Comm. ACM* **8** (1965), 89–101, 158–165.

This began a long and fruitful line of research — see for example this:

13) H. Barendregt, *The Lambda Calculus, its Syntax and Semantics*, North-Holland, 1984.

The power of the λ calculus is evident in the textbook developed for MIT's introductory course in computer science, which is available online:

14) H. Abelson, G. J. Sussman and J. Sussman, *Structure and Interpretation of Computer Programs*, available at `http://www-mitpress.mit.edu/sicp/`

It cites pioneers like Haskell Curry, and it even has a big "λ" on the cover!

Students call it "the wizard book", because the cover also features a picture of a wizard. It's used at over 100 colleges and universities, and it has spawned a semi-mythical secret society called The Knights of the Lambda Calculus, whose self-referential emblem celebrates the ability of the λ calculus to do recursion.

In 1980, Lambek made a great discovery:

15) Joachim Lambek, "From lambda calculus to Cartesian closed categories", in *To H. B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism*, eds. J. P. Seldin and J. Hindley, Academic Press, 1980, pp. 376-402.

He showed that just as algebraic theories can be regarded as cartesian categories, theories formulated in the λ calculus can be regarded as cartesian closed categories (or CCCs, for short).

Lambek's discovery introduced a semantics for the λ calculus, since it lets us to speak of "models" of theories formulated in the λ calculus, just as we could for algebraic theories. In computer programming, the importance of a model is that it gives a picture of what a program actually accomplishes. A model in the category of sets, for example, sends any program to an actual function between sets.

There's no way to list all the interesting references to CCCs and the λ-calculus, but here are some online places to get going on them, starting out easy and working up to the harder ones. This Wikipedia article is quite good:

16) Wikipedia, "Lambda calculus", available at `http://en.wikipedia.org/wiki/Lambda_calculus`.

These blog entries by Mark Chu-Carroll are *lots* of fun — just the kind of readable, informal exposition I aspire to:

17) Mark Chu-Carroll, "Lambda calculus", available at `http://goodmath.blogspot.com/2006/06/lamda-calculus-index.html`

Mark Chu-Carroll, "Category theory", available at `http://scienceblogs.com/goodmath/goodmath/category_theory/`

These go deeper:

18) Peter Selinger, "Lecture notes on the lambda calculus", available at `http://www.mscs.dal.ca/~selinger/papers.html#lambdanotes`

and deeper:

19) Phil Scott, "Some aspects of categories in computer science", available at http://www.site.uottawa.ca/~phil/papers/handbook.ps

and here's a classic:

20) Joachim Lambek and Phil Scott, *Introduction to Higher Order Categorical Logic*, volume **7** of Cambridge Studies in Advanced Mathematics, Cambridge U. Press, 1986.

Dolan and Trimble are far from the first to study the relation between games and categories. In the 1970s, Conway invented a wonderful theory of games and surreal numbers:

21) John H. Conway, *On Numbers and Games*, Academic Press, New York, 1976. Second edition: A. K. Peters, Wellesley, Massachusetts, 2001.

22) Elwyn Berlekamp, John H. Conway, Richard Guy, *Winning Ways*, vols. 1-2, Aadmic Press, New York, 1982. Second edition, vols. 1-4, A. K. Peters, Wellelsey, Massachusetts, 2001-2004.

23) Dierk Schleicher and Michael Stoll, "An introduction to Conway's games and numbers", available as math.CO/0410026.

In 1977, Joyal modified Conway's work a bit and related it explicitly to category theory:

24) Andr Joyal, "Remarques sur la theorie des jeux a deux personnes", *Gazette des Sciences Mathematiques du Quebec*, Vol **I** no 4 (1977), 46–52.

For an online version in English, try:

25) Andr Joyal, trans. Robin Houston, "Remarks on the theory of two-person games", 2003. Available at http://www.ma.man.ac.uk/~rhouston/Joyal-games.ps

I don't know the subsequent history very well — I'm no expert on any of this stuff! — but by 1990 Martin Hyland was giving lectures on Conway games and logic. In 1992, Andreas Blass published an influential paper on "game semantics" for logic, where propositions are interpreted as games and winning strategies are proofs:

26) Andreas R. Blass, "Game semantics and linear logic", *Annals of Pure and Applied Logic* **56** (1992), 183–220.

Then came these important papers:

27) Samson Abramsky and Radha Jagadeesan, "Games and full completeness for multiplicative linear logic", *Journal of Symbolic Logic* **59** (1994), 543–574. Also available at http://citeseer.ist.psu.edu/564168.html

28) Martin Hyland and C.-H. Luke Ong, "Fair games and full completeness for multiplicative linear logic without the MIX-rule", available at http://citeseer.ist.psu.edu/hyland93fair.html

According to Samson Abramsky,

> *After these results, it was clear that the most notorious issue in programming language semantics, the "full abstraction problem for PCF", was in range. Remarkably enough, two different teams:*
>
> - *Abramsky, Jagadeesan and Malacaria*
> - *Hyland and Ong*
>
> *produced really quite different constructions which yielded in the end the same result: a synthetic construction of the fully abstract model. (The technical issue in both cases was how to accomodate the linear exponentials, i.e. the possibility to copy and delete inputs to functions. It turned out there are two very different approaches which can be taken. The HO approach (also independently found by Hanno Nickau, incidentally) is quite related to the ideas of the Lorenzen school, but, crucially, done compositionally. The AJM approach is related to the Geometry of Interaction — but takes the quite demanding step of making an honest CCC out of it.)*
>
> *After that, the next key step was to see that the whole space of programming languages and computational features opened up to a game theoretic analysis in a very systematic way, by varying the conditions on strategies. This step was taken by myself and my students, and has led to a substantial further development. More recently, Luke Ong, Dan Ghica, Andrzej Murawski and myself have developed algorithmic game semantics, as a basis for compositional program analysis and verification, and — in Luke and Andrzej's hands — as a beautiful meeting point between semantics and algorithmics.*
>
> *There have of course been many other developments too, and many people have contributed. There have been recent workshops on these topics, e.g. in Seattle as part of the Federated Logic Conference in August.*

"PCF" is a souped-up version of the typed λ calculus that allows one to do arithmetic and full-fledged computation. Here are the papers on PCF mentioned above:

29) Samson Abramsky, R. Jagadeesan, and P. Malacaria, "Full abstraction for PCF", *Information and Computation* **163** (2000), 409–470. Available at `http://web.comlab.ox.ac.uk/oucl/work/samson.abramsky/pubs.html`

30) Martin Hyland and C.-H. Luke Ong, "On full abstraction for PCF", *Information and Computation* **163** (2000), 285–408.

Luke Ong has also written other papers using game theory to study the λ calculus:

31) A. D. Ker, H. Nickau, and C.-H. Luke Ong, "A universal innocent game model for the Bhm tree λ theory", in *Computer Science Logic: Proceedings of the 8th Annual Conference on the EACSL Madrid, Spain, September 1999*, LNCS Volume **1683**, Springer-Verlag, 1999, pp. 405–419.

32) A. D. Ker, H. Nickau, and C.-H. Luke Ong, "Innocent game models of untyped λ-calculus", *Theoretical Computer Science* **272** (2002), 247–292.

For a good introduction to all this work, try these:

33) Robin Houston, *Categories of Games*, M.Sc. thesis, U. Manchester, 2003. Available at `http://www.cs.man.ac.uk/~houstorx/msc.pdf`

    Robin Houston, *Mathematics of Games*, continuation report, U. Manchester, 2004. Available at `http://www.cs.man.ac.uk/~houstorx/continuation.pdf`

Finally, for more on categories, intuitionistic logic, and linear logic, see "Week 227".

Addenda: I thank Samson Abramsky, James Dolan, Dominic Hughes, Tom Payne, Esa Peuha and Vaughn Pratt for helpful corrections. When I wrote the first version of this Week's Finds, I was ignorant of work before Dolan and Trimble's that also described the free cartesian closed category on one object in terms of games. In addition to Abramsky's corrections (some of which are above), I was gently set straight by Dominic Hughes, who has permitted me to attach this post of his from the category theory mailing list:

> *This "backtracking game" characterisation has been known since around '93–'94, in the work of Hyland and Ong:*
>
> - *M. Hyland and L. Ong. "On full abstraction for PCF".* Information and Computation*, Volume* ***163****, pp. 285–408, December 2000. [Under review for 6 years!]* `ftp://ftp.comlab.ox.ac.uk/pub/Documents/techpapers/Luke.Ong/pcf.ps.gz`
>
> *(PCF is an extension of typed λ calculus.) My D.Phil. thesis extended the λ calculus (free CCC) characterisation to second-order, published in:*
>
> - *"Games and Definability for System F".* Logic in Computer Science*, 1997* `http://boole.stanford.edu/~dominic/papers/`
>
> *To characterise the free CCC on an arbitrary set $\{Z, Y, X, \ldots\}$ of generators (rather than a single generator, as you discuss), one simply adds the following Copycat Condition:*
>
> > *Whenever first player plays an occurrence of X, the second player must play an occurrence of X.*
>
> *[Try it: see how $X \to Y \to X$ has just one winning strategy.] Although the LICS'97 paper cited above appears to be the first place the Copycat Condition appears in print, I like to think it was already understood at the time by people working in the area. Technically speaking, winning strategies correspond to η-expanded β-normal forms. See pages 5–7 of my thesis for an informal description of the correspondence.*
>
> *It sounds like you've reached the point of trying to figure out how composition should work. Proving associativity is fiddly. Hyland and Ong give a very elegant treatment, via a larger CCC of games in which both players can backtrack. The free CCC subcategory is carved out as the so-called innocent strategies. This composition is almost identical to that presented by Coquand in:*

> - *"A semantics of evidence for classical arithmetic". Thierry Coquand.* Proceedings of the CLICS workshop*, Aarhus, 1992.*
>
> *Dominic*
>
> *PS A game-theoretic characterisation with an entirely different flavour (winning strategies less "obviously" corresponding to η-long β-normal forms) is:*
>
> - *Abramsky, S., Jagadeesan, R. and Malacaria, P., "Full Abstraction for PCF".* Info. & Comp. ***163** (2000), 409–470.* *`http://web.comlab.ox.ac.uk/oucl/work/samson.abramsky/pcf.pdf`* *[Announced concurrently with Hyland-Ong, around '93-'94.]*

On a different subject, James Dolan had this to say:

> *you describe holodeck strategies for "lady or tiger" where you take back "just when the tiger is about to eat you", but that's not the way it works. you take back just* after *the tiger has eaten you.*
>
> *(i guess that this is partially because of your lack of experience with computer games with a "saved game" feature. typically you die in the game and the computer plays some sort of funeral or at least funereal music; then you're taken to the reincarnation gallery where you select one to return to from your catalog of previous lives. or something like that.)*

In the first version of this Week's Finds I claimed that all systematic ways of picking an element of $(X^X)^{X^X}$ could be defined using the λ calculus. I was disabused of this notion by Vaughan Pratt, who wrote:

> *Hi, John,*
>
> *In "Week 240", you said*
>
> > *The moral of this game is that all systematic methods for picking an element of $(X^X)^{X^X}$ for an unknown set X can be written using the λ calculus.*
>
> *What is unsystematic about the contagious-fixpoint functional? This is the functional that maps those functions that have any fixpoints to the identity function (the function that makes every element a fixpoint) and functions without fixpoints to themselves (thus preserving the absence of fixpoints). It's a perfectly good functional that is equally well defined for all sets X, its statement in no way depends on X, and conceptually the concept of contagious fixpoints is even intuitively natural, but how do you write it using the λ calculus?*
>
> *Many more examples in this vein at* JPAA ***128**, 33–92 (Pare and Roman, "Dinatural numbers", 1998). The above is the case $K = \{0\}$ of Freyd's (proper) class of examples.*
>
> *Vaughan*

Here Pratt uses "functional" to mean what I was calling an "operator".

For more discussion, go to the n-Category Caf.

Unlike chess or astrology, mathematics has the curious property of being an intellectual game that really matters.

— *Rudy Rucker*

Week 241

November 20, 2006

I've been working too hard, and running around too much, to write This Week's Finds for a while. A bunch of stuff has built up that I want to explain. Luckily I've been running around explaining stuff — higher gauge theory, and tales of the dodecahedron.

This weekend I went to Baton Rouge. I was invited to Louisiana State University by Jorge Pullin of loop quantum gravity fame, and I used the opportunity to get a look at LIGO — the Laser Interferometry Gravitational-Wave Observatory!

I described this amazing experiment back in "Week 189", so I won't rehash all that. Suffice it to say that there are two installations: one in Hanford Washington, and one in Livingston Louisiana. Each consists of two evacuated tubes 4 kilometers long, arranged in an L shape.

Laser beams bounce back and forth between mirrors suspended at the ends of the tubes, looking for tiny changes in their distance that would indicate a gravitational wave passing through, stretching or squashing space. And when I say "tiny", I mean smaller than the radius of a proton! This is serious stuff.

Jorge drove me in his SUV to Livingston, a tiny town about 20 minutes from Baton Rouge. While he runs the gravity program at Louisiana State University, which has links to LIGO, he isn't officially part the LIGO team. His wife is. When I first met Gabriela Gonzalez, she was studying the Brownian motion of torsion pendulums. The mirrors in LIGO are hung on pendulums made of quartz wire, to minimize the effect of vibrations. But, the random jittering of atoms due to thermal noise still affects these pendulums. She was studying this noise to see its effect on the accuracy of the experiment.

This was way back when LIGO was just being planned. Now that LIGO is a reality, she's doing data analysis, helping search for gravitational waves produced by pairs of neutron stars and/or black holes as they spiral down towards a sudden merger. Together with an enormous pageful of authors, she helped write this paper, based on data taken from the "first science run" — the first real LIGO experiment, back in 2002:

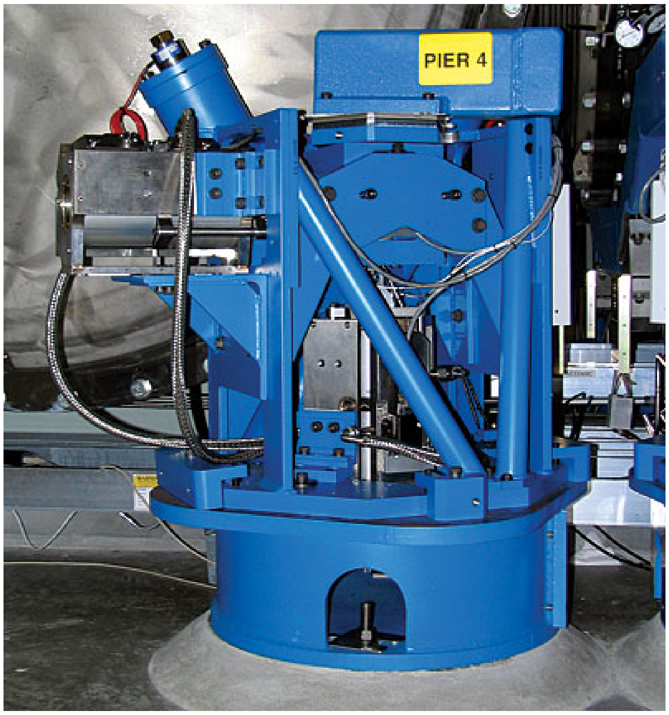
1) The LIGO Scientific Collaboration, "Analysis of LIGO data for gravitational waves from binary neutron stars", *Phys. Rev.* **D69** (2004), 122001. Also available at gr-qc/0308069.

She's one of the folks with an intimate knowledge of the experimental setup, who keeps the theorists' feet on the ground while they stare up into the sky.

On the drive to Livingston, Jorge pointed out the forests that surround the town.

These forests are being logged:

I asked him about this — when I last checked, the vibrations from falling trees were making it impossible to look for gravitational waves except at night! He said they've added a "hydraulic external pre-isolator" to shield the detector from these vibrations — basically a super-duper shock absorber:



Now they can operate LIGO day and night.

I also asked him how close LIGO had come to the sensitivity levels they were seeking. When I wrote "Week 189", during the first science run, they still had a long way to go. That's why the above paper only sets upper limits on neutron star collisions within 180 kiloparsecs. This only reaches out to the corona of the Milky Way — which includes the Small and Large Magellanic Clouds. We don't expect many neutron star collisions in this vicinity: maybe one every 3 years or so. The first science run didn't see any, and the set an upper limit of about 170 per year: the best experimental upper limit so far, but definitely worth improving, and nowhere near as fun as actually *seeing* gravitational waves.

But Jorge said the LIGO team has now reached its goals: they should be able to see collisions out to 15 megaparsecs! By comparison, the center of the Virgo cluster is about 20 megaparsecs away. In fact, they should already be able to see about half the galaxies in this cluster.

They're now on their seventh science run, and they'll keep upping the sensitivity in future projects called "Enhanced LIGO" and "Advanced LIGO". The latter should see neutron star collisions out to 300 megaparsecs:

2) Advanced LIGO, `http://www.ligo.caltech.edu/advLIGO/`

When we arrived at the gate, Jorge spoke into the intercom and got us let in. Our guide, Joe Giaime, was running a bit a late, so we walked over and looked at the interferometer's arms, each of which stretched off beyond sight, 2.5 kilometers of concrete tunnel surrounding the evacuated piping — the world's largest vacuum facility:

Schoolkids have been invited to paint pictures on some of these pipes:

One can tell this is the South. The massive construction caused pools of water to form in the boggy land near the facility, and these pools then attracted alligators. These have been dealt with firmly. The game hunters who occasionally fired potshots at the facility were treated more forgivingly: instead of feeding them to the alligators, the LIGO folks threw a big party and invited everyone from the local hunting club. Hospitality works wonders down here.

The place was pretty lonely. During the week lots of scientists work here, but this was Saturday, and on weekends there's just a skeleton crew of two. There's usually not much to do now that the experiment is up and running. As Joe Giaime later said, there have been no "Jodie Foster moments" like in the movie Contact, where the scientists on duty suddenly see a signal, turn on the suspenseful background music, and phone the President. There's just too much data analysis required to see any signal in real time: data from both Livingston and Hanford is sent to Caltech, and then people grind away at it. So, about the most exciting thing that happens is when the occasional tiny earthquake throws the laser beam out of phase lock.

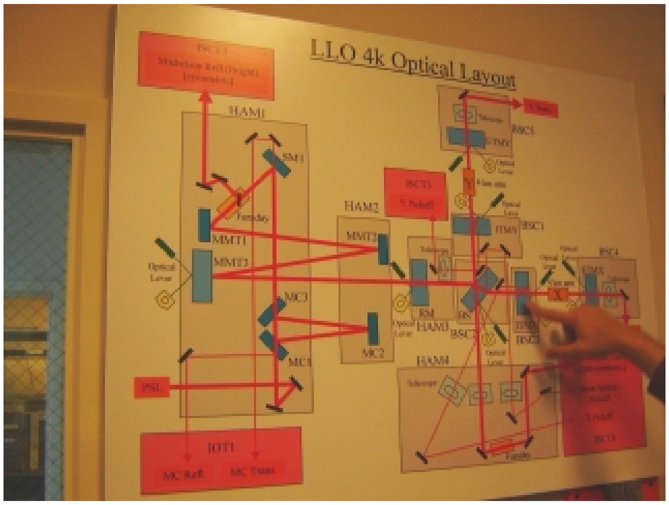
When Joe showed up, I got to see the main control room, which is dimly lit, full of screens indicating noise and sensitivity levels of all sorts — and even some video monitors showing the view down the laser tube:

This is where the people on duty hang out — you can see those video monitors on top:

One of them had brought his sons, in a feeble attempt to dispose of the huge supply of Halloween candy that had somehow collected here.

I also got to see a sample of the 400 "optical baffles" which have been installed to absorb light spreading out from the main beam before it can bounce back in and screw things up. The interesting thing is that these baffles and their placement were personally designed by Kip Thorne and some other godlike LIGO figure. Moral: unless they've gone soft, even bigshot physicists like to actually think about physics now and then, not just manage enormous teams.

But overall, there was surprisingly little to see, since the innermost workings are all sealed off, in vacuum. The optics are far more complicated than my description — "a laser bouncing between two suspended mirrors" — could possibly suggest. But, all I got to see was a chart showing how they work:



Oh well. I'm glad I don't need to understand this stuff in detail. It was fun to get a peek.

By the way, I wasn't invited to Louisiana just to tour LIGO and eat beignets and alligator sushi. My real reason for going there was to talk about higher gauge theory — a generalization of gauge theory which studies the parallel transport not just of point particles, but also strings and higher-dimensional objects:

3) John Baez, "Higher gauge theory", `http://math.ucr.edu/home/baez/highergauge`

This is a gentler introduction to higher gauge theory than my previous talks, some of which I inflicted on you in "Week 235". It explains how BF theory can be seen as a higher gauge theory, and briefly touches on Urs Schreiber's work towards exhibiting Chern-Simons theory and 11-dimensional supergravity as higher gauge theories. The webpage has links to more details.

I was also travelling last weekend — I went to Dartmouth and gave this talk:

4) John Baez, "Tales of the Dodecahedron: from Pythagoras through Plato to Poincare", `http://math.ucr.edu/home/baez/dodecahedron/`

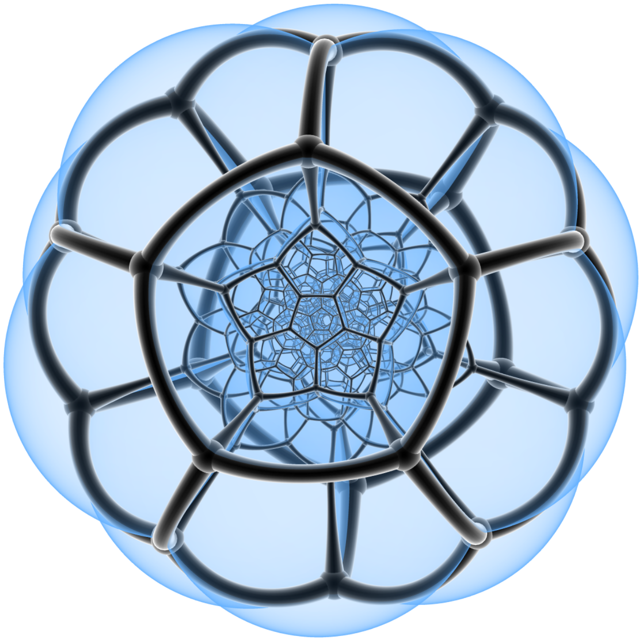
It's full of pictures and animations — fun for the whole family!

I started with the Pythagorean fascination with the pentagram, and how you can use the pentagram to give a magical picture proof of the irrationality of the golden ratio.

I then mentioned how Plato used four of the so-called Platonic solids to serve as atoms of the four elements — earth, air, water and fire - leaving the inconvenient fifth solid, the dodecahedron, to play the role of the heavenly sphere. This is what computer scientists call a "kludge" — an awkard solution to a pressing problem. Yes, there are twelve constellations in the Zodiac — but unfortunately, they're arranged quite differently than the faces of the dodecahedron.

This somehow led to the notion of the dodecahedron as an atom of "aether" or "quintessence" — a fifth element constituting the heavenly bodies. If you've ever seen the science fiction movie The Fifth Element, now you know where the title came from! But once upon a time, this idea was quite respectable. It shows up as late as Kepler's Mysterium Cosmographicum, written in 1596.

I then went on to discuss the 120-cell, which gives a way of chopping a spherical universe into 120 dodecahedra.



This leads naturally to the Poincare homology sphere, a closely related 3-dimensional manifold made by gluing together opposite sides of *one* dodecahedron.

The Poincare homology sphere was briefly advocated as a model of the universe that could explain the mysterious weakness of the longest-wavelength ripples in the cosmic background radiation — the ripples that only wiggle a few times as we scan all around the sky:

5) J.-P. Luminet, J. Weeks, A. Riazuelo, R. Lehoucq, and J.-P. Uzan, "Dodecahedral space topology as an explanation for weak wide-angle temperature correlations in the cosmic microwave background", *Nature* **425** (2003), 593. Also available as astro-ph/0310253.

The idea is that if we lived in a Poincare homology sphere, we'd see several images of each very distant point in the universe. So, any ripple in the background radiation would repeat some minimum number of times: the lowest-frequency ripples would be suppressed.

Alas, this charming idea turns out not to fit other data. We just don't see the same distant galaxies in several different directions:

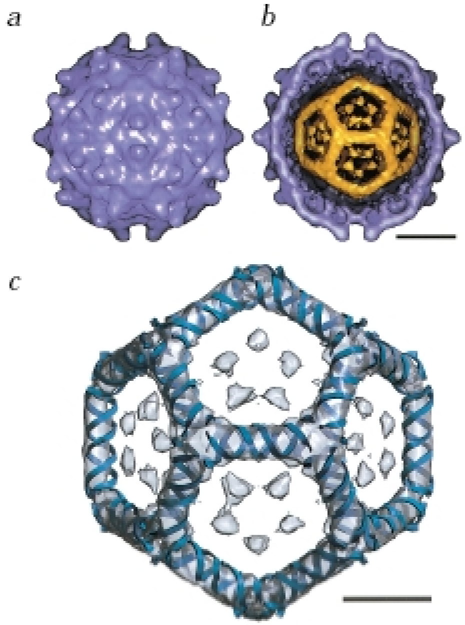
6) Neil J. Cornish, David N. Spergel, Glenn D. Starkman and Eiichiro Komatsu, "Constraining the topology of the universe", *Phys. Rev. Lett.* **92** (2004) 201302. Also available as astro-ph/0310233.

For a good review of this stuff, see:

7) Jeffrey Weeks, “The Poincare dodecahedral space and the mystery of the missing fluctuations”, *Notices of the AMS* **51** (2004), 610–619. Also available at `http://www.ams.org/notices/200406/fea-weeks.pdf`

In the abstract of my talk, I made the mistake of saying that the regular dodecahedron doesn’t appear in nature — that instead, it was invented by the Pythagoreans. You should never say things like this unless you want to get corrected!

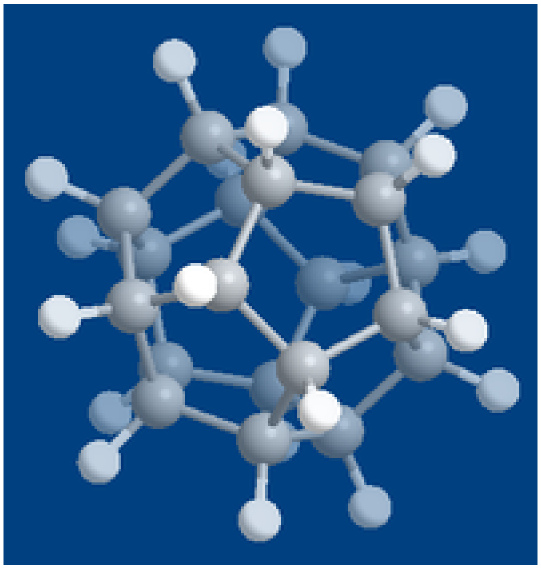
Dan Piponi pointed out this dodecahedral virus:



8) Liang Tang et al, “The structure of Pariacoto virus reveals a dodecahedral cage of duplex RNA”, *Nature Structural Biology* **8** (2001), 77–83. Also available at `http://www.nature.com/nsmb/journal/v8/n1/pdf/nsb0101_77.pdf`

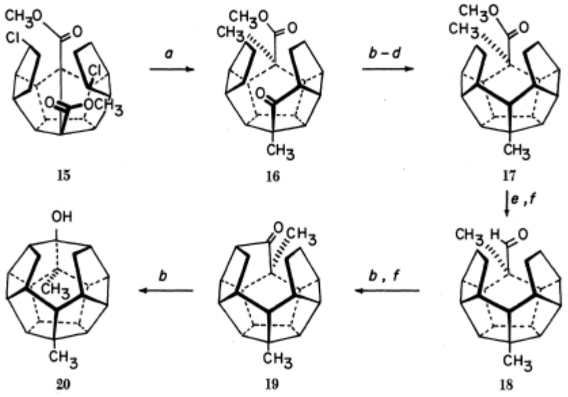
The first black line is 100 angstroms long (10^{-8} meters), while the second is 50 angstroms long.

Garett Leskowitz pointed out the molecule “dodecahedrane”, with 20 carbons at the vertices of a dodecahedron and 20 hydrogens bonded to these:



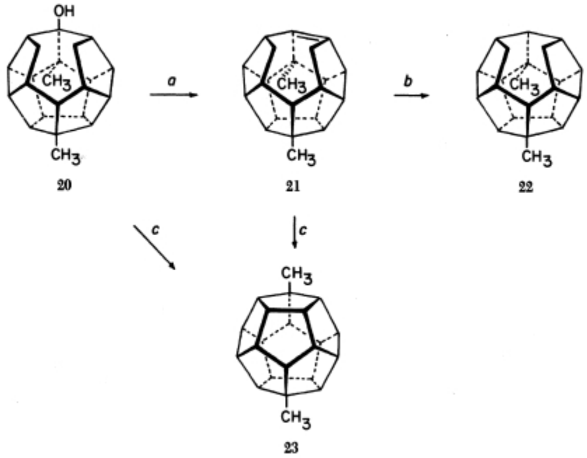
9) Wikipedia, "Dodecahedrane", `http://en.wikipedia.org/wiki/Dodecahedrane`

This molecule hasn't been found in nature yet, but chemists can synthesize it using reactions like these:



a, Li/NH$_3$, then CH$_3$I; *b*, hν; *c*, TsOH, C$_6$H$_6$, *d*, HN═NH, CH$_3$OH; *e*, (*i*-Bu)$_2$AlH; *f*, PCC, CH$_2$Cl$_2$.

Scheme IV



a, TsOH, C_6H_6; *b*, H_2NNH_2, H_2O_2; *c*, CF_3SO_3H, GH_2Cl_2.

Scheme V

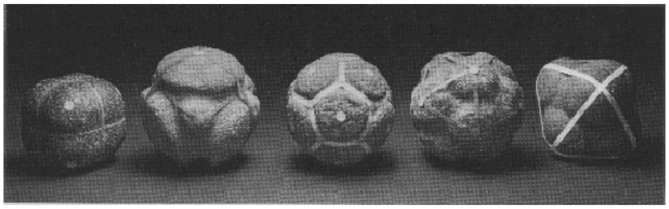
10) Robert J. Ternansky, Douglas W. Balogh and Leo A. Paquette, "Dodecahedrane", *J. Am. Chem. Soc.* **104** (1982), 4503–4504.

11) Leo A. Paquette, "Dodecahedrane — the chemical transliteration of Plato's universe (a review)", *Proc. Nat. Acad. Sci. USA* **14** part 2 (1982), 4495–4500. Also available at `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=346698`

So, there's probably a bit somewhere in our galaxy.

Of course, what I *meant* was that people didn't come up with regular dodecahedra after seeing them in nature — that instead, the Pythagoreans dreamt them up, possibly after seeing pyrite crystals that look sort of similar. These crystals are called "pyritohedra".

But, even here I made a mistake. The Pythagoreans seem not to have been the first to discover the dodecahedron. John McKay told me that stone spheres with Platonic solids carved on them have been found in Scotland, dating back to around 2000 BC!



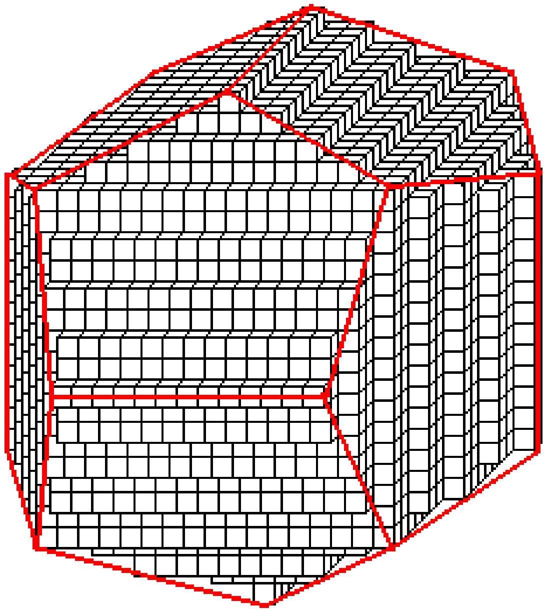
13) Michael Atiyah and Paul Sutcliffe, "Polyhedra in physics, chemistry and geometry", available as `math-ph/0303071`.

14) Dorothy N. Marshall, "Carved stone balls", *Proc. Soc. Antiq. Scotland* **108** (1976/77), 40–72. Available at `http://ads.ahds.ac.uk/catalogue/library/psas/`

Indeed, stone balls with geometric patterns on them have been found throughout Scotland, and occasionally Ireland and northern England. They date from the Late Neolithic to the Early Bronze age: 2500 BC to 1500 BC. For comparison, the megaliths at Stonehenge go back to 2500–2100 BC.

Nobody knows what these stone balls were used for, though the article by Marshall presents a number of interesting speculations.

The pyritohedron is interesting in itself, so before I turn to some really fancy math, let me talk a bit about this guy. Since pyrite is fundamentally a cubic crystal, the pyritohedron is basically made out of little cubic cells, as shown here:



12) Steven Dutch, "Building isometric crystals with unit cells", `http://www.uwgb.edu/dutchs/symmetry/isometuc.htm`

It has 12 pentagonal faces, orthogonal to these vectors:

$$\begin{array}{cccc}
(2,1,0) & (2,-1,0) & (-2,1,0) & (-2,-1,0) \\
(1,0,2) & (-1,0,2) & (1,0,-2) & (-1,0,-2) \\
(0,2,1) & (0,2,-1) & (0,-2,1) & (0,-2,-1)
\end{array}$$

You can see how this works by going here:

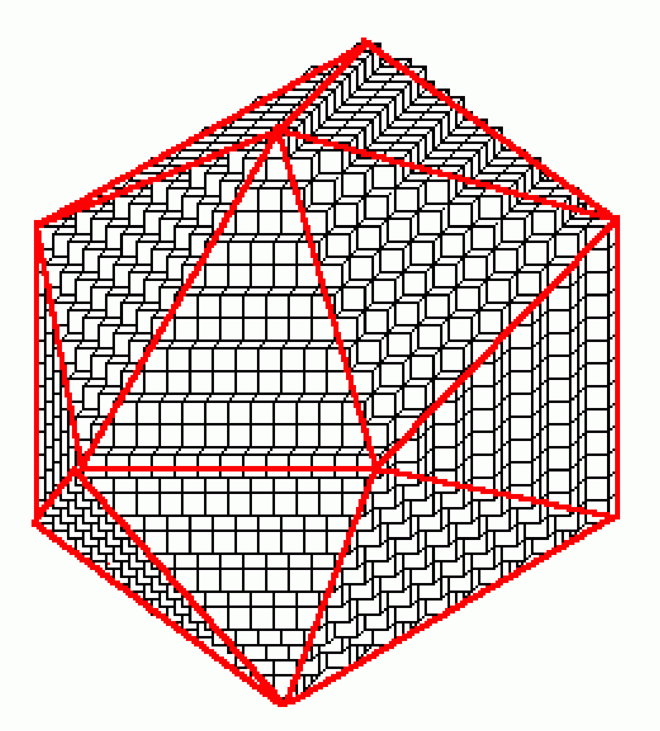
13) mindat.org, "Pyrite", `http://www.mindat.org/min-3314.html`

If your webbrowswer can handle Java, go to this webpage and click on "Pyrite no. 3" to see a rotating pyritohedron. Then, while holding your left mouse button down when the cursor is over the picture of the pyritohedron, type "m" to see the vectors listed above.

Why “m”? These vectors are called “Miller indices”. In general, Miller indices are outwards-pointing vectors orthogonal to the faces of a crystal; we can use them to classify crystals.

The Miller indices for the pyritohedron have a nice property. If you think of these 12 vectors as points in space, they’re the corners of three 2×1 rectangles: a rectangle in the xy plane, a rectangle in the xz plane, and a rectangle in the yz plane.

These points are also corners of an icosahedron! It’s not a regular icosahedron, though. It’s probably the “pseudoicosahedron” shown in Steven Dutch’s site above:

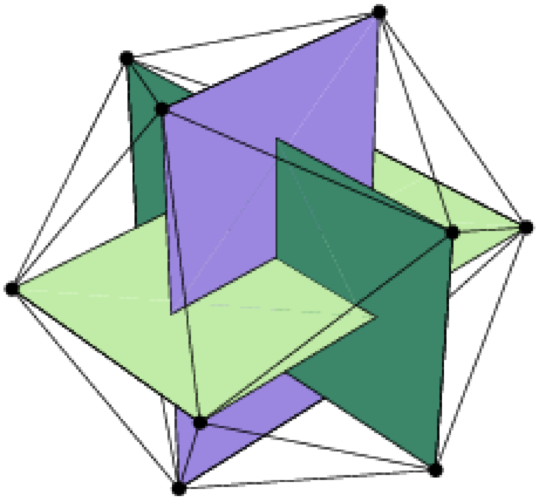


Apparently iron pyrite can also form a pseudoicosahedron — see “Pyrite No. 7” on the mindat.org website above. Does anyone have actual photos?

To get the corners of a regular icosahedron, we just need to replace the number 2 by the golden ratio $\Phi = (\sqrt{5}+1)/2$:

$$\begin{array}{cccc} (\Phi,1,0) & (\Phi,-1,0) & (-\Phi,1,0) & (-\Phi,-1,0) \\ (1,0,\Phi) & (-1,0,\Phi) & (1,0,-\Phi) & (-1,0,-\Phi) \\ (0,\Phi,1) & (0,\Phi,-1) & (0,-\Phi,1) & (0,-\Phi,-1) \end{array}$$

Now our rectangles are golden rectangles:



Since the pseudoicosahedron does a cheap imitation of this trick, with the number 2 replacing the golden ratio, the number 2 deserves to be called the "fool's golden ratio". I thank Carl Brannen for explaining this out to me!

The regular docahedron is "dual" to the regular icosahedron: the vertices of the icosahedron are Miller indices for the dodecahedron. Similarly, I bet the pyritohedron is dual to the pseudoicosahedron.

So, we could call the pyritohedron the "fool's dodecahedron", and the pseudoicosahedron the "fool's icosahedron". Fool's gold may have fooled the Greeks into inventing the regular dodecahedron, by giving them an example of a fool's dodecahedron.

As pointed out by Noam Elkies and James Dolan, there is a sequence of less and less foolish dodecahedra whose faces have normal vectors

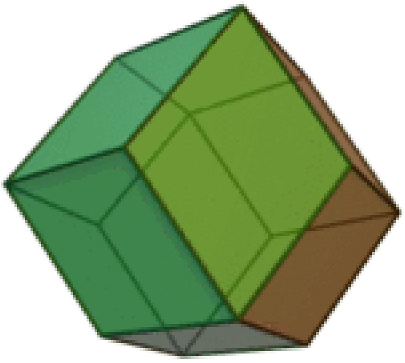
$$\begin{array}{cccc} (B,A,0) & (B,-A,0) & (-B,A,0) & (-B,-A,0) \\ (A,0,B) & (-A,0,B) & (A,0,-B) & (-A,0,-B) \\ (0,B,A) & (0,B,-A) & (0,-B,A) & (0,-B,-A) \end{array}$$

where A and B are the nth and $(n+1)$st Fibonacci numbers, respectively. As $n \to \infty$, these dodecahedra approach a regular dodecahedron in shape, because the ratio of successive Fibonacci numbers approaches the golden ratio.

When $A = 1$ and $B = 2$, we get the fool's dodecahedron, since only a fool would think $2/1$ is the golden ratio.

However, this is not the most foolish of all dodecahedra! The case $A = 1$ and $B = 1$

gives the rhombic dodecahedron, which doesn't even have pentagonal faces:



14) Wikipedia, "Rhombic dodecahedron", `http://en.wikipedia.org/wiki/Rhombic_dodecahedron`

So, the rhombic dodecahedron deserves to be called the "moron's dodecahedron" — at least for people who think it's actually a regular dodecahedron.

But actually, even this dodecahedron isn't the dumbest. The Fibonacci numbers start with 0:

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots$$

So, even more foolish is the case $A = 0$ and $B = 1$. Here our 12 vectors reduce to just 6 different ones:

$$\begin{array}{cccc} (1,0,0) & (1,-0,0) & (-1,0,0) & (-1,-0,0) \\ (0,0,1) & (-0,0,1) & (0,0,-1) & (-0,0,-1) \\ (0,1,0) & (0,1,-0) & (0,-1,0) & (0,-1,-0) \end{array}$$

These are normal to the faces of a cube. So, the cube deserves to be called the "half-wit's dodecahedron": it doesn't even have 12 faces, just 6.

Moving in the direction of increasing wisdom, we can consider the case $A = 2$, $B = 3$. This gives a dodecahedron which is closer to regular than the pyritohedron. And, apparently it exists in nature! It shows up as number 12 in this list of crystals:

15) Ian O. Angell and Moreton Moore, "Projections of cubic crystals", section 4: The diagrams, `http://www.iucr.org/education/pamphlets/12/full-text`

They also call this guy a pyritohedron, so presumably some pyrite forms these less foolish crystals! You can compare it with the $A = 1$, $B = 2$ case here:

16) Ian O. Angell and Moreton Moore, "Projections of cubic crystals, Graphical index of figures", `http://www.iucr.org/education/pamphlets/12/graphical-index`

The $A = 1$, $B = 2$ pyritohedron is figure 9, while the $A = 2$, $B = 3$ pyritohedron is figure 12. It's noticeably better!

Let me wrap up by mentioning a fancier aspect of the dodecahedron which has been intriguing me lately. I already mentioned it in "Week 230", but in such a general setting that it may have whizzed by too fast. Let's slow down a bit and enjoy it.

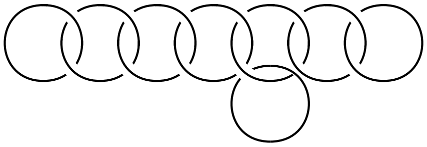
The rotational symmetries of the dodecahedron form a 60-element subgroup of the rotation group $\mathrm{SO}(3)$. So, the "double cover" of the rotational symmetry group of the dodecahedron is a 120-element subgroup of $\mathrm{SU}(2)$. This is called the "binary dodecahedral group". Let's call it G.

The group $\mathrm{SU}(2)$ is topologically a 3-sphere, so G acts as left translations on this 3-sphere, and we can use a dodecahedron sitting in the 3-sphere as a fundamental domain for this action. This gives the 120-cell. The quotient $\mathrm{SU}(2)/G$ is the Poincare homology sphere!

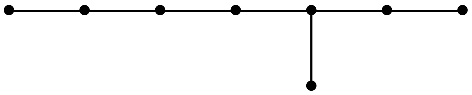
But, we can also think of G as acting on $\mathbb{C}^2$. The quotient $\mathbb{C}^2/G$ is not smooth: it has an isolated singular coming from the origin in $\mathbb{C}^2$. But as I mentioned in "Week 230", we can form a "minimal resolution" of this singularity. This gives a holomorphic map

$$p\colon M \to \mathbb{C}^2/G$$

where M is a complex manifold. If we look at the points in M that map to the origin in $\mathbb{C}^2/G$, we get a union of 8 Riemann spheres, which intersect each other in this pattern:



Here I've drawn linked circles to stand for these intersecting spheres, for a reason soon to be clear. But, already you can see that we've got 8 spheres corresponding to the dots in this diagram:

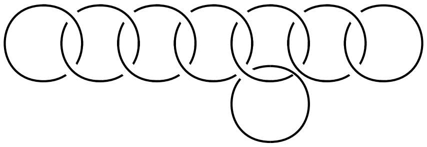


where the spheres intersect when there's an edge between the corresponding dots. And, this diagram is the Dynkin diagram for the exceptional Lie group E_8!

I already mentioned the relation between the E_8 Dynkin diagram and the Poincare homology sphere in "Week 164", but now maybe it fits better into a big framework. First, we see that if we take the unit ball in $\mathbb{C}^2$, and see what points it gives in $\mathbb{C}^2/G$, and then take the inverse image of these under

$$p\colon M \to \mathbb{C}^2/G,$$

we get a 4-manifold whose boundary is the Poincare homology 3-sphere. So, we have a cobordism from the empty set to the Poincare homology 3-sphere! Cobordisms can be described using "surgery on links", and the link that describes this particular cobordism is:



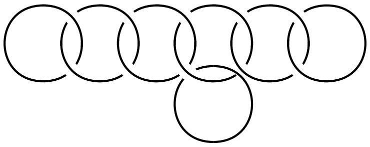
Second, by the "McKay correspondence" described in "Week 230", all this stuff also works for other Platonic solids! Namely:

If G is the "binary octahedral group" — the double cover of the rotational symmetry group of the octahedron — then we get a minimal resolution

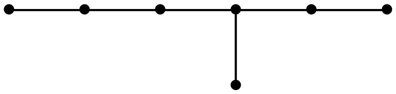
$$p\colon M \to \mathbb{C}^2/G$$

which yields, by the same procedure as above, a cobordism from the empty set to the 3-manifold $\mathrm{SU}(2)/G$.

This cobordism can be described using surgery on this link:



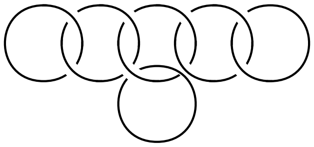
which encodes the Dynkin diagram of E_7:



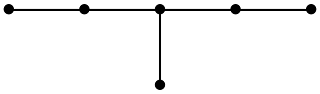
And, if G is the "binary tetrahedral group" — the double cover of the rotational symmetry group of the tetrahedron — then a minimal resolution

$$p\colon M \to \mathbb{C}^2/G$$

yields, by the same procedure as above, a cobordism from the empty set to the 3-manifold $\mathrm{SU}(2)/G$. This cobordism can be described using surgery on this link:



which encodes the Dynkin diagram of E_6:



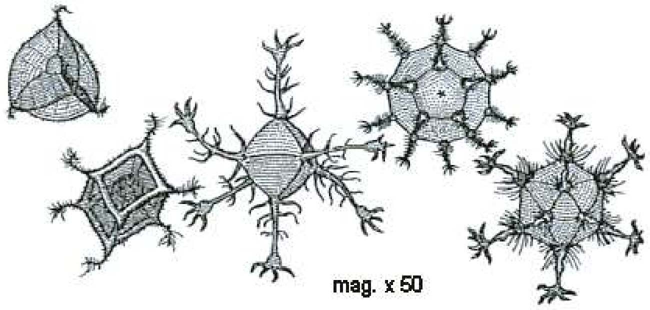
I don't fully understand this stuff, that's for sure. But, I want to. The Platonic solids are still full of mysteries.

Addenda: Someone with the handle "Dileffante" has found another nice example of the dodecahedron in nature — and even in Nature:

> *While perusing a Nature issue I found this short notice on a paper, and I remembered that in your talk (which I saw online) you mentioned that the dodecahedron was not found in nature. Now I see in "Week 241" that there are some things dodecahedral after all, but nevertheless, I send this further dodecahedron which was missing there.*
>
> *Nature commented in issue 7075:*
>
> 15) *"The complete Plato",* Nature ***439*** *(26 January 2006), 372–373.*
>
> > *According to Plato, the heavenly ether and the classical elements — earth, air, fire and water — were composed of atoms shaped like polyhedra whose faces are identical, regular polygons. Such shapes are now known as the Platonic solids, of which there are five: the tetrahedron, cube, octahedron, icosahedron and dodecahedron. Microscopic clusters of atoms have already been identified with all of these shapes except the last.*
> >
> > *Now, researchers led by Jos Luis Rodrguez-Lpez of the Institute for Scientific and Technological Research of San Luis Potos in Mexico and Miguel Jos-Yacamn of the University of Texas, Austin, complete the set. They find that clusters of a gold-palladium alloy about two nanometres across can adopt a dodecahedral shape.*
>
> *The article is in:*
>
> 16) *Juan Martn Montejano-Carrizales, Jos Luis Rodrguez-Lpez, Umapada Pal, Mario Miki-Yoshida and Miguel Jos-Yacamn, "The completion of the Platonic atomic polyhedra: the dodecahedron",* Small ***2*** *(2006), 351–355.*
>
> *Here's the abstract:*
>
> > *Binary AuPd nanoparticles in the 1–2 nm size range are synthesized. Through HREM imaging, a dodecahedral atomic growth pattern of five fold axis is identified in the round shaped (85%) particles. Our results demonstrate the first experimental evidence of this Platonic atomic solid at this size range of metallic nanoparticles. Stability of such Platonic structures are validated through theoretical calculations.*
>
> *Either there is some additional value in the construction, or the authors (and Nature editors) were unaware of dodecahedrane.*

Dodecahedrane is a molecule built from carbon and hydrogen — a bit different from an "atomic cluster" of the sort discussed here. It's a matter of taste whether that's important, but I bet these gold-palladium nanoparticles occur in nature, while dodecahedrane seems to be unstable.

My friend Geoffrey Dixon contributed these pictures of Platonic life forms:



They look a bit like Ernst Haeckel's pictures from his book "Kunstformen der Natur" (artforms of nature).

Finally, here's a really important addendum: in March 2009, Lieven le Bruyn discovered that the ancient Scots did *not* carve stone balls to look like Platonic solids! The whole story is something between a hoax and a series of misunderstandings:

17) Lieven le Bruyn, "The Scottish solids hoax", from his blog *neverendingbooks*, March 25, 2009, `http://www.neverendingbooks.org/index.php/the-scottish-solids-hoax.html`

18) John Baez, "Who discovered the icosahedron?", talk at the *Special Session on History and Philosophy of Mathematics*, 2009 Fall Western Section Meeting of the AMS, November 7, 2009. Available at `http://math.ucr.edu/home/baez/icosahedron/`

You can read a bunch of freewheeling discussions triggered by this Week's Finds at the n-Category Caf.

The essence of mathematics lies in its freedom.

— *Georg Cantor*

Week 242

December 17, 2006

This week I'd like to talk about a paper by Jeffrey Morton. Jeff is a grad student now working with me on topological quantum field theory and higher categories. I've already mentioned his work on categorified algebra and quantum mechanics in "Week 236". He'll be be finishing his Ph.D. thesis in the spring of 2007 — and as usual, that means he's already busy applying for jobs.

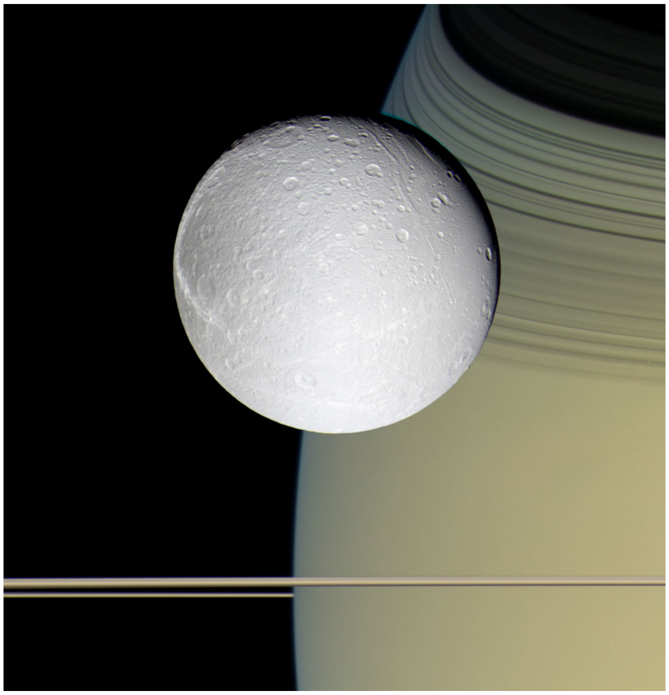
As all you grad students reading this know, applying for jobs is pretty scary the first time around: there are some tricks involved, and nobody prepares you for it. I remember myself, wondering what I'd do if I didn't succeed. Would I have to sell ice cream from one of those trucks that plays a little tune as it drives around the neighborhood? A job in the financial industry seemed scarcely more appealing: less time to think about math, and less ice cream too.

Luckily things worked out for me. . . and I'm sure they'll work out for Jeff and my other student finishing up this year — Derek Wise, who is working on Cartan geometry and MacDowell-Mansouri gravity.

But, to help them out a bit, I'd like to talk about their work. This has been high on my list of interests for the last few years, of course, but I've mostly been keeping it under wraps.

This time I'll talk about Jeff's thesis; next time Derek's. But first, let's start with some cool astronomy pictures!

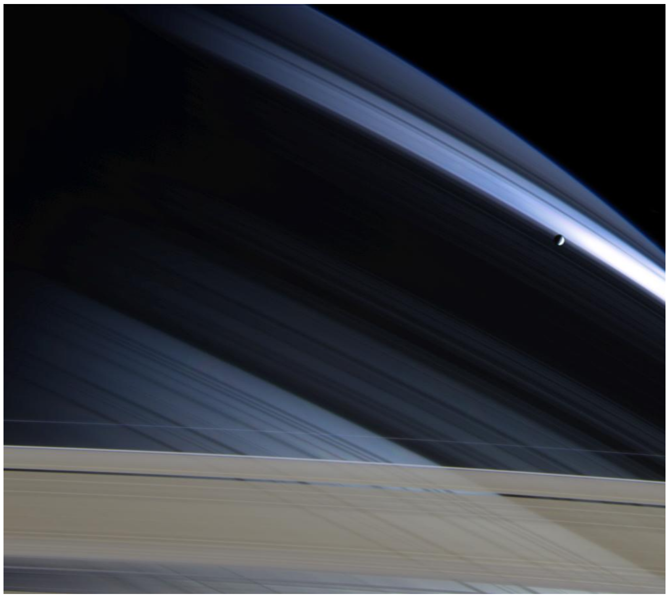
Here's a photo of Saturn, Saturn's rings, and its moon Dione, taken by the Cassini orbiter in October last year:



1) NASA, "Ringside with Dione", `http://solarsystem.nasa.gov/multimedia/display.cfm?IM_ID=4163`

It's so vivid it seems like a composite fake, but it's not! With the Sun shining from below, delicate shadows of the B and C rings cover Saturn's northern hemisphere. Dione seems to hover nearby. Actually it's 39,000 kilometers away in this photo. It's 1,200 kilometers in diameter, about the third the size of our Moon.

Here's a photo of Saturn, its rings, and its moon Mimas, taken in November 2004:



2) NASA, "Nature's canvas", `http://saturn.jpl.nasa.gov/multimedia/images/image-details.cfm?imageID=1088`

It's gorgeous, but it takes some work to figure out what's going on!

The blue stuff in the background is Saturn, with lines created by shadows of rings. The bright blue-white stripe near Mimas is sunlight shining through a break in the rings called the "Cassini division". The brownish stuff near the bottom is the A ring — you can see right through it. Above it there's a break and a thinner ring called the F ring. Below it is the Cassini division itself.

This is just one of many photos taken by Cassini and Huyghens, the probe that Cassini dropped onto Saturn's moon Titan — see "Week 210" for more on that. You can see more of these photos here:

3) NASA, "Cassini-Huygens", `http://saturn.jpl.nasa.gov/`

I hope you see from these beautiful images, and others on This Week's Finds, that we are *already in space*. We don't need people up there for us to effectively *be there*.

Alas, not everyone recognizes this. An expensive American program to set up a base on the Moon, perhaps as a stepping stone to a manned mission to Mars, is starting to drain money from more exciting unmanned missions. NASA guesses this program will cost $104 billion up to the time when we land on the Moon — again — in 2020. By 2024, the Government Accounting Office guesses the price will be $230 billion. By comparison, the Cassini-Huygens mission cost just about $3.3 billion.

And what will be benefits of a Moon base be? It's unclear: at best, some vague dream of "space colonization".

Mind you, I'm in favor of space exploration, and even colonization. But, these are very different things!

Colonies are usually about making money. Governments support them in hopes of turning a profit: think Columbus and Isabella, or other adventurers funded by colonial powers.

Right now most of the money lies in near-earth orbit, not on the Moon and Mars. Telecommunication satellites and satellite photos are established businesses. The next step may be tourism. Dennis Tito, Gregory Olsen and Mark Shuttleworth have already paid the Russian government $20 million each to visit the International Space Station. This orbits at an altitude of about 350 kilometers, in the upper "thermosphere" — the layer of the Earth's atmosphere where gases get ionized by solar radiation.

If this is too pricey for you, wait a few years. Richard Branson's company Virgin Galactic plans to give 500 people per year a 7-minute experience of weightlessness at a cost of just $200,000 each. Alas, you'll only go up 100 kilometers, near the bottom of the thermosphere.

Some competition may lower the price. Jeff Bezos, the founder of Amazon, has bought a lot of land in Texas to a build space port for his company Blue Origins. He wants to do test flights by next year, and he eventually wants 50 flights a year in a vehicle that holds 3. If you've always looked forward to using your seat cushion as a flotation device in the event of a water landing, you'll love this:

> *"During an abort situation, the crew capsule would separate, using small solid-rocket motors to safely recover the space flight participants. The abort module containing the solid-rocket motors would then jettison from the crew capsule."*

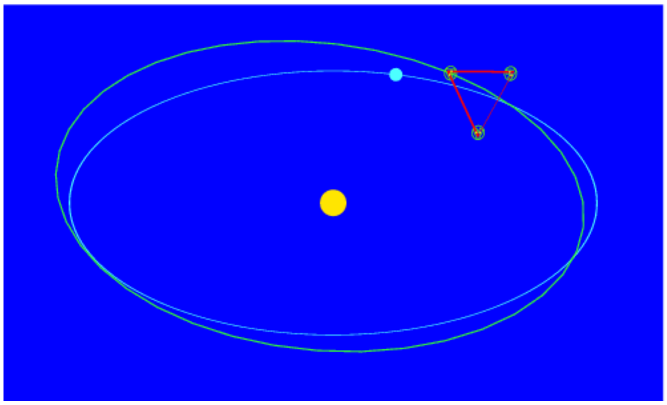
None of this stuff requires any taxpayer funding. It's a bit self-indulgent and silly, but it may eventually grow and merge with other profit-making forms of space colonization.

Exploration is a bit different: seeing what's out there, mainly for the sake of adventure and understanding. For this we should send machines, not people. Machines can be designed to do well in vacuum. People can't - not yet. This will probably change when nanotech, AI and cyborg technologies kick in. But for now, unmanned probes are the way to go.

Here are some of the wonderful things we could do, all for less than setting up a Moon base:

4) The Laser Interferometer Space Antenna (LISA), http://saturn.jpl.nasa.gov/

The idea of LISA is to put 3 satellites in a huge equilateral triangle following the Earth in its orbit around the Sun, and bounce lasers between them to detect gravitational waves (see "Week 143"):



This would avoid the ground noise that plagues LIGO (see "Week 241"), and it could detect waves of much lower frequencies. If all works well, it could see gravitational waves from the *very* early Universe, long before the hot gas enough cooled to let light through. We're talking times like $10^{-38} seconds after the Big Bang! That's the biggest adventure I can imagine. . . back to the birth pangs of the Universe.

Right now LISA is scheduled for launch around 2016. But as you'll soon see, this may not happen.

5) Constellation-X, http://constellation.gsfc.nasa.gov

This would be a team of X-ray telescopes, combining forces to be 100 times more powerful than any previous single one. Among many other things, Constellation-X could study the X-rays emitted by matter falling into things that look like black holes. The redshift of these X-rays is our best test of general relativity for very strong gravitational fields. So, it's our best way of checking that these black hole candidates really do have event horizons!

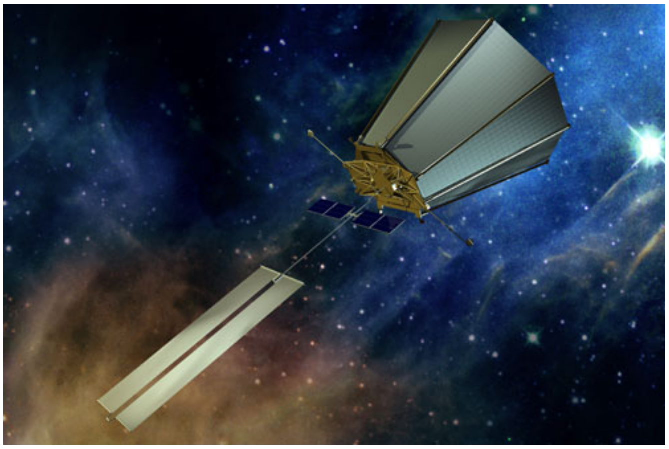
In February 2006, when NASA put out their latest budget, they said Constellation-X would be "delayed indefinitely". And in September 2006, a National Research Council committee was formed to pick *one* of NASA's five "Beyond Einstein" programs for the first shot at funding: LISA, Constellation-X, the Joint Dark Energy Mission, the Inflation Probe and the Black Hole Finder. Currently the Joint Dark Energy Mission seems to be in the lead:

6) Steinn Sigursson, "NASA: double down on science", Dynamics of Cats, September 16, 2006, `http://scienceblogs.com/catdynamics/2006/09/nasa_double_down_on_science.php`

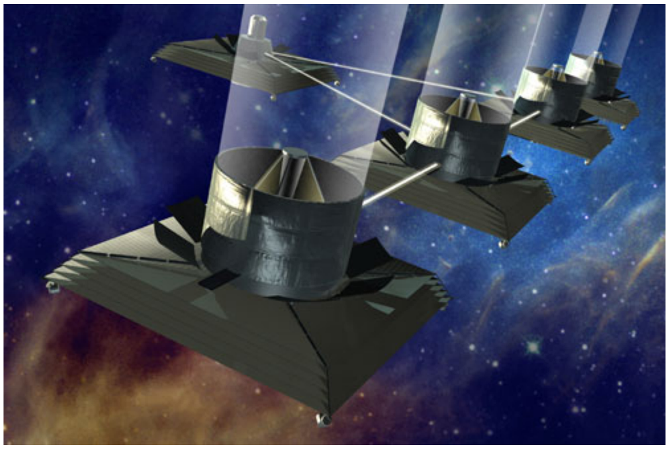
A decision is expected around September 2007.

7) The Terrestrial Planet Finder (TPF), `http://planetquest.jpl.nasa.gov/TPF/`

This could study Earth-like planets orbiting stars up to 45 light years away. It would consist of two observatories: a visible-light "coronagraph" that blocks out the light from a star so it can see nearby fainter objects:



and an infrared interferometer made of several units flying in formation:



In February 2006, NASA halted work on the TPF. In June 2006, thanks to public pressure, Congress reinstated funding for this program and also a mission to Jupiter's moon Europa, which could have oceans underneath its icy crust. However, at last report, NASA was continuing to fight *against* reinstating these missions:

8) Louis D. Friedman, "Congressional inaction leaves science still devastated", *The Planetary Society*, November 26, 2006, `http://planetary.org/programs/projects/sos/20061122.html`

The constantly shifting situation makes it hard to know what's going on.

9) The Nuclear Spectroscopic Telescope Array (NuStar), `http://www.nustar.caltech.edu/`

This is an orbiting observatory with three telescopes, designed to see hard X-rays. It could conduct a thorough survey of black hole candidates throughout the universe. It could study relativistic jets of particles from the cores of active galaxies (which are probably also black holes). And, it could study young supernova remnants — hot new neutron stars.

NASA suddenly canceled work on NuStar in February 2006.

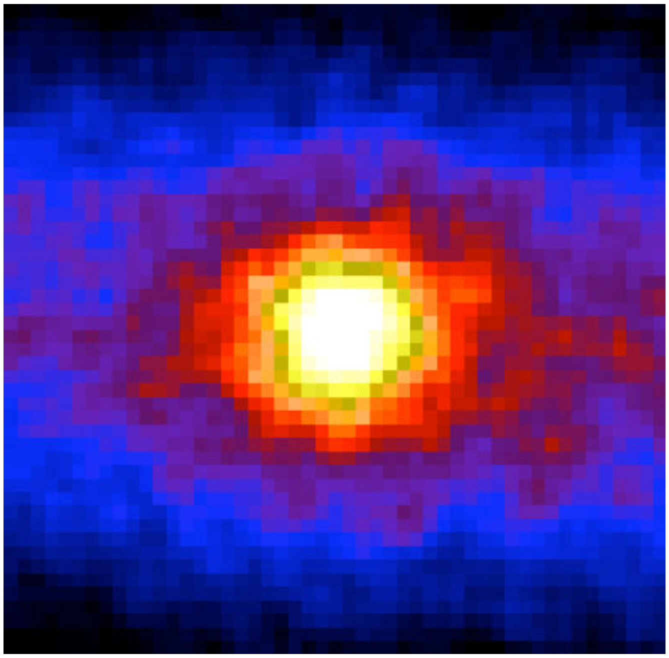
10) Dawn, `http://dawn.jpl.nasa.gov/`

The Dawn mission seeks to understand the early Solar System by probing the asteroid belt and taking a good look at Ceres and Vesta. Ceres is the largest asteroid of all, 950 kilometers in diameter. It seems have a rocky core, a thick mantle of water ice, and a thin dusty outer crust. Vesta is the second largest, about 530 kilometers in diameter. It's very different from Ceres: it's not round, and it's all rock. A certain group of stony meteorites called "HED meteorites" are believed to be pieces of Vesta!

NASA cancelled the Dawn mission in March 2006 — but later that month, they changed their minds.

It's depressing to contemplate all the wonderful things we could miss while spending hundreds of billions to "send canned primates to Mars", as Charles Stross so cleverly put it in his novel "Accelerando" (see "Week 222"). I'm all for humanity spreading through space. I just don't think we should do it in a clunky, low-tech way like setting up a base on the Moon where astronauts sit around and. . . what, play golf? It's like something out of old science fiction!

To cheer myself up again, here's a picture of the Sun:



11) Joanne Hewett, "Sun Shots", `http://cosmicvariance.com/2006/10/13/sun-shots/`

It was taken not with light, but with *neutrinos*. It was made at the big neutrino observatory in Japan, called Super-Kamiokande. It took about 504 days and nights to make.

That's right — nights! Neutrinos go right through the Earth.

As you probably know, neutrinos oscillate between three different kinds, but only electron neutrinos are easy to detect, so we see about third as many neutrinos from the Sun as naively expected. That's the kind of thing they're studying at Super-Kamiokande.

But what I want to know is: what's the "glare" in this picture? Neutrinos are made by the process of fusion, which involves this reaction:

$$\text{proton} + \text{electron} \to \text{neutron} + \text{electronneutrino}$$

Fusion mostly happens in the Sun's core, which has a density of 160 grams per cubic centimeter (10 times denser than lead) and a temperature of 15 million kelvin (300 thousand times hotter than the "broil" setting on an American oven).

So, what's the disk in this picture: the whole Sun, or the Sun's core? And what's the glare?

Okay, now for some serious mathematical physics:

12) Jeffrey Morton, "A double bicategory of cobordisms with corners", available as math.CT/0611930v1.

People have been talking a long time about topological quantum field theory and higher categories. The idea is that categories, 2-categories, 3-categories and the like can describe how manifolds can be chopped into little pieces — or more precisely, how these little pieces can be glued together to form manifolds. Then the problem of doing quantum field theory on some manifold can be reduced to the problem of doing it on these pieces and gluing the results together. This works easiest if the theory is "topological", not requiring a background metric.

There's a lot of evidence that this is a good idea, but getting the details straight has proved tough, even at the 2-category level. This is what Morton does, in a rather clever way. Very roughly, his idea is to use something I'll call a "weak double category", and prove that these:

- $(n-2)$-dimensional manifolds
- $(n-1)$-dimensional manifolds with boundary
- n-dimensional manifolds with corners

give a weak double category called $n\mathsf{Cob}_2$. The proof is a cool mix of topology and higher category theory. He then shows that this particular weak double category can be reinterpreted as something a bit more familiar — a "weak 2-category".

In the rest of his thesis, Jeff will use this formalism to construct some examples of "extended TQFTs", which are roughly maps of weak 2-categories

$$Z\colon n\mathsf{Cob}_2 \to 2\mathsf{Vect}$$

where $2\mathsf{Vect}$ is the weak 2-category of "2-vector spaces". He's focusing on some extended TQFTs called the Dijkgraaf-Witten models, coming from finite groups.

But, he's also thought about the case where the finite group is replaced by a compact Lie group. In this case we get something called BF theory, which is a lot like an extended TQFT, but not quite, because there are some divergences (infinities) that arise. In this case of 3d spacetime with the Lie group $\mathrm{SU}(2)$, BF theory gives a nice theory of quantum gravity called the Ponzano-Regge model. And, as I hinted back in "Week 232", we can let 2d space in this model be a manifold with *boundary* by poking little holes in space. Then these holes wind up acting like particles!

So, we get a relation like this:

$(n-2)$-dimensional manifolds	MATTER
$(n-1)$-dimensional manifolds with boundary	SPACE
n-dimensional manifolds with corners	SPACETIME

I like this a lot: it reminds me of the title of Weyl's famous book "Raum, Zeit, Materie", meaning "Space, Time, Matter". He never guessed this trio was related to the objects, morphisms and 2-morphisms in a weak 2-category! It's too bad we can't seem to get something like this to work for full-fledged quantum gravity.

It would be fun to talk more about this. However, to understand Morton's work more deeply, you need to understand a bit about "weak double categories". He explains them quite nicely, but I think I'll spend the rest of this Week's Finds giving a less detailed introduction, just to get you warmed up.

This chart should help:

	Bigons	**Squares**
Laws holding as equations	strict 2-categories	strict double categories
Laws holding up to isomorphism	weak 2-categories	weak double categories

2-categories are good for describing how to glue together 2-dimensional things that, at least in some abstract sense, are shaped like *bigons*. A "bigon" is a disc with its boundary divided into two halves:

$$\begin{tikzcd}
X \arrow[r, bend left=60, "f"{name=U}] \arrow[r, bend right=60, "g"'{name=D}] \arrow[r, dash, dashed] & Y \arrow[from=U, to=D, Rightarrow, "B"]
\end{tikzcd}$$

The big arrow indicates that we think of the bigon B as "going from" the top semicircle, f, to the bottom semicircle, g. Similarly, we think of the arcs f and g as going from the point X to the point Y.

Similarly, double categories are good for describing how to glue together 2-dimensional gadgets that are shaped like *squares*:

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[d, "g"'] \arrow[dr, phantom, "S"] & X' \arrow[d, "g'"] \\
Y \arrow[r, "f'"'] & Y'
\end{tikzcd}$$

Both 2-categories and double categories come in "strict" and "weak" versions. The strict versions have operations satisfying a bunch of laws "on the nose", as equations. In the weak versions, these laws hold up to isomorphism whenever possible.

A few more details might help. . . .

A 2-category has a set of objects, a set of morphisms $f\colon X \to Y$ going from any object X to to any object Y, and a set of 2-morphisms $T\colon f \Rightarrow g$ going from any morphism $f : X \to Y$ to any morphism $g\colon X \to Y$. We can visualize the objects as dots:

$$\underset{X}{\bullet}$$

the morphisms as arrows:

$$X \; \bullet \xrightarrow{f} \bullet \; Y$$

and the 2-morphisms as bigons:

$$\begin{tikzcd}
X \arrow[r, bend left=50, "f"{name=U}] \arrow[r, bend right=50, "g"'{name=D}] \arrow[r, dash, dashed] & Y
\arrow[Rightarrow, from=U, to=D, "B"]
\end{tikzcd}$$

We can compose morphisms like this:

$$\begin{tikzcd}
X \arrow[r, "f"] & Y \arrow[r, "g"] & Z
\end{tikzcd} \quad \text{gives} \quad \begin{tikzcd}
X \arrow[r, "fg"] & Z
\end{tikzcd}$$

We can also compose 2-morphisms vertically:

$$\begin{tikzcd}
X \arrow[r, bend left=60, "f"{name=U}] \arrow[r, "g"{name=M, description}] \arrow[r, bend right=60, "h"'{name=D}] & Y
\arrow[Rightarrow, from=U, to=M, "S"]
\arrow[Rightarrow, from=M, to=D, "T"]
\end{tikzcd} \quad \text{gives} \quad \begin{tikzcd}
X \arrow[r, bend left=50, "f"{name=U}] \arrow[r, bend right=50, "h"'{name=D}] \arrow[r, dash, dashed] & Y
\arrow[Rightarrow, from=U, to=D, "ST"]
\end{tikzcd}$$

and horizontally:

$$\begin{tikzcd} X \arrow[r, bend left=50, "f"] \arrow[r, bend right=50, "g"'] \arrow[r, phantom, "\Downarrow S"] & Y \arrow[r, bend left=50, "f'"] \arrow[r, bend right=50, "g'"'] \arrow[r, phantom, "\Downarrow T"] & Z & \text{gives} & X \arrow[r, bend left=50, "f"] \arrow[r, bend right=50, "g"'] \arrow[r, phantom, "\Downarrow S\cdot T"] & Y \end{tikzcd}$$

There are also a bunch of laws that need to hold. I don't want to list them; you can find them in Jeff's paper (also see "Week 80"). I just want to emphasize how a strict 2-category is different from a weak one.

In a strict 2-category, the composition of morphisms is associative on the nose:

$$(fg)h = f(gh)$$

and there are identity morphisms that satisfy these laws on the nose:

$$1f = f = f1$$

In a weak 2-category, these equations are replaced by 2-isomorphisms — that is, invertible 2-morphisms. And, these 2-isomorphisms need to satisfy new equations of their own!

What about double categories?

Double categories are like 2-categories, but instead of bigons, we have squares.

More precisely, a double category has a set of objects:

$$\begin{tikzcd} \bullet \\ X \end{tikzcd}$$

a set of horizontal arrows:

$$\begin{tikzcd} X \arrow[r, "f"] & X' \end{tikzcd}$$

a set of vertical arrows:

$$\begin{tikzcd} X \arrow[d, "g"'] \\ Y \end{tikzcd}$$

and a set of squares:

$$\begin{tikzcd} X \arrow[r, "f"] \arrow[d, "g"'] \arrow[dr, phantom, "S"] & X' \arrow[d, "g'"] \\ Y \arrow[r, "f'"'] & Y' \end{tikzcd}$$

We can compose the horizontal arrows like this:

$$\begin{tikzcd}
X \arrow[r, "f"] & Y \arrow[r, "f'"] & Z & \text{gives} & X \arrow[r, "f \cdot f'"] & Z
\end{tikzcd}$$

We can compose the vertical arrows like this:

$$\begin{tikzcd}
X \arrow[d, "g"'] & & \\
Y \arrow[d, "g"'] & \text{gives} & X \arrow[d, "gg'"'] \\
Z & & Z
\end{tikzcd}$$

And, we can compose the squares both vertically:

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[d, "g"'] \arrow[dr, phantom, "SS'"] & X' \arrow[d, "g'"] & & & \\
Y \arrow[r] \arrow[d, "h"'] \arrow[dr, phantom, "S'"] & Y' \arrow[d, "h'"] & \text{gives} & X \arrow[r, "f"] \arrow[d, "gh"'] & X' \arrow[d, "g'h'"] \\
Z \arrow[r, "f'"'] & Z' & & Z \arrow[r, "f'"'] & Z'
\end{tikzcd}$$

and horizontally:

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[d, "h"'] \arrow[dr, phantom, "S"] & Y \arrow[r, "g"] \arrow[d] \arrow[dr, phantom, "S'"] & Z \arrow[d, "h'"] & \text{gives} & X \arrow[r, "f \cdot g"] \arrow[d, "h"'] \arrow[dr, phantom, "S \cdot S'"] & Z \arrow[d, "h'"] \\
X' \arrow[r, "f'"'] & Y' \arrow[r, "g'"'] & Z' & & X' \arrow[r, "f' \cdot g'"'] & Z'
\end{tikzcd}$$

In a strict double category, both vertical and horizontal composition of morphisms is associative on the nose:

$$(fg)h = f(gh)$$

$$(f \cdot g) \cdot h = f \cdot (g \cdot h)$$

and there are identity morphisms for both vertical and horizontal composition, which satisfy the usual identity laws on the nose.

In a weak double category, we want these laws to hold only up to isomorphism. But, it turns out that this requires us to introduce bigons as well! The reason is fascinating but too subtle to explain here. I didn't understand it until Jeff pointed it out. But, it turns out that Dominic Verity had already introduced the right concept of weak double category — a gadget with both squares and bigons — in *his* Ph.D. thesis a while back:

13) Dominic Verity, *Enriched categories, internal categories, and change of base*, Ph.D. dissertation, University of Cambridge, 1992.

Interestingly, if you weaken *only* the laws for vertical composition, you don't need to introduce bigons. The resulting concept of "horizontally weak double category" has been studied by Grandis and Pare:

14) Marco Grandis and Bob Par, "Limits in double categories", *Cah. Top. Geom. Diff. Cat.* **40** (1999), 162–220. Also available at `http://www.dima.unige.it/~grandis/Dbl.Cahiers.pdf`

    Marco Grandis and Bob Par, "Adjoints for double categories", *Cah. Top. Geom. Diff. Cat.* **45** (2004), 193–240. Also available at `http://www.dima.unige.it/~grandis/Dbl.Adj.pdf`

and more recently by Martin Hyland's student Richard Garner:

15) Richard Garner, "Double clubs", available as `math.CT/0606733`

and Tom Fiore:

16) Thomas M. Fiore, "Pseudo algebras and pseudo double categories", available as `math.CT./0608760`.

At this point I should admit that the terminology in this whole field is a bit of a mess. I've made up simplified terminology for the purposes of this article, but now I should explain how it maps to the terminology most people use:

ME	THEM
strict 2-category	2-category
weak 2-category	bicategory
strict double category	double category
weak double category	double bicategory
horizontally weak double category	pseudo double category

Verity used the term "double bicategory" to hint that his gadgets have both squares and bigons, so they're like a blend of double categories and bicategories. It's a slightly unfortunate term, since experts know that a double category is a category object in Cat, but Verity's double bicategories are not bicategory objects in BiCat. Morton mainly uses Verity's double bicategories — but in the proof of his big theorem, he also uses bicategory objects in BiCat.

There's a lot more to say, but I'll stop here and let you read the rest in Jeff's paper!

Addenda: I thank Charlie Clingen for catching some typos in my diagrams, and Nathan Urban and Torbjrn Larsson for helping me update some of my information on the funding of NASA programs. I won't attempt to keep this information up to date, since it's changing too often. But, I'd like to it be correct as of the date I wrote it!

Sean Carroll writes:

> *Hi John–*
>
> *Just a couple of comments on This Week's Finds —*
>
> *You mention a bunch of missions that could "probably" be funded for the cost of a Moon base. That's being quite conservative! Each of those missions is about $1 billion or less, while the Moon base is upwards of $200 billion.*
>
> *And you asked about the neutrino image of the Sun. The "haze" is just an imaging problem, not a feature of the Sun; the resolution of this image is worse than 10 degrees (I forget the exact number), so we're certainly not looking at any substructure inside the Sun (whose entire disk is only half a degree wide).*
>
> *Sean*

I've deleted the word "probably". According to a comment on Joanne Hewett's blog entry, each pixel in the neutrino picture of the Sun is one degree in size. The Sun itself is just half a degree wide.

For more discussion, go to the n-Category Caf.

Week 243

December 25, 2006

Today I'd like to talk a bit about the first stars in the Universe, and some hotly contested possible observations of these stars. Then I want to describe a new paper by my student Derek Wise. But first — if anyone gave you a gift certificate for a bookstore this holiday season, here are two suggestions.

The first one is really easy and fun:

1) William Poundstone, *Fortune's Formula: The Untold Story of the Scientific Betting System that Beat the Casinos and Wall Street*, Farrar, Strauss and Giroux, New York, 2005.

Packed with rollicking tales of gangsters, horse-racing, blackjack, and insider trading, this is secretly the story of how Claude Shannon developed information theory — and how he and his sidekick John Kelly Jr. used it to make money in casinos and Wall Street. I'd known about Shannon's work on information… but not that he beat 99.9% of mutual fund managers, making an average compound return of 28% for many years — as compared to 27% for Warren Buffett!

This book has just a few equations in it. I was delighted by one discovered by Kelly, which I'd never seen before. Translating into my own favorite notation, it goes like this:

$$S = \log M$$

It's the fundamental equation relating gambling to information! Let me explain it — in language far more complicated than you'll see in Poundstone's book.

What's M? It's the best possible average growth of a gambler's money. For example, if his best possible strategy lets him triple his money on average, then $M = 3$.

What's S? This is the amount of "inside information" the gambler has: information he has, that the people he's betting against don't.

Some technical stuff: First, the above "average" is a geometric mean, not an arithmetic mean. Second, if we measure information in bits, we need to use base 2 in the logarithm. Physicists would probably prefer to use base e, which means measuring information in "nits". It doesn't really matter, but let's use base 2 for now.

To get a feeling for why Kelly's theorem is true, it's best to start with the simplest example. If $S = 1$, then $M = 2$. So, if a gambler receives one bit of inside information, he can double his money!

This sounds amazing, but it's also obvious.

Suppose you have one bit of inside information: for example, whether a flipped coin will land heads up or tails up. Then you can make a bet with somebody where they give you $1,000,000 if you guess the coin flip correctly, and you give them $1,000,000 if you guess wrong. This is a fair bet, so they will accept. That is, they'll *think* it's fair if they don't suspect you have inside information! But since you do have this information, you'll win the bet, and double your money on this coin flip.

Kelly's equation is usually phrased in terms of the *rate* at which the gambler gets inside information, and the *rate* at which his money grows. So, for example, to earn

12% interest annually, you only need to receive

$$\log(1.12) = 0.163$$

bits of inside information — and find some dupe willing to make bets with you about this.

The last part is the hard part: the "inside information" really needs to be information people don't believe you have. I must learn hundreds of bits of information about math each year — stuff only I know — but I haven't found anyone simultaneously smart enough to understand it and dumb enough to make bets with me about it!

Still, I like this relation between information theory and gambling, because one stream of Bayesian probability theory says probabilities are subjectively defined in terms of the bets you would accept.

The argument for this is called the "Dutch book argument". It basically shows how you can make money off someone who makes bets in ways that correspond to stupid probabilities that don't add to 1, or fail to be coherent in other ways:

2) Carlton M. Caves, *Probabilities as betting odds and the Dutch book*, available at `http://info.phys.unm.edu/~caves/reports/dutchbook.pdf`

So, there's a deep relation between gambling and probability — no news here, really.

But, there's also a deep relation between probability and information theory, discovered by Shannon. Briefly, it goes like this: the information you obtain by learning the value of a random variable is

$$S = -\sum_i p_i \log(p_i)$$

where the sum is taken over all the possible values of this random variable, and p_i is the probability that it takes its ith value. So, for example, if you flip a fair coin, where $p_1 = p_2 = 1/2$, the information you get by looking at the coin is

$$-\left[\frac{1}{2}\log\left(\frac{1}{2}\right) + \frac{1}{2}\log\left(\frac{1}{2}\right)\right] = 1$$

One bit!

So: gambling is related to probability, and probability is related to information. Kelly's result closes the circle by providing a direct relation between gambling and information!

But, apparently some of Kelly's ideas are still controversial in the world of economics and stock trading. If you read Poundstone's book, you'll learn why.

The next book takes more persistence to read:

3) Avner Ash and Robert Gross, *Fearless Symmetry: Exposing the Hidden Patterns of Numbers*, Princeton U. Press, Princeton, 2006.

The authors do a creditable job of what might at first seem utterly impossible: explaining heavy-duty modern number theory to ordinary mortals. The formal prerequisites are little more than high school algebra, and the style is expository, but anyone except an expert will need to stop and think at times.

They start by explaining modular arithmetic — you know, stuff like adding and multiplying " $\mod 7$". Then they tackle groups, and permutations, since the main theme of the book is symmetry. Then they move on to algebraic varieties, in a simple no-nonsense style cleverly adapted from Grothendieck's later work (without terrifying the reader by mentioning this fact).

Next they tackle some serious number theory: quadratic reciprocity, Galois groups, and elliptic curves. Then they describe more general forms of reciprocity, leading up to a taste of the Langlands program. They conclude with a sketch of how Fermat's last theorem was proved.

These days mathematical physicists are all excited about a variant of the Langlands program: the so-called "geometric" Langlands program, which is related to string theory. Drinfeld has been running a seminar on this at Chicago for years, but that's not what got the physicists interested — it's these papers by Witten that did it:

4) Anton Kapustin and Edward Witten, "Electric-magnetic duality and the geometric Langlands program", 225 pages, available as hep-th/0604151.

5) Sergei Gukov and Edward Witten, "Gauge theory, ramification, and the geometric Langlands program", 160 pages, available as hep-th/0612073.

So, if you're trying to learn this geometric Langlands stuff, and you want to fit it into the grand landscape of mathematics, the book Fearless Symmetry could be a fun way to learn some the math underlying the ordinary Langlands stuff.

I started girding myself for a discussion of the Langlands program in "Week 217", "Week 218" and "Week 221", but then I got distracted. I'll get back to it someday, but right now I'm in the mood for lighter stuff... so let me tell you a bit about the first stars.

The story starts around 380,000 years after the Big Bang, when the hot hydrogen and helium forming our Universe cooled down to 3000 kelvin - just cool enough for the electrons to stick to the atomic nuclei instead of zipping around on their own.

When the electrons in a gas are hot enough for some to zip around on their own, we say the gas is "ionized". When a *lot* of them are zipping around, we call it a "plasma". Because charged particles interact with the electromagnetic field, light doesn't pass through plasma cleanly: it keeps getting absorbed and re-emitted.

So, before our story started, you couldn't see very far: it would be like trying to look through a wall of fire. But, around 380,000 years after the Big Bang, the gas became transparent!

What would it have looked like? Nobody ever seems to talk about this. So, I'll just guess, and hope some expert corrects me.

Back when the gas filling the Universe was 5000 kelvin in temperature, just a bit cooler than the surface of the Sun, everything was yellow. You couldn't see far at all: you would have been blinded by a yellow glare.

But when it cooled to 4000 kelvin in temperature, the Universe became orange.

And when it cooled to 3000 kelvin, the Universe became red.

And when it cooled a tiny bit further, it became infrared. As far as visible light goes, the Universe became transparent!

This would happen everywhere more or less at once. But since light takes time to travel, you'd see a transparent sphere around you, expanding outwards at the speed of light, with reddish walls.

It's been sort of like this ever since.

So, when we look far away with our best telescopes, we look back in time to the time when the Universe became transparent — but no further. We're surrounded by a distant, ancient wall of fire. It's now about 13.3 billion light-years away — or 13.3 billion year back in time, if you prefer. And, it's receding at a rate of one light-year per year.

But by now, the light from this wall of fire has been severely redshifted. In other words, it's been stretched along with the expansion of the Universe — stretched by a factor of 1100, in fact!

So, what had been the hot infrared glow of 3000-kelvin plasma is now a feeble microwave glow corresponding to an icy temperature of 2.7 kelvin. This is the famous "cosmic microwave background radiation".

But let's go back in time. . . .

From the moment the hot gas became transparent to the time when the first stars formed, the Universe was dark except for the dimming infrared glow of that distant wall of fire. This era is called the "Dark Ages".

During the Dark Ages, gas cooled down and clumped under its own gravity — apparently with a lot of help from cold dark matter of some unknown sort. Without postulating this matter, nobody can figure out how galaxies formed as soon as they did.

As befits their name, the Dark Ages are still shrouded in mystery. There are a lot of unanswered questions besides the nature of dark matter. Which formed first — individual stars, or galaxies? And, when did the Dark Ages end?

It's currently believed that the first stars formed sometime between 150 million and 1 billion years after the Big Bang.

At the later end of that range, the Universe could have gotten quite cold before starlight warmed up the interstellar gas and reionized it. There's even a spooky theory that the Universe was full of hydrogen snowflakes near the end of the Dark Ages — see "Week 196" for more on this, and a timeline of the earlier history of the Universe.

But, the current best guess, based on data from the Wilkinson Microwave Anisotopy Probe, says that reionization happened 400 million years after the Big Bang:

6) Marcelo A. Alvarez, Paul R. Shapiro, Kyungjin Ahn and Ilian T. Iliev, "Implications of WMAP 3 year data for the sources of reionization", *Astrophys. J.* **644** (2006), L101–L104. Also available as astro-ph/0604447.

This would be too early for hydrogen snow, since my rough calculation says the microwave background radiation was 30 kelvin then, while hydrogen freezes at 14 kelvin.

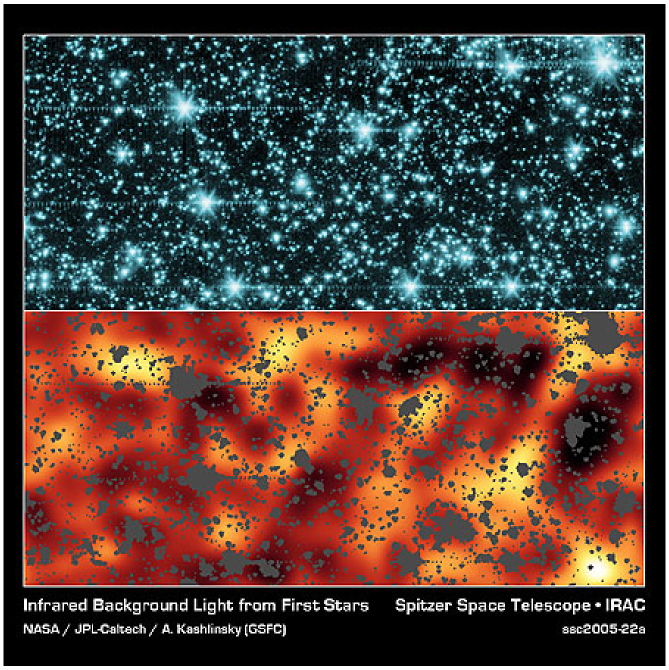
What were the first stars like? Without heavier elements to catalyze nuclear fusion, they could have been larger than current-day stars: perhaps hundreds of times the size of our Sun! These so-called Population III stars have not actually been seen. But, it's possible that we've finally caught a glimpse of them, not individually but in a sort of statistical sense:

7) A. Kashlinsky, R. G. Arendt, J. Mather and S. H. Moseley, "New measurements of cosmic infrared background fluctuations from early epochs", to appear in *Ap. J. Letters*. Available as astro-ph/0612445.

8) A. Kashlinsky, R. G. Arendt, J. Mather and S. H. Moseley, "On the nature of the sources of the cosmic infrared background", to appear in *Ap. J. Letters*. Available as astro-ph/0612447.

Using delicate techniques to carefully sift through the *infrared* (not microwave) background radiation, the authors claim to find radiation not accounted for by previously known sources. Assuming the standard cosmological scenario, the sources of this radiation date back to less than 1 billion years after the Big Bang, and were individually much brighter than current-day stars.

Here's a picture of their data:


Infrared Background Light from First Stars
Spitzer Space Telescope • IRAC
NASA / JPL-Caltech / A. Kashlinsky (GSFC)
ssc2005-22a

9) NASA / JPL-Caltech / A. Kashlinsky, "Infrared background light from first stars", `http://www.spitzer.caltech.edu/Media/releases/ssc2005-22/`

On top is a photograph taken by the Spitzer Space Telescope: a 10-hour infrared exposure of a tiny patch of sky, 6×12 arcminutes across, chosen for having a bare minimum of foreground stars, galaxies and dust. (For comparison, the Moon is 30 arcminutes across.) On the bottom is the same picture with known sources of infrared subtracted. What's left may be the severely redshifted light from early stars!

Or, it may not. In the following news story, Ned Wright of UCLA said, "I'm very skeptical of this result. I think it's wrong. I think what they're seeing is incompletely subtracted residuals from nearby sources."

10) Dinesh Ramde, Associated Press, "Hints of early stars may have been found", `http://www.usatoday.com/tech/science/space/2005-11-02-early-stars_x.htm`

So, we'll have to see how it goes. . . .

But in the meantime, we can think about mathematical physics. My student Derek Wise is graduating this year, and he's doing his thesis on Cartan geometry, MacDowell-Mansouri gravity and BF theory. Let me say a little about this paper of his:

11) Derek Wise, "MacDowell-Mansouri gravity and Cartan geometry", available as gr-qc/0611154.

Elie Cartan is one of the most influential of 20th-century geometers. At one point he had an intense correspondence with Einstein on general relativity. His "Cartan geometry" idea is an approach to the concept of parallel transport that predates the widely used Ehresmann approach (connections on principal bundles). It simultaneously generalizes Riemannian geometry and Klein's Erlangen program (see "Week 213"), in which geometries are described by their symmetry groups:

$$\begin{tikzcd}
\text{Euclidean geometry} \arrow[r] \arrow[d] & \text{Klein geometry} \arrow[d] \\
\text{Riemannian geometry} \arrow[r] & \text{Cartan geometry}
\end{tikzcd}$$

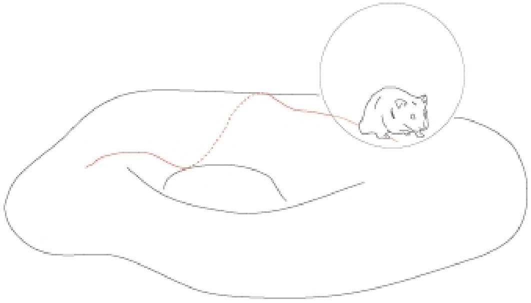
Given all this, it's somewhat surprising how few physicists know about Cartan geometry!

Recognizing this, Derek explains Cartan geometry from scratch before showing how it underlies the so-called MacDowell-Mansouri approach to general relativity. This plays an important role both in supergravity and Freidel and Starodubtsev's work on quantum gravity (see "Week 235") — but until now, it's always seemed like a "trick".

What's the basic idea? Derek explains it all very clearly, so I'll just provide a quick sketch. Cartan describes the geometry of a lumpy bumpy space by saying what it would be like to roll a nice homogeneous "model space" on it. Homogeneous spaces are what Klein studied; now Cartan takes this idea and runs with it... or maybe we should say he *rolls* with it!

For example, we could study the geometry of a lumpy bumpy surface by rolling a *plane* on it. If our surface is itself a plane, this rolling motion is trivial, and we say the surface is "flat" in the sense of Cartan geometry. But in general, the rolling motion is interesting and serves to probe the geometry of the surface.

Alternatively, we could study the geometry of the same surface by rolling a *sphere* on it. Derek illustrates this with a picture of a hamster crawling around in a plastic "hamster ball", which is something you can actually buy for your pet hamster to let it explore your house without escaping or getting in trouble:



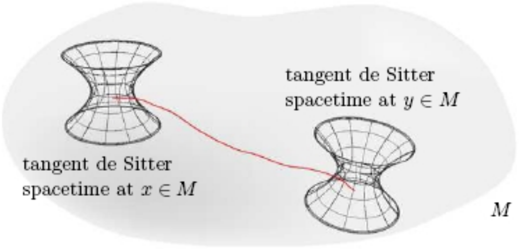
(I've read about falling cats in papers on gauge theory, but this is the first mathematical physics paper I've read containing the word "hamster".)

If our surface is itself a sphere of the same radius, this rolling motion is trivial, and we say the surface is flat in the sense of Cartan geometry — but now it's a different sense than when we used a plane as our "model geometry"!

Which model geometry should we use in a given problem? It depends on which one best approximates the lumpy bumpy space we're studying!

The ordinary formulation of general relativity fits into this framework, with a little work. Two well-known mathematical gadgets called the "Lorentz connection" and "coframe field" fit together to describe what would happen if we rolled a copy of Minkowski spacetime over the lumpy bumpy spacetime we live in.

That's great if Minkowski spacetime is the best homogeneous approximation to the spacetime we live in. But nowadays we think the cosmological constant is nonzero, so the Universe is expanding in a roughly exponential way. This makes another model geometry, "deSitter spacetime", the best one to use!



So, if we know Cartan geometry, we can use that... and we get something called the MacDowell-Mansouri formulation of gravity. Or, if we don't want our spacetime to have lumps and bumps — if we want it to look locally just like the Klein model geometry — we can use a different theory, a topological field theory called BF theory (see "Week 232").

In short, the passage from a topological field theory describing a "locally homogeneous" spacetime to full-fledged gravity with all its lumps and bumps is nicely understood in terms of how Cartan's approach to geometry generalizes Klein's!

For more details, you'll just have to read Derek's paper. You might also try these:

12) Michel Biesunski, "Inside the coconut: the Einstein-Cartan discussion on distant parallelism", in *Einstein and the History of General Relativity*, eds. D. Howard and J. Stachel, Birkhauser, Boston, 1989.

This describes the correspondence between Cartan and Einstein. I believe this centered, not on Cartan geometry per se, but on the "teleparallel" formulation of gravity (see "Week 176"). But, they're somewhat related.

13) Richard W. Sharpe, *Differential Geometry: Cartan's Generalization of Klein's Erlangen Program*, Springer-Verlag, New York, 1997.

This is the main textbook on Cartan geometry. But, it's probably best to read a few chapters of Derek's paper first, since the key ideas are presented more intuitively.

My friend the geometer and analyst Rafe Mazzeo, whom I recently saw at Stanford, told me that Cartan geometry was all the rage these days. I'm embarrassed to say I hadn't known this! I think the kinds of Cartan geometry being intensively studied are related to conformal geometry, CR structures and stuff like that. . .

Merry Christmas!

Addenda: I thank Chris Weed for catching typos. For more discussion, go to the n-Category Caf.

The Universe has as many different centers as there are living beings in it.

— *Alexander Solzhenitsyn*

Week 244

February 2, 2007

In January I spent a week at this workshop at the Fields Institute in Toronto:

1) *Higher Categories and Their Applications*, `http://math.ucr.edu/home/baez/fields/`

It was really fun — lots of people working on n-categories were there. I'll talk about it next time. But as usual, more happens at a fun conference than can possibly be reported. So, this time I'll only talk about a conversation I had in a caf before the conference started!

But first, here's a fun way to challenge your math pals:

Q: When the first calculus textbook was written — and in what language?

A: In 1530, in Malayalam — a south Indian language!

This book is called the *Ganita Yuktibhasa*, or "compendium of astronomical rationales". It was written by Jyesthadeva, an astronomer and mathematician from Kerala — a state on the southwest coast of India. It summarizes and explains the work of many researchers of the Kerala school, which flourished from the 1400's to the 1600's. But it's unique for its time, since it contains proofs of many results.

For example, it has a proof that

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \dots$$

Of course, this result isn't stated in modern notation! It's actually stated as a poem — a recipe for the circumference of a circle, which in translation goes something like this:

> Multiply the diameter by four. Subtract from it and add to it alternately the quotients obtained by dividing four times the diameter by the odd numbers 3, 5, etc.

The proof sounds nice! Jyesthadeva starts with something like this:

$$\frac{\pi}{4} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \frac{1}{1+\left(\frac{n}{N}\right)^2}$$

In modern terms, the right-hand side is just the integral

$$\int_0^1 \frac{dx}{1+x^2}$$

You can use geometry to see this equals $\pi/4$. Then, as far as I can tell, he writes

$$\frac{1}{1+\left(\frac{n}{N}\right)^2} = 1 - \left(\frac{n}{N}\right)^2 + \left(\frac{n}{N}\right)^4 - \dots$$

and notes that

$$1^k + 2^k + \ldots + N^k \sim \frac{N^{k+1}}{k+1}$$

for large N. This gives

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots$$

Voila!

In fact, this result goes back to Madhava, an amazing mathematician from Kerala who lived much earlier, from 1350 to 1425. What's even more impressive is that Madhava also knew a formula equivalent to the more general result

$$\arctan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \ldots$$

He used this to compute π to 11 decimal places!

It's an interesting question whether any of the results of the Kerala school found their way west and influenced the development of mathematics in Europe. There's been a lot of speculation, but nobody seems to know for sure. For more info, try these:

2) The MacTutor History of Mathematics Archive, "Madhava of Sangamagramma", `http://www-history.mcs.st-andrews.ac.uk/Biographies/Madhava.html`

3) The MacTutor History of Mathematics Archive, "Jyesthadeva", `http://www-history.mcs.st-andrews.ac.uk/Biographies/Jyesthadeva.html`

4) Wikipedia, "Yuktibhasa", `http://en.wikipedia.org/wiki/Yuktibhasa`

Before the conference started, I spent a nice morning talking with Tom Leinster in a caf on Bloor Street. There's nothing like talking about math in a nice warm caf when it's cold outside! At some point my former grad student Toby Bartels showed up — he'd just taken a long Greyhound bus from Nebraska — and joined in the conversation. We talked about this paper:

5) Tom Leinster, "The Euler characteristic of a category", available as `math.CT/0610260`.

Everyone know how to measure the size of a set — by its number of elements, or "cardinality". But what's the size of a category? That's the question this paper tackles!

Some categories are just sets in disguise: the "discrete" categories, whose only morphisms are identity morphisms. We'd better define the size of such a category to be the cardinality of its set of objects.

For example, the category with just one object and its identity morphism is called 1. It looks sort of like this:

$$\bullet$$

where I've drawn the object but not its identity morphism. Clearly, its size should be 1.

We could also have a category with just two objects and their identity morphisms. It looks like this:

$$\bullet \qquad\qquad \bullet$$

and its size should be 2.

But what about this?

$$\bullet \longleftrightarrow \bullet$$

Here we have a category with two objects and an invertible morphism between them, which I've drawn as an arrow pointing both ways. Again, I won't draw the identity morphisms.

In other words, we have two objects that are *isomorphic* — and in a unique way. How big should this category be?

Any mathematician worth her salt knows that having two things that are isomorphic in a unique way is just like having one: you can't do anything more with them — or less. So, the size of this category:

$$\bullet \longleftrightarrow \bullet$$

should equal the size of this one:

$$\bullet$$

namely, 1.

More technically, we say these categories are "equivalent". We'll demand that equivalent categories have the same size. This is a powerful principle. If we didn't insist on this, we'd be insane.

But what about this category:

$$\bullet \longrightarrow \bullet$$

Now we have two objects and a morphism going just one way! This is *not* equivalent to a discrete category, so we need a new idea to define its size.

If we were willing to make up new kinds of numbers, we could make up a new number for the size of this category. But let's suppose that this is against the rules.

There's a cute way to turn any category into a space, which I described in "Week 70" — and in more detail in items J and K of "Week 117", back when I was giving a minicourse on homotopy theory. If we do this to the category

$$\bullet \longrightarrow \bullet$$

what do we get? The unit interval, of course! It's a pretty intuitive notion, at least in this example.

We also get the unit interval if we turn this guy

$$\bullet \longleftrightarrow \bullet$$

into a space. So, even though these categories aren't equivalent, they give the same space. So, let's declare that they have the same size — namely, 1.

In fact, let's adopt this as a new principle! We'll demand that two categories have the same size whenever they give the same space.

Whenever categories are equivalent, they give the same space (where "the same" means "homotopy equivalent"). So, our new principle includes our previous principle as a special case. But, we can say more. If you like adjoint functors, you'll enjoy this:

whenever there's a pair of adjoint functors going between two categories, they give the same space. For example, these categories

$$\begin{tikzcd}\bullet \arrow[r] & \bullet\end{tikzcd}$$

and

$$\begin{tikzcd}\bullet \arrow[r, leftrightarrow] & \bullet\end{tikzcd}$$

aren't equivalent, but there's a pair of adjoint functors going between them. (If you don't like adjoint functors, oh well — just ignore this.)

Next, what's the size of this category?

$$\begin{tikzcd}\bullet \arrow[r, bend left] \arrow[r, bend right] & \bullet\end{tikzcd}$$

This is my feeble attempt to draw a category with two objects, and two morphisms going from the first object to the second.

If we turn this category into the space, what do we get? The circle, of course! But what's the "size", or "cardinality", of a circle?

That's a tricky puzzle, because it's hard to know what counts as a right answer. It turns out the right answer is zero. Why? Because the "Euler characteristic" of the circle is zero!

As you may know, Euler lived in Knigsberg, a city with lots of islands and bridges:

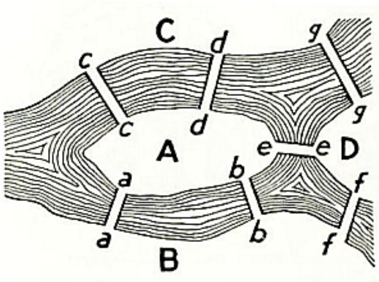


Figure 98. *Geographic Map: The Königsberg Bridges.*

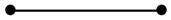
In fact, he published a paper in 1736 showing that you can't walk around Knigsberg and cross each bridge exactly once, winding up where you started. My crazy theory is that living there also helped him invent the concept of Euler characteristic. I have no evidence for this, except for this apocryphal story I just made up:

Once upon a time, Euler was strolling along one of the bridges of Knigsberg. He looked across the river, and noticed that workers were building a bridge to a small island that had previously been unconnected to the rest. He noticed that this reduced the number of isolated islands by one. Of course, anyone could have seen that! But in a burst of genius, Euler went further — he realized this meant a bridge was like a "negative island". And so, he invented the concept of "Euler characteristic". In its simplest form, it's just the number of islands minus the number of bridges.

For example, if you have two islands in the sea:

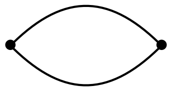
the land has Euler characteristic 2.

If you build a bridge:



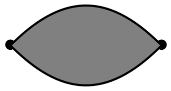
the land now has Euler characteristic 1. This makes sense, because the land is now effectively just one island. So, a bridge acts as a "negative island"!

But now, if you build a *second* bridge:



the land has Euler characteristic 0. This is sort of weird. But, Euler saw it was a good idea.

To understand why, you have to go further and imagine building a "bridge between bridges" — filling in the space between the bridges with an enormous deck:



This reduces the number of bridges by one. We've effectively got one island again, though much bigger now. So, we're back to having Euler characteristic 1.

In short, adding a "bridge between bridges" should add 1 to the Euler characteristic. Just as a bridge counts as a negative island, a bridge between bridges counts as a negative bridge — or an island:

$$-(-1) = 1.$$

It's all consistent, in its own weird way.

So, Euler defined the Euler characteristic to be

$$V - E + F$$

where V is the number of islands (or "vertices"), E is the number of bridges (or "edges") and F is the number of bridges between bridges (or "faces").

At least that's how the story goes.

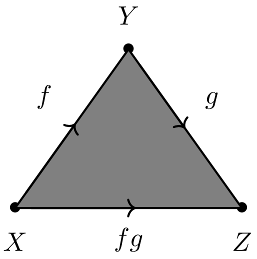
By the way, you must have noticed that the number 1 looks like an interval, while the number 0 looks like a circle. But did you notice that the Euler characteristic of the interval is 1, and the Euler characteristic of the circle is 0? I can never make up my mind whether this is a coincidence or not.

Anyway, we can easily generalize the Euler characteristic to higher dimensions, and define it as an alternating sum. And that turns out to be important for us now, because it turns out that often when we turn a category into a space, we get something higher-dimensional!

This shouldn't be obvious, since I haven't told you the rule for turning a category into a space. You might think we always get something 1-dimensional, built from vertices (objects) and edges (morphisms). But the rule is more subtle. Whenever we have 2 morphisms end to end, like this:

$$\begin{tikzcd} \underset{X}{\bullet} \arrow[r, "f"] & \underset{Y}{\bullet} \arrow[r, "g"] & \underset{Z}{\bullet} \end{tikzcd}$$

we can compose them and get a morphism fg going all the way from x to z. We should draw this morphism too... so the space we get is a *triangle*:



More importantly, the triangle is filled in, just like Euler's "bridge between bridges", to show that it's *solid*, not hollow.

Simlarly, when we have 3 morphisms laid end to end we get a tetrahedron, and so on.

Using these rules, it's not hard to find a category that gives a sphere, or a torus, or an n-holed torus, when you turn it into a space. I'll leave that as a puzzle.

In fact, for *any* manifold, you can find a category that gives you that manifold when you turn it into a space! In fact we can get any space at all this way, up to "weak homotopy equivalence" — whatever that means. So, let's adopt a new principle: whenever our category gives a space whose Euler characteristic is well-defined, we should define the size of our category to be that.

I say "when it's well-defined", because it's also possible for a category — even one with just finitely many objects and morphisms — to give an infinite-dimensional space whose Euler characteristic is a divergent series:

$$n_0 - n_1 + n_2 - n_3 + n_4 - \ldots$$

Okay. At this point it's time for me to say what Leinster actually did: he came up with a *formula* that you can use to compute the size of a category, without using any topology. Sometimes it gives divergent answers — which is no shame: after all, some categories are infinitely big. But when it converges, it satisfies all the principles I've mentioned.

Even better, it works for a lot of categories that give spaces whose Euler chacteristic diverges! For example, we can take any group G and think of it as a category with one object, with the group elements as morphisms. When we turn this category into a space, it becomes something famous called the "classifying space" of G. This is often an infinite-dimensional monstrosity whose Euler characteristic diverges. But, Leinster's formula still works — and it gives

$$1/|G|$$

the reciprocal of the usual cardinality of G.

Now we're getting fractions!

For example, suppose we take G to be the group with just 2 elements, called $\mathbb{Z}/2$. If we think of it as a category, and then turn that into a space, we get a huge thing usually called "infinite-dimensional real projective space", or $\mathbb{RP}^\infty$ for short. This is built from one vertex, one edge, one triangle, and so on. So, if we try to work out its Euler characteristic, we get the divergent series

$$1 - 1 + 1 - 1 + 1 - \ldots$$

But, if we use Leinster's formula, we get $1/2$. And that's cute, because once there were heated arguments about the value of

$$1 - 1 + 1 - 1 + 1 - \ldots$$

Some mathematicians said it was 0:

$$(1 - 1) + (1 - 1) + (1 - 1) + \ldots = 0$$

while others said it was 1:

$$1 + (-1 + 1) + (-1 + 1) + (-1 + 1) + \ldots = 1$$

Some said "it's divergent, so forget it!" But others wisely compromised and said it equals $1/2$. This can be justified using "Abel summation".

All this may seem weird — and it is; that's part of the fun. But, Leinster's answer matches what you'd expect from the theory of "homotopy cardinality":

6) John Baez, "The mysteries of counting: Euler characteristic versus homotopy cardinality", `http://math.ucr.edu/home/baez/counting/`

This webpage has transparencies of a talk I gave on this, and lots of links to papers that generalize the concepts of cardinality and Euler characteristic. I'm obsessed with this topic. It's really exciting to think about new ways to extend the simplest concepts of math, like counting.

That's why I invented a way to compute the cardinality of a groupoid — a category where every morphism has an inverse, so all the morphisms describe "symmetries". The idea is that the more symmetries an object has, the smaller it is. Applying this to the above example, where our category has one object, and this object has 2 symmetries, one gets $1/2$. If this seems strange, try the explanation in "Week 147".

Later James Dolan took this idea, generalized it to a large class of spaces that don't necessarily come from groupoids, and called the result "homotopy cardinality". We wrote a paper about this.

What Leinster has done is generalize the idea in another direction: from groupoids to categories. The cool thing is that his generalization matches the Euler characteristic of spaces coming from categories (when that's well-defined, without divergent series) and the homotopy cardinality of spaces coming from groupoids (when that's well-defined).

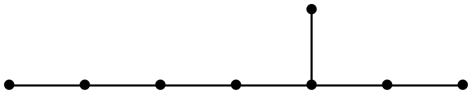
Of course he doesn't call his thing the "size" of a category; he calls it the "Euler characteristic" of a category.

Our conversation over coffee was mainly about me trying to understand the formula he used to define this Euler characteristic. One thing I learned is that the "category algebra" idea plays a key role here.

It's a simple idea. Given a category X, the category algebra $C[X]$ consists of all formal complex linear combinations of morphisms in X. To define the multiplication in this algebra, it's enough to define the product fg whenever f and g are morphisms in our category. If the composite of f and g is defined, we just let fg be this composite. If it's not, we set $fg = 0$.

Mathematicians seem to be most familiar with the category algebra idea when our category happens to be a group (a category with one object, all of whose morphisms are invertible). Then it's called a "group algebra".

Category algebras are also pretty familiar when our category is a "quiver" (a category formed from a directed graph by freely throwing in formal composites of edges). Then it's called a "quiver algebra". These are really cool — especially if our graph becomes a Dynkin diagram, like this:



when we ignore the directions of the edges. To see what I mean, try item E in "Week 230", where I sketch how these quiver algebras are related to quantum groups. There's a lot more to say about this, but not today!

In combinatorics, category algebras are familiar when our category is a "partially ordered set", or "poset" for short (a category with at most one morphism from any given object to any other). These category algebras are usually called "incidence algebras".

In physics, Alain Connes has given a nice explanation of how Heisenberg invented "matrix mechanics" when he was trying to understand how atoms jump from one state to another, emitting and absorbing radiation. In modern language, Heisenberg took a groupoid with n objects, each one isomorphic to each other in a unique way. He called the objects "states" of a quantum system, and he called the morphisms "transitions". Then, he formed its category algebra. The result is the algebra of $n \times n$ matrices!

(This might seem like a roundabout way to get to $n \times n$ matrices, but Heisenberg *didn't know about matrices* at this time. They weren't part of the math curriculum for physicists back then!)

Connes has generalized the heck out of Heisenberg's idea, studying the "groupoid algebras" of various groupoids.

So, category algebras are all over the place. But for some reason, few people study all these different kinds of category algebra in a unified way — or even *realize* they're all category algebras! I feel sort of sorry for this neglected concept. That's one reason I was happy to see it plays a role in Leinster's definition of the Euler characteristic for categories.

Suppose our category X is finite. Then, we can define an element of the category algebra $C[X]$ which is just the sum of all the morphisms in X. This is called ζ, or the "zeta function" of our category. Sometimes ζ has an inverse, and then this inverse is called μ, or the "Mbius function" of our category.

Actually, these terms are widely used only when our category is a poset, thanks to the work of Gian-Carlo Rota, who used these ideas in combinatorics:

7) Gian-Carlo Rota, "On the foundations of combinatorial theory I: Theory of Mbius Functions", *Zeitschrift fr Wahrscheinlichkeitstheorie und Verwandte Gebiete* **2** (1964), 340–368.

If you want to know what these ideas are good for, check this out:

8) Wikipedia, "Incidence algebra", `http://en.wikipedia.org/wiki/Incidence_algebra`

See the stuff about Euler characteristics in this article? That's a clue! The relation to the Riemann zeta function and its inverse (the original "Mbius function") are clearer here:

9) Wikipedia, "Mbius inversion formula", `http://en.wikipedia.org/wiki/M%C3%B6bius_inversion_formula`

These show up when we think of the whole numbers $1, 2, 3, \ldots$ as a poset ordered by divisibility.

Anyway, Leinster has wisely generalized this terminology to more general categories. And when $\zeta^{-1} = \mu$ exists, it's really easy to define his Euler characteristic of the category X. You just write μ as a linear combination of morphisms in your category, and sum all the coefficients in this linear combination!

Unfortunately, there are lots of important categories whose zeta function is not invertible: for example, any group other than the trivial group. So, Leinster needs a somewhat more general definition to handle these cases. I don't feel I deeply understand it, but I'll explain it, just for the record.

Besides the category algebra $C[X]$, consisting of linear combinations of morphisms in X, there's also a vector space consisting of linear combinations of *objects* in X. Heisenberg would probably call this "the space of states", and call $C[X]$ the "algebra of observables", since that's what they were in his applications to quantum physics. Let's do that.

The algebra of observables has an obvious left action on the vector space of states, where a morphism $f\colon x \to y$ acts on x to give y, and it acts on every other object to give 0. In Heisenberg's example, this is precisely how he let the algebra of observables act on states.

The algebra of observables also has an obvious *right* action on the vector space of states, where $f\colon x \to y$ acts on y to give x, and it acts on every other object to give 0.

Leinster defines a "weighting" on X to be an element w of the vector space of states with

$$\zeta w = 1$$

Here "1" is the linear combination of objects where all the coefficients equal 1. He also defines a "coweighting" to be an element w^* in the vector space of states with

$$w^* \zeta = 1$$

If ζ has an inverse, our category has both a weighting and a coweighting, since we can solve both these equations to find w and w^*. But often there will be a weighting and coweighting even when ζ doesn't have an inverse. When both a weighting and coweighting exist, the sum of the coefficients of w equals the sum of coefficients of w^* — and this sum is what Leinster takes as the "Euler characteristic" of the category X!

This is a bit subtle, and I don't deeply understand it. But, Leinster proves so many nice theorems about this "Euler characteristic" that it's clearly the right notion of the size of a category — or, with a further generalization he mentions, even an n-category! And, it has nice relationships to other ideas, which are begging to be developed further.

We're still just learning to count.

Addendum: For more discussion, go to the n-Category Caf.

Week 245

February 11, 2007

The University of Toronto is an urban campus, rather grey and chilly at this time of year. Nestled amid other buildings at the southern edge of campus, the Fields Institute doesn't stand out.

But inside, you'll find a spacious and peaceful atrium, with a fireplace to keep you cozy. A spiral staircase winds up three or four stories. Hanging from the ceiling far above is a 3d model of the "120-cell": a beautiful 4-dimensional solid with 120 regular dodecahedra as faces.

This is a tribute to the great geometer H. S. M. Coxeter, master of polyhedra, who worked for 60 years at the University of Toronto after studying philosophy at Cambridge under Wittgenstein. You'll also find Coxeter's piano sitting at the base of the spiral staircase.

It's out of tune, but resting on it there's a wonderful strange portrait of him playing the very same piano — at the age of three. He looks a bit like the child Mozart. And indeed, at the age of 12 Coxeter composed an opera!

The Fields Institute specializes in having conferences, and it's a great place for that. A friendly and efficient staff, public workstations, wireless internet everywhere, a nice little cafe in the back, and the centerpiece: a large lecture room with 3 double blackboards. Unfortunately the middle blackboard doesn't stay up — it's needed that repair for years, old-timers say. But apart from that, everything is as close to mathematician's heaven as could be expected.

Eugenia Cheng, Peter May and I ran a workshop at the Fields Institute from January 9th to 13th:

1) "Higher Categories and Their Applications", `http://math.ucr.edu/home/baez/fields/`

You can see photos of people and abstracts of their talks at this site. You can also see PDF files of many of their talks — and even listen to talks!

The first day, Tuesday, was all about 2-categories and 3-categories - "lower category theory", you might say. While some are eagerly sailing into the stratosphere of n-categories for general n, or even $n = \infty$, there's still a lot to understand for $n = 2$ and 3.

For starters, Tom Leinster spoke about strict 2-categories versus weak ones (also known as bicategories). It's a famous fact — a generalization of Mac Lane's coherence theorem — that every weak 2-category $\mathcal{C}$ is equivalent to a strict one $\mathrm{st}(\mathcal{C})$. However, this is true *if* your notion of equivalence is suitably weak! In short, what we've got is an inclusion of weak 3-categories:

$$i\colon \mathsf{Strict2Cat} \to \mathsf{Weak2Cat}$$

where

$$\begin{aligned}\mathsf{Strict2Cat} = [&\text{strict 2-categories},\\ &\text{strict 2-functors},\\ &\text{strict natural transformations},\\ &\text{modifications}]\end{aligned}$$

and

$$\begin{aligned}\mathsf{Weak2Cat} = [&\text{weak 2-categories},\\ &\text{weak 2-functors},\\ &\text{weak natural transformations},\\ &\text{modifications}]\end{aligned}$$

Every object in Weak2Cat is equivalent to one in the image of this inclusion. But, the inclusion is not itself an equivalence!

Steve Lack spoke about Gray-categories, also known as "semistrict" 3-categories — a convenient middle ground between the strict 3-categories and the weak ones (also known as tricategories).

The idea here goes back to John Gray. In the usual Cartesian product of categories, whenever we have a morphism

$$f\colon A \to B$$

in the first category and a morphism

$$f'\colon A' \to B'$$

in the second, we get a commuting square:

$$\begin{tikzcd}
(A,A') \arrow[r, "(f,1)"] \arrow[d, "(1,f')"'] & (B,A') \arrow[d, "(1,f')"] \\
(A,B') \arrow[r, "(f,1)"'] & (B,B')
\end{tikzcd}$$

in their Cartesian product. The same is true for the Cartesian product of 2-categories. But in the "Gray" tensor product of 2-categories, these squares commute only up to 2-isomorphism. And, we can use this weakening of the Cartesian product to weaken the concept of strict 3-category, and obtain the concept of "semistrict" 3-category, or "Gray-category".

Here's how. A strict 3-category is a gizmo with:

- a bunch of objects,
- for any pair of objects x, y, a 2-category $\mathrm{Hom}(x, y)$,

and

- for any triple of objects x, y, z, a 2-functor

$$\circ\colon\ \mathrm{Hom}(x,y) \times \mathrm{Hom}(y,z) \to \mathrm{Hom}(x,z)$$

  such that

- associativity and the unit laws hold.

A semistrict 3-category is a gizmo with:

- a bunch of objects,
- for any pair of objects x, y, a 2-category $\mathrm{Hom}(x, y)$,

and

- for any triple of objects x, y, z, a 2-functor

$$\circ\colon\ \mathrm{Hom}(x,y) \otimes \mathrm{Hom}(y,z) \to \mathrm{Hom}(x,z)$$

  where $\otimes$ is the Gray tensor product, such that

- associativity and the unit laws hold.

The slight difference is very important. Not every weak 3-category is equivalent to a strict one. But, they're all equivalent to semistrict ones!

There are, alas, some deficiencies in the semistrict world, which Steve Lack has recently noted:

2) Steve Lack, “Bicat is not triequivalent to Gray”, available as math.CT/0612299.

To understand this, you may need a little warmup. Given strict 2-categories $\mathcal{B}$ and $\mathcal{C}$ there's a strict 2-category $\mathrm{Hom}(\mathcal{B},\mathcal{C})$ such that strict 2-functors

$$\mathcal{A}\times\mathcal{B}\to\mathcal{C}$$

are in natural 1-1 correspondence with strict 2-functors

$$\mathcal{A}\to\mathrm{Hom}(\mathcal{B},\mathcal{C})$$

Here's what $\mathrm{Hom}(\mathcal{B},\mathcal{C})$ is like:

- $\mathrm{Hom}(\mathcal{B},\mathcal{C})$ has strict 2-functors from $\mathcal{B}$ to $\mathcal{C}$ as objects,
- strict natural transformations between these as morphisms,
- modifications between these as 2-morphisms.

We can pose the same question with the Gray tensor product replacing the Cartesian product. Given 2-categories $\mathcal{B}$ and $\mathcal{C}$ there's a 2-category $[\mathcal{B},\mathcal{C}]$ such that strict 2-functors

$$\mathcal{A}\otimes\mathcal{B}\to\mathcal{C}$$

are in natural 1-1 correspondence with strict 2-functors

$$\mathcal{A}\to[\mathcal{B},\mathcal{C}]$$

Here's what $[\mathcal{B},\mathcal{C}]$ is like:

- $[\mathcal{B},\mathcal{C}]$ has strict 2-functors from $\mathcal{B}$ to $\mathcal{C}$ as objects,
- weak natural transformations between these as morphisms,
- modifications between these as 2-morphisms.

This suggests that we consider a 3-category intermediate between Strict2Cat and Weak2Cat. It's called Gray, and it goes like this:

Gray = [strict 2-categories,
strict 2-functors,
weak natural transformations,
modifications]

We have inclusions of weak 3-categories:

$$\mathsf{Strict2Cat}\to\mathsf{Gray}\to\mathsf{Weak2Cat}$$

and Lack shows, not only that the second inclusion fails to be an equivalence, but that there's *no* equivalence between Gray and Weak2Cat.

All this suggests that for some purposes we really need to face up to weak 2-categories: the strict and semistrict setups aren't flexible enough for every job. The same is undoubtedly true at the 3-category level — and that's where the next talk comes in!

In the next talk, Nick Gurski spoke about weak 3-categories. He wrote his thesis about these, and I'm starting to really wish he'd put his thesis on the arXiv, so everyone can see how cool it is and learn more about 3-categories. But, I guess he wants to perfect it.

In his talk, Nick not only explained the definition of weak 3-category, which is famously complicated — he did his best to convince us that we could reinvent this definition ourselves if we tried! Then he went ahead and discussed various proofs that every weak 3-category is equivalent to a semistrict one.

An interesting theme of all three talks was the idea of treating the "strictification" functor implicit in Mac Lane's coherence theorem:

$$\mathsf{st}\colon \mathsf{Weak2Cat} \to \mathsf{Strict2Cat}$$

as the left adjoint of the inclusion

$$i\colon \mathsf{Strict2Cat} \to \mathsf{Weak2Cat}$$

where now we think of both Strict2Cat and Weak2Cat as mere 1-categories. You can read more about this idea here:

3) Miles Gould, "Coherence for categorified operadic theories", available as `math.CT/0607423`.

On Tuesday night, Mike Shulman gave an introduction to model categories, which are a tool developed by Quillen in the late 1960s to unify homotopy theory and homological algebra. If you want to understand the basics of model categories, you should probably start by listening to his talk, and then read this:

4) W. G. Dwyer and J. Spalinski, "Homotopy theories and model categories", available at `http://hopf.math.purdue.edu/Dwyer-Spalinski/theories.pdf`

For more references, try "Week 170".

Here's the rough idea:

In homotopy theory we study topological spaces; in homological algebra we study chain complexes. But, in both cases we study them in a funny way. There's a category of topological spaces and continuous maps, and there's a category of chain complexes and chain maps, but these categories are not everything that counts. Normally, we say two objects in a category are "the same" if they're isomorphic. But in this case we often use a weaker concept of equivalence!

In homotopy theory, we say a map between spaces

$$f\colon X \to Y$$

is a "weak homotopy equivalence" if it induces isomorphisms on homotopy groups:

$$\pi_n(f)\colon \pi_n(X) \to \pi_n(Y)$$

In homological algebra, we say a map between chain complexes

$$f\colon X \to Y$$

is a "quasi-isomorphism" if it induces isomorphisms on homology groups:

$$H_n(f)\colon H_n(X) \to H_n(Y)$$

Model category theory formalizes this by speaking of a category $\mathcal{C}$ equipped with a classes of morphisms called "weak equivalences". We can formally invert these and get a new category $\mathrm{Ho}(\mathcal{C})$ where the weak equivalences are isomorphisms: this is called the "homotopy category" or "derived category" of our model category. But this loses information, so it's often good *not* to do this.

In a model category, we also have a class of morphisms called "fibrations", which you should imagine as being like fiber bundles. Dually, we have a class of morphisms called "cofibrations", which you should imagine as well-behaved inclusions, like the inclusion of the closed unit interval in the real line — not the inclusion of the rationals into the real line.

Finally, the weak equivalences, fibrations and cofibrations satisfy some axioms that make them interlock in a powerful way. These axioms are a bit mind-numbing at first glance, so I won't list them. But, they encapsulate a lot of wisdom about homotopy theory and homological algebra!

On Wednesday the talks were about n-categories and homotopy theory. I kicked them off with a general introduction to the "Homotopy Hypothesis": Grothendieck's idea that homotopy theory was secretly about ∞-groupoids — that is, ∞-categories where all the j-morphisms have weak inverses.

5) John Baez, "The homotopy hypothesis", `http://math.ucr.edu/home/baez/homotopy/`

Part of the idea is that if you hand me a space X, I can cook up an ∞-groupoid which has:

- points of X as objects,
- paths in X as morphisms,
- homotopies between paths in X as 2-morphisms,
- homotopies between homotopies between paths in X as 3-morphisms,
- etc…

This is called the "fundamental ∞-groupoid of X".

But another part of the idea is that if you hand me a model category $\mathcal{C}$, I can cook up an ∞-category which has:

- nice objects of $\mathcal{C}$ as objects,
- morphisms in $\mathcal{C}$ as morphisms,

- homotopies between morphisms in $\mathcal{C}$ as 2-morphisms,
- homotopies between homotopies between morphisms in $\mathcal{C}$ as 3-morphisms,
- etc. . . .

The basic idea here is simple: we're studying homotopies between homotopies between. . . and so on.

(But, there's a little technicality — this "nice object" business. An object of $\mathcal{C}$ is "fibrant" if its unique morphism from the initial object is a fibration, and "cofibrant" if its unique morphism to terminal object is a cofibration. Objects with both properties are what I'm calling "nice". For example, in the category of topological spaces, the "cell complexes" (made by gluing balls together) are nice. In the category of chain complexes, the "projective" chain complexes are nice. Only for these nice objects do homotopies work as well as you'd hope. Luckily, every object in $\mathcal{C}$ is weakly equivalent to one of these nice ones.)

The interesting thing about the above ∞-category is that it's an "$(\infty, 1)$-category", meaning that all its j-morphisms are weakly invertible for $j > 1$. For example, maps between spaces aren't necessarily invertible, even up to homotopy — but homotopies are always invertible.

We can define "(∞, k)-categories" for any k in the same way, and we see that $(\infty, 0)$-categories are just ∞-groupoids. So, the Homotopy Hypothesis reveals the beginning of what might be a very nice pattern. Roughly:

- Topological spaces, as studied in homotopy theory, are secretly $(\infty, 0)$-categories.
- Model categories, as studied in homotopy theory, are secretly $(\infty, 1)$-categories.
- ????, as studied in homotopy theory (not yet?), are secretly $(\infty, 2)$-categories.
- Etcetera. . . .

Presumably the ???? should be filled in with something like "model 2-categories", with the primordial example being the 2-category of model categories, just as the primordial example of a model category is the category of spaces.

But, there's only been a little study of this sort of "meta-homotopy theory" so far. For example:

6) Julie Bergner, "Three models for the homotopy theory of homotopy theories", available as `math.AT/0504334`.

After my talk, Simona Paoli spoke about her work on turning the homotopy hypothesis from a dream into a reality:

7) Simona Paoli, "Semistrict models of connected 3-types and Tamsamani's weak 3-groupoids", available as `math.AT/0607330`.

8) Simona Paoli, "Semistrict Tamsamani n-groupoids and connected n-types", available as `math.AT/0701655`.

Eugenia Cheng then spent the afternoon leading us through another approach:

9) Clemens Berger, "A cellular nerve for higher categories", available at http://citeseer.ist.psu.edu/383423.html

10) Denis-Charles Cisinski, "Batanin higher groupoids and homotopy types", available as math.AT/0604442.

I would love to explain this stuff, mainly as an excuse for learning it better! But alas, I'm getting a bit tired and we're only on the second day of the workshop... I must hurry on.

On Wednesday evening, Peter May spoke about some applications of weak 2-categories that appear in his new book:

11) Peter May and J. Sigurdsson, *Parametrized Homotopy Theory*, American Mathematical Society, 2006.

The rough idea is that we have a weak 2-category with:

- spaces as objects,
- spectra over $X \times Y$ as morphisms from X to Y,
- maps between spectra over $X \times Y$ as 2-morphisms.

Lots of ideas from "parametrized" stable homotopy theory are neatly encoded as results about this 2-category.

Thursday was all about $(\infty, 1)$-categories. The first talk was by Mike Shulman, who gave a nice intuitive treatment of Andr Joyal's approach to $(\infty, 1)$-categories.

In 1957, Daniel Kan figured out a nice way to describe ∞-groupoids as simplicial sets with a certain property: now they're called "Kan complexes". They're very popular among homotopy theorists. You can read about them here:

12) Paul G. Goerss and J. F. Jardine, *Simplicial Homotopy Theory*, Birkhuser, Basel, 1999.

Given this, it's not so surprising that we can describe $(\infty, 1)$-categories as simplicial sets with some more general property. In fact this was done by Boardmann and Vogt back in 1973. In the last decade or so, Joyal has developed an enormous body of results about these $(\infty, 1)$-categories, which he calls "quasicategories". He is writing a book on the subject, which is not quite done yet — but it's already enormously influenced the state of higher category theory, and I expect it will continue to do so.

Next Julie Bergner compared different approaches to $(\infty, 1)$-categories. I mentioned a while back that she's one of the few people who has worked hard on "meta-homotopy theory". That was very much in evidence in her talk.

She began by describing a bunch of different definitions of $(\infty, 1)$-category. But then she showed these definitions weren't really so different! For each definition, she constructed a model category of all $(\infty, 1)$-categories of that type. And then, she sketched the proof that all these model categories were "Quillen equivalent".

For details, listen to her talk or try this paper:

13) Julie Bergner, A survey of $(\infty, 1)$-categories, available as math.AT/0610239.

In the afternoon, Andr Joyal spoke about quasicategories. I urge you to listen to his talk and also the minicourse he later gave on this subject:

14) Andr Joyal, "Graduate course on basic aspects of quasicategories", http://www.fields.utoronto.ca/audio/#crs-quasibasic

I can't possibly summarize this stuff! It basically amounts to taking the whole of category theory and extending it to quasicategories.

(Well, I guess I just summarized it, but. . . .)

After Joyal's talk, Joshua Nichols-Barrer spoke about using quasicategories as an approach to understanding "stacks", which are like sheaves, only categorified.

In the evening, Kathryn Hess spoke about some work she's doing with Steve Lack, on parallel transport in bundles of bicategories. Sounds like physics, but they came to the subject from a completely different motivation!

Finally, Dorette Pronk spoke about weak 2-categories and weak 3-categories of fractions. The notion of a "calculus of fractions" goes back at least to the work of Gabriel and Zisman in 1967:

15) P. Gabriel and M. Zisman, *Categories of Fractions and Homotopy Theory*, Springer-Verlag, Berlin, 1967.

Say you have a category and you want to throw in formal inverses to some morphisms. Well, you can do it! But in general, the morphisms in the resulting category will be arbitrarily long "zig-zag" diagrams in your original category, like this:

$$X_1 \to X_2 \leftarrow X_3 \to X_4 \leftarrow X_5 \to X_6$$

The arrows pointing backwards are the ones you threw in formal inverses for.

This is a nuisance! But luckily, in nice cases, you only need to use zig-zags of length two. This is what a "calculus of fractions" achieves. A classic example is when you start with a model category $\mathcal{C}$, and you throw in formal inverses for the weak equivalences to get the "homotopy category" $\mathrm{Ho}(\mathcal{C})$.

Dorette Pronk has been looking at how all this generalizes when you have a weak 2-category or weak 3-category and you throw in *weak* inverses to some morphisms. This has some interesting applications to stacks:

16) Dorette A. Pronk, "Etendues and stacks as bicategories of fractions", *Compositio Mathematica* **102** (1996), 243–303. Also available at http://www.numdam.org/numdam-bin/recherche?h=nc&id=CM_1996__102_3_243_0

Dorette's talk ended at 9pm, and everyone went home and collapsed after a hard day's work. Actually not: a bunch of us went out and partied! One of the great things about working on n-categories is the sense of camaraderie among the small crowd that does this.

Friday's talks were about higher gauge theory. Since I've discussed this many times here, I'll be terse. Alissa Crans explained Lie 2-groups and Lie 2-algebras, and then Danny Stevenson explained his work on connections, 2-connections and Schreier theory (see "Week 223"). In the afternoon, Urs Schreiber described his ideas on higher-dimensional parallel transport and local trivializations, with a little help from Toby Bartels.

Friday evening, we heard talks from Simon Willerton (on Hopf monads) and Igor Bakovic (on 2-bundles). Quite an evening! Bakovic is an impressive young Croatian fellow who seems to have taught himself n-categories. We were all horrified when it became clear he had over 30 pages of transparencies, but his talk was actually quite nice.

And if you like higher-dimensional diagrams anywhere near as much as I do, you've got to take a look at Willerton's slides:

17) Simon Willerton, "The diagrammatics of Hopf monads", http://math.ucr.edu/home/baez/fields/willerton/

Again the talks ended at 9pm.

Finally, on Saturday morning, Aaron Lauda spoke about Frobenius algebras and their relation to Khovanov homology:

18) Aaron Lauda, "Frobenius algebras, quantum topology and higher categories", available at http://www.math.columbia.edu/~lauda/talks/Fields/

Urs Schreiber then wrapped things up with a talk about the quantization of strings from a higher category viewpoint. You can get a good feeling for this from his blog entries at the n-Category Caf, which are all listed on my webpage for this workshop.

Speaking of the n-Category Caf — after the workshop ended, Bruce Bartlett interviewed Urs and me about this blog, which we run together with David Corfield. You can see the interview here:

19) John Baez and Urs Schreiber, "Interview by Bruce Bartlett", http://math.ucr.edu/home/baez/interview2.html

---

Addendum: For more discussion, go to the n-Category Caf.

Week 246

February 25, 2007

I've been gearing up to tell a big, wonderful story about the quest to generalize quantum knot invariants to higher dimensions by categorifying the theory of quantum groups. This story began at least 14 years ago! I talked about it way back in "Week 2".

At the time, Louis Crane and Igor Frenkel had just come out with a draft of a paper called "Hopf categories and their representations", which began tackling this problem. This is roughly when Crane invented the word "categorification" — and their paper is a big part of why I got interested in n-categories.

The subject moved rather slowly until Frenkel's student Mikhail Khovanov got into the game and categorified the Jones polynomial — a famous invariant of knots related to the very simplest quantum group, the one called "quantum $\mathrm{SU}(2)$". Now categorifying knot theory is a hot topic.

James Dolan, Todd Trimble and I have been chewing away on this subject from a quite different angle, which may ultimately turn out to be the same — or at least related. In the process, we've needed to learn, reinvent or remodel a lot of classical work on group theory, incidence geometry, and combinatorics. It's been a great adventure, and it's far from over.

I'm dying to explain some of this stuff, and I'll start soon. But first I need to talk about something less pleasant: the troubles with fundamental physics.

If you care at all about physics, you've probably heard about these:

1) Peter Woit, *Not Even Wrong: The Failure of String Theory and the Continuing Challenge to Unify the Laws of Physics*, Basic Books, New York, 2006.

2) Lee Smolin, *The Trouble With Physics: The Rise of String Theory, the Fall of a Science, and What Comes Next*, Houghton Mifflin, New York, 2006.

I won't "review" these books. I'll just talk about some points they raise — in a very nontechnical way.

Their importance is that they explain the problems of string theory to the large audience of people who get their news about fundamental physics from magazines and popular books. Experts were already aware of these problems, but in the popular media there's always been a lot of hype, which painted a much rosier picture. So, casual observers must have gotten the impression that physics was always on the brink of a Theory of Everything... but mysteriously never reaching it. These books correct that impression.

In fact, string theory still hasn't reached the stage of making any firm predictions. For the last few decades, astrophysicists have been making amazing discoveries in fundamental physics: dark matter, dark energy, neutrino oscillations, maybe even cosmic inflation in the very early universe! Soon the Large Hadron Collider will smash particles against each other hard enough to see the Higgs boson — or not. With luck, it may even see brand new particles. But about all this, string theory has had little to say.

To get actual predictions, practical physicists sometimes build "string-inspired" scenarios. These scenarios aren't *derived* from string theory: to get specific predictions, they have to throw in lots of extra assumptions. For example, since string theory involves supersymmetry, string theorists resort to supersymmetric versions of the Standard Model to guess what the Large Hadron Collider might see. But the simplest supersymmetric version of the Standard Model involves over 100 undetermined parameters! Even the particles we actually see are put in by hand, not derived from string theory. If it turns out we see some other particles, we can just stick those in too.

Someday this situation may change, but it's dragged on for a while now. There's no reason why theoretical physics should always move fast. The universe has taken almost 14 billion years to reach its current state of self-knowledge — what's a few more decades? But, coming after an era of incredibly rapid progress stretching from 1905 to 1983, the current period of stagnation feels like an eternity. So, physicists are getting a bit desperate. This has led to some strange behavior.

For example, some people have tried to refute the claim that string theory makes no testable predictions by arguing that it predicts the existence of gravity! This is better known as a "retrodiction".

Others say that since string theory requires extra assumptions to make definite predictions about our universe, we should — instead of making some assumptions and using them to predict something — study the space of *all possible* extra assumptions. For example, there are lots of Calabi-Yau manifolds that could serve as the little curled-up dimensions of spacetime, and lots of ways we could stick D-branes here or there, etcetera.

This space of all possible extra assumptions is called the "Landscape". Since it's vaguely defined, the main things we know about it are:

a) it's big,

b) it keeps growing as string theorists come up with new ideas,

c) nobody has yet found a point in it that matches our universe.

Despite this, or perhaps because of it, the Landscape has been the subject of many discussions. Often these devolve into arguments about the "anthropic principle". Roughly, this says that if the universe were really different, we wouldn't be having this argument — so it must be like it is!

One can in fact draw some conclusions from the anthropic principle. But it's really just the low-budget limit of experimental physics. You can always get more conclusions from doing more experiments. The experiment where you just check to see if you're alive is really cheap — but you don't learn much from it.

(Of course I'm oversimplifying things for comic effect, but usually people take the opposite approach, overcomplicating this stuff to make it sound more profound than it is.)

Serious string theorists are mostly able to work around this tomfoolery, but it exerts a demoralizing effect. So, when Woit and Smolin came out with their books, a lot of tempers snapped, and a lot of strange arguments were applied against them.

For example, one popular argument was "Okay, buster — can you do better?" The idea here seems to be that until you know a solution to the problems faced by string theory, you shouldn't point out these problems — at least not publicly. This goes against my experience: hard problems tend to get solved only *after* lots of people openly admit they exist.

Another closely related argument was "String theory is the only game in town." Until some obviously better theory shows up, we should keep working on string theory.

It's true there's no obviously better theory than string theory. Loop quantum gravity, in particular, has problems that are just as serious as string theory.

But, the "only game in town" argument is still flawed.

Once I drove through Las Vegas, where there really *is* just one game in town: gambling. I stopped and took a look. I saw the big fancy casinos. I saw the glazed-eyed grannies feeding quarters into slot machines, hoping to strike it rich someday. It was clear: the odds were stacked against me. But, I didn't respond by saying "Oh well — it's the only game in town" and starting to play.

Instead, I *left* that town.

It's no good to work on string theory with a glum attitude like "it's the only game in town." There are lots of other wonderful things for theoretical physicists to do. Things where your work has a good chance of matching experiment... or things where you take a huge risk by going out on your own and trying something new.

Indeed, if following the crowd were the name of the game, string theory might never have been invented in the first place. It didn't fall from the sky fully formed, obviously better than its competitors. A handful of people took a big chance by working on it for many years before it proved its worth.

In his book, Lee Smolin argues that physics is in the midst of a scientific revolution, and that these times demand people who don't just follow fashion:

> *The point is that different kinds of people are important in normal and revolutionary science. In the normal periods, you need only people who, regardless of their degree of imagination (which may well be high), are really good at working with the technical tools — let us call them master craftspeople. During revolutionary periods, you need seers, who can peer ahead into the darkness.*

He later regretted this way of putting it, and I think rightly so. The term "seer" suggests that some people have a better-than-average ability to see the right answers to profound questions. This may be true, but it's hard to tell ahead of time who is a seer and who is not. Smolin later wrote:

> *Here is a metaphor due to Eric Weinstein that I would have put in the book had I heard it before. Let us take a different twist on the landscape of theories and consider the landscape of possible ideas about post standard model or quantum gravity physics that have been proposed. Height is proportional to the number of things the theory gets right. Since we dont have a convincing case for the right theory yet, that is a high peak somewhere off in the distance. The existing approaches are hills of various heights that may or may not be connected across some ridges and high valleys to the real peak. We assume the landscape is covered by fog so we cant see where the real peak is, we can only feel around and detect slopes and local maxima.*
>
> *Now to a rough approximation, there are two kinds of scientists — hill climbers and valley crossers. Hill climbers are great technically and will always advance*

> *an approach incrementally. They are what you want once an approach has been defined, i.e. a hill has been discovered, and they will always go uphill and find the nearest local maximum. Valley crossers are perhaps not so good at those skills, but they have great intuition, a lot of serendipity, the ability to find hidden assumptions and look at familiar topics new ways, and so are able to wander around in the valleys, or cross exposed ridges, to find new hills and mountains.*
>
> *I used craftspeople vs. seers for this distinction, Kuhn referred to normal science vs. revolutionary science, but the idea was the same.*
>
> *With the scene set, here is my critique. First, to progress, science needs a mix of hill climbers and valley crossers. The balance needed at any one time depends on the problem. The more foundational and risky a problem is the more the balance needs to be shifted towards valley crossers. If the landscape is too rugged, with too many local maxima, and there are too many hill climbers vs. valley crossers, you will end up with a lot of hill climbers camped out on the tops of hills, each group defending their hills, with not enough valley crossers to cross those perilous ridges and swampy valleys to find the real mountain.*
>
> *This is what I believe is the situation we are in. And — and this is the point of Part IV [of the book] — we are in it, because science has become professionalized in a way that takes the characteristics of a good hill climber as representative of what is a good, or promising scientist. The valley crossers we need have been excluded, or pushed to the margins where they are not supported or paid much attention to.*
>
> *My claim is then 1) we need to shift the balance to include more valley crossers, and 2) this is easy to do, if we want to do it, because there also are criteria that can allow us to pick out who is worthy of support. They are just different criteria.*

This is a good analysis, but it leaves out one thing: most "valley crossers" get stuck wandering around in valleys. Even those who succeed once are likely to fail later: think of Einstein's long search for a unified field theory, or Schrdinger's "unitary field theory" involving a connection with torsion, or Heisenberg's nonlinear spinor field theory, or Kelvin's vortex atoms. It's not surprising these geniuses spent a lot of time on failed theories — what's surprising is their successes.

So, failure is an unavoidable cost of doing business, and encouraging more "valley crossers" or "risk takers" will inevitably look like encouraging more failures.

Unfortunately, the alternative is even more risky. If everyone pursues the same approach, we'll all succeed or fail together — and chances are we'll fail. The reason for backing some risk takers is that it "diversifies our portfolio". It reduces overall risk by increasing the chance that *someone* will succeed.

(It's no coincidence that Eric Weinstein, mentioned above by Smolin, works as an investment banker. He's also a student of Raoul Bott — but that's another story!)

Near the end of his book, Woit quotes the mathematican Michael Atiyah, who also seems to raise the possibility that we need some more risk-taking:

> *If we end up with a coherent and consistent unified theory of the universe, involving extremely complicated mathematics, do we believe that this represents*

> *"reality"? Do we believe that the laws of nature are laid down using the elaborate algebraic machinery that is now emerging in string theory? Or is it possible that nature's laws are much deeper, simple yet subtle, and that the mathematical description we use is simply the best we can do with the tools we have? In other words, perhaps we have not yet found the right language or framework to see the ultimate simplicity of nature.*

Most people who read these words and try to find this "right framework" will fail. But, we can hope that someday a few succeed.

For the fascinating tale of Schrdinger's "unitary field theory", see this nice book:

3) Walter Moore, *Schrdinger: His Life and Thought*, Cambridge U. Press, Cambridge, 1989.

For more about the search for unified field theories in early 20th century, see:

4) Hubert F. M Goenner, "On the history of unified field theories", *Living Reviews of Relativity* **7**, (2004), 2. Available at `http://www.livingreviews.org/lrr-2004-2`

Addenda: I thank Eugenia Cheng and Eugene Lerman for catching mistakes. For more discussion, go to the n-Category Caf.

> Apart from agreeing with reality it is certainly a magnificent achievement of pure thought.
>
> — *Einstein, to Hermann Weyl, about Weyl's attempt to unify gravity and electromagnetism.*

Week 247

March 23, 2007

Symmetry has fascinated us throughout the ages. Greek settlers in Sicily may have seen irregular 12-sided crystals of pyrite in Sicily and dreamt up the regular dodecahedron simply because it was more beautiful, more symmetrical.

The Alhambra, a Moorish palace in Granada built around 1300, has tile patterns with at least 13 of the 17 possible symmetry groups:

1) Branko Grnbaum, "What symmetry groups are present in the Alhambra?", *Notices of the AMS* **53** (2006), 670–673. Also available at `http://www.ams.org/notices/200606/comm-grunbaum.pdf`

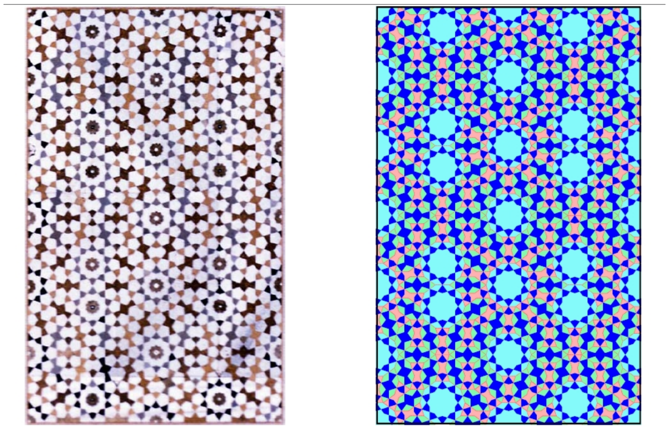
You can see some of these patterns here:

2) "Moresque tiles", `http://www.spsu.edu/math/tile/grammar/moor.htm`

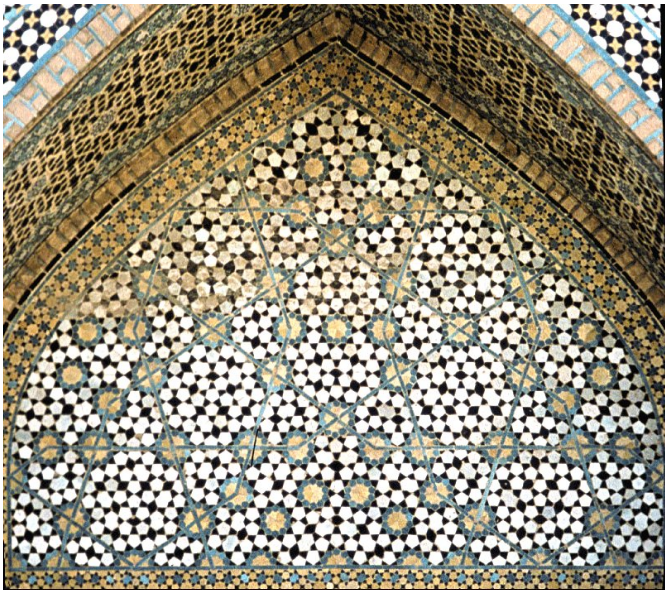
Recently, Peter Lu and Paul Steinhardt discovered that Islamic tile designs also include "quasicrystals". A perfectly repetitive tiling can't have 5-fold symmetry. Nor can a 3-dimensional crystal: that's why the dodecahedra formed by pyrite aren't regular. But by using patterns that never quite repeat, the Islamic artists achieved *approximate* 5-fold symmetry:

3) Peter J. Lu and Paul J. Steinhardt, "Decagonal and quasi-crystalline tilings in medieval Islamic architecture", *Science* **315** (2007), 1106–1110.

Here's an example from the I'timad al-Daula mausoleum in the Indian city of Agra, built by Islamic conquerors in 1622 - together with a more mathematical version constructed by Lu and Steinhardt:



Here's another, from the Darb-i Imam shrine in Isfahan, Iran, also built in the 1600s:



This came as a big surprise, since everyone had *thought* that the math behind quasicrystals was first discovered by Penrose around 1974, then seen in nature by Shechtman, Blech, Gratias and Cahn in 1983. It goes to show that the appeal of symmetry, even in its subtler forms, is very old! It also goes to show that you can make big discoveries just by looking carefully at what's in front of you.

For more on quasicrystals, try this:

4) Steven Webber, "Quasicrystals", `http://www.jcrystal.com/steffenweber/`

Of course, the appeal of symmetry didn't end with ancient Greeks or medieval Islamic monarchs. It also seems to have gotten ahold of John Fry, chief executive of Fry's

Electronics — a chain of retail shops whose motto is "Your best buys are always at Fry's". In 1994 he set up something called the American Institute of Mathematics. The headquarters was in a Fry's store in Palo Alto — not very romantic. But last year, this institute announced plans to move to a full-scale replica of the Alhambra!

5) Associated Press, "Silicon valley will get Alhambra-like castle", August 18, 2006. Available at http://www.jcrystal.com/steffenweber/

And this week, the institute flexed its mighty PR muscles and coaxed reporters from the New York Times, BBC, Le Monde, Scientific American, Science News, and so on to write about a highly esoteric advance in our understanding of symmetry — a gargantuan calculation involving the Lie group E_8:

6) American Institute of Mathematics, Mathematicians map E_8, `http://aimath.org/E8`

The calculation is indeed huge. The *answer* takes up 60 gigabytes of data: the equivalent of 45 days of music in MP3 format. If this information were written out on paper, it would cover Manhattan!

But what's the calculation *about?* It almost seems a good explanation of that would *also* cover Manhattan. I took a stab at it here:

7) John Baez, "News about E_8", `http://golem.ph.utexas.edu/category/2007/03/news_about_e8.html`

but I only got as far as sketching a description of E_8 and some gadgets called R-polynomials. Then come Kazhdan-Lusztig polynomials, and Kazhdan-Lusztig-Vogan polynomials. . . . For more details, follow the links, especially to the page written by Jeffrey Adams, who led the project.

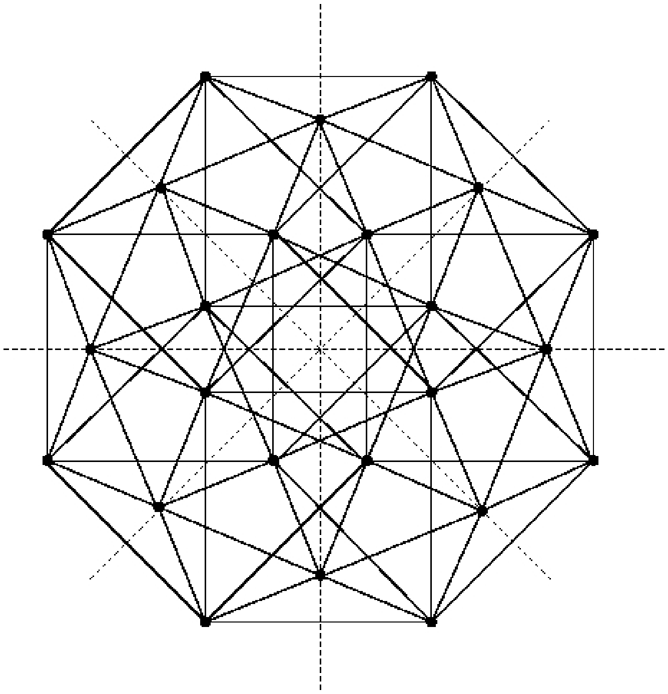
In weeks to come, I'll say more about some topics tangentially related to this calculation — especially flag varieties, representation theory and the Weil conjectures. I may even talk about Kazhdan-Lusztig polynomials!

For starters, though, let's just look at some pretty pictures by John Stembridge that hint at the majesty of E_8. Then I'll sketch the real subject of Weeks to come: symmetry, geometry, and "groupoidification".

To warm up to E_8, let's first take a look at D_4, D_5, E_6, and E_7.

In "Week 91" I spoke about the D_4 lattice. To get this, first take a bunch of equal-sized spheres in 4 dimensions. Stack them in a hypercubical pattern, so their centers lie at the points with integer coordinates. A bit surprisingly, there's a lot of room left over — enough to fit in another copy of this whole pattern: a bunch of spheres whose centers lie at the points with *half-integer* coordinates!

If you stick in these extra spheres, you get the densest known packing of spheres in 4 dimensions. Their centers form the "D_4 lattice". It's an easy exercise to check that each sphere touches 24 others. The centers of these 24 are the vertices of a marvelous shape called the "24-cell" — one of the six 4-dimensional Platonic solids. It looks like this:



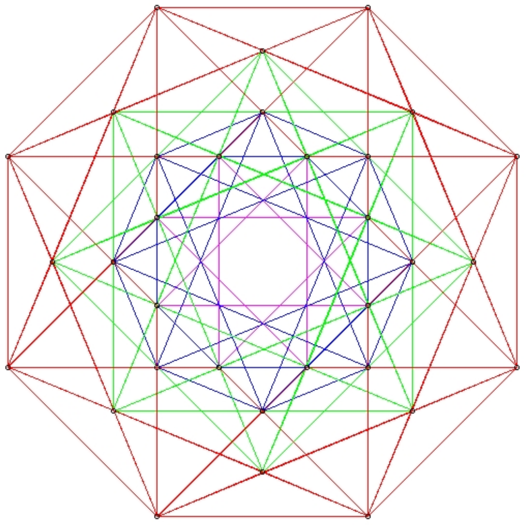
8) John Baez, picture of 24-cell, in "a review of On Quaternions and Octonions: Their Geometry, Arithmetic and Symmetry, by John H. Conway and Derek A. Smith', available at http://math.ucr.edu/home/baez/octonions/conway_smith/

Here I'm using a severe form of perspective to project 4 dimensions down to 2. The coordinate axes are drawn as dashed lines; the solid lines are the edges of the 24-cell.

How about in 5 dimensions? Here the densest known packing of spheres uses the "D_5 lattice". This is a lot like the D_4 lattice... but only if you think about it the right way.

Imagine a 4-dimensional checkerboard with "squares" — really hypercubes! — alternately colored red and black. Put a dot in the middle of each black square. Voila! You get a rescaled version of the D_4 lattice. It's not instantly obvious that this matches my previous description, but it's true.

If you do the same thing with a 5-dimensional checkerboard, you get the "D_5 lattice", by definition. This gives the densest known packing of spheres in 5 dimensions. In this packing, each sphere has 40 nearest neighbors. The centers of these nearest neighbors are the vertices of a solid that looks like this:



9) John Stembridge, "D_5 root system", available at `http://www.math.lsa.umich.edu/~jrs/data/coxplanes/`

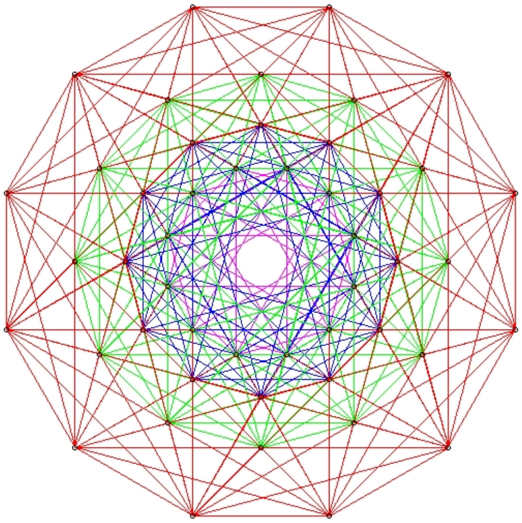
If you do the same thing with a 6-dimensional checkerboard, you get the "D_6 lattice"… and so on.

However, in 8 dimensions something cool happens. If you pack spheres in the pattern of the D_8 lattice, there's enough room left to stick in an extra copy of this whole pattern! The result is called the "E_8 lattice". It's twice as dense as the D_8 lattice.

If you then take a well-chosen 7-dimensional slice through the origin of the E_8 lattice, you get the E_7 lattice. And if you take a well-chosen 6-dimensional slice of this, you get the E_6 lattice. For precise details on what I mean by "well-chosen", see "Week 65".

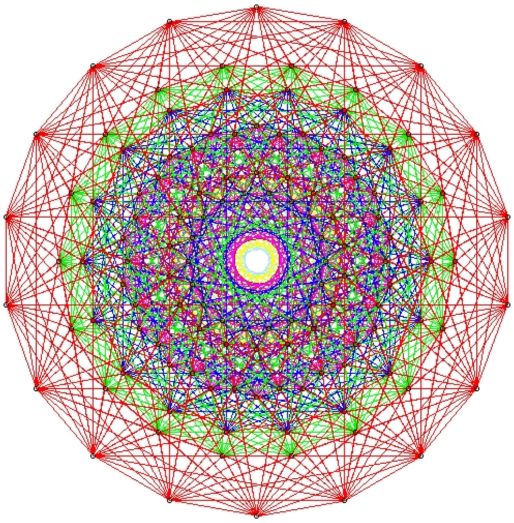
E_6 and E_7 give denser packings of spheres than D_6 and D_7. In fact, they give the densest known packings of spheres in 6 and 7 dimensions!

In the E_6 lattice, each sphere has 72 nearest neighbors. They form the vertices of a solid that looks like this:



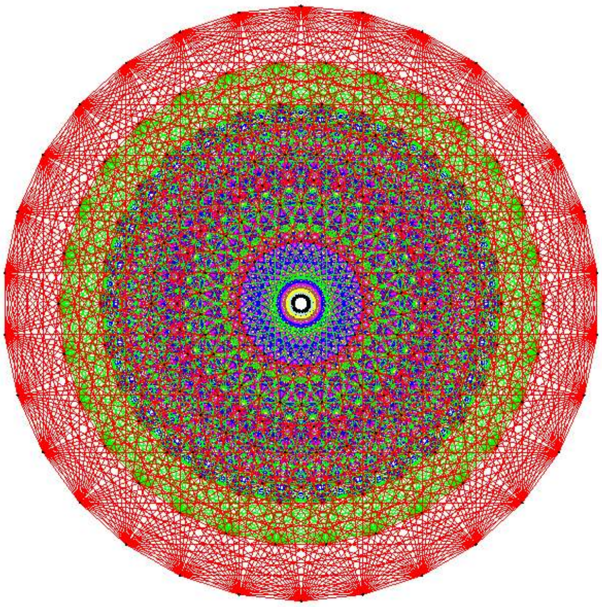
10) John Stembridge, "E_6 root system", available at `http://www.math.lsa.umich.edu/~jrs/data/coxplanes/`

In the E_7 lattice, each sphere has 126 nearest neighbors. They form the vertices of a solid like this:



11) John Stembridge, "E_7 root system", available at `http://www.math.lsa.umich.edu/~jrs/data/coxplanes/`

In the E_8 lattice, each sphere has 240 nearest neighbors. They form the vertices of a solid like this:



12) John Stembridge, "E_8 root system", available at `http://www.math.lsa.umich.edu/~jrs/data/coxplanes/`

Faithful readers will know I've discussed these lattices often before. For how they give rise to Lie groups, see "Week 63". For more about "ADE classifications", see "Week 64" and "Week 230". I haven't really added much this time, except Stembridge's nice pictures. I'm really just trying to get you in the mood for a big adventure involving all these ideas: the Tale of Groupoidification!

If we let this story lead us where it wants to go, we'll meet all sorts of famous and fascinating creatures, such as:

- Coxeter groups, buildings, and the quantization of logic
- Hecke algebras and Hecke operators
- categorified quantum groups and Khovanov homology
- Kleinian singularities and the McKay correspondence
- quiver representations and Hall algebras
- intersection cohomology, perverse sheaves and Kazhdan-Lusztig theory

However, the charm of the tale is how many of these ideas are unified and made simpler thanks to a big, simple idea: groupoidification.

So, what's groupoidification? It's a method of exposing the combinatorial underpinnings of linear algebra — the hard bones of set theory underlying the flexibility of the continuum.

Linear algebra is all about vector spaces and linear maps. One of the lessons that gets drummed into you when you study this subject is that it's good to avoid picking bases for your vector spaces until you need them. It's good to keep the freedom to do coordinate transformations. . . and not just keep it in reserve, but keep it *manifest!*

As Hermann Weyl wrote, "The introduction of a coordinate system to geometry is an act of violence".

This is a deep truth, which hits many physicists when they study special and general relativity. However, as Niels Bohr quipped, a deep truth is one whose opposite is also a deep truth. There are some situations where a vector space comes equipped with a god-given basis. Then it's foolish not to pay attention to this fact!

The most obvious example is when our vector space has been *defined* to consist of formal linear combinations of the elements of some set. Then this set is our basis.

This often happens when we use linear algebra to study combinatorics.

But if sets give vector spaces, what gives linear operators? Your first guess might be *functions*. And indeed, functions between sets do give linear operators between their vector spaces. For example, suppose we have a function

$$f\colon \{\mathrm{livecat}, \mathrm{deadcat}\} \to \{\mathrm{livecat}, \mathrm{deadcat}\}$$

which "makes sure the cat is dead":

$$f(\mathrm{livecat}) = \mathrm{deadcat}$$
$$f(\mathrm{deadcat}) = \mathrm{deadcat}$$

Then, we can extend f to a linear operator defined on formal linear combinations of cats:

$$F(a\mathrm{livecat} + b\mathrm{deadcat}) = a\mathrm{deadcat} + b\mathrm{deadcat}$$

Written as a matrix in the $\{\mathrm{livecat}, \mathrm{deadcat}\}$ basis, this looks like

$$\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}$$

(The relation to quantum mechanics here is just a vague hint of themes to come. I've deliberately picked an example where the linear operator is *not* unitary.)

So, we get some linear operators from functions. . . but not all! We only get operators whose matrices have exactly a single 1 in each column, the rest of the entries being 0. That's because a function $f\colon X \to Y$ sends each element of X to a single element of Y.

This is very limiting. We can do better if we get operators from *relations* between sets. In a relation between sets X and Y, an element of X can be related to any number of elements of Y, and vice versa. For example, let the relation

$$R\colon \{1,2,3,4\} \nrightarrow \{1,2,3,4\}$$

be "is a divisor of". Then 1 is a divisor of everybody, 2 is a divisor of itself and 4, 3 is only a divisor of itself, and 4 is only a divisor of itself. We can encode this in a matrix:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}$$

where 1 means “is a divisor of” and 0 means “is not a divisor of”.

We can get any matrix of 0’s and 1’s this way. Relations are really just matrices of truth values. We’re thinking of them as matrices of numbers. Unfortunately we’re still far from getting *all* matrices of numbers!

We can do better if we get matrices from *spans* of sets. A span of sets, written $S\colon X \nrightarrow Y$, is just a set S equipped with functions to X and Y. We can draw it like this:

$$\begin{tikzcd}
& S \arrow[dl, "G"'] \arrow[dr, "F"] & \\
X & & Y
\end{tikzcd}$$

It’s supposed to look like a bridge — hence the term “span”.

Spans of sets are like relations, but where you can be related to someone more than once!

For example, X could be the set of Frenchman and Y could be the set of Englishwomen. S could be the set of Russians. As you know, every Russian has exactly one favorite Frenchman and one favorite Englishwoman. So, F could be the function “your favorite Frenchman”, and G could be “your favorite Englishwoman”.

Then, given a Frenchman x and an Englishwoman y, they’re related by the Russian s whenever s has x as their favorite Frenchman and y as their favorite Englishwoman:

$$F(s) = xandG(s) = y.$$

Some pairs (x, y) will be related by no Russians, others will be related by one, and others will be related by more than one! I bet the pair

$$(x, y) = (\text{Grard Depardieu}, \text{Emma Thompson})$$

is related by at least 57 Russians.

This idea lets us turn spans of sets into matrices of natural numbers. Given a span of finite sets:

$$\begin{tikzcd}
& S \arrow[dl, "G"'] \arrow[dr, "F"] & \\
X & & Y
\end{tikzcd}$$

we get an $X \times Ymatrix$ whose (x, y) entry is the number of Russians — I mean elements s of S — such that

$$F(s) = xandG(s) = y.$$

We can get any finite-sized matrix of natural numbers this way.

Even better, there’s a way to “compose” spans that nicely matches the usual way of multiplying matrices. You can figure this out yourself if you solve this puzzle:

> *Let X be the set of people on Earth. Let T be the $X \times X$ matrix corresponding to the relation “is the father of”. Why does the matrix T^2 correspond to the relation “is the paternal grandfather of”? Let S correspond to the relation “is a friend of”. Why doesn’t the matrix S^2 correspond to the relation “is a friend of a friend of”? What span does this matrix correspond to?*

To go further, we need to consider spans, not of sets, but of groupoids!

I'll say more about this later — I suspect you're getting tired. But for now, briefly: a groupoid is a category with inverses. Any group gives an example, but groupoids are more general — they're the modern way of thinking about symmetry.

There's a way to define the cardinality of a finite groupoid:

12) John Baez and James Dolan, "From finite sets to Feynman diagrams", in *Mathematics Unlimited — 2001 and Beyond*, vol. 1, eds. Bjorn Engquist and Wilfried Schmid, Springer, Berlin, 2001, pp. 29–50. Also available as math.QA/0004133.

And, this can equal any nonnegative *rational* number! This lets us generalize what we've done from finite sets to finite groupoids, and get rational numbers into the game.

A span of groupoids is a diagram

$$\begin{tikzcd} & S \arrow[dl, "G"'] \arrow[dr, "F"] & \\ X & & Y \end{tikzcd}$$

where X, Y, S are groupoids and F, G are functors. If all the groupoids are finite, we can turn this span into a finite-sized matrix of nonnegative rational numbers, by copying what we did for spans of finite sets.

There's also a way of composing spans of groupoids, which corresponds to multiplying matrices. However, there's a trick involved in getting this to work — I'll have to explain this later. For details, try:

13) Jeffrey Morton, "Categorified algebra and quantum mechanics", *Theory and Application of Categories* **16** (2006), 785–854. Available at `http://www.emis.de/journals/TAC/volumes/16/29/16-29abs.html`; also available as math.QA/0601458.

14) Simon Byrne, *On Groupoids and Stuff*, honors thesis, Macquarie University, 2005, available at `http://www.maths.mq.edu.au/~street/ByrneHons.pdf` and `http://math.ucr.edu/home/baez/qg-spring2004/ByrneHons.pdf`

Anyway: the idea of "groupoidification" is that in many cases where mathematicians think they're playing around with linear operators between vector spaces, they're *actually* playing around with spans of groupoids!

This is especially true in math related to simple Lie groups, their Lie algebras, quantum groups and the like. While people usually study these gadgets using linear algebra, there's a lot of combinatorics involved — and where combinatorics and symmetry show up, one invariably finds groupoids.

As the name suggests, groupoidification is akin to categorification. But, it's a bit different. In categorification, we try to boost up mathematical ideas this way:

sets $\to$ *categories*
functions $\to$ *functors*

In groupoidification, we try this:

vector spaces → groupoids
linear operators → spans of groupoids

Actually, it's "decategorification" and "degroupoidification" that are systematic processes. These processes lose information, so there's no systematic way to reverse them. But, as I explained in "Week 99", it's still fun to try! If we succeed, we discover an extra layer of structure beneath the math we thought we understood. . . and this usually makes that math *clearer* and *less technical*, because we're not seeing it through a blurry, information-losing lens.

Okay, that's enough for now. On a completely different note, here's a book on "structural realism" and quantum mechanics:

15) Dean Rickles, Steven French, and Juha Saatsi, *The Structural Foundations of Quantum Gravity*, Oxford University Press, Oxford, 2006. Containing:

    - Dean Rickles and Steven French, "Quantum gravity meets structuralism: interweaving relations in the foundations of physics". Also available at `http://fds.oup.com/www.oup.co.uk/pdf/0-19-926969-6.pdf`
    - Tian Yu Cao, "Structural realism and quantum gravity".
    - John Stachel, "Structure, individuality, and quantum gravity". Also available as gr-qc/0507078.
    - Oliver Pooley, "Points, particles, and structural realism". Also available at `http://philsci-archive.pitt.edu/archive/00002939/`
    - Mauro Dorato and Massimo Pauri, "Holism and structuralism in classical and quantum general relativity". Also available at `http://philsci-archive.pitt.edu/archive/00001606/`
    - Dean Rickles, "Time and structure in canonical gravity". Also available at `http://philsci-archive.pitt.edu/archive/00001845/`
    - Lee Smolin, "The case for background independence". Also available as hep-th/0507235.
    - John Baez, "Quantum quandaries: A category-theoretic perspective". Also available at `http://math.ucr.edu/home/baez/quantum/` and as quant-ph/0404040.

Very loosely speaking — I ain't no philosopher — structural realism is the idea that what's "real" about mathematics, or the abstractions in physical theories, are not individual entities but the structures, or patterns, they form. So, instead of asking tired questions like "What is the number 2, really?" or "Do points of spacetime really exist?", we should ask more global questions about the roles that structures like "natural numbers" or "spacetime" play in math and physics. It's a bit like how in category theory, we can only understand an object in the context of the category it inhabits.

Finally, here's a puzzle for lattice and Lie group fans. The dots in Stembridge's pictures are the shortest nonzero vectors in the D_5, E_6, E_7, and E_8 lattices — or in technical terms, the "roots". Of course, only for ADE Dynkin diagrams are the roots all of equal length — but those are the kind we have here. Anyway: in the D_5 case, only 32 of the 40 roots are visible. The other 8 are hidden in back somewhere. Where are they?

I asked John Stembridge about this and he gave a useful clue. His planar pictures show projections of the roots into what he calls the "Coxeter plane".

Recall from "Week 62" that the "Coxeter group" associated to a Dynkin diagram acts as rotation/reflection symmetries of the roots; it's generated by reflections through the roots. There's a basis of roots called "simple roots", one for each dot in our Dynkin diagram, and the product of reflections through all these simple roots is called the "Coxeter element" of our Coxeter group — it's well-defined up to conjugation. The "Coxeter plane" is the canonical plane on which the Coxeter element acts as a rotation.

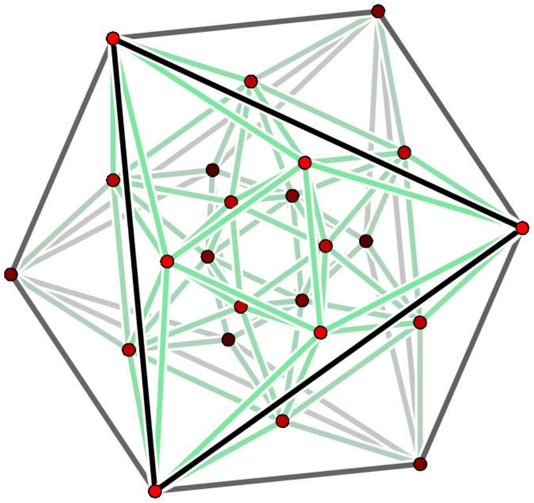
A rotation by how much? The order of the Coxeter element is called the "Coxeter number" and denoted h, so the Coxeter element acts on the Coxeter plane as a rotation of $2\pi/h$. The Coxeter number is important for other reasons, too! Here's how it goes:

Coxeter group	Coxeter number
A_n	$n+1$
B_n	$2n$
C_n	$2n$
D_n	$2n-2$
E_6	12
E_7	18
E_8	30
F_4	12
G_2	6

For D_5 the Coxeter number is 8, which accounts for the 8-fold symmetry of Stembridge's picture in that case. The E_8 picture has 30-fold symmetry! My D_4 picture has 8-fold symmetry, so I must not have been projecting down to the Coxeter plane.

Anyway, this stuff should help answer my puzzle. I don't know the answer, though.

Addenda: I thank David Corfield and James Dolan for catching mistakes. Tony Smith found a nice picture created by Gnter Ziegler of the D_4 root system (that is, the 24-cell) viewed from the Coxeter plane. The D_4 root system is 4-dimensional, but it's been drawn with a bit of 3d perspective. The 6-fold symmetry is evident:



16) Gnter M. Ziegler, "picture of 24-cell", `http://www.math.tu-berlin.de/~ziegler/24-cell.jpeg`

For more discussion, go to the n-Category Caf.

The true spirit of delight, the exaltation, the sense of being more than Man, which is the touchstone of the highest excellence, is to be found in mathematics as surely as poetry.

— *Bertrand Russell*

Week 248

March 28, 2007

This week I'll continue the Tale of Groupoidification. But first: some new views of the Sun!

Here's a cool movie of the Moon passing in front of the Sun, as viewed from the "STEREO B" spacecraft. Click on it:

2) Astronomy Picture of the Day, March 3 2007, "Lunar transit from STEREO", `http://antwrp.gsfc.nasa.gov/apod/ap070303.html`

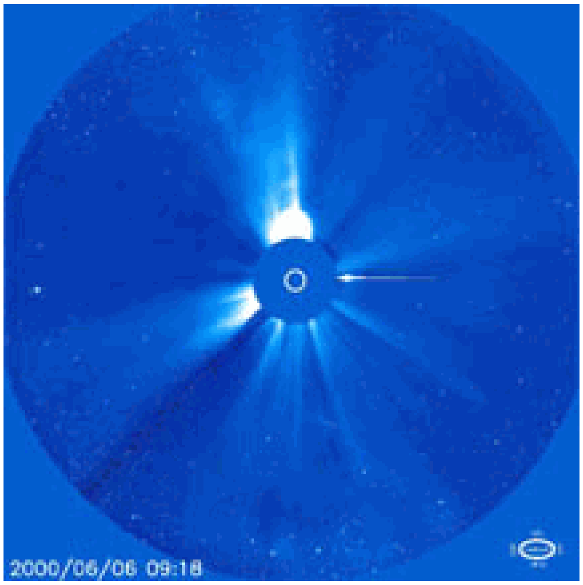
As the name hints, there's a pair of STEREO satellites in orbit around the Sun. One is leading the Earth a little, the other lagging behind a bit, to provide a stereoscopic view of coronal mass ejections.

What's a "coronal mass ejection"? It's an event where the Sun shoots off a blob of ionized gas — billions of tons of it — at speeds around 1000 kilometers per second!

That sounds cataclysmic... but it happens between once a day and 5-6 times a day, depending on where we are in the 11-year solar cycle, also known as the "sunspot cycle". Right now we're near the minimum of this cycle. Near the maximum, coronal mass ejections can really screw up communication systems here on Earth. For example, in 1998 a big one seems to have knocked out a communication satellite called Galaxy 4, causing 45 million people in the US to lose their telephone pager service:

3) Gordon Holman and Sarah Benedict, "Solar Flare Theory: Coronal mass ejections, solar flares, and the Earth-Sun connection", `http://www.agu.org/sci_soc/articles/eisbaker.html`

So, it's not only fun but also practical to understand coronal mass ejections. Here's a movie of one taken by the Solar and Heliospheric observatory (SOHO):



4) NASA, "Cannibal coronal mass ejections", `http://science.nasa.gov/headlines/y2001/ast27mar_1.htm`

As I mentioned in "Week 150", SOHO is a satellite orbiting the Sun right in front of the Earth, at an unstable equilibrium — a "Lagrange point" — called L1. SOHO is bristling with detectors and telescopes of all sorts, and this movie was taken by a coronagraph, which is a telescope specially designed to block out the Sun's disk and see the fainter corona.

If a coronal mass ejection hits the Earth, it does something like this:

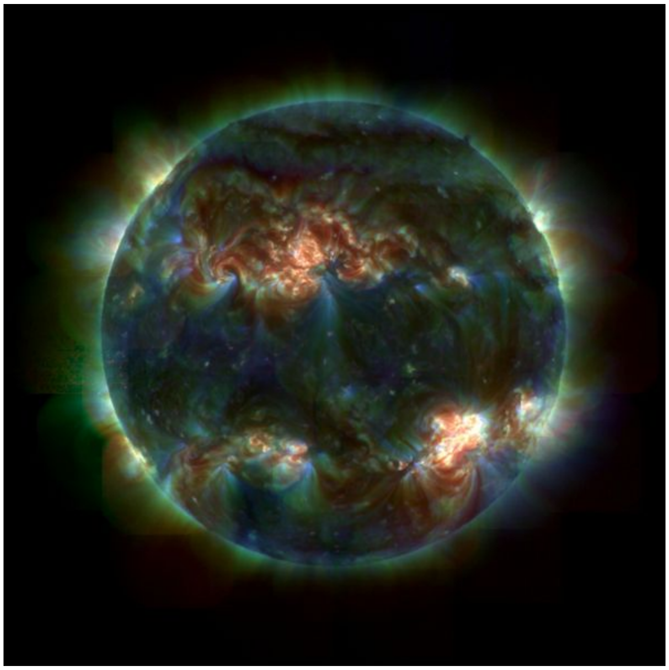
5) NASA, "What is a CME?", `http://www.nasa.gov/mpg/111836main_what_is_a_cme_NASA%20WebV_1.mpg`

In this artist's depiction you can see the plasma shoot off from the Sun, hit the Earth's magnetic field — this actually takes one to five days — and squash it, pushing field lines around to the back side of the Earth. When the magnetic field lines reconnect in back, trillions of watts of power come cascading down through the upper atmosphere, producing auroras. Here's a nice movie of what *those* can look like:

6) YouTube, "Aurora (Northern Lights)", `http://www.youtube.com/watch?v=qIXs6Sh0DKs`

I wish I understood this magnetic field line trickery better! Magnetohydrodynamics — the interactions between electromagnetic fields and plasma — is a branch of physics that always gave me the shivers. The Navier-Stokes equations describing fluid flow are bad enough — if you can prove they have solutions, you'll win $1,000,000 from the Clay Mathematics Institute. Throw in Maxwell's equations and you get a real witches' brew of strange phenomena.

In fact, this subject is puzzling even to experts. For example, why is the Sun's upper atmosphere — the corona — so hot? Here's a picture of the Sun in X-rays taken by another satellite:



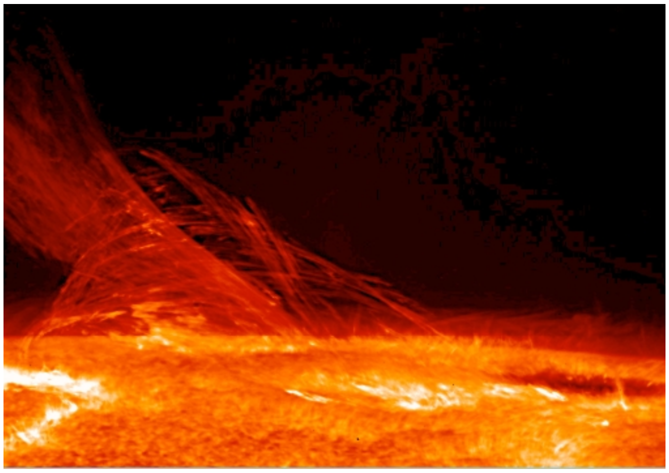
7) Transition Region and Coronal Explorer (TRACE), "Images of the sun", `http://trace.lmsal.com/POD/TRACEpodarchive26.html`

This lets you see plasma in the corona with temperatures between 1 million kelvin (shown as blue) and 2 million kelvin (red). By comparison, the visible surface of the Sun is a mere 5800 kelvin!

Where does the energy come from to heat the corona? There are lots of competing theories. It could even be due to "magnetic field reconnection", the same topological phenomenon that triggers auroras when coronal mass ejections smash into the Earth's magnetic field, as in that movie above. For more, try this:

8) Andrew L. Haynes, Clare E. Parnell, Klaus Galsgaard and Eric R. Priest, "Magnetohydrodynamic evolution of magnetic skeletons", *Proc. Roy. Soc. Lond. A* **463** (2007) 1097–1115. Also available as `astro-ph/0702604`.

A new satellite called Hinode is getting a good look at what's going on, and it seems the magnetic field on the Sun's surface is much more dynamic than before thought:



9) NASA, "Hinode: investigating the Sun's magnetic field", `http://www.nasa.gov/mission_pages/solar-b/`

In fact, weather on the Sun may be more complex than on the Earth. There's "rain" when plasma from the corona cools and falls back down to the Sun's surface... and sometimes there are even tornados! You think tornados on Earth are scary? Check out this movie made during an 8-hour period in August 2000, near the height of the solar cycle:

10) TRACE, "Tornados and fountains in a filament on 2 Aug. 2000", movie 13, `http://trace.lmsal.com/POD/`

Besides the tornados, near the end you can see glowing filaments of plasma following magnetic field lines!

Now for something simpler: the Tale of Groupoidification.

I don't want this to be accessible only to experts, since a bunch of it is so wonderfully elementary. So, I'm going to proceed rather slowly. This may make the experts impatient, so near the end I'll zip ahead and sketch out a bit of the big picture.

Last time I introduced spans of sets. A span of sets is just a set S equipped with functions to X and Y:

$$\begin{tikzcd} & S \arrow[dl, "G"'] \arrow[dr, "F"] & \\ X & & Y \end{tikzcd}$$

Simple! But the important thing is to understand this thing as a "witnessed relation".

Have you heard how computer scientists use the term "witness"? They say the number 17 is a "witness" to the fact that the number 221 isn't prime, since 17 evenly divides 221.

That's the idea here. Given a span S as above, we can say an element x of X and an element y of Y are "related" if there's an element s of S with

$$F(s) = x \quad \text{and} \quad G(s) = y$$

The element s is a "witness" to the relation.

Last week, I gave an example where a Frenchman x and an Englishwoman y were related if they were both the favorites of some Russian s.

Note: there's more information in the span than the relation it determines. The relation either holds or fails to hold. The span does more: it provides a set of "witnesses". The relation holds if this set of witnesses is nonempty, and fails to hold if it's empty.

At least, that's how mathematicians think. When I got married last month, I discovered the state of California demands *two* witnesses attend the ceremony and sign the application for a marriage license. Here the relation is "being married", and the witnesses attest to that relation — but for the state, one witness is not enough to prove that the relation holds! They're using a more cautious form of logic.

To get the really interesting math to show up, we need to look at other examples of "witnessed relations" — not involving Russians or marriages, but geometry and symmetry.

For example, suppose we're doing 3-dimensional geometry. There's a relation "the point x and the line y lie on a plane", but it's pretty dull, since it's always true. More interesting is the witnessed relation "the point x and the line y lie on the plane z". The reason is that sometimes there will be just *one* plane containing a point and a line, but when the point lies on the line, there will be *lots*.

To think of this "witnessed relation" as a span

$$\begin{tikzcd}
& S \arrow[dl, "G"'] \arrow[dr, "F"] & \\
X & & Y
\end{tikzcd}$$

we can take X to be the set of points and Y to be the set of lines.

Can we take S to be the set of planes? No! Then there would be no way to define the functions F and G, because the same plane contains lots of different points and lines. So, we should take S to be the set of triples (x, y, z) where x is a point, y is a line, and z is a plane containing x and y. Then we can take

$$F(x, y, z) = x$$

and

$$G(x, y, z) = y$$

A "witness" to the fact that x and y lie on a plane is not just a plane containing them, but the entire triple.

(If you're really paying attention, you'll have noticed that we need to play the same trick in the example of witnesses to a marriage.)

Spans like this play a big role in "incidence geometry". There are lots of flavors of incidence geometry, with "projective geometry" being the most famous. But, a common feature is that we always have various kinds of "figures" — like points, lines, planes, and so on. And, we have various kinds of "incidence relations" involving these figures. But to really understand incidence geometry, we need to go beyond relations and use spans of sets.

Actually, we need to go beyond spans of sets and use spans of groupoids! The reason is that incidence geometries usually have interesting symmetries, and a groupoid is like a "set with symmetries". For example, consider lines in 3-dimensional space. These form a set, but there are also symmetries of 3-dimensional space mapping one line to another. To take these into account we need a richer structure: a groupoid!

Here's the formal definition: a groupoid consists of a set of "objects", and for any objects x and y, a set of "morphisms"

$$f\colon x \to y$$

which we think of as symmetries taking x to y. We can compose a morphism $f\colon x \to y$ and a morphism $g\colon y \to z$ to get a morphism $fg\colon x \to z$. We think of fg as the result of doing first f and then g. So, we demand the associative law

$$(fg)h = f(gh)$$

whenever either side is well-defined. We also demand that every object x has an identity morphism

$$1_x\colon x \to x$$

We think of this as the symmetry that doesn't do anything to x. So, given any morphism $f\colon x \to y$, we demand that

$$f1_y = f = 1_x f$$

So far this is the definition of a "category". What makes it a "groupoid" is that every morphism $f\colon x \to y$ has an "inverse"

$$f^{-1}\colon y \to x$$

with the property that

$$ff^{-1} = 1_x$$

and

$$f^{-1}f = 1_y$$

In other words, we can "undo" any symmetry.

So, in our spans from incidence geometry:

$$\begin{tikzcd} & S \arrow[dl, "G"'] \arrow[dr, "F"] & \\ X & & Y \end{tikzcd}$$

X, Y and S will be groupoids, while F and G will be maps between groupoids: that is, "functors"!

What's a functor? Given groupoids A and B, clearly a functor

$$F\colon A \to B$$

should send any object x in A to an object $F(x)$ in B. But also, it should send any morphism in A:

$$f\colon x \to y$$

to a morphism in B:

$$F(f)\colon F(x) \to F(y)$$

And, it should preserve all the structure that a groupoid has, namely composition:

$$F(fg) = F(f)F(g)$$

and identities:

$$F(1_x) = 1_{F(x)}$$

It then automatically preserves inverses too:

$$F(f^{-1}) = F(f)^{-1}$$

Given this, what's the meaning of a span of groupoids? You could say it's a "invariant" witnessed relation — that is, a relation with witnesses that's *preserved* by the symmetries at hand. These are the very essence of incidence geometry. For example, if we have a point and a line lying on a plane, we can rotate the whole picture and get a new point and a new line lying on a new plane. Indeed, a "symmetry" in incidence geometry is precisely something that preserves all such "incidence relations".

For those of you not comfy with groupoids, let's see how this actually works. Suppose we have a span of groupoids:

$$\begin{tikzcd} & S \arrow[dl, "G"'] \arrow[dr, "F"] & \\ X & & Y \end{tikzcd}$$

and the object s is a witness to the fact that x and y are related:

$$F(s) = x \quad \text{and} \quad G(s) = y$$

Also suppose we have a symmetry sending s to some other object of S:

$$f\colon s \to s'$$

This gives morphisms

$$F(f)\colon F(s) \to F(s')$$

in X and

$$G(f)\colon G(s) \to G(s')$$

in Y. And if we define

$$F(s') = x' \quad \text{and} \quad G(s') = y'$$

we see that s' is a witness to the fact that x' and y' are related.

Let me summarize the Tale so far:

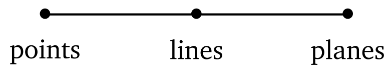
- Spans of groupoids describe "invariant witnessed relations".
- Invariant witnessed relations are the essence of incidence geometry.

- There's a way to turn spans of groupoids into matrices of numbers, so that multiplying matrices corresponds to some nice way of "composing" spans of groupoids (which I haven't really explained yet).

From all this, you should begin to vaguely see that starting from any sort of incidence geometry, we should be able to get a bunch of matrices. Facts about incidence geometry will give facts about linear algebra!

"Groupoidification" is an attempt to reverse-engineer this process. We will discover that lots of famous facts about linear algebra are secretly facts about incidence geometry!

To prepare for what's to come, the maniacally diligent reader might like to review "Week 178", "Week 180", "Week 181", "Week 186" and "Week 187", where I explained how any Dynkin diagram gives rise to a flavor of incidence geometry. For example, the simplest-looking Dynkin diagrams, the A_n series, like this for $n = 3$:



give rise to n-dimensional projective geometry. I may have to review this stuff, but first I'll probably say a bit about the theory of group representations and Hecke algebras.

(There will also be other ways to get spans of groupoids, that don't quite fit into what's customarily called "incidence geometry", but still fit very nicely into our Tale. For example, Dynkin diagrams become "quivers" when we give each edge a direction, and the "groupoid of representations of a quiver" gives rise to linear-algebraic structures related to a quantum group. In fact, I already mentioned this in item E of "Week 230". Eventually this will let us groupoidify the whole theory of quantum groups! But, I don't want to rush into that, since it makes more sense when put in the right context.)

By the way, some of you have already pointed out how unfortunate it is that *last* Week was devoted to E_8, instead of *this* one. Sorry.

Addendum: I thank logopetria for catching typos. For more discussion, go to the n-Category Caf.

> Science is the only news. When you scan through a newspaper or magazine, all the human interest stuff is the same old he-said-she-said, the politics and economics the same sorry cyclic dramas, the fashions a pathetic illusion of newness, and even the technology is predictable if you know the science.
>
> — *Stewart Brand*

Week 249

April 8, 2007

As you may recall, I'm telling a long story about symmetry, geometry, and algebra. Some of this tale is new work done by James Dolan, Todd Trimble and myself. But a lot of it is old work by famous people which deserves a modern explanation.

A great example is Felix Klein's "Erlangen program" — a plan for reducing many sorts of geometry to group theory. Many people tip their hat to the Erlanger program, but few seem to deeply understand it, and even fewer seem to have read what Klein actually wrote about it!

The problem goes back a long ways. In 1871, while at Gttingen, Klein worked on non-Euclidean geometry, and showed that hyperbolic geometry was consistent if and only if Euclidean geometry was. In the process, he must have thought hard about the role of symmetry groups in geometry. When he was appointed professor at Erlangen in 1872, he wrote a lecture outlining his "Erlanger Programm" for reducing geometry to group theory.

But, he didn't actually give this lecture as his inaugural speech! He spoke about something else.

So, nobody ever heard him announce the Erlangen program. And, until recently, the lecture he wrote was a bit hard to find. Luckily, now you can get it online:

1) Felix Klein, "Vergleichende Betrachtungen ueber neuere geometrische Forschungen", Verlag von Andreas Deichert, Erlangen, 1872. Also available at the University of Michigan Historical Mathematics Collection, `http://www.hti.umich.edu/cgi/t/text/text-idx?c=umhistmath;idno=ABN7632`

Even better, Johan Ernst Mebius has recently prepared an HTML version, with links to the above version:

2) Johan Ernst Mebius, "Felix Klein's Erlanger Programm", `http://www.xs4all.nl/~jemebius/ErlangerProgramm.htm`

But what if you have the misfortune of only reading English, not German? Until now the only translation was quite hard to obtain:

3) Felix Klein, "A comparative review of recent researches in geometry", trans. M. W. Haskell, *Bull. New York Math. Soc.* **2**, (1892-1893), 215–249.

In case you're wondering, the "New York Mathematical Society" no longer exists! It was founded in 1888, but in 1894 it went national and became the American Mathematical Society.

Luckily, after Thomas Love pointed out the existence of this old translation, Chris Hillman was able to get ahold of it and scan it in! Then Robin Houston created a PDF file of the whole thing, and Lukas-Fabian Moser created a DjVu file. Then Nitin C. Rughoonauth took the marvelous step of putting it into LaTeX! So now, you can read Klein's paper in English here:

4) "The Erlangen program", `http://math.ucr.edu/home/baez/erlangen/`

English-speakers can read more about the Erlangen program here:

5) Felix Klein, *Elementary Mathematics from an Advanced Standpoint: Geometry*, part 3: "Systematic discussion of geometry and its foundations", Dover, New York, 1939.

Luckily Dover keeps its books in print!

For more on the Erlangen program, try these:

6) Garrett Birkhoff and M. K. Bennett, 'Felix Klein and his "Erlanger Programm"', in *History and Philosophy of Modern Mathematics*, eds. W. Aspray and P. Kitcher, Minnesota Stud. Philos. Sci. XI, University of Minnesota Press, Minneapolis, 1988, pp. 145–176.

7) Hans A. Kastrup, "The contributions of Emmy Noether, Felix Klein and Sophus Lie to the modern concept of symmetries in physical systems", in *Symmetries in Physics (1600–1980)*, ed. M. G. Doncel, World Scientific, Singapore, 1987, pp. 113–163.

8) I. M. Yaglom, *Felix Klein and Sophus Lie: Evolution of the Idea of Symmetry in the Nineteenth Century*, trans. S. Sossinsky, Birkhauser, Boston, 1988.

For more about Klein, try "Week 213" and this little biography:

9) MacTutor History of Mathematics Archive, "Felix Klein", `http://www-history.mcs.st-andrews.ac.uk/Biographies/Klein.html`

But what does the Erlangen program actually amount to, in the language of modern mathematics? This will take a while to explain, so the best thing is to dive right in.

Last week in the Tale of Groupoidification I tried to explain two slogans:

GROUPOIDS ARE LIKE 'SETS WITH SYMMETRIES'

SPANS OF GROUPOIDS ARE LIKE 'INVARIANT WITNESSED RELATIONS'

They're a bit vague; they're mainly designed to give you enough intuition to follow the next phase of the Tale, which is all about how:

GROUPOIDS GIVE VECTOR SPACES

SPANS OF GROUPOIDS GIVE LINEAR OPERATORS

But before the next phase, I need to say a bit about how groupoids and spans of groupoids fit into Klein's Erlangen program.

Groupoids are a modern way to think about symmetries. A more traditional approach would use a group acting as symmetries of some set. And the most traditional approach of all, going back to Galois and Klein, uses a group acting *transitively* on a set.

So, let me explain the traditional approach, and then relate it to the modern one.

I hope you know what it means for a group G to "act" on a set X. It means that for any element x of X and any guy g in G, we get a new element gx in X. We demand that

$$1x = x$$

and

$$g(hx) = (gh)x.$$

More precisely, this is a "left action" of G on X, since we write the group elements to the left of x. We can also define right actions, and someday we may need those too.

We say an action of a group G on a set X is "transitive" if given any two elements of X, there's some guy in G mapping the first element to the second. In this case, we have an isomorphism of sets

$$X = G/H$$

for some subgroup H of G.

For example, suppose we're studying a kind of geometry where the symmetry group is G. Then X could be the set of figures of some sort: points, or lines, or something fancier. If G acts transitively on X, then all figures of this sort "look alike": you can get from any one to any other using a symmetry. This is often the case in geometry... but not always.

Suppose G acts transitively on X. Pick any figure x of type X and let H be its "stabilizer": the subgroup consisting of all guys in G that map x to itself. Then we get a one-to-one and onto map

$$f \colon X \to G/H$$

sending each figure gx in X to the equivalence class $[g]$ in G/H.

If you haven't seen this fact before, you should definitely prove it — it's one of the big ways people use symmetry!

Here's one kind of thing people do with this fact. The 3d rotation group $G = \mathrm{SO}(3)$ acts on the sphere $X = S^2$, and the stabilizer of the north pole is the 2d rotation group $H = \mathrm{SO}(2)$, so the sphere is isomorphic to $G/H = \mathrm{SO}(3)/\mathrm{SO}(2)$. The same sort of result holds in any dimension, and we can use it to derive facts about spheres from facts about rotation groups, and vice versa.

A grander use of this fact is to set up a correspondence between sets on which G acts transitively and subgroups of G. This is one of the principles lurking behind Galois theory.

Galois applied this principle to number theory — see "Week 201" for details. But, it really has nothing particular to do with number theory! In his Erlangen program, Klein applied it to geometry.

Klein's goal was to systematize a bunch of different kinds of non-Euclidean geometry. Each kind of geometry he was interested in had a different group of symmetries. For example:

- n-dimensional spherical geometry has the rotation group $\mathrm{SO}(n+1)$ as symmetries. (Or, if you want to include reflections, the bigger group $\mathrm{O}(n+1)$.)
- n-dimensional Euclidean geometry has the Euclidean group $\mathrm{ISO}(n)$ as symmetries. (This group is built from rotations in $\mathrm{SO}(n)$ together with translations in $\mathbb{R}^n$.)
- n-dimensional hyperbolic geometry has the group $\mathrm{SO}(n,1)$ as symmetries. (This group also shows up in special relativity under the name of the "Lorentz group": it acts on the "mass hyperboloid", and that's how hyperbolic geometry shows up in special relativity.)
- n-dimensional projective geometry has the group $\mathrm{SL}(n+1)$ as symmetries. (This group consists of $(n+1) \times (n+1)$ matrices with determinant 1. Scalar multiples of the identity act trivially on projective space, so it's actually better to use the "projective general linear group" $\mathrm{PGL}(n+1)$, consisting of invertible matrices mod scalars. But, this has the same Lie algebra as $\mathrm{SL}(n+1)$, so people are often a bit slack about which group they use.)

The details here don't matter much yet; the point is that there are lots of interesting kinds of geometry, with interesting symmetry groups!

Klein realized that in any kind of geometry like this, a "type of figure" corresponds to a set on which G acts transitively. Here a "figure" could be a point, a line, a plane, or something much fancier. Regardless of the details, the set of all figures of the same type can be written as G/H, and G acts transitively on this set.

The really cool part is that we can use Klein's idea to *define* a geometry for any group G. To do this, we just say that *every* subgroup H of G gives rise to a type of figure. So, we work out all the subgroups of G. Then, we work out all the incidence relations — relations like "a point lies on a line". To do this, we take two sets of figures, say

$$X = G/H$$

and

$$Y = G/K$$

and find all the invariant relations between them: that is, subsets of $X \times Y$ preserved by all the symmetries. I'll say more about how to do this next time — we can use something called "double cosets". In nice cases, like when G is a simple Lie group and H and K are so-called "parabolic" subgroups, these let us express all the invariant relations in terms of finitely many "atomic" ones! So, we can really carry out Klein's program of thoroughly understanding geometry starting from groups — at least in nice cases.

In short, group actions — especially transitive ones — are a traditional and very powerful way of using symmetry to tackle lots of problems.

So, to bridge the gap between the traditional and the new, I should explain how group actions give groupoids. I'll show you that:

> *A GROUPOID EQUIPPED WITH CERTAIN EXTRA STUFF IS*
> *THE SAME AS A GROUP ACTION*

It's not very hard to get a groupoid from a group action. Say we have a group G acting on a set X. Then the objects of our groupoid are just elements of X, and a morphism

$$g\colon x \to y$$

is just a group element g with

$$gx = y.$$

Composing morphisms works the obvious way — it's basically just multiplication in the group G.

Some people call this groupoid an "action groupoid". I often call it the "weak quotient" $X//G$, since it's like the ordinary quotient X/G, but instead of declaring that x and y are *equal* when we have a group element g sending x to y, we instead declare they're *isomorphic* via a specified isomorphism $g\colon x \to y$.

But for now, let's call $X//G$ the "action groupoid".

So, group actions give action groupoids. But, these groupoids come with extra stuff!

First of all, the action groupoid $X//G$ always comes equipped with a functor

$$X//G \xrightarrow{p} G$$

sending any object of $X//G$ to the one object of G, and any morphism $g\colon x \to y$ to the corresponding element of G. Remember, a group is a groupoid with one object: this is the 21st century!

Second of all, this functor p is always "faithful": given two morphisms from x to y, if p maps them to the same morphism, then they were equal.

And that's all! Any groupoid with a faithful functor to G is equivalent to the action groupoid $X//G$ for some action of G on some set X. This takes a bit of proving... let's not do it now.

So: in my slogan

> *A GROUPOID EQUIPPED WITH CERTAIN EXTRA STUFF IS*
> *THE SAME AS A GROUP ACTION*

the "certain extra stuff" was precisely a faithful functor to G.

What if we have a *transitive* group action? Then something nice happens.

First of all, saying that G acts transitively on X is the same as saying there's a morphism between any two objects of $X//G$. In other words, all objects of $X//G$ are isomorphic. Or in other words, there's just one isomorphism class of objects.

Just as a groupoid with one object is a group, a groupoid with one *isomorphism class* of objects is *equivalent* to a group. Here I'm using the usual notion of "equivalence" of categories, as explained back in "Week 76".

So, G acts transitively on X precisely when $X//G$ is equivalent to a group!

And what group? Well, what could it possibly be? It's just the stabilizer of some element of X! So, in the case of a transitive group action, our functor

$$X//G \xrightarrow{p} G$$

is secretly equivalent to the inclusion

$$H \xrightarrow{i} G$$

of the stabilizer group of this element.

So, we see how Klein's old idea of geometrical figures as subgroups of G is being generalized. We can start with any groupoid Y of "figures" and "symmetries between figures", and play with that. It becomes an action groupoid if we equip it with a faithful functor to some group G:

$$Y \to G$$

Then the action is transitive if all the objects of Y are isomorphic. In that case, our functor is equivalent to an inclusion

$$H \to G$$

and we're back down to Klein's approach to geometry. But, it's actually good to generalize what Klein did, and think about arbitrary "groupoids over G" — that is, groupoids equipped with functors to G.

So, when we blend our ideas on spans of groupoids with Klein's ideas, we'll want to use spans of groupoids "over G" — that is, commutative diamonds of groupoids and functors, like this:

$$\begin{tikzcd} & S \arrow[dl] \arrow[dr] & \\ X \arrow[dr] & & Y \arrow[dl] \\ & G & \end{tikzcd}$$

There's much more to say about this, but not today!

I'll say one last thing before quitting. It's a bit more technical, but I feel an urge to see it in print.

People often talk about "the" stabilizer group of a transitive action of some group G on some set X. This is a bit dangerous, since every element of X has its own stabilizer, and they're not necessarily all equal!

However, they're all *conjugate*: if the stabilizer of x is H, then the stabilizer of gx is gHg^{-1}.

So, when I say above that

$$X//G \xrightarrow{p} G$$

is equivalent to

$$H \xrightarrow{i} G$$

I could equally well have said it's equivalent to

$$H \xrightarrow{i'} G$$

where the inclusion i' is the inclusion i conjugated by g. If you know some category theory, you'll see that i and i' are naturally isomorphic: a natural isomorphism between functors between groups is just a "conjugation". Picking the specific inclusion i requires picking a specific element x of X.

Of course, I'll try to write later issues in a way that doesn't force you to have understood all these nuances!

Addendum: For more discussion, go to the n-Category Caf.

> There is no benefit today in arithmetic in Roman numerals. There is also no benefit today in insisting that the group concept is more fundamental than that of groupoid.
>
> — *Ronald Brown*

Week 250

April 26, 2007

Right now I'm in a country estate called Les Treilles in southern France, at a conference organized by Alexei Grinbaum and Michel Bitbol:

1) "Philosophical and Formal Foundations of Modern Physics", `http://www-drecam.cea.fr/Phocea/Vie_des_labos/Ast/ast_visu.php?id_ast=762`

It's very beautiful here, but about 20 philosophers, physicists and mathematicians have agreed to spend six days indoors discussing quantum gravity, the history of relativity, quantum information theory and the like. And guess what? Now it's our afternoon off, and I'm spending my time writing This Week's Finds! Some people just don't know how to enjoy life.

In fact, I want to continue telling you The Tale of Groupoidification. But before I do, here's a puzzle that Jeffrey Bub raised the other night at dinner. It's not hard, but it's still a bit surprising.

You and your friend each flip a fair coin and then look at it. You can't look at your friend's coin; they can't look at yours. You can't exchange any information while the game is being played, though you can choose a strategy beforehand. Each of you must guess whether the other's coin lands heads up or tails up. Your goal, as a team, is to maximize the chance that you're both correct.

What's the best strategy, and what's the probability that you both guess correctly?

Here's an obvious line of thought.

Since you don't have any information about your friend's coin flip, it doesn't really matter what you guess. So, you might as well guess "heads". You'll then have a $1/2$ chance of being right. Similarly, your friend might as well guess "heads" — or for that matter, "tails". They'll also have a $1/2$ chance of being right. So, the chance that you're both right is $1/2 \times 1/2 = 1/4$.

I hope that sounds persuasive — but you can actually do much better!

How? I'll give away the answer at the end.

Jeffrey Bub is famous for his work on the philosophy of quantum mechanics, and in his talk today he mentioned a similar but more sophisticated game, the Popescu-Rohrlich game. Here you and your friend each flip coins as before. But now, after looking at your coin, you each write either "yes" or "no" on a pad of paper. Your goal, as a team, is to give the same response when at least one coin lands heads up, but different responses otherwise.

Classically the best you can do is both say "yes" — or, if you prefer, both say "no". Then you'll have a $3/4$ chance of winning. But, if before playing the game you and your friend prepare a pair of spin-$1/2$ particles in the Bell state, and you each keep one, you can use these to boost your chance of winning to about 85%!

I think the underlying idea first appeared here:

1) S. Popescu and D. Rohrlich, "Nonlocality as an axiom", *Found. Phys.* **24** (1994), 379–385.

For the "game" version, try this:

2) Nicolas Gisin, "Can relativity be considered complete? From Newtonian nonlocality to quantum nonlocality and beyond", available as quant-ph/0512168.

There's a lot more to say about this — especially about the "Popescu-Rohrlich box", a mythical device which would let you win all the time at this game, but still not allow signalling. The existence of such a box is logically possible, but forbidden by quantum mechanics. It can only exist in certain "supra-quantum theories" which allow even weirder correlations than quantum mechanics.

But, I don't understand this stuff, so you should just read this:

3) Valerio Scarani, "Feats, features and failures of the PR-box", available as quant-ph/0603017.

Okay — now for our Tale. I want to explain double cosets as spans of groupoids... but it's best if I start with some special relativity.

Though Newton seems to have believed in some form of "absolute space", the idea that motion is relative predates Einstein by a long time. In 1632, in his Dialogue Concerning the Two Chief World Systems, Galileo wrote:

> *Shut yourself up with some friend in the main cabin below decks on some large ship, and have with you there some flies, butterflies, and other small flying animals. Have a large bowl of water with some fish in it; hang up a bottle that empties drop by drop into a wide vessel beneath it. With the ship standing still, observe carefully how the little animals fly with equal speed to all sides of the cabin. The fish swim indifferently in all directions; the drops fall into the vessel beneath; and, in throwing something to your friend, you need throw it no more strongly in one direction than another, the distances being equal; jumping with your feet together, you pass equal spaces in every direction.*
>
> *When you have observed all these things carefully (though doubtless when the ship is standing still everything must happen in this way), have the ship proceed*

> *with any speed you like, so long as the motion is uniform and not fluctuating this way and that. You will discover not the least change in all the effects named, nor could you tell from any of them whether the ship was moving or standing still.*

As a result, the coordinate transformation we use in Newtonian mechanics to switch from one reference frame to another moving at a constant velocity relative to the first is called a "Galilei transformation". For example:

$$(t, x, y, z) \mapsto (t, x + vt, y, z)$$

By the time Maxwell came up with his equations describing light, the idea of relativity of motion was well established. In 1876, he wrote:

> *Our whole progress up to this point may be described as a gradual development of the doctrine of relativity of all physical phenomena. Position we must evidently acknowledge to be relative, for we cannot describe the position of a body in any terms which do not express relation. The ordinary language about motion and rest does not so completely exclude the notion of their being measured absolutely, but the reason of this is, that in our ordinary language we tacitly assume that the earth is at rest.... There are no landmarks in space; one portion of space is exactly like every other portion, so that we cannot tell where we are. We are, as it were, on an unruffled sea, without stars, compass, sounding, wind or tide, and we cannot tell in what direction we are going. We have no log which we can case out to take a dead reckoning by; we may compute our rate of motion with respect to the neighboring bodies, but we do not know how these bodies may be moving in space.*

So, the big deal about special relativity is *not* that motion is relative. It's that this is possible while keeping the speed of light the same for everyone — as Maxwell's equations insist, and as we indeed see! This is what forced people to replace Galilei transformations by "Lorentz transformations", which have the new feature that two coordinate systems moving relative to each other will disagree not just on where things are, but *when* they are.

As Einstein wrote in 1905:

> *Examples of this sort, together with the unsuccessful attempts to discover any motion of the earth relative to the "light medium", suggest that the phenomena of electrodynamics as well as mechanics possess no properties corresponding to the idea of absolute rest. They suggest rather that, as has already been shown to the first order of small quantities, the same laws of electrodynamics and optics will be valid for all frames of reference for which the equations of mechanics are valid. We will elevate this conjecture (whose content will be called the "principle of relativity") to the status of a postulate, and also introduce another postulate, which is only apparently irreconcilable with it, namely, that light is always propagated in empty space with a definite velocity c which is independent of the state of motion of the emitting body. These two postulates suffice for attaining a simple and consistent theory of the electrodynamics of moving bodies based on Maxwell's theory for stationary bodies.*

So, what really changed with the advent of special relativity? First, our understanding of precisely which transformations count as symmetries of spacetime. These transformations form a *group*. Before special relativity, it seemed the relevant group was a 10-dimensional gadget consisting of:

- 3 dimensions of spatial translations
- 1 dimension of time translations
- 3 dimensions of rotations
- 3 dimensions of Galilei transformations

Nowadays this is called the “Galilei group”:
With special relativity, the relevant group became the “Poincare group”:

- 3 dimensions of spatial translations
- 1 dimension of time translations
- 3 dimensions of rotations
- 3 dimensions of Lorentz transformations

It’s still 10-dimensional, not any bigger. But, it acts differently as transformations of the spacetime coordinates (t, x, y, z).

Another thing that changed was our appreciation of the importance of symmetry! Before the 20th century, group theory was not in the toolkit of most theoretical physicists. Now it is.

Okay. Now suppose you’re the only thing in the universe, floating in empty space, not rotating. To make your stay in this thought experiment a pleasant one, I’ll give you a space suit. And for simplicity, suppose special relativity holds true exactly, with no gravitational fields to warp the geometry of spacetime.

Would the universe be any different if you were moving at constant velocity? Or translated 2 feet to the left or right? Or turned around? Or if it were one day later?

No! Not in any observable way, at least! It would seem exactly the same.

So in this situation, it doesn’t really make much sense to say “where you are”, or “which way you’re facing”, or “what time it is”. There are no “invariant propositions” to make about your location or motion. In other words, there’s nothing to say whose truth value remains unchanged after you apply a symmetry.

Well, *almost* nothing to say! The logicians in the crowd will note that you can say “T”: the tautologously true statement. You can also say “F”: the tautologously false statement. But, these aren’t terribly interesting.

Next, suppose you have a friend also floating through space. Now there are more interesting invariant propositions. There’s nothing much invariant to say about just you, and nothing to say about just your friend, but there are invariant *relations*. For example, you can measure your friend’s speed relative to you, or your distance of closest approach.

Mathematicians study invariant relations using a tool called “double cosets”. I want to explain these today, since we’ll need them soon in the Tale of Groupoidification.

"Double cosets" sound technical, but that's just to keep timid people from understanding the subject. A double coset is secretly just an "atomic" invariant relation: one that can't be expressed as "P or Q" where P and Q are themselves invariant relations — unless precisely one of P or Q is tautologously false.

So, atomic invariant relations are like prime numbers: they can't be broken down into simpler bits. And, as we'll see, every invariant relation can be built out of atomic ones!

Here's an example in the case we're considering:

> *"My friend's speed relative to me is 50 meters/second, and our distance of closest approach is 10 meters."*

This is clearly an invariant relation. It's atomic if we idealize the situation and assume you and your friends are points — so we can't ask which way you're facing, whether you're waving at each other, etc.

To see *why* it's atomic, note that we can always find a frame of reference where you're at rest and your friend is moving by like this:

——— friend ——→

you

If you and your friend are points, the situation is *completely described* (up to symmetries) by the relative speed and distance of closest approach. So, the invariant relation quoted above can't be written as "P or Q" for other invariant relations.

The same analysis shows that in this example, *every* atomic invariant relation is of this form:

> *"My friend's speed relative to me is s, and our distance of closest approach is d."*

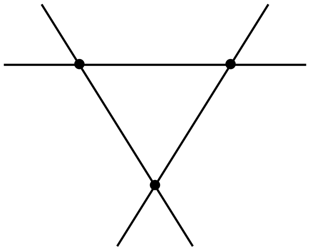
for some nonnegative numbers s and d.

(Quiz: why don't we need to let s be negative if your friend is moving to the left?)

From this example, it's clear there are often infinitely many double cosets. But there are some wonderful examples with just *finitely many* double cosets — and these are what I'll focus on in our Tale.

Here's the simplest one. Suppose we're doing projective plane geometry. This is a bit like Euclidean plane geometry, but there are more symmetries: every transformation that preserves lines is allowed. So, in addition to translations and rotations, we also have other symmetries.

For example, imagine taking a blackboard with some points and lines on it:



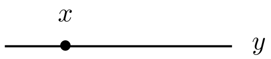
We can translate it and rotate it. But, we can also view it from an angle: that's another symmetry in projective geometry! This hints at how projective geometry arose from the study of perspective in painting.

We get even more symmetries if we use a clever trick. Suppose we're standing on the blackboard, and it extends infinitely like an endless plain. Points on the horizon aren't really points on the blackboard. They're called "points at infinity". But, it's nice to include them as part of the so-called "projective plane". They make things simpler: now every pair of lines intersects in a unique point, just as every pair of points lies on a unique line. You've probably seen how parallel railroad tracks seem to meet at the horizon — that's what I'm talking about here. And, by including these extra points at infinity, we get extra symmetries that map points at infinity to ordinary points, and vice versa.

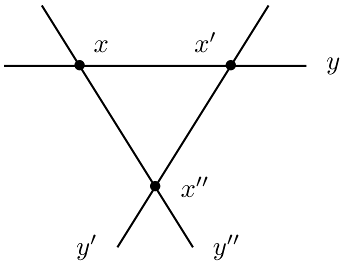
I gave a more formal introduction to projective geometry in "Week 106" and "Week 145", and "Week 178". If you read these, you'll know that points in the projective plane correspond to lines through the origin in a 3d space. And, you'll know a bit about the group of symmetries in projective geometry: it's the group $G = \mathrm{PGL}(3)$, consisting of 3×3 invertible matrices, modulo scalars.

(I actually said $\mathrm{SL}(3)$, but I was being sloppy — this is another group with the same Lie algebra.)

For some great examples of double cosets, let F be the space of "flags". A "flag" is a very general concept, but in projective plane geometry a flag is just a point x on a line y:



An amazing fact is that there are precisely 6 atomic invariant relations between a pair of flags. You can see them all in this picture:



There are six flags here, and each exemplifies a different atomic invariant relation to our favorite flag, say (x, y).

For example, the flag (x', y') has the following relation to (x, y):

> *"The point of (x', y') lies on the line of (x, y), and no more."*

By "no more" I mean that no further incidence relations hold.

There's a lot more to say about this, and we'll need to delve into it much deeper soon. . . but not yet. For now, I just want to mention that all this stuff generalizes from $G = \mathrm{PGL}(3)$ to any other simple Lie group! And, the picture above is an example of a general concept, called an "apartment". Apartments are a great way to visualize atomic invariant relations between flags.

This "apartment" business is part of a wonderful theory due to Jacques Tits, called the theory of "buildings". The space of *all* flags is a building; a building has lots of apartments in it. Buildings have a reputation for being scary, because in his final polished treatment, Tits started with a few rather unintuitive axioms and derived everything from these. But, they're actually lots of fun if you draw enough pictures!

Next, let me explain why people call atomic invariant relations "double cosets".

First of all, what's a relation between two sets X and Y? We can think of it as a subset S of $X \times Y$: we say a pair (x, y) is in S if the relation holds.

Next, suppose some group G acts on both X and Y. What's an "invariant" relation? It's a subset S of $X \times Y$ such that whenever (x, y) is in S, so is (gx, gy). In other words, the relation is preserved by the symmetries.

Now let's take these simple ideas and make them sound more complicated, to prove we're mathematicians. Some of you may want to take a little nap right around now — I'm just trying to make contact with the usual way experts talk about this stuff.

First, let's use an equivalent but more technical way to think of an invariant relation: it's a subset of the quotient space $G\backslash(X \times Y)$.

Note: often I'd call this quotient space $(X \times Y)/G$. But now I'm writing it with the G on the left side, since we had a *left* action of G on X and Y, hence on $X \times Y$ — and in a minute we're gonna need all the sides we can get!

Second, recall from last Week that if G acts *transitively* on both X and Y, we have isomorphisms

$$X \cong G/H$$

and

$$Y \cong G/K$$

for certain subgroups H and K of G. Note: here we're really modding out by the *right* action of H or K on G.

Combining these facts, we see that when G acts transitively on both X and Y, an invariant relation is just a subset of

$$G\backslash(X \times Y) \cong G\backslash(G/H \times G/K)$$

Finally, if you lock yourself in a cellar and think about this for a few minutes (or months), you'll realize that this weird-looking set is isomorphic to

$$H\backslash G/K$$

This notation may freak you out at first — I know it scared me! The point is that we can take G, mod out by the right action of K to get G/K, and then mod out by the left action of H on G/K, obtaining

$$H\backslash(G/K).$$

Or we can take G, mod out by the left action of H to get $H\backslash G$, and then mod out by the right action of K on $H\backslash G$, obtaining

$$(H\backslash G)/K.$$

And, these two things are isomorphic! So, we relax and write

$$H\backslash G/K$$

A point in here is called a "double coset": it's an equivalence class consisting of all guys in G of the form

$$hgk$$

for some fixed g, where h ranges over H and k ranges over K.

Since subsets of $H\backslash G/K$ are invariant relations, we can think of a point in $H\backslash G/K$ as an "atomic" invariant relation. Every invariant relation is the union — the logical "or" — of a bunch of these.

So, just as any hunk of ordinary matter can be broken down into atoms, every invariant statement you can make about an entity of type X and an entity of type Y can broken down into "atomic" invariant relations — also known as double cosets!

So, double cosets are cool. But, it's good to fit them into the "spans of groupoids" perspective. When we do this, we'll see:

> *A SPAN OF GROUPOIDS EQUIPPED WITH CERTAIN EXTRA STUFF IS*
> *THE SAME AS A DOUBLE COSET.*

This relies on the simpler slogan I mentioned last time:

> *A GROUPOID EQUIPPED WITH CERTAIN EXTRA STUFF IS*
> *THE SAME AS A GROUP ACTION.*

Let's see how it goes. Suppose we have two sets on which G acts transitively, say X and Y. Pick a figure x of type X, and a figure y of type Y. Let H be the stabilizer of x, and let K be the stabilizer of y. Then we get isomorphisms

$$X \cong G/H$$

and

$$Y \cong G/K$$

The subgroup $H \cap K$ stabilizes both x and y, and

$$Z = G/(H \cap K)$$

is another set on which G acts transitively. How can we think of this set? It's the set of all pairs of figures, one of type X and one of type Y, which are obtained by taking the pair (x, y) and applying an element of G. So, it's a subset of $X \times Y$ that's invariant under the action of G. In other words, it's an invariant relation between X and Y!

Furthermore, it's the smallest invariant subset of $X \times Y$ that contains the pair (x, y). So, it's an *atomic* invariant relation — or in other words, a double coset!

Now, let's see how to get a span of groupoids out of this. We have a commutative diamond of group inclusions:

$$\begin{tikzcd}
 & H \cap K \arrow[dl, hook] \arrow[dr, hook] & \\
H \arrow[dr, hook] & & K \arrow[dl, hook] \\
 & G &
\end{tikzcd}$$

This gives a commutative diamond of spaces on which G acts transitively:

$$\begin{tikzcd}
& G/(H \cap K) \arrow[dl] \arrow[dr] & \\
G/H \arrow[dr] & & G/K \arrow[dl] \\
& G/G &
\end{tikzcd}$$

We already have names for three of these spaces — and G/G is just a single point, say $*$:

$$\begin{tikzcd}
& Z \arrow[dl] \arrow[dr] & \\
X \arrow[dr] & & Y \arrow[dl] \\
& * &
\end{tikzcd}$$

Now, in "Week 249" I explained how you could form the "action groupoid" $X//G$ given a group G acting on a space X. If I were maniacally consistent, I would write it as $G\backslash\backslash X$, since G is acting on the left. But, I'm not. So, the above commutative diamond gives a commutative diamond of groupoids:

$$\begin{tikzcd}
& Z//G \arrow[dl] \arrow[dr] & \\
X//G \arrow[dr] & & Y//G \arrow[dl] \\
& *//G &
\end{tikzcd}$$

The groupoid on the bottom has one object, and one morphism for each element of G. So, it's just G! So we have this:

$$\begin{tikzcd}
& Z//G \arrow[dl] \arrow[dr] & \\
X//G \arrow[dr] & & Y//G \arrow[dl] \\
& G &
\end{tikzcd}$$

So — voila! — our double coset indeed gives a span of groupoids

$$\begin{tikzcd}
& Z//G \arrow[dl] \arrow[dr] & \\
X//G & & Y//G
\end{tikzcd}$$

$X//G$ is the groupoid of figures just like x (up to symmetry), $Y//G$ is the groupoid of figures just like y, and $Z//G$ is the groupoid of *pairs* of figures satisfying the same atomic invariant relation as the pair (x, y). For example, point-line pairs, where the point lies on the line! For us, a pair of figures is just a more complicated sort of figure.

But, this span of groupoids is a span "over G", meaning it's part of a commutative diamond with G at the bottom:

$$\begin{tikzcd} & Z//G \arrow[dl] \arrow[dr] & \\ X//G \arrow[dr] & & Y//G \arrow[dl] \\ & G & \end{tikzcd}$$

If you remember everything in "Week 249" — and I bet you don't — you'll notice that this commutative diamond is equivalent to diamond we started with:

$$\begin{tikzcd} & H \cap K \arrow[dl] \arrow[dr] & \\ H \arrow[dr] & & K \arrow[dl] \\ & G & \end{tikzcd}$$

We've just gone around in a loop! But that's okay, because we've learned something en route.

To tersely summarize what we've learned, let's use the fact that a groupoid is equivalent to a group precisely when it's "connected": that is, all its objects are isomorphic. Furthermore, a functor between connected groupoids is equivalent to an inclusion of groups precisely when it's "faithful": one-to-one on each homset. So, when I said that:

> *A SPAN OF GROUPOIDS EQUIPPED WITH CERTAIN EXTRA STUFF IS THE SAME AS A DOUBLE COSET.*

what I really meant was:

> *A SPAN OF CONNECTED GROUPOIDS FAITHFULLY OVER G IS THE SAME AS A DOUBLE COSET.*

If that's too terse, let me elaborate for you: a "span of connected groupoids faithfully over G" is a commutative diamond

$$\begin{tikzcd} & C \arrow[dl] \arrow[dr] & \\ A \arrow[dr] & & B \arrow[dl] \\ & G & \end{tikzcd}$$

where A, B, C are connected groupoids and the arrows are faithful functors.

This sounds complicated, but it's mainly because we're trying to toss in extra conditions to make our concepts match the old-fashioned "double coset" notion. Here's a simpler, more general fact:

A SPAN OF GROUPOIDS FAITHFULLY OVER G
IS THE SAME AS A SPAN OF G-SETS.

where a "G-set" is a set on which G acts. This is the natural partner of the slogan I explained last Week, though not in this language:

A GROUPOID FAITHFULLY OVER G
IS THE SAME AS A G-SET.

Things get even simpler if we drop the "faithfulness" assumption, and simply work with groupoids over G, and spans of these. This takes us out of the traditional realm of group actions on sets, and into the 21st century! And that's where we want to go.

Indeed, for the last couple weeks I've just been trying to lay out the historical context for the Tale of Groupoidification, so experts can see how the stuff to come relates to stuff that's already known. In some ways things will get simpler when I stop doing this and march ahead. But, I'll often be tempted to talk about group actions on sets, and double cosets, and other traditional gadgets. . . so I feel obliged to set the stage.

Okay — here's the answer to the puzzle. Close your eyes if you want to think about it more.

An optimal strategy is for you and your friend to each look at your own coin, and then guess that the other coin landed the other way: heads if yours was tails, and tails if yours was heads. With this strategy, the chance you're both correct is $1/2$.

Or, you can both guess that the other coin landed the *same* way. This works just as well.

The point is: you and your friend can do twice as well at this game if you each use the result of your own coin toss to guess the result of the other's coin toss!

It seems paradoxical that using this random and completely uncorrelated piece of information — the result of your own coin toss — helps you guess what your friend's coin will do, and vice versa.

But of course it *doesn't*. You each still have just a $1/2$ chance of guessing the other's coin toss correctly. What the trick accomplishes is correlating your guesses, so you both guess right or both guess wrong together. This improves the chance of winning from $1/2 \times 1/2$ (the product of two independent probabilities) to $1/2$.

By the way, the translation of the passage by Einstein is due to Michael Friedman, a philosopher at Stanford; he used it in his talk at this conference. There's a lot more to say about talks at this conference. Let's see if I get around to it.

Also by the way: if you fix a collection of n G-sets, there's always a Boolean algebra of n-ary invariant relations. Only the case $n = 2$ is related to double cosets, but everything else I said generalizes easily to higher n using "n-legged spans" of groupoids: an obvious generalization of the 2-legged spans I've been discussing so far. In Boolean algebra people often use the term "atom" to stand for an element that can't be written as "P or Q" unless exactly one of P or Q is tautologously false.

Although I am a typical loner in daily life, my consciousness of belonging to the invisible community of those who strive for truth, beauty and justice has preserved me from feeling isolated.

— *Albert Einstein*