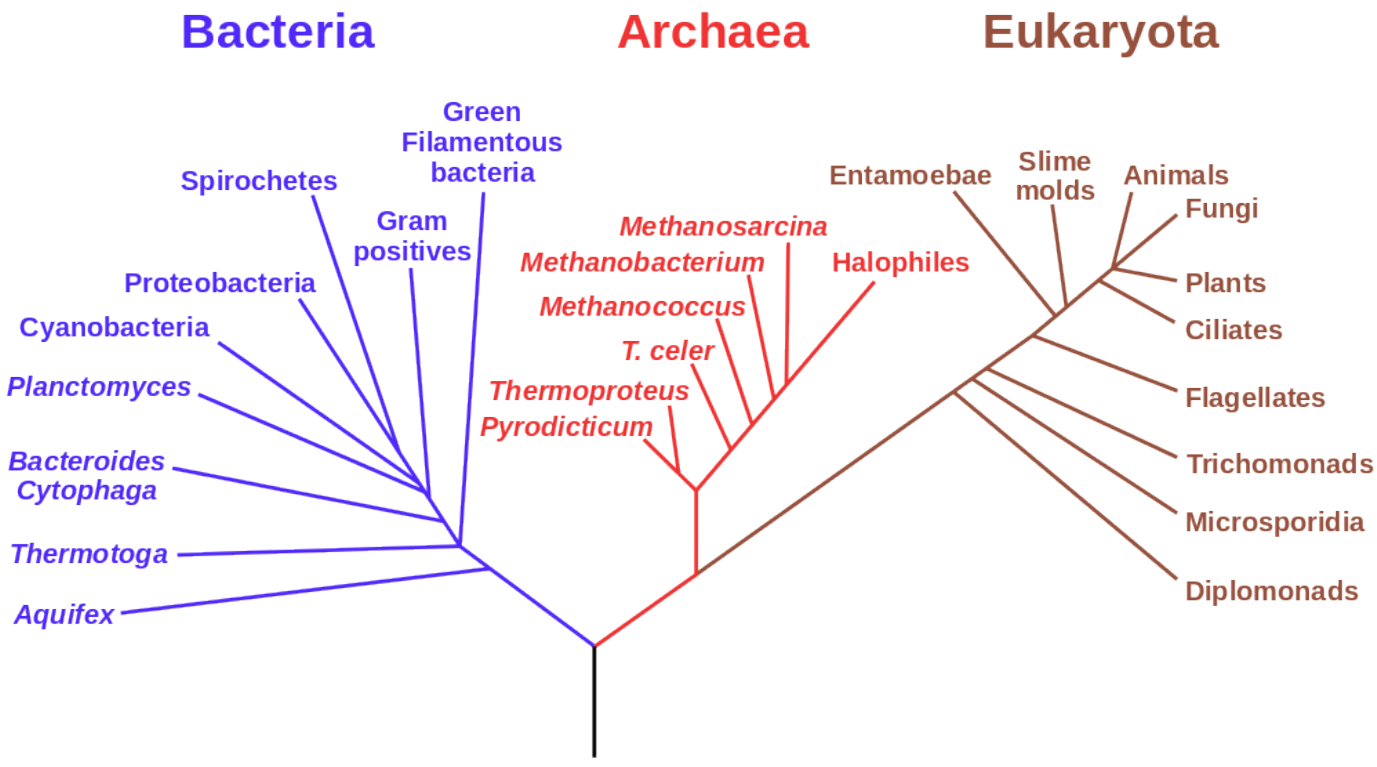
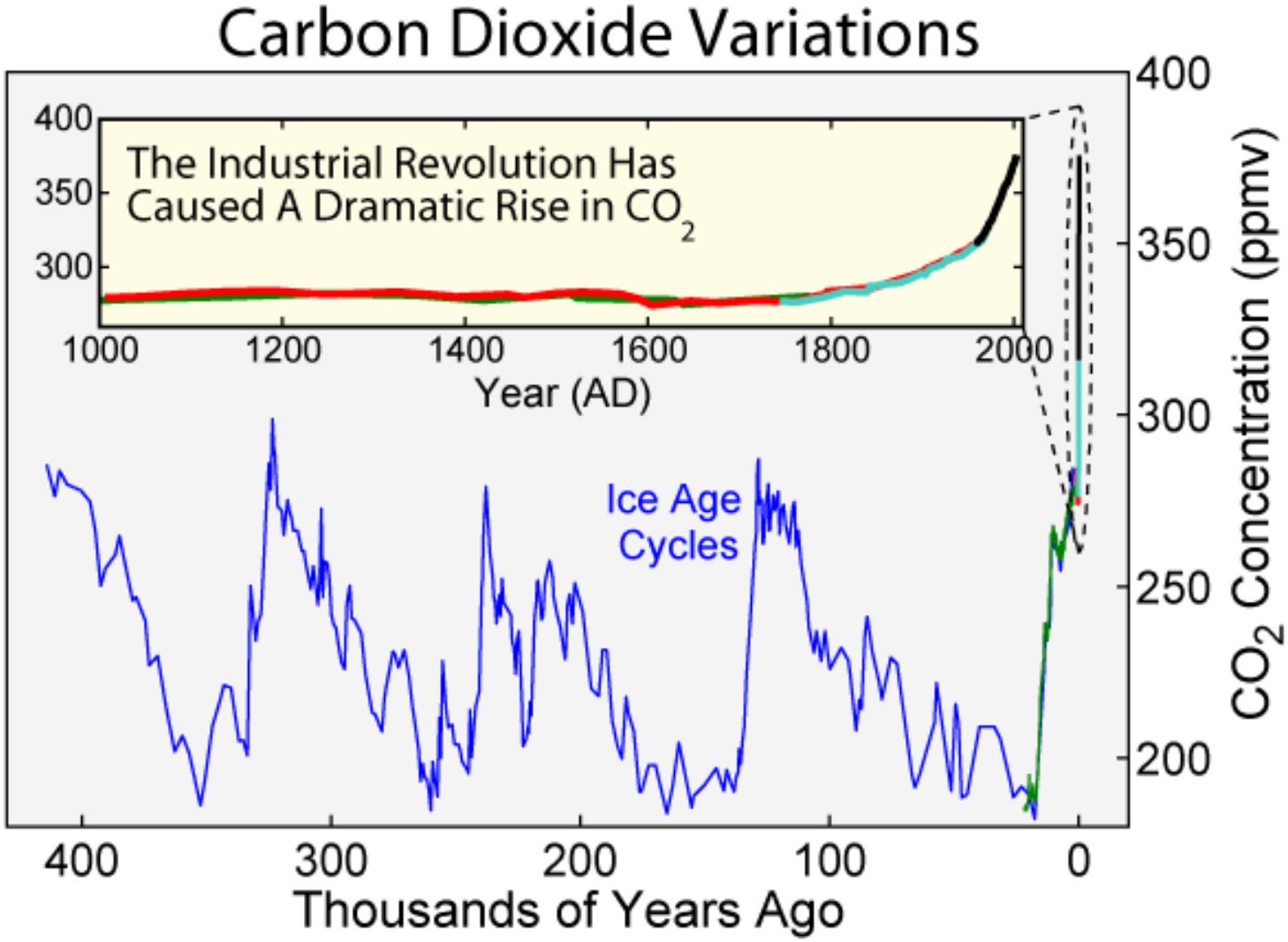
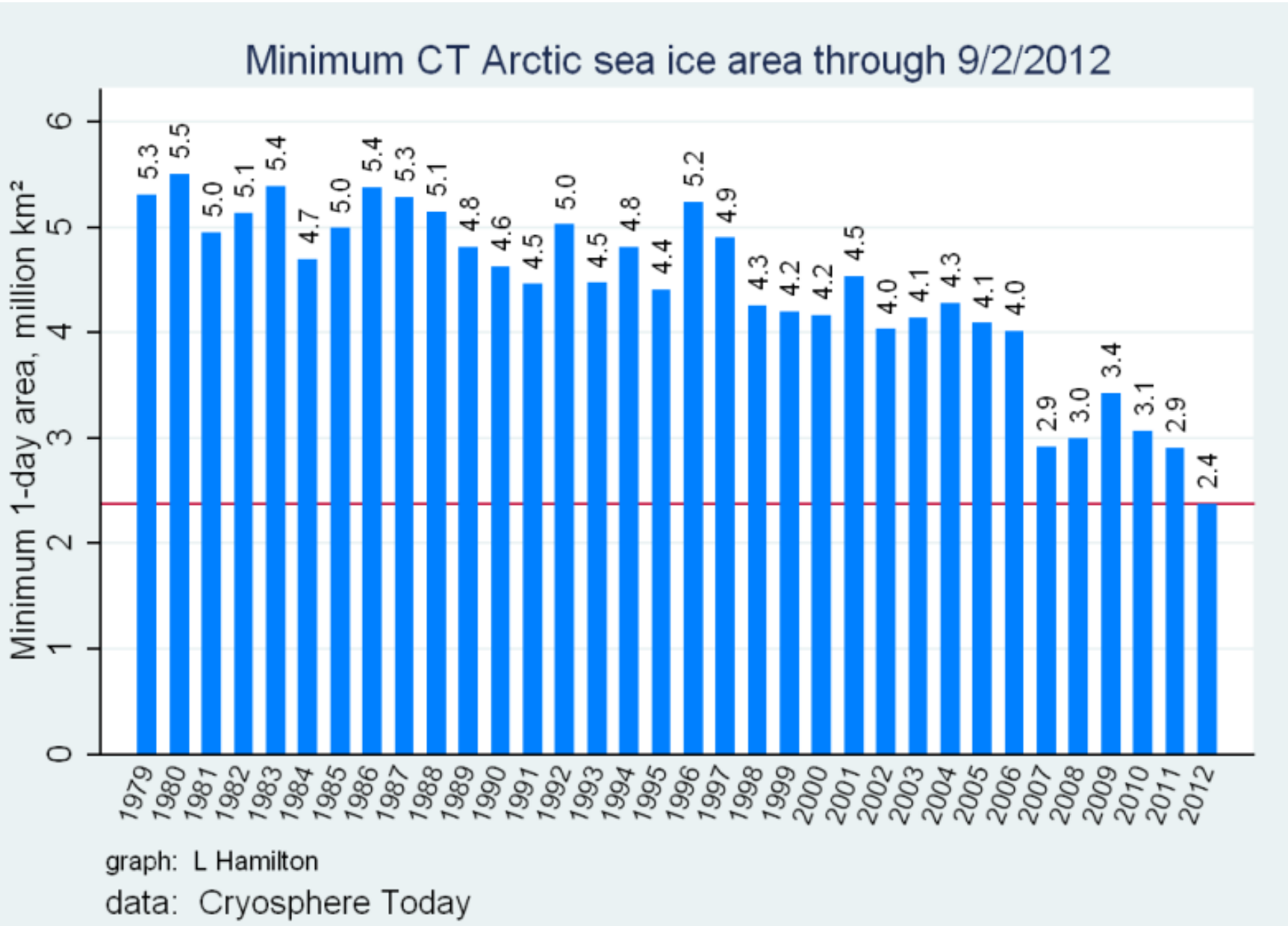
Operads and the Tree of Life

John Baez and Nina Otter

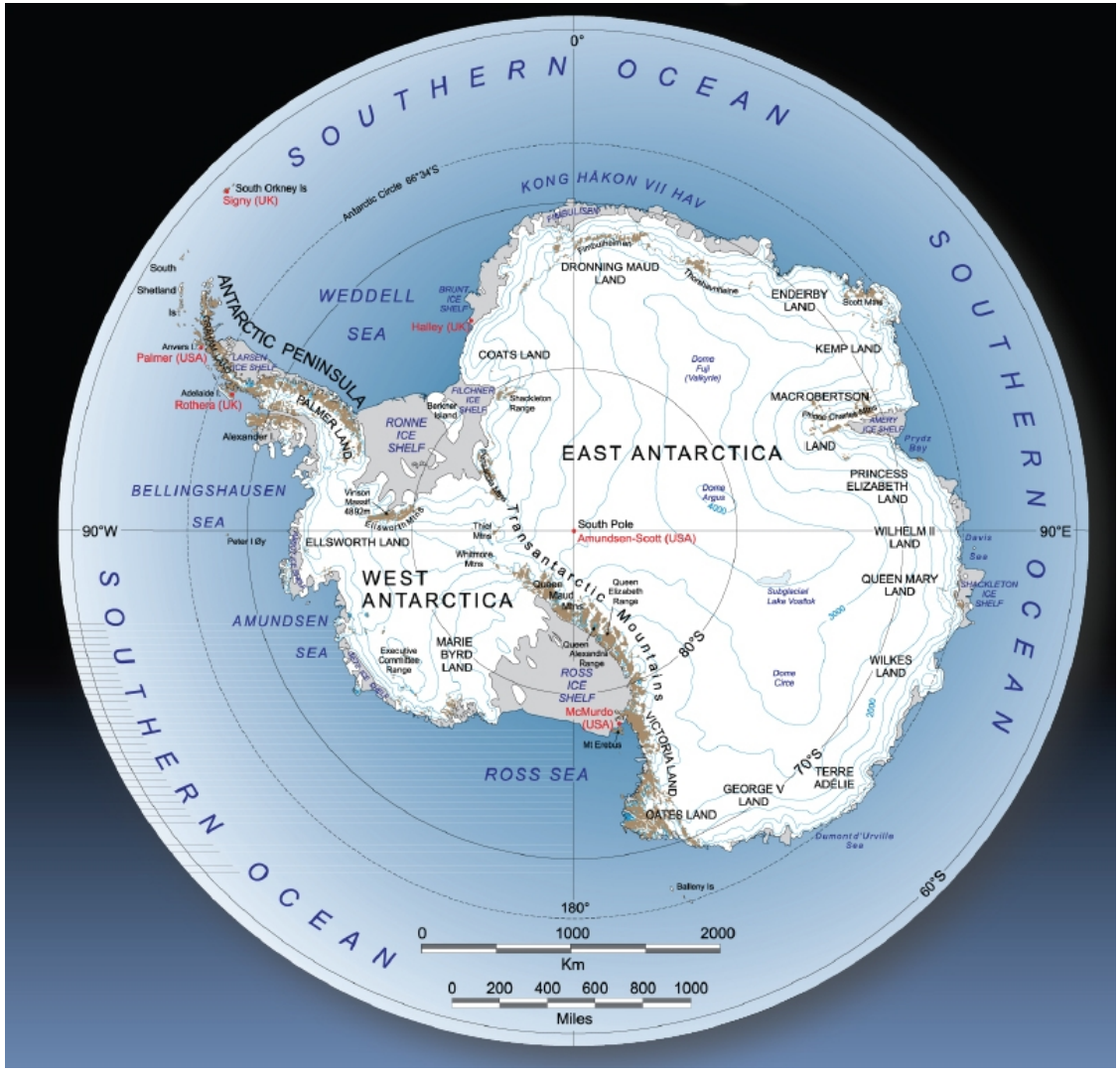


We have entered a new geological epoch, the Anthropocene, in which the biosphere is rapidly changing due to human activities.




Minimum CT Arctic sea ice area through 9/2/2012
Minimum 1-day area, million km²
0
1
2
3
4
5
6
5.3
5.5
5.0
5.1
5.4
4.7
5.0
5.4
5.3
5.1
4.8
4.6
4.5
5.0
4.5
4.8
4.4
5.2
4.9
4.3
4.2
4.2
4.5
4.0
4.1
4.3
4.1
4.0
2.9
3.0
3.4
3.1
2.9
2.4
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
graph: L Hamilton
data: Cryosphere Today

Last week two teams of scientists claimed the Western Antarctic Ice Sheet has been irreversibly destabilized, and will melt causing $\sim$ 3 meters of sea level rise in the centuries to come.

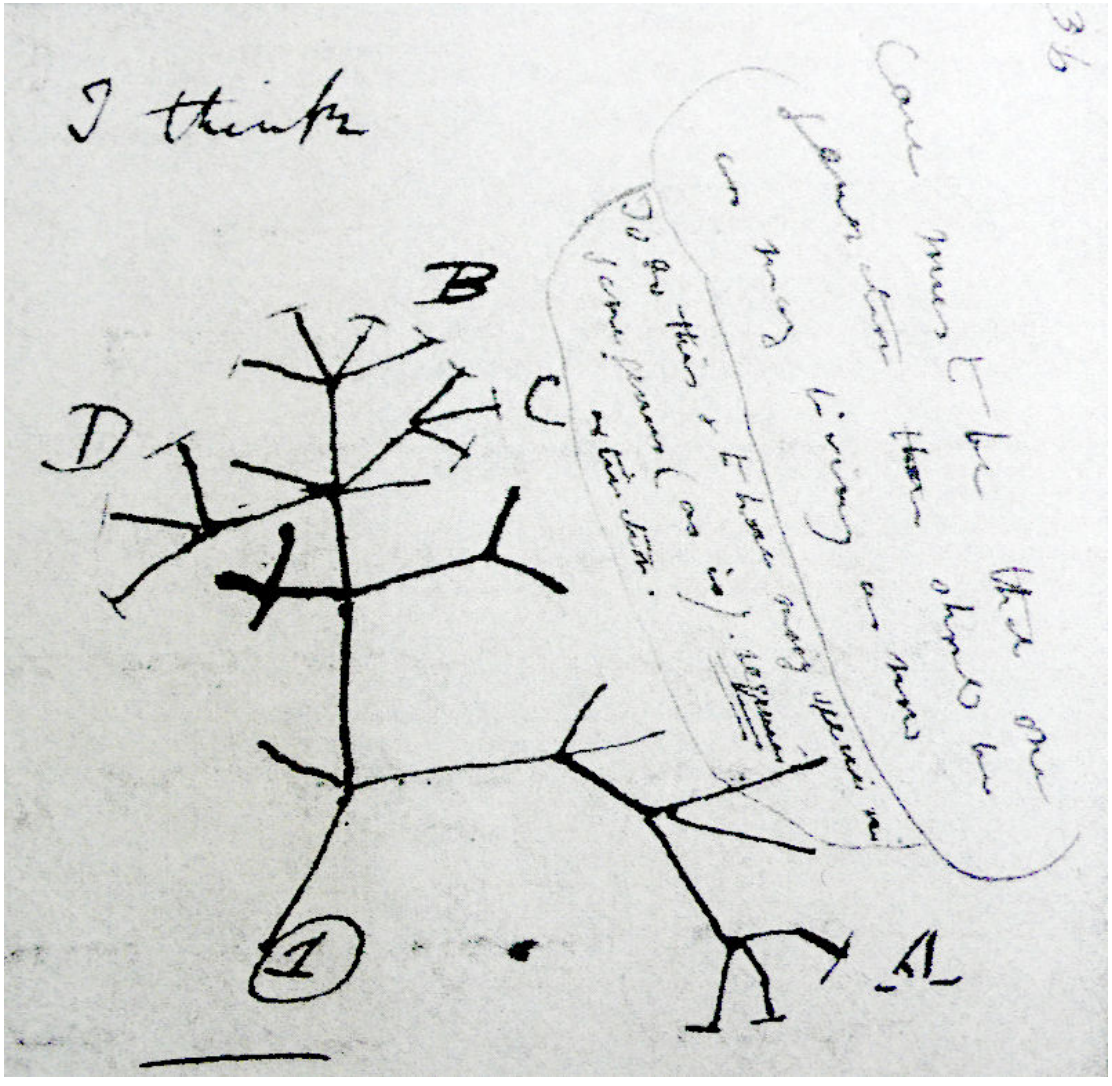


So, we can expect that mathematicians will be increasingly focused on *biology*, *ecology* and *complex systems* — just as last century's mathematics was dominated by fundamental physics.

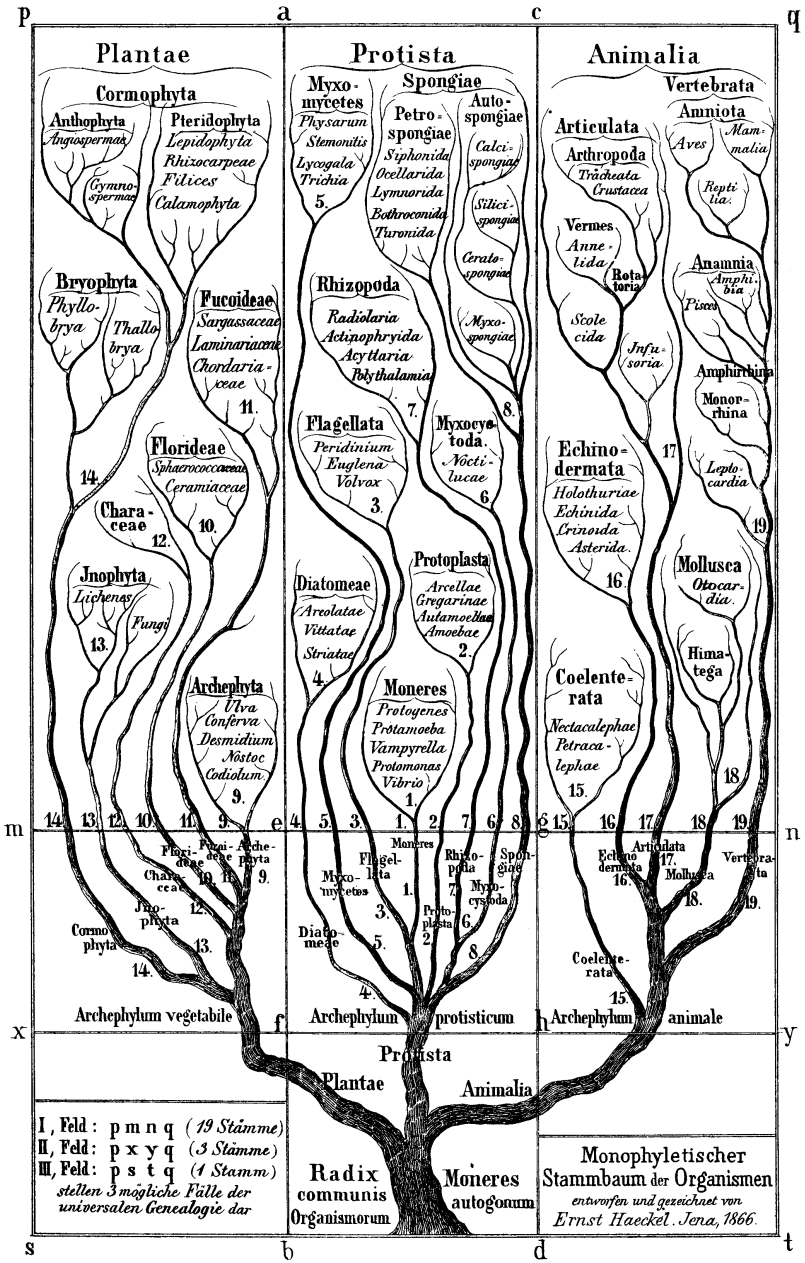
Luckily, these new topics are full of fascinating mathematical structures—and while mathematics takes time to have an effect, it can do truly amazing things. Think of Church and Turing's work on computability, and computers today!

Trees are important, not only in mathematics, but also biology.

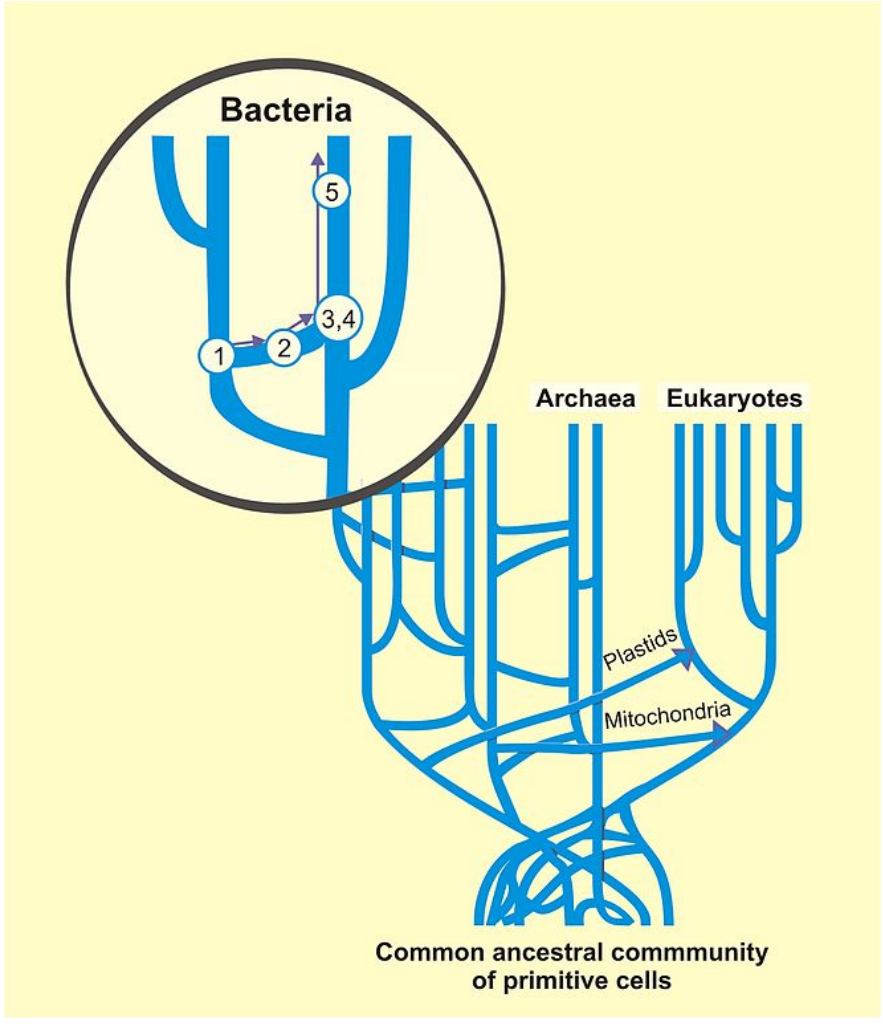
The most important is the ‘tree of life’. Darwin thought about it:



In the 1860s, the German naturalist Haeckel drew it:



Now we know that the ‘tree of life’ is not really a tree, due to endosymbiosis and horizontal gene transfer:



But a tree is often a good approximation. Biologists who try to infer phylogenetic trees from present-day genetic data often use simple models where:

- the genotype of each species follows a random walk, but
- species branch in two at various times.

These are called Markov models.

The simplest Markov model for DNA evolution is the Jukes–Cantor model. Consider a genome of fixed length: that is, one or more pieces of DNA having a total of N base pairs, each taken from the set $\{\mathrm{A},\mathrm{T},\mathrm{C},\mathrm{G}\}$:

$$\cdots \mathbf{ATCGATTGAGCTCTAGCG} \cdots$$

As time passes, the Jukes–Cantor model says the genome changes randomly, with each base pair having the same constant rate of randomly flipping to any other.

So, we get a 'Markov process' on the set of genomes,

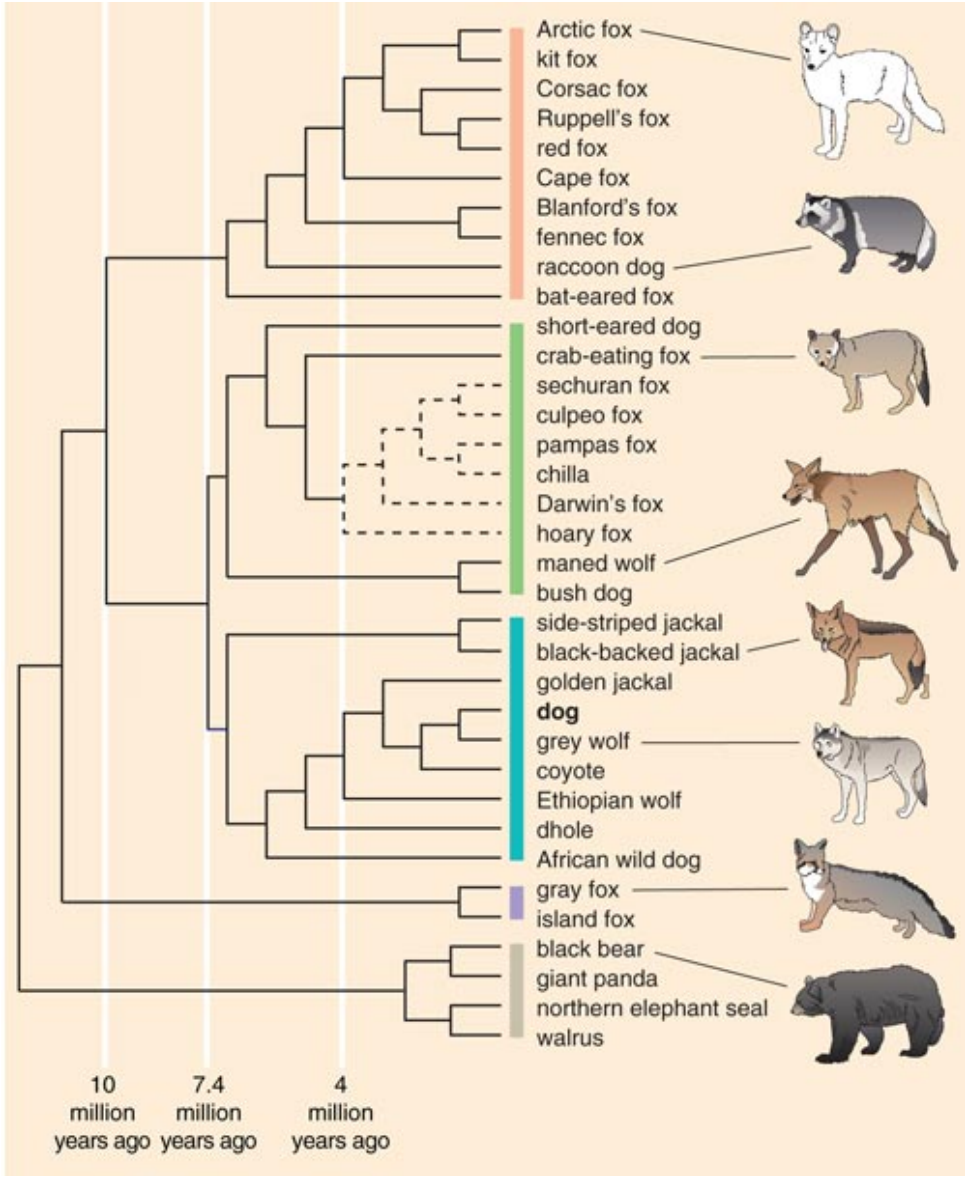
$$X = \{\mathrm{A},\mathrm{T},\mathrm{C},\mathrm{G}\}^N$$

I'll explain Markov processes later!

However, a species can also split in two!

So, given current-day genome data from various species, biologists try to infer the most probable tree where, starting from a common ancestor, the genome undergoes a random walk most of the time but branches in two at certain times.

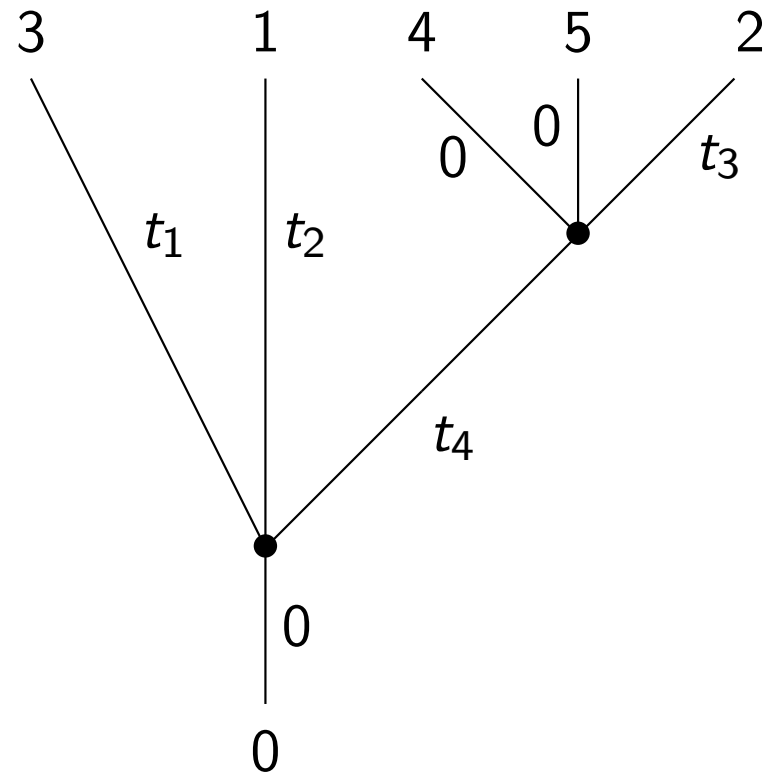
For example, here is Elaine Ostrander's reconstruction of the tree for canids:



Define a **phylogenetic tree** to be a rooted tree with leaves labelled by numbers $1, 2, \ldots, n$ and edges labelled by **times** or **lengths** in $[0, \infty)$. We require that:

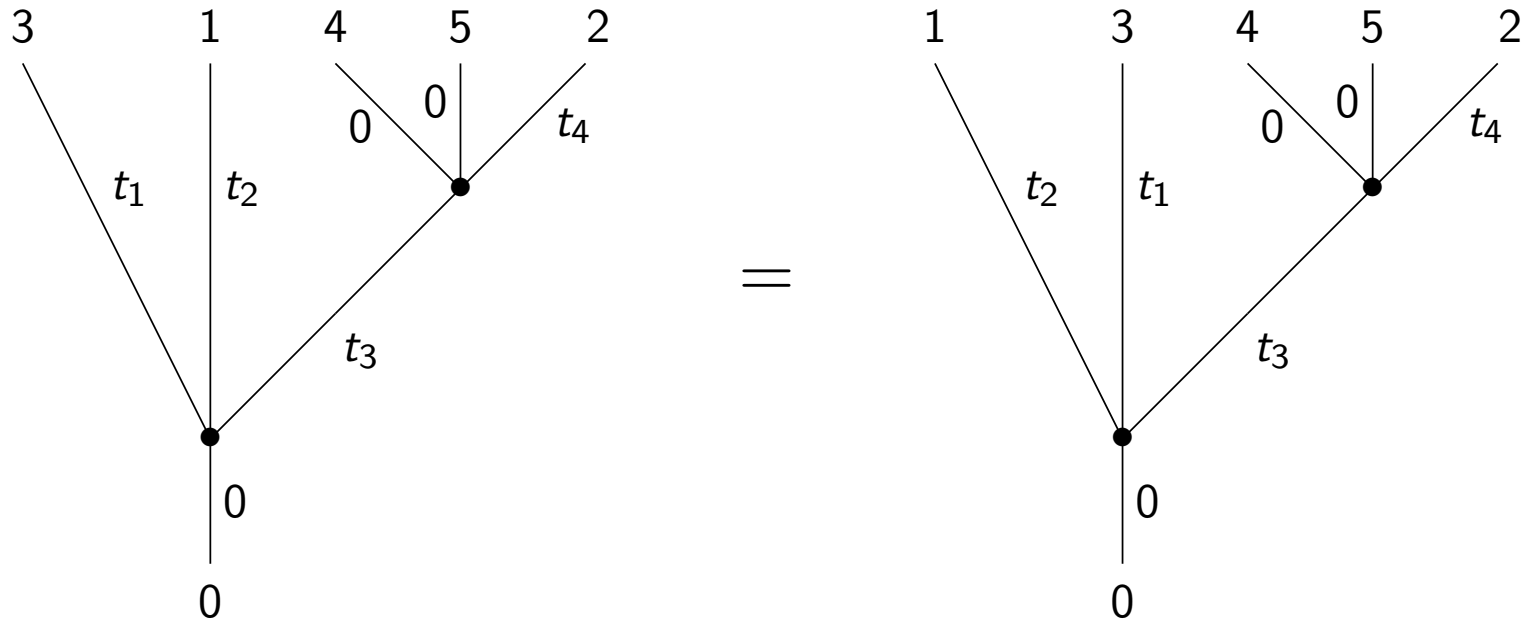
- the length of every edge is positive, except perhaps for edges incident to a leaf or the root;
- a vertex that is an only child cannot have only one child.

For example, here is a phylogenetic tree with 5 leaves:



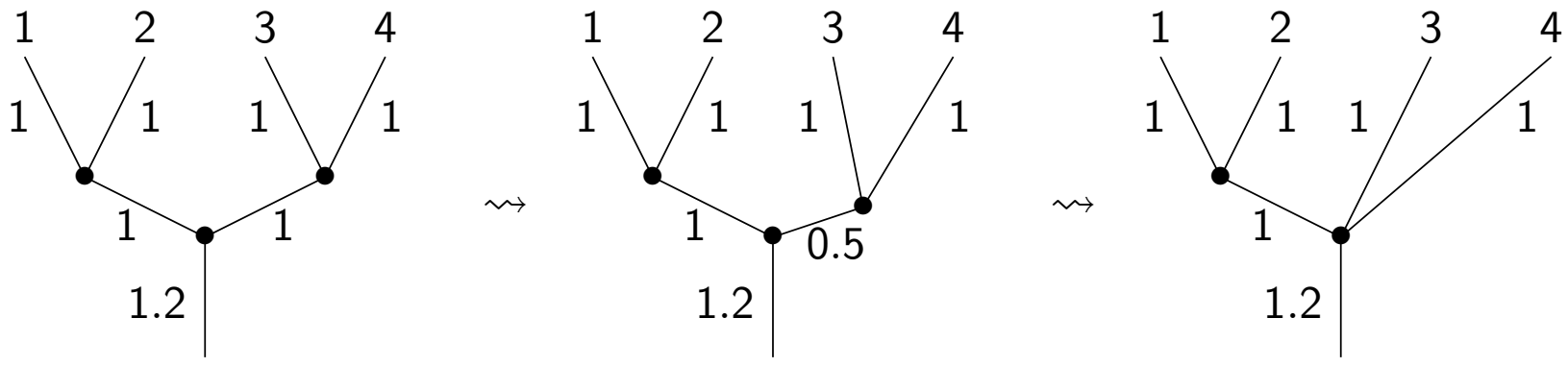
where $t_1, t_2, t_3 \geq 0$ and $t_4 > 0$ The root is labelled 0, the leaves $1, 2, 3, 4, 5$.

The embedding of the tree in the plane is irrelevant, so these are the same phylogenetic tree:



Let $\mathbf{Phyl}_n$ be the set of phylogenetic trees with n leaves. This has an 'obvious' topology.

Here is a continuous path in $\mathbf{Phyl}_4$:



Phylogenetic trees reconstructed by biologists are typically binary. There's a heated debate about trees of higher arity: *can a species split into 3 or more species simultaneously?*

Generically, no. Binary trees form an open dense set of $\mathbf{Phyl}_n$, except for $\mathbf{Phyl}_1$. But trees of higher arity matter when we consider paths, paths of paths,... etc. in $\mathbf{Phyl}_n$.

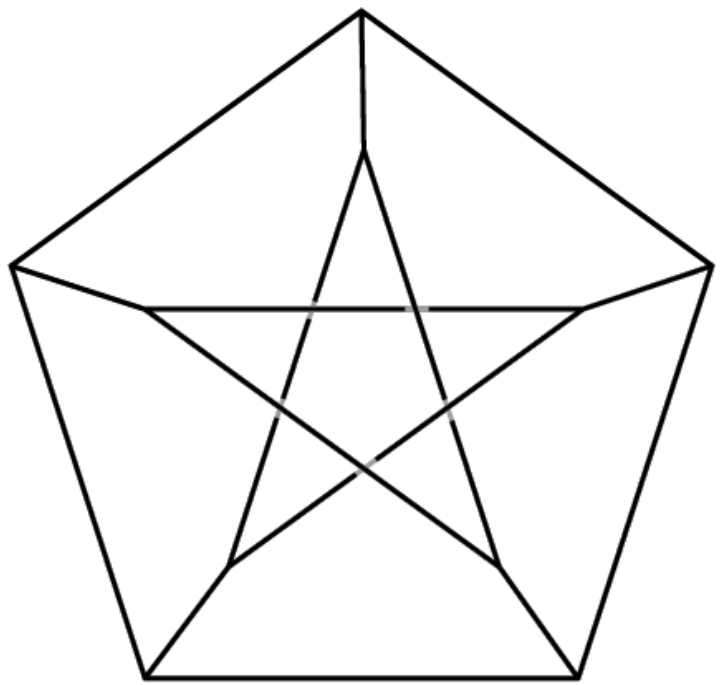
This was emphasized here:

- Louis Billera, Susan Holmes and Karen Vogtmann, Geometry of the space of phylogenetic trees, *Advances in Applied Mathematics* **27** (2001), 733–767.

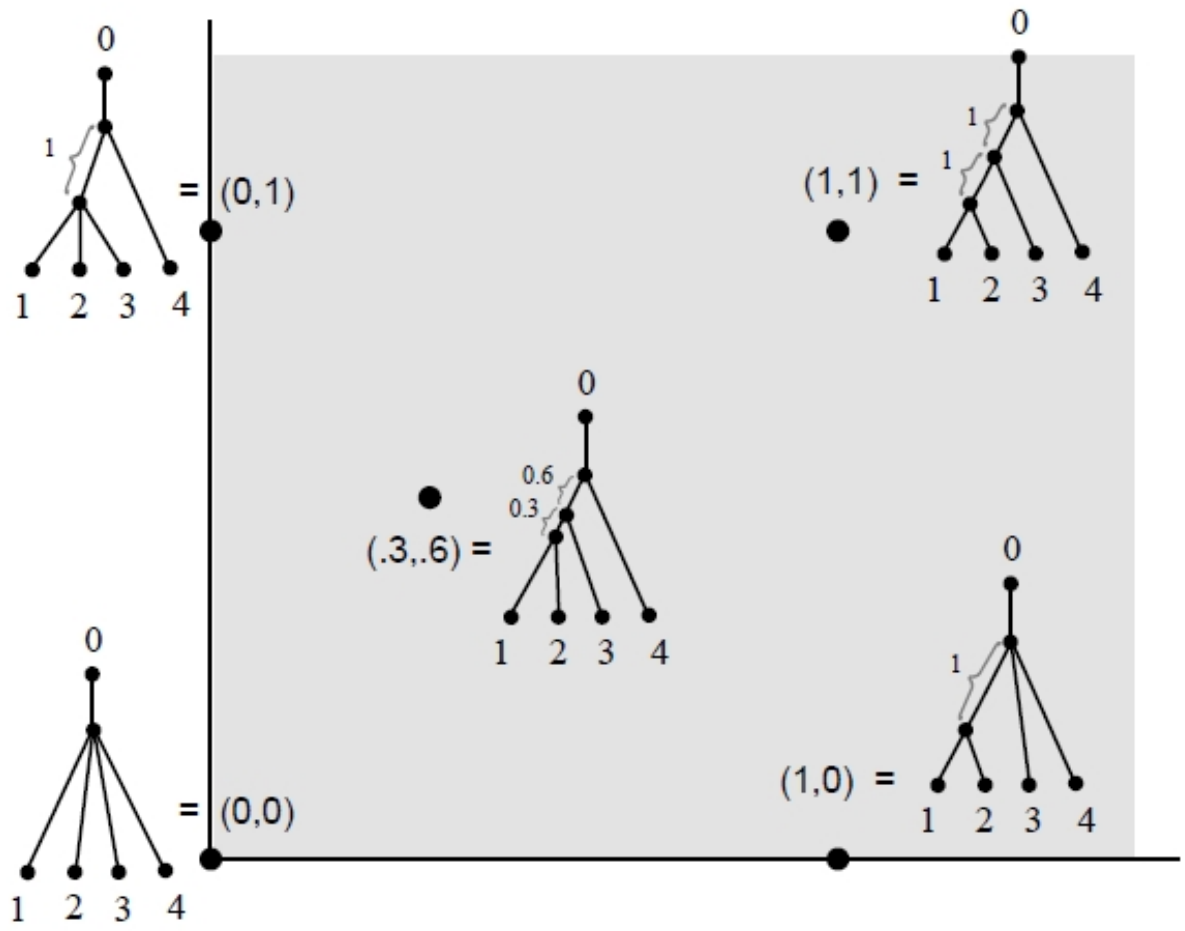
Billera, Holmes and Vogtmann study the set $\mathcal{T}_n$ of phylogenetic trees with n leaves where the lengths of **external edges** — edges incident to the root and leaves — are fixed to a constant value. The reason:

$$\mathbf{Phyl}_n \cong \mathcal{T}_n \times [0, \infty)^{n+1}$$

For example, they note $\mathcal{T}_4$ is the cone on the Petersen graph:

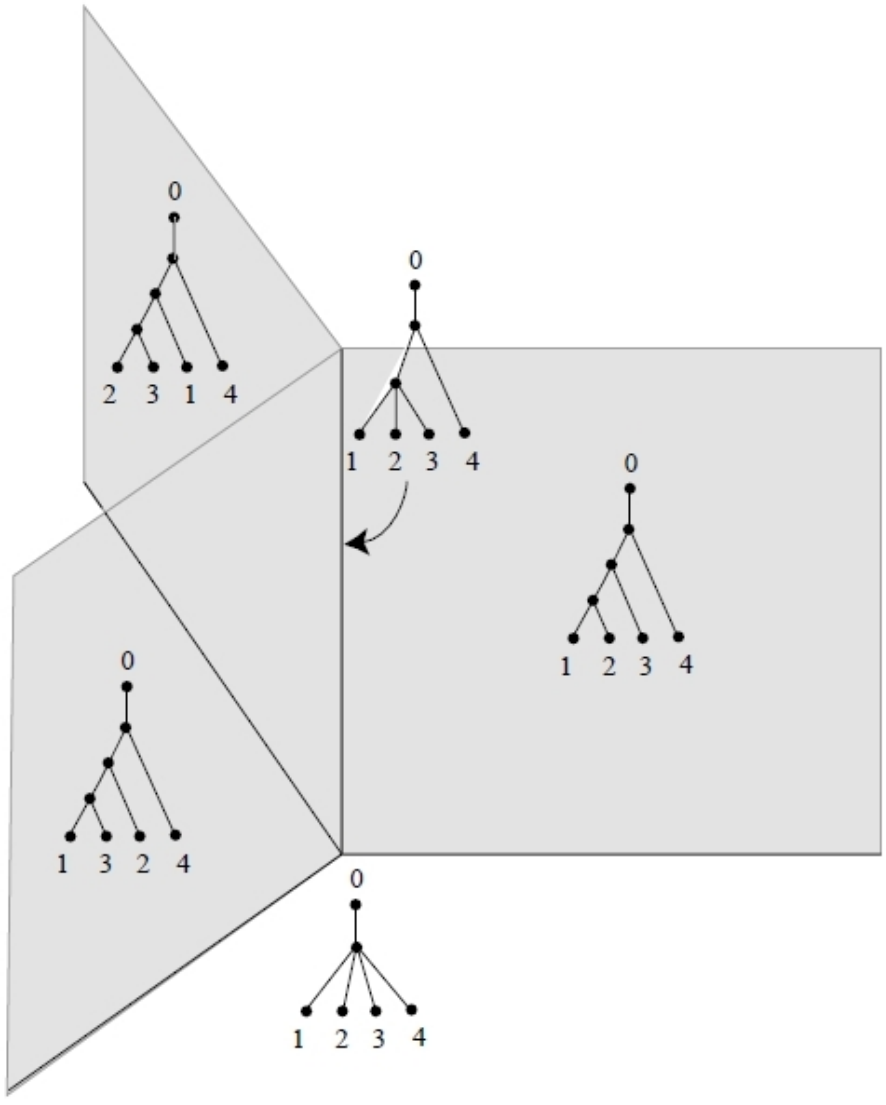


The cone on any edge of the Petersen graph is a quadrant:

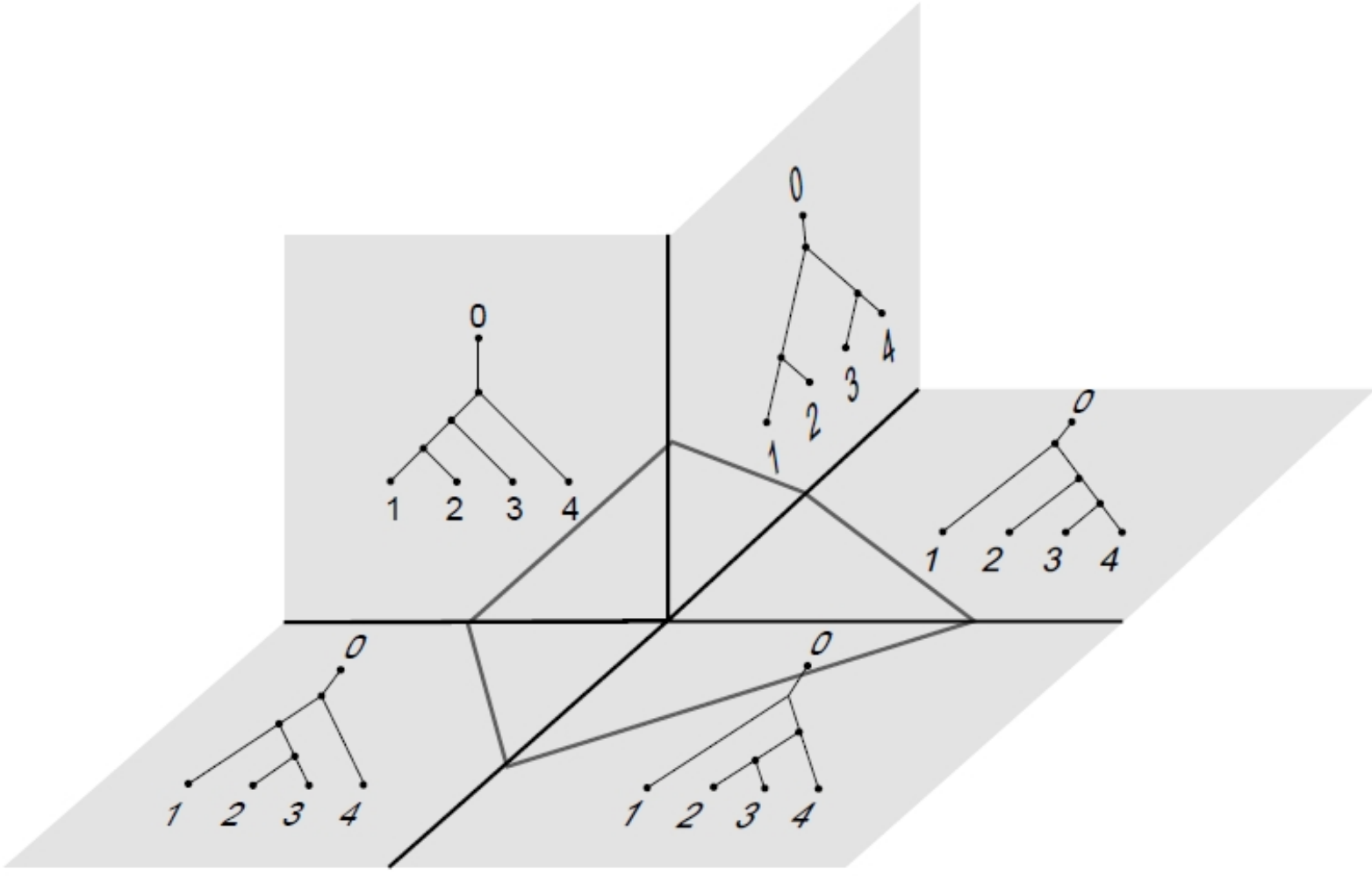


(All these drawings with upside-down trees are theirs!)

The cone on any vertex of the Petersen graph is a ray, at which 3 quadrants meet:



The cone on any pentagon in the Petersen graph looks like this:



This is a copy of the famous Stasheff pentagon!

This should remind us of the connection between trees and operads!

Today my operads will always be 'permutative', meaning we have an action of S_n on $\mathbf{O}_n$, compatible with composition.

They will also be 'topological', meaning that $\mathbf{O}_n$ is a topological space, and composition and permutations act as continuous maps.

Proposition. There is an operad **Phyl**, the **phylogenetic operad**, whose space of n-ary operations is $\mathbf{Phyl}_n$. Composition and permutations are defined in the visually evident way.

So:

- What is the mathematical nature of this operad?
- How is it related to 'Markov processes with branching'?
- How is it related to known operads in topology?

Answer: **Phyl** is the coproduct of **Com**, the operad for commutative semigroups, and $\mathbf{[0, \infty)}$, the operad having only unary operations, one for each $t \in [0, \infty)$. The first describes branching, the second describes Markov processes. **Phyl** is closely related to the Boardman–Vogt **W** construction applied to **Com**. Let's see how this works...

The point of operads is that they have 'algebras'. An **algebra** of $\mathbf{O}$ is a topological space X on which each operation $f \in \mathbf{O}_n$ acts as a map

$$\alpha(f) \colon X^n \to X$$

obeying some plausible conditions.

These conditions simply say there's an operad homomorphism $\alpha \colon \mathbf{O} \to \mathbf{End}(X)$, where $\mathbf{End}(X)$ is the operad whose n-ary operations are maps $X^n \to X$.

More generally we can consider an algebra of $\mathbf{O}$ in any symmetric monoidal category C enriched over Top. This is an object $X \in C$ with an operad homomorphism $\alpha \colon \mathbf{O} \to \mathbf{End}(X)$.

A **coalgebra** of $\mathbf{O}$ in C is an algebra of $\mathbf{O}$ in C^{op}. We'll see that the phylogenetic operad has interesting coalgebras in $\mathrm{FinStoch}$, the category of finite sets and 'stochastic maps'.

I'll use $\mathbf{[0,\infty)}$ as the name for the operad having only unary operations, one for each $t \in [0,\infty)$, with composition of operations given by addition.

The category $\mathrm{FinStoch}$ has finite sets as objects, and a morphism $f : X \to Y$ is a map sending each point in X to a probability distribution on Y.

Alternatively, a morphism $f : X \to Y$ is a $Y \times X$-shaped matrix of real numbers where:

- the entries are nonnegative,
- each column sums to 1.

Such a matrix is called **stochastic**. Composition of morphisms is matrix multiplication.

$\mathrm{FinStoch}$ becomes a symmetric monoidal category enriched over Top, where the tensor product of X and Y is $X \times Y$.

A **Markov process** is an algebra of $\mathbf{[0, \infty)}$ in $\mathrm{FinStoch}$.

Concretely, a **Markov process** is a finite set X together with a stochastic map $\alpha(t) \colon X \to X$ for each $t \geq 0$, such that:

- $\alpha(s + t) = \alpha(s)\alpha(t)$,
- $\alpha(0) = 1$
- $\alpha(t)$ depends continuously on t.

Since $[0, \infty)$ is commutative, a coalgebra of $\mathbf{[0, \infty)}$ in $\mathrm{FinVect}$ is the same thing!

If X is a set of possible genomes, a Markov process on X describes the random changes of the genome with the passage of time.

There's a unique operad **Com** with one n-ary operation for each $n > 0$, and none for $n = 0$.

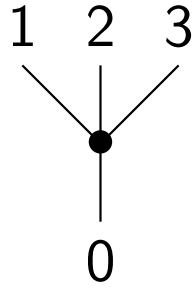
Algebras of **Com** in Top are commutative topological semigroups: there is just one way to multiply n elements for $n > 0$.

Any finite set X becomes a cocommutative coalgebra in $\mathrm{FinStoch}$. The unique n-ary operation of **Com** acts as the diagonal

$$\Delta_n \colon \mathbb{R}^X \to \mathbb{R}^X \times \cdots \times \mathbb{R}^X$$

This is a map, a special case of a stochastic map.

This map describes the 'n-fold duplication' of a probability distribution f on the set X of genomes when a species branches!



Any pair of operads $\mathbf{O}$ and $\mathbf{O}'$ has a coproduct $\mathbf{O} + \mathbf{O}'$.

By general abstract nonsense, an algebra of $\mathbf{O} + \mathbf{O}'$ is an object X that is both an algebra of $\mathbf{O}$ and an algebra of $\mathbf{O}'$, with no compatibility conditions imposed.

In fact:

Theorem. The operad **Phyl** is the coproduct $\mathbf{Com} + \mathbf{[0, \infty)}$.

And thus:

Corollary. Given any Markov process, its underlying finite set X naturally becomes a coalgebra of **Phyl** in $\mathrm{FinStoch}$.

Proof. X is automatically a coalgebra of **Com**, and the Markov process makes it into a coalgebra of $\mathbf{[0,\infty)}$. Thus, it becomes a coalgebra of $\mathbf{Phyl} \cong \mathbf{Com} + \mathbf{[0,\infty)}$.

How is **Phyl** related to $\mathbf{W}(\mathbf{Com})$, where $\mathbf{W}$ is the construction that Boardman and Vogt used to get an operad for loop spaces?

Define addition on $[0, \infty]$ in the obvious way, where

$$\infty + t = t + \infty = \infty$$

Then $[0, \infty]$ becomes a topological monoid, so there's an operad with only unary operations, one for each $t \in [0, \infty]$.

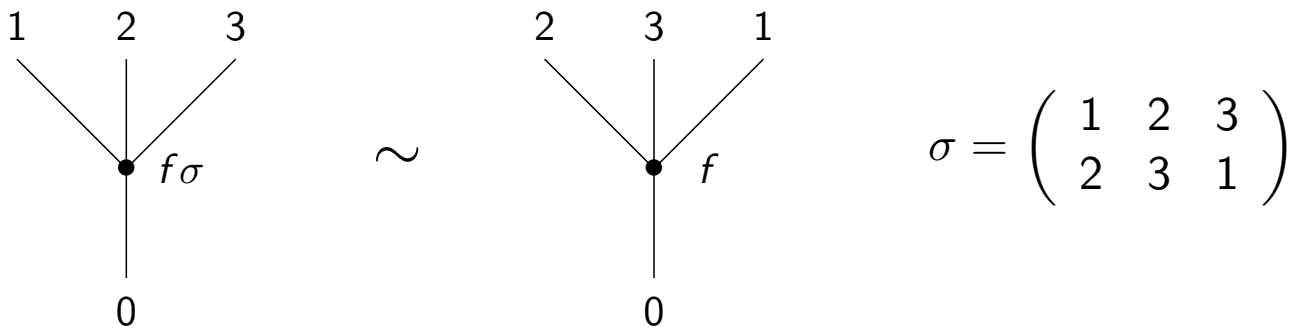
Let's call this operad $\mathbf{[0, \infty]}$.

Theorem. For any operad $\mathbf{O}$, Boardman and Vogt's operad $\mathbf{W}(\mathbf{O})$ is isomorphic to a suboperad of $\mathbf{O} + \mathbf{[0, \infty]}$.

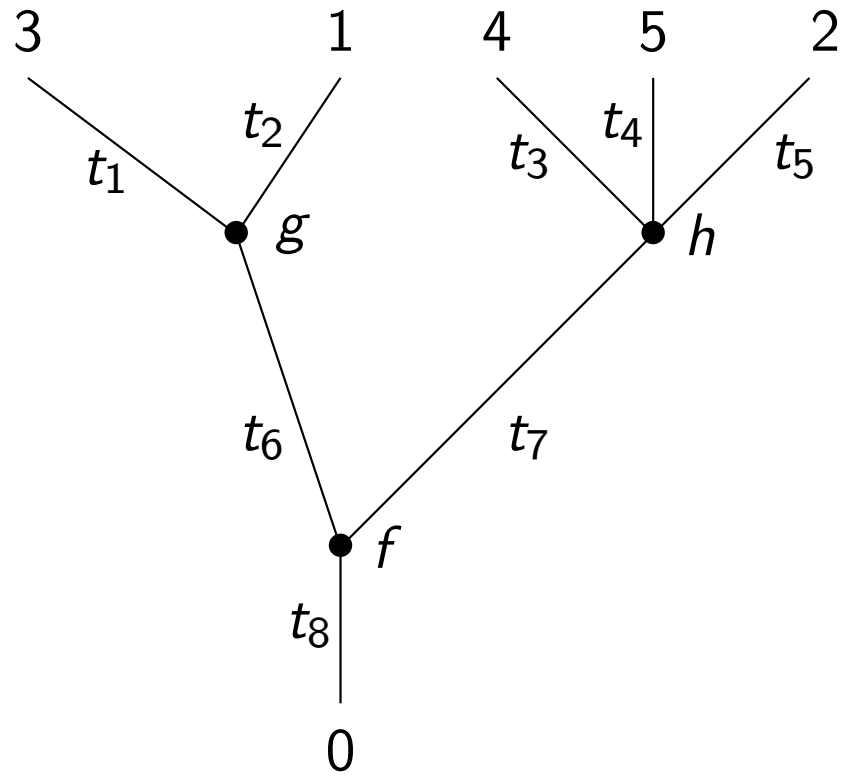
Operations in $\mathbf{O} + \mathbf{[0, \infty]}$ are equivalence classes of planar rooted trees with:

- vertices except for leaves and the root labelled by operations in $\mathbf{O}$,
- edges labelled by lengths in $[0, \infty]$, such that
- only external edges can have length 0.

The equivalence relation comes from permuting edges coming into a vertex, e.g.:



Here is an operation in $\mathbf{O} + \mathbf{[0, \infty]}$:

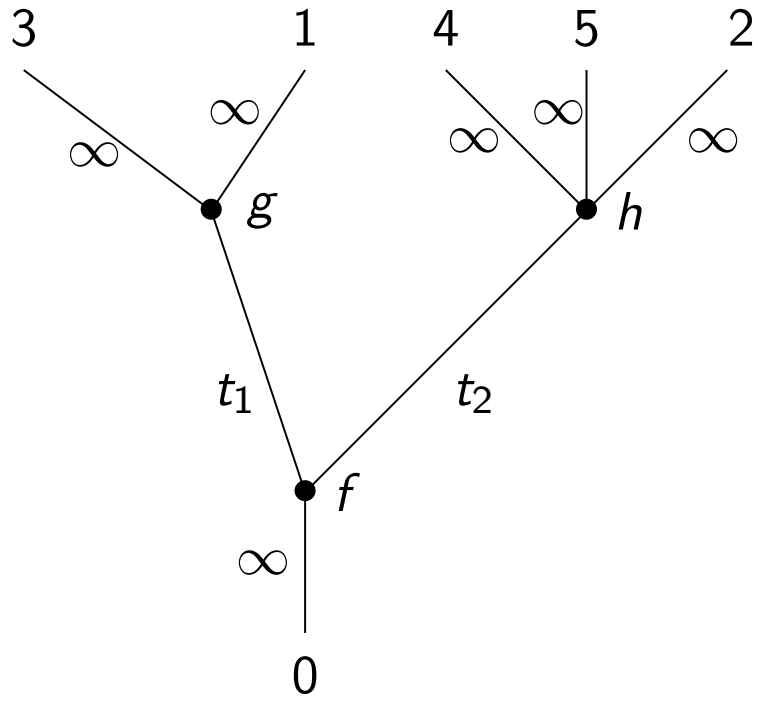


where $t_1, t_2, t_3, t_4, t_5, t_8 \geq 0$ and $t_6, t_7 > 0$.

An operation is in $\mathbf{W}(\mathbf{O})$ iff either

- it is the identity operation or
- all external edges have length ∞.

Here is an operation in $\mathbf{W}(\mathbf{O})$:



where $t_1, t_2 > 0$.

Berger and Moerdijk showed: if S_n acts freely on $\mathbf{O}_n$ and $\mathbf{O}_1$ is well-pointed, $\mathbf{W}(\mathbf{O})$ is a cofibrant replacement for $\mathbf{O}$.

This is true for $\mathbf{O} = \mathbf{Assoc}$, the operad whose algebras are topological semigroups, with $n!$ operations of arity $n > 0$. This is why Boardman and Vogt could use $\mathbf{W}(\mathbf{Assoc})$ as an operad for loop spaces.

But S_n does *not* act freely on $\mathbf{Com}_n$. $\mathbf{W}(\mathbf{Com})$ is *not* a cofibrant replacement for $\mathbf{Com}$. It is *not* an operad for infinite loop spaces.

Nonetheless $\mathbf{W}(\mathbf{Com})$ is interesting because

$$\mathbf{W}(\mathbf{Com})_n \cong \mathcal{T}_n$$

where $\mathcal{T}_n$ is Billera, Holmes and Vogtmann's space of phylogenetic trees with external edges having fixed lengths.

And the larger operad $\mathbf{Com} + \mathbf{[0, \infty]}$, a compactification of $\mathbf{Phyl} \cong \mathbf{Com} + \mathbf{[0, \infty)}$, is also interesting.

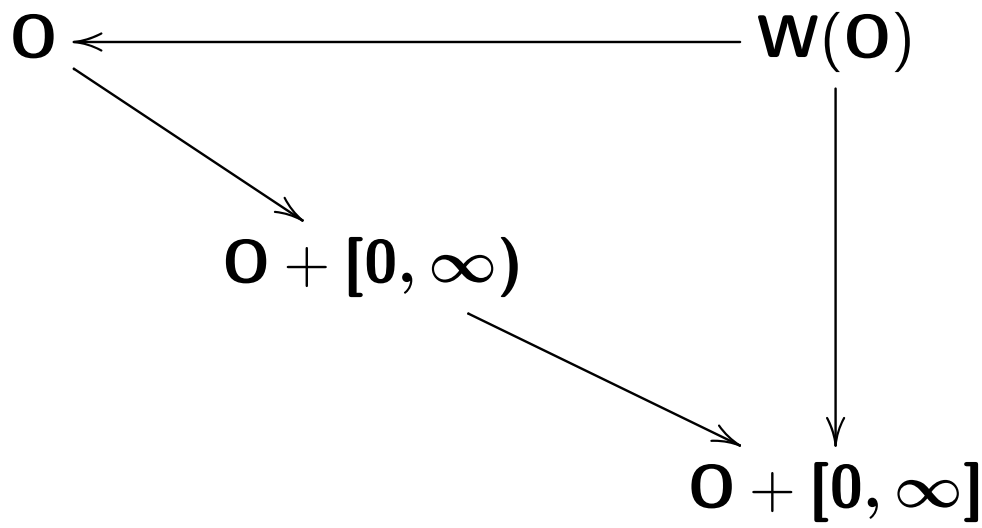
The reason is that any Markov process $\alpha \colon [0, \infty) \to \mathrm{End}(X)$ approaches a limit as $t \to \infty$. Indeed, it extends uniquely to a homomorphism of topological monoids $\alpha \colon [0, \infty] \to \mathrm{End}(X)$.

We thus get:

Proposition. Given any Markov process, its underlying finite set X naturally becomes a coalgebra of $\mathbf{Com} + \mathbf{[0, \infty]}$ in $\mathrm{FinStoch}$.

Summary for topologists who don't care about applications:

For any operad $\mathbf{O}$ we have weak equivalences



and if $\mathbf{O}$ is well-pointed and S_n acts freely on $\mathbf{O}_n$, $\mathbf{W}(\mathbf{O})$ is cofibrant.

Finally: *the mystery of tropical trees!*

The **tropical rig** is $(-\infty, \infty]$ with minimization as $+$ and addition as $\times$. We can do algebraic geometry over this rig and define 'tropical curves'.

In 2007, Gathmann, Kerber and Markwig showed that a certain moduli space of genus 0 tropical curves with $n+1$ marked points is the space of trees $\mathcal{T}_n$ studied by Billera, Holmes and Vogtmann.

Also in 2007, Mikhalkin showed this moduli space has a compactification that is a smooth compact tropical variety.

Nina Otter has shown this compactification is $\mathbf{W}(\mathbf{Com})_n$. The operad structure on $W(\mathbf{Com})$ corresponds to the 'tropical clutching map' described by Abramovich, Caporaso and Payne in 2012.

The mystery: why are tropical curves related to phylogenetic trees? Is this connection good for something?