A Compositional and Statistical Approach to Natural Language

Tai-Danae Bradley
CUNY and Tunnel

Goal:

Understand meaning of words, phrases, sentences, and concepts in natural language.

You shall know a word by the company it keeps.

— John Firth

(the Yoneda lemma for linguistics)

Language is compositional and statistical.

Compositional:
red + firetruck

Statistical:
frequency count (red vs. blue firetruck)

Red contributes to the meaning of *firetruck*.

An inside-out approach:

Define a **monoidal functor** from
a grammar category (pregroups) to a
meaning category (vector spaces).

An outside-in approach:

Let **statistics** serve as a proxy for grammar.

That is, learn "what goes with what" in the language
given some samples of that language.

Goal (rephrased): Infer a probability distribution on a set of text data.

View language as a quantum-many body problem.

I. classical to quantum probability

II. a tensor network language model

Let π be a probability distribution on a finite set S.

$$\pi\colon S \to [0,1]$$
$$\sum_{s} \pi(s) = 1$$

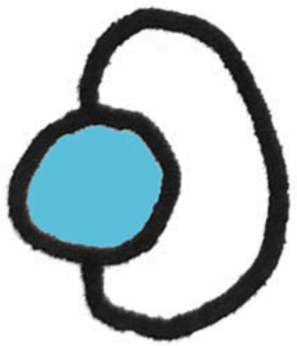
Let π be a probability distribution on a finite set S.

$$\pi\colon S \to [0,1]$$
$$\sum_s \pi(s) = 1$$

Pass from S to the free vector space $\mathbb{C}^S$ by $s \mapsto |s\rangle$.

$$|s_i\rangle = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \qquad \langle s_i | s_j \rangle = \delta_{ij}$$

The quantum version of a probability distribution is a **density operator** ρ, which is a self-adjoint, positive semidefinite operator with trace one. A density operator is also called a **quantum state.**



$$= 1$$

Every density ρ on $\mathbb{C}^S$ defines a probability distribution π_ρ on S by

$$\pi_\rho(s) := \langle s|\rho|s\rangle$$

Every density ρ on $\mathbb{C}^S$ defines a probability distribution π_ρ on S by

$$\pi_\rho(s) := \langle s|\rho|s\rangle$$

Given a distribution π on S, there is more than one way to define a density so that $\pi_\rho = \pi$.

Here are two.

1. A diagonal operator

$$\rho_{\text{diag}} = \begin{bmatrix} \pi(s_1) & 0 & \cdots & 0 \\ 0 & \pi(s_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi(s_n) \end{bmatrix}$$

Note: $\pi(s) = \langle s|\rho_{\text{diag}}|s\rangle = \pi_{\rho_{\text{diag}}}(s)$

2. Projection onto a single unit vector $|\psi\rangle$

$$\rho = |\psi\rangle\langle\psi|$$

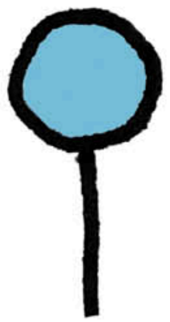
where

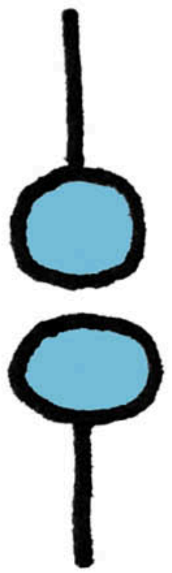
$$|\psi\rangle = \sum_{s \in S} \sqrt{\pi(s)}|s\rangle$$

Again: $\pi(s) = \langle s|\rho|s\rangle = \pi_\rho(s)$.

We **always** use the density $\rho = |\psi\rangle\langle\psi|$ where

$$|\psi\rangle = \sum_{s \in S} \sqrt{\pi(s)}|s\rangle$$

$|\psi\rangle =$ 

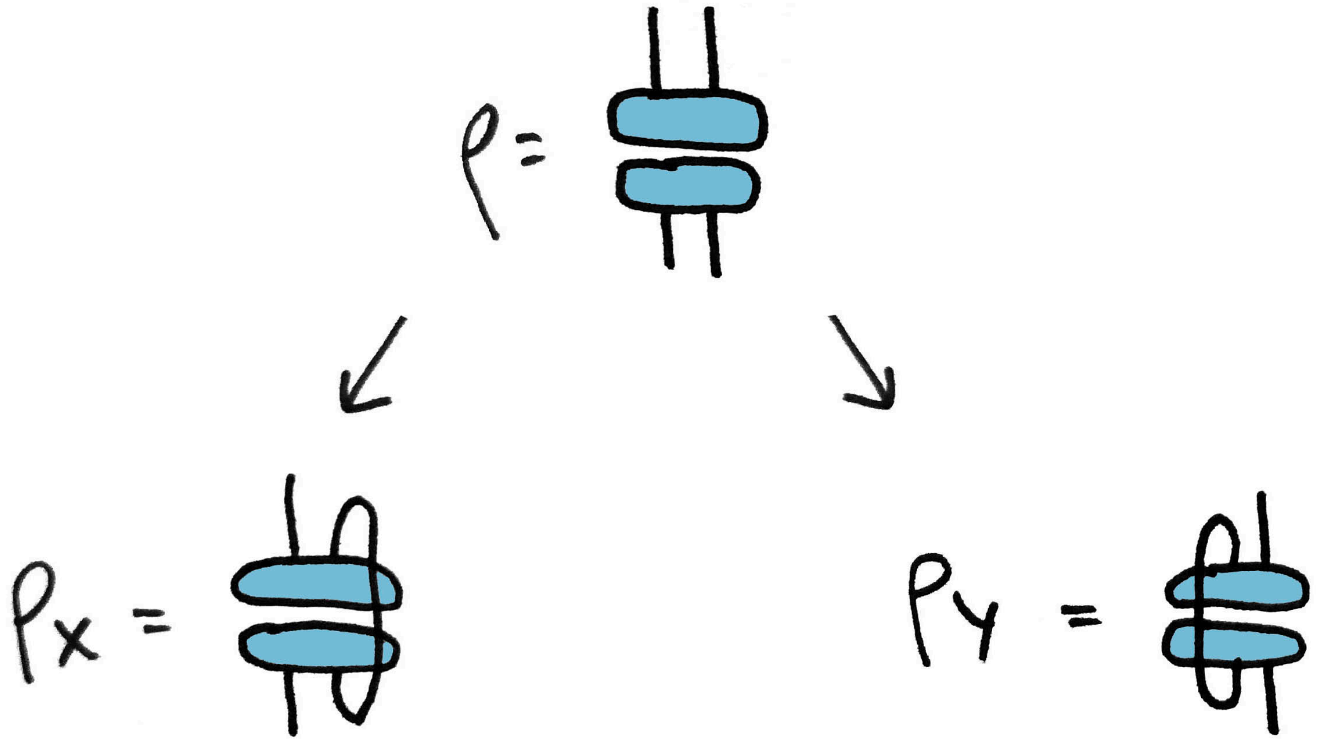
$\rho =$ 

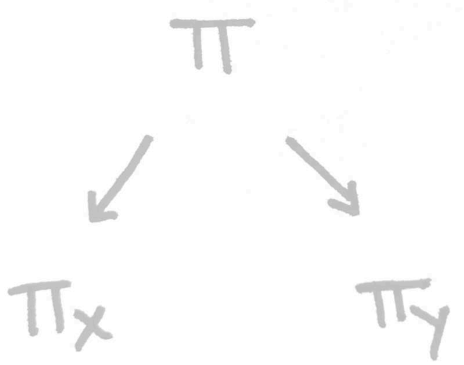
Why bother?

Consider a **joint** distribution $\pi\colon X \times Y \to [0, 1]$. We have **marginal** distributions on X and Y by "integrating out."

- The quantum version of π is a density operator on $\mathbb{C}^X \otimes \mathbb{C}^Y$.
- The quantum version of marginalizing is the **partial trace.**

The partial trace gives rise to **reduced density operators,** which are the quantum analogues of marginal distributions.





Reduced densities contain the marginal distributions **and more.**

$$\rho_X = \begin{bmatrix} \pi_X(x_1) & * & \cdots & * \\ * & \pi_X(x_2) & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & \pi_X(x_m) \end{bmatrix}$$

The ijth entry is proportional to the number of **shared continuations** in Y between x_i and x_j.

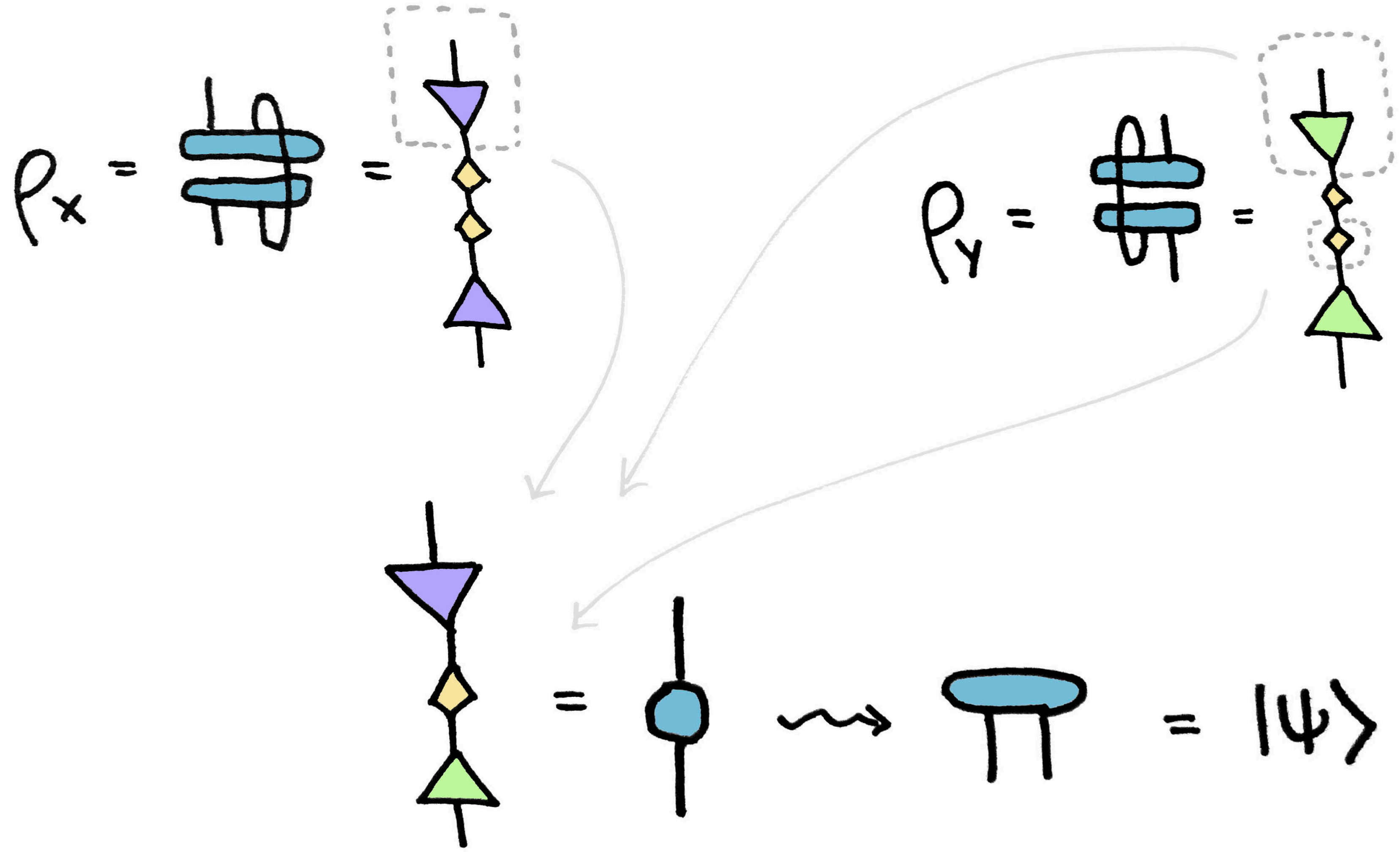
$$(x_i, y) \qquad (x_j, y)$$

This "extra information" captures subsystem interactions.

It is *lost* by classical marginalization!

It is encoded in the **spectral information** of the reduced densities.

The "extra information" contained in reduced densities is encoded in their spectral decompositions and is akin to conditional probability:



What can we do with this?

An unsupervised machine learning problem:

Infer a probability distribution π on a set S of text data given samples.

Here's the main idea:

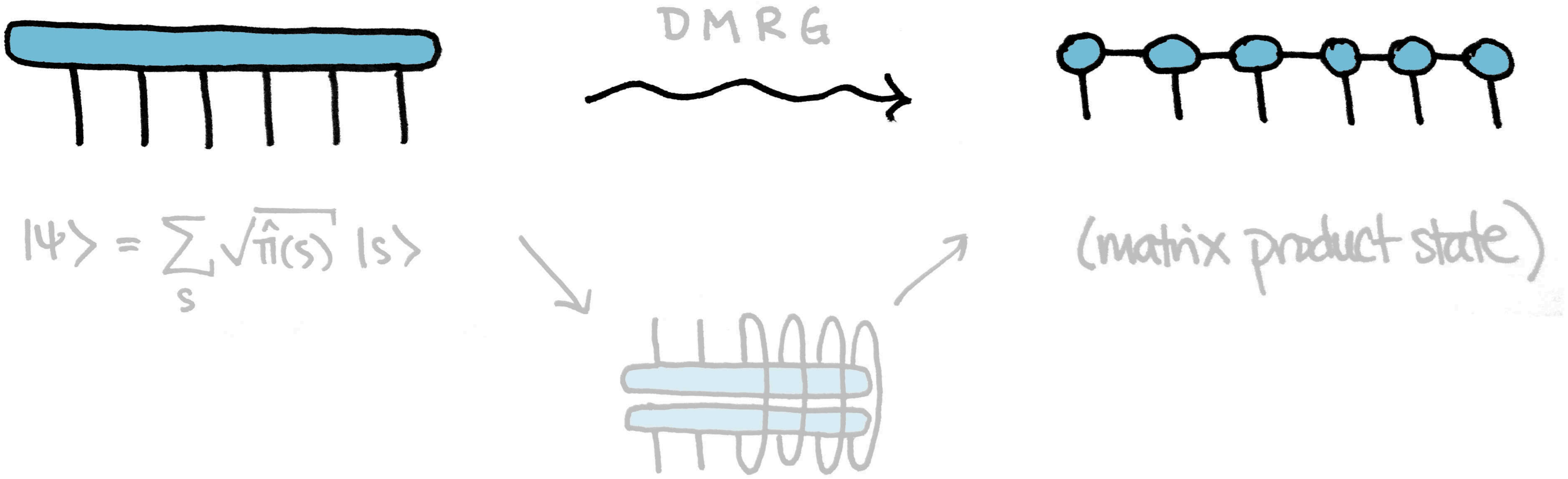
Let S be a set of sequences of length N from a finite alphabet A.

$$s = (a_1, a_2, \ldots, a_N)$$

Let $T \subset S$ be a set of sequences with empirical probability distribution $\hat{\pi}: T \to [0,1]$.

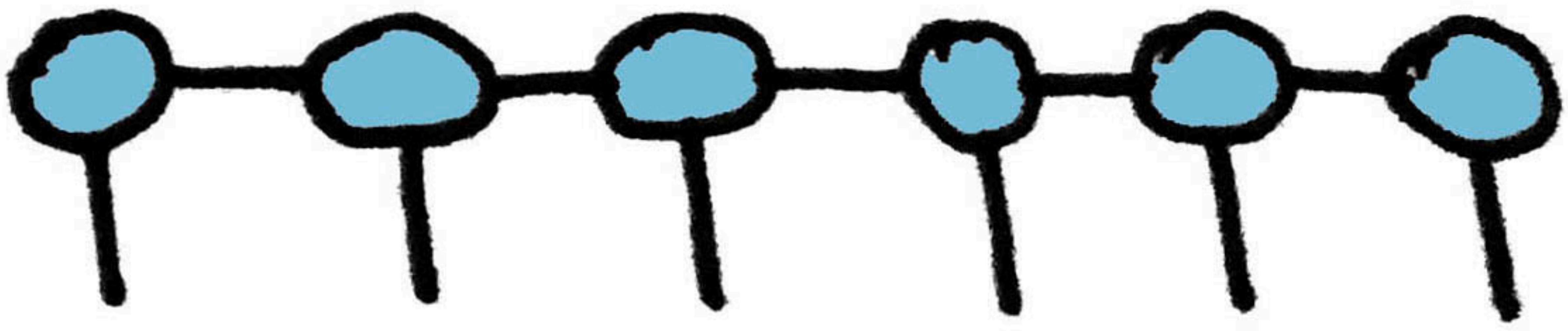
Consider the state $|\psi\rangle = \sum_{s \in T} \sqrt{\hat{\pi}(s)} |s\rangle$ in $\mathbb{C}^A \otimes \cdots \otimes \mathbb{C}^A$.

Apply a physics-inspired deterministic algorithm[1] to produce a low rank tensor factorization of $|\psi\rangle$ called a **matrix product state**.



[1] The *density matrix renormalization group* (DMRG) procedure.

Each tensor is comprised of eigenvectors of reduced densities from $\rho = |\psi\rangle\langle\psi|$. As a result, **the model knows which "words" go together to form meaningful "expressions"** based on the statistics of the data.



Modeling Sequences with Quantum States: A Look Under the Hood

arXiv:1910.07425

with Miles Stoudenmire (Flatiron Institute)
John Terilla (CUNY and Tunnel)