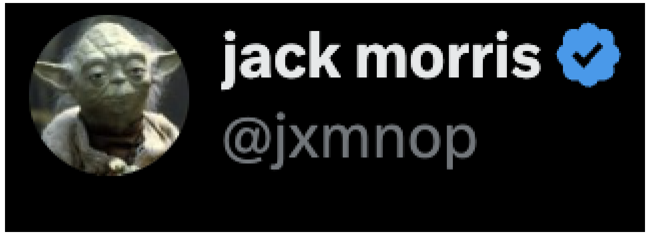
All AI Models Might Be the Same

Harnessing the Universal Geometry of Embeddings

Rishi Jha & Jack Morris

Cornell Tech

Computational Platonic Space



About Jack

2020 – 2021: AI Resident at Google

2021 – Present: PhD Student
@ Cornell Tech

2024 – Present: Researcher at Meta

About Rishi

Cornell Tech

- Third-Year PhD Student working with **Vitaly Shmatikov**

University of Washington, Seattle

Microsoft & Google (soon)

vec2text
Volodymyr Kuleshov
Vitaly Shmatikov
Sasha Rush

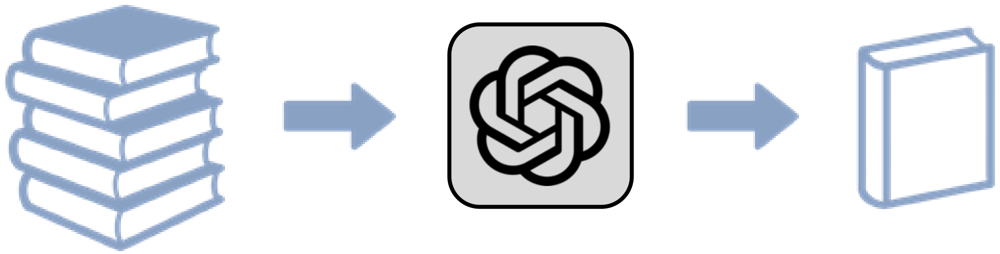
vec2vec
Collin Zhang
Vitaly Shmatikov

In collaboration with…

Gameplan

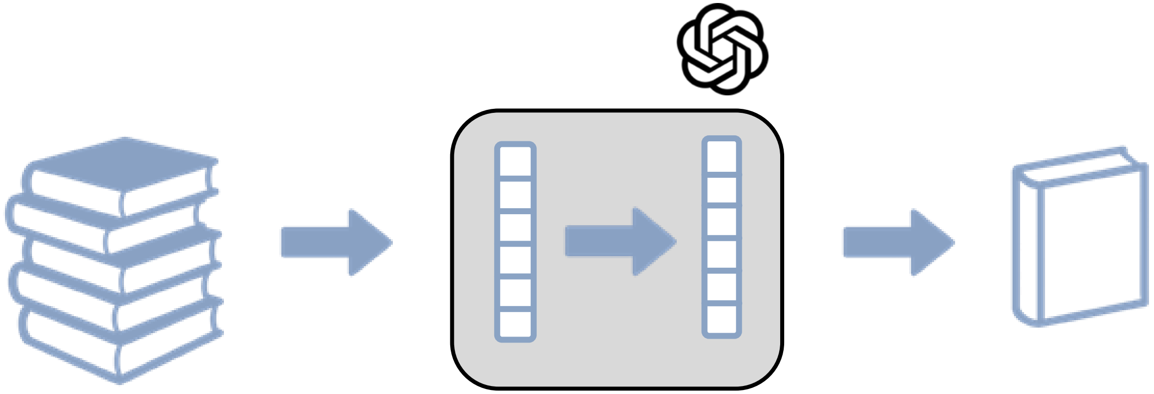
1. **Background: what are embeddings?**
2. vec2text: How much information do embeddings leak?
3. vec2vec: Translating embeddings with no help
4. Conclusion

From language models to embeddings



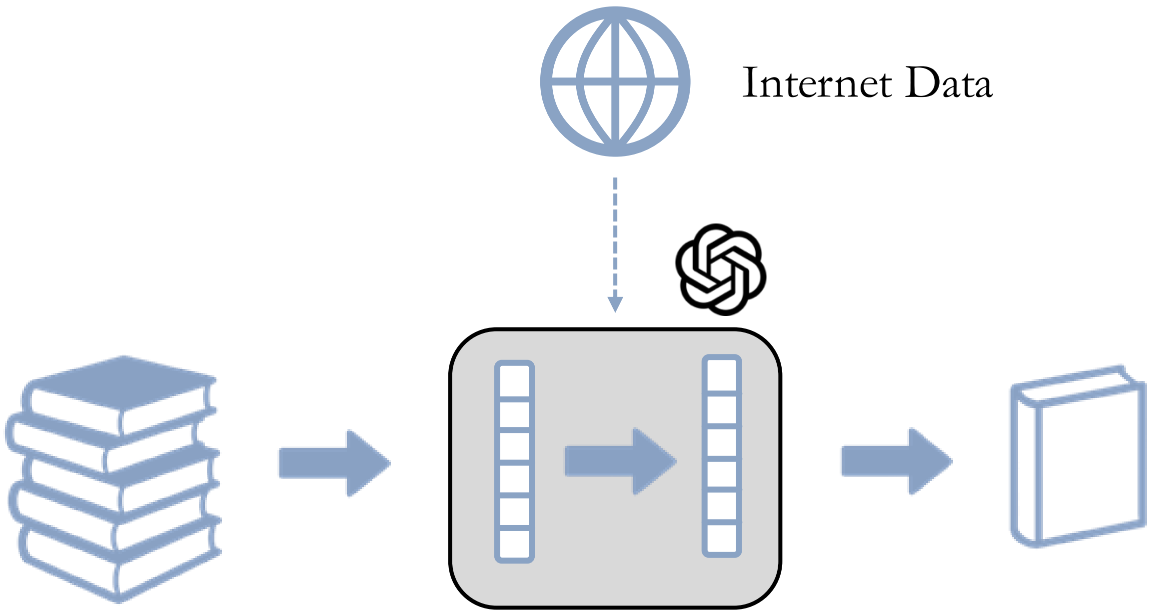
When most people think of language models, they think of text-to-text models.

From language models to embeddings



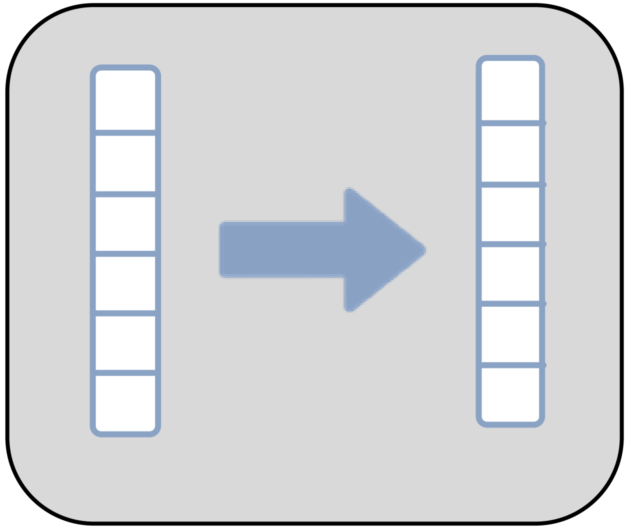
Despite **emitting** text, LMs operate on vector representations of the text called **embeddings.**

From language models to embeddings



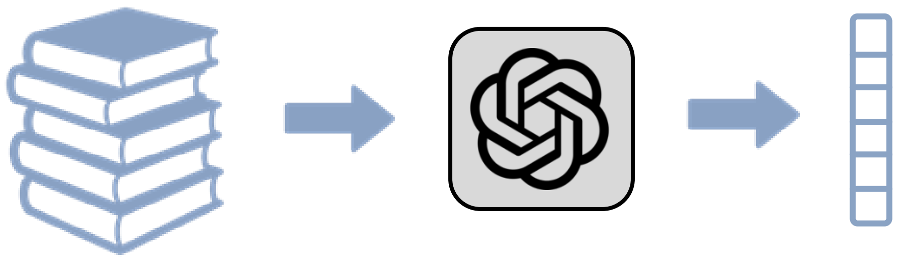
LM weights are usually trained with data from the entire internet…

From language models to embeddings



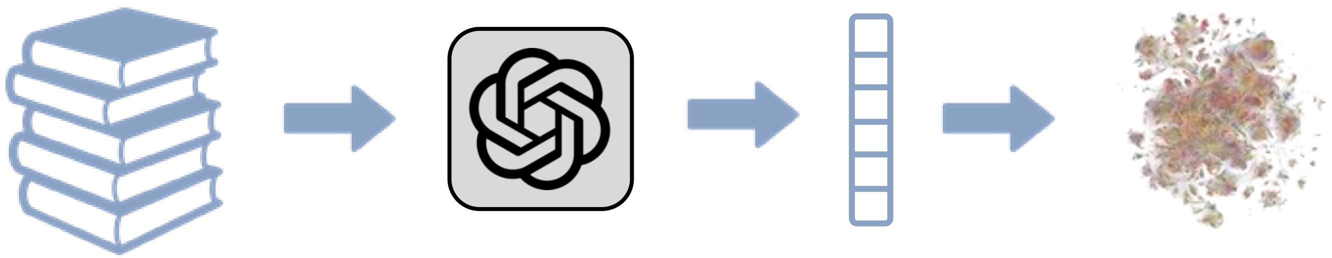
… imbuing these vector representations with **strong semantic priors.**

From language models to embeddings

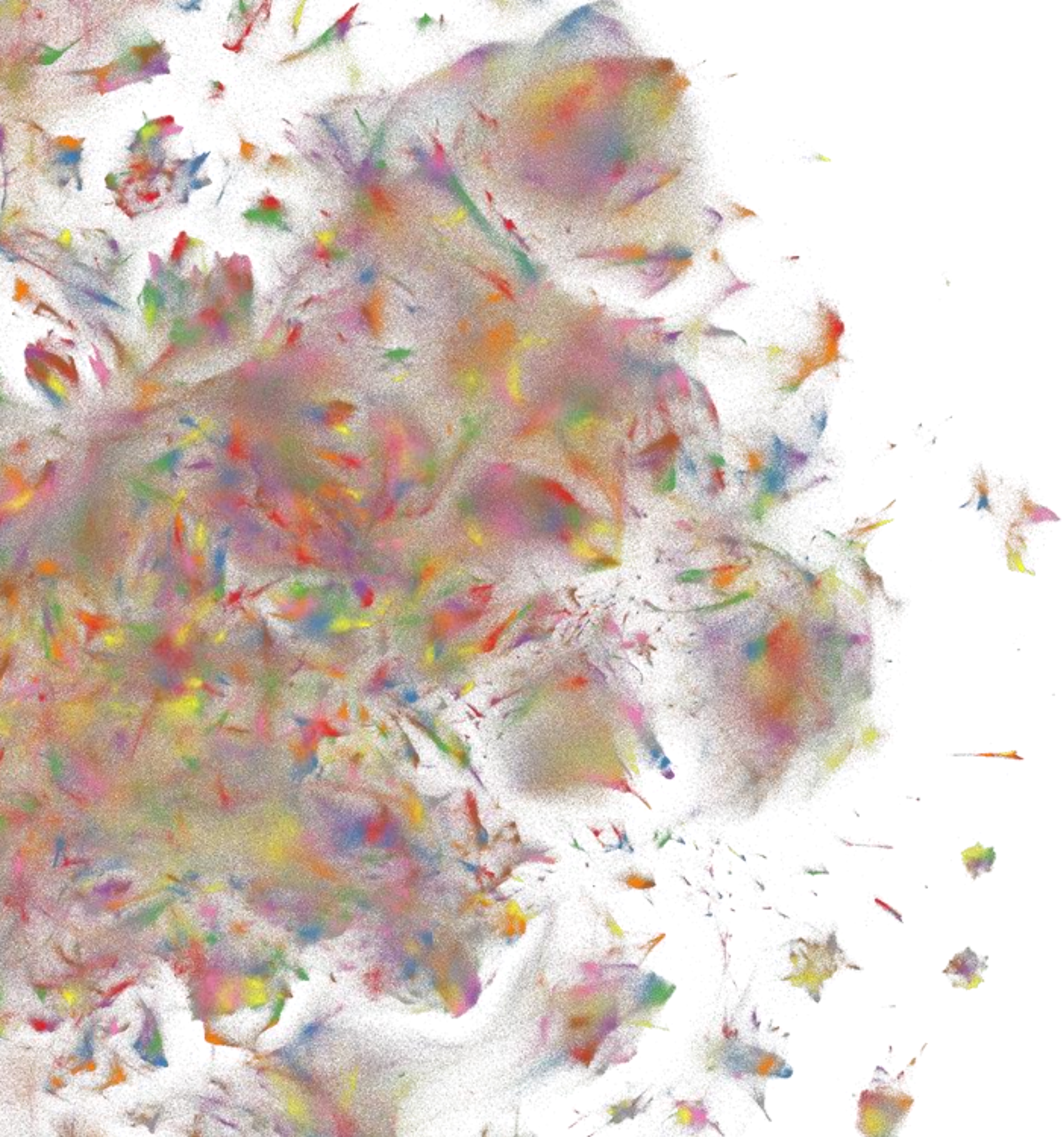


Language models that just emit embeddings are called **encoders**

From language models to embeddings



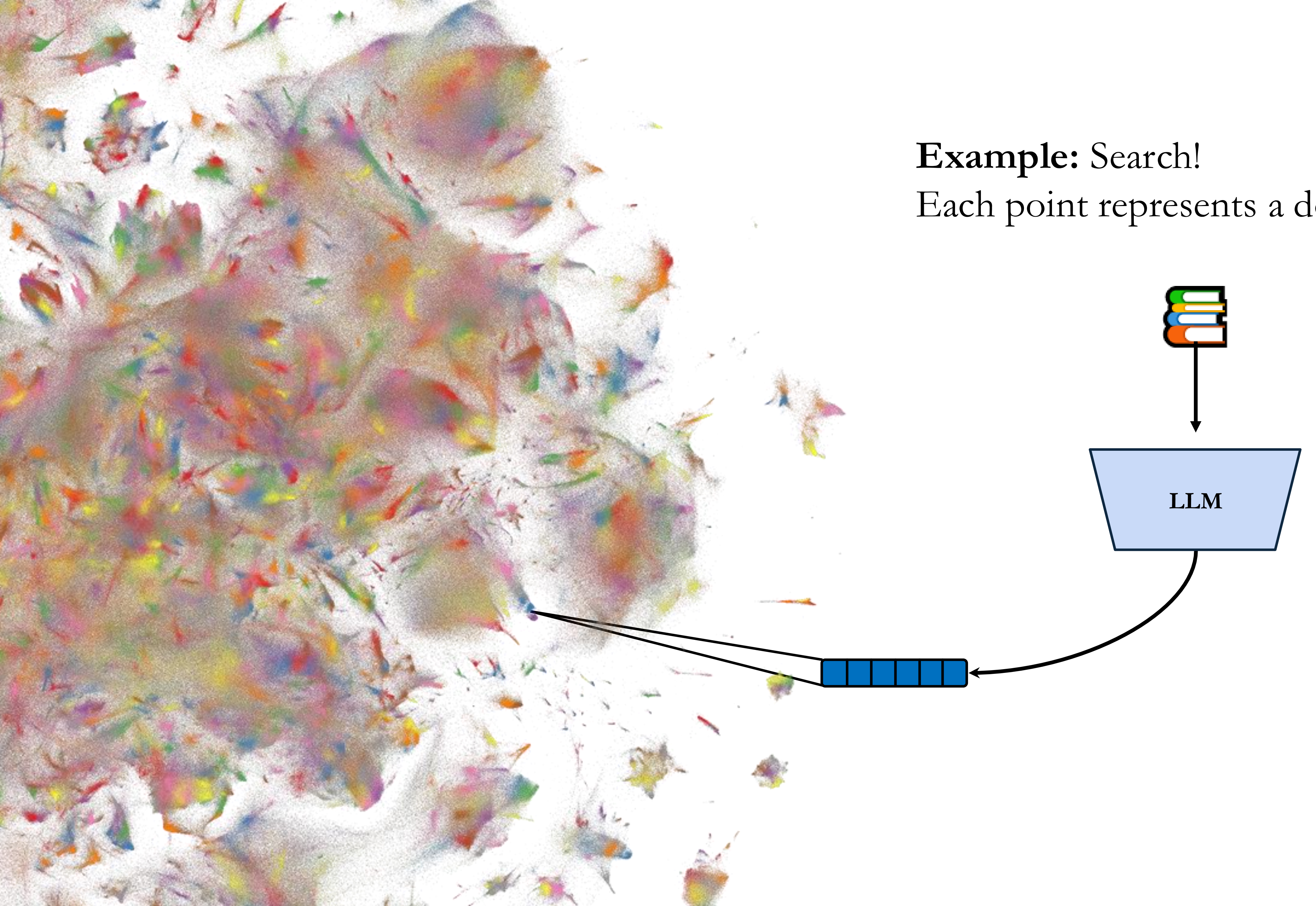
Semantic priors (+ some post-training) make encoders useful!



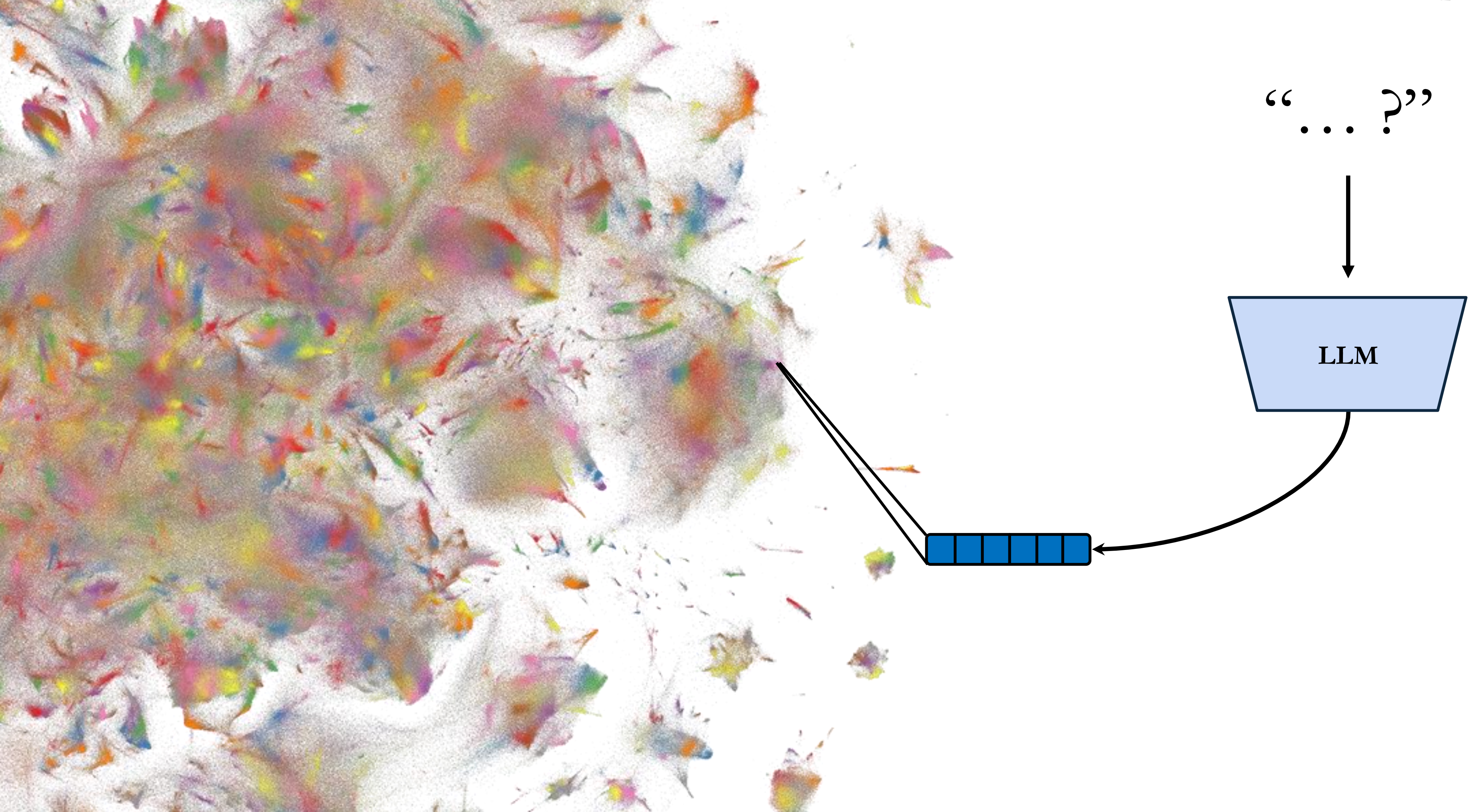
Example: Search!
Each point represents a document.

Example: Search!

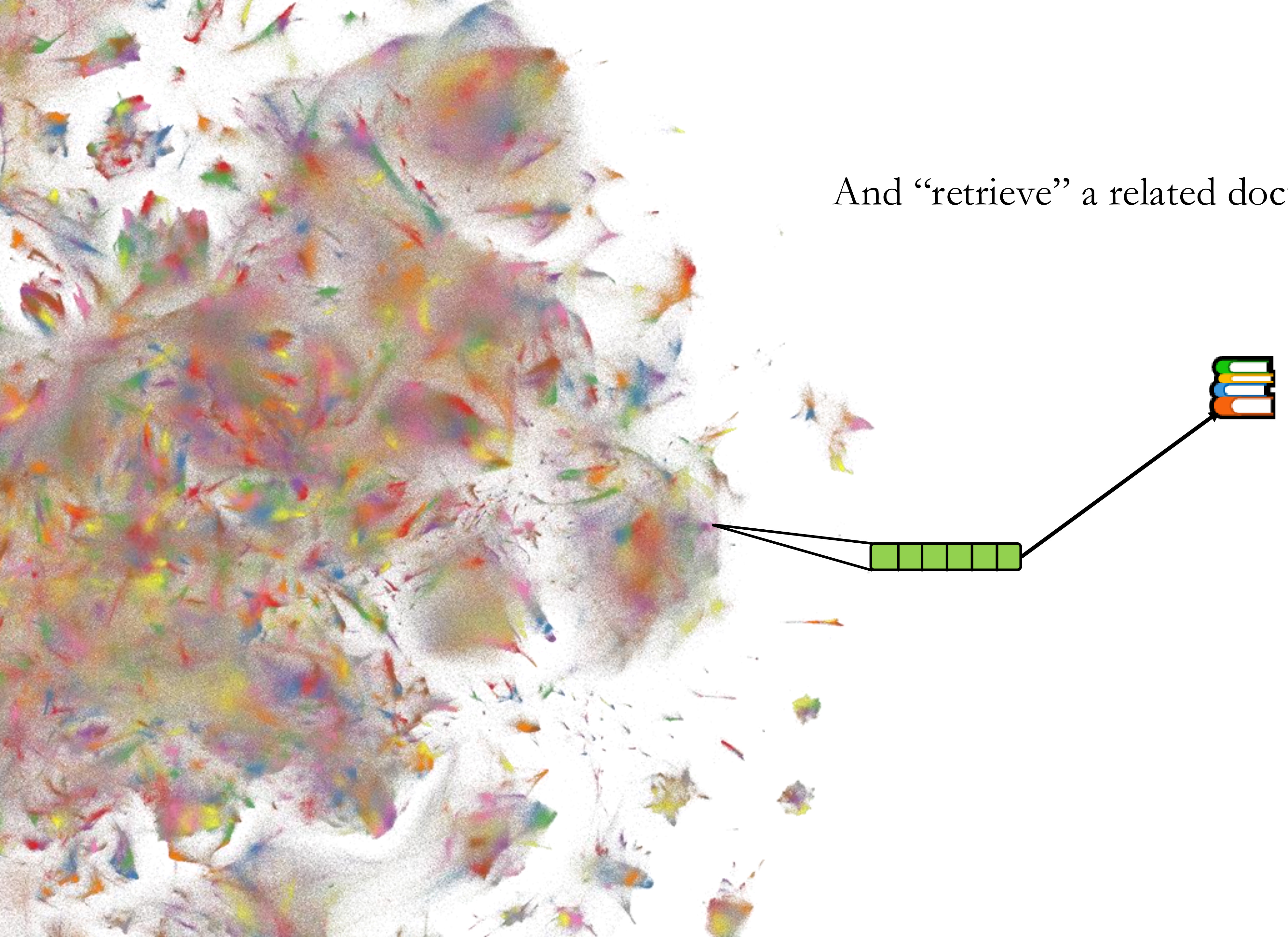
Each point represents a document.



A user can ask a question…

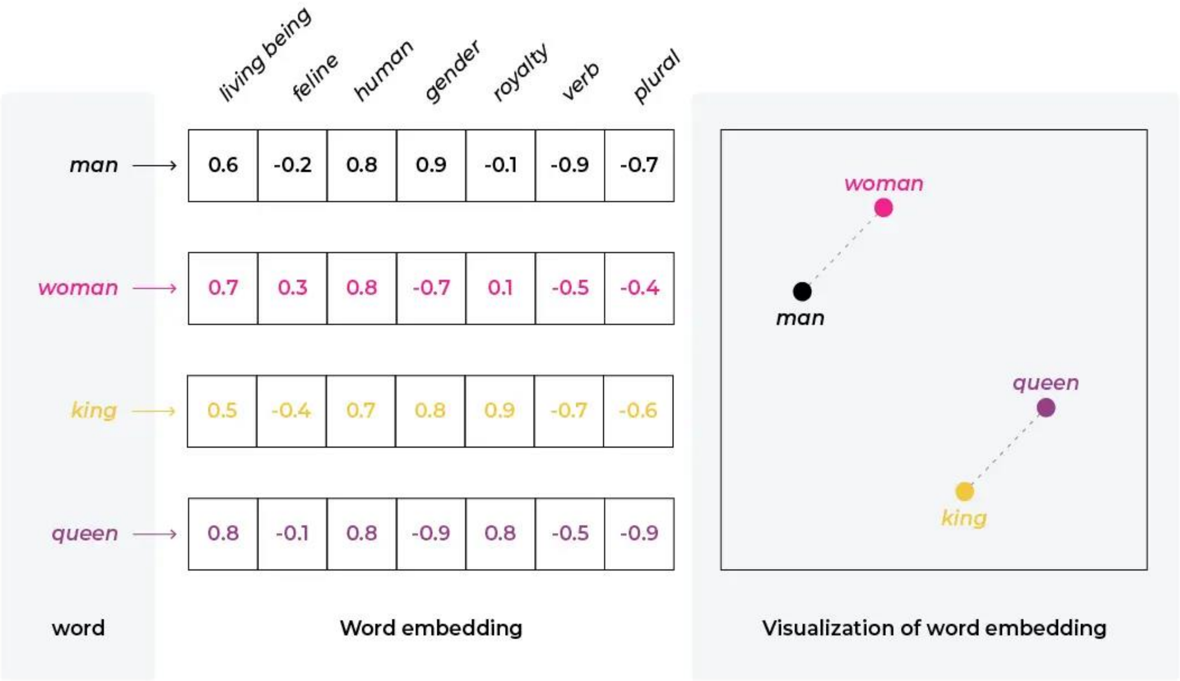


And “retrieve” a related document as an answer.



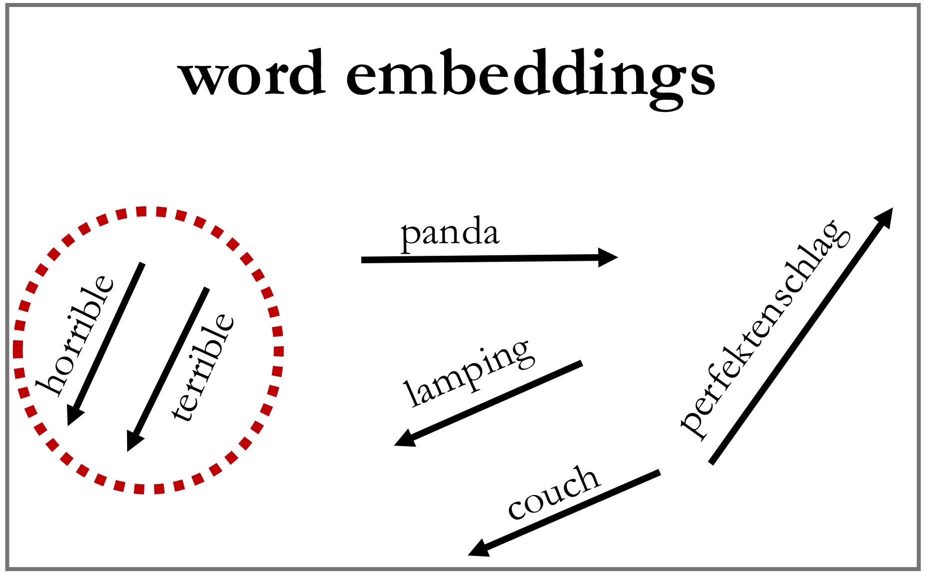
Question: Why are embeddings so useful for search?

Simpler setup: Word embeddings



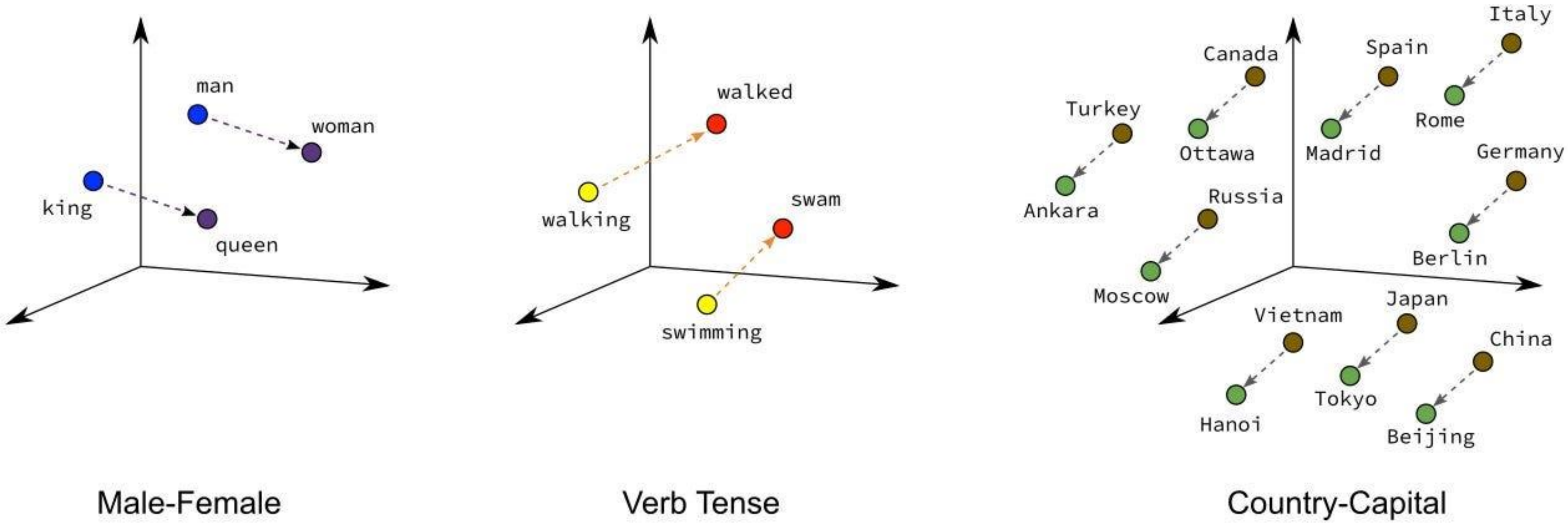
Predecessor to document embeddings: each word → vector!

Similar vectors = similar meaning



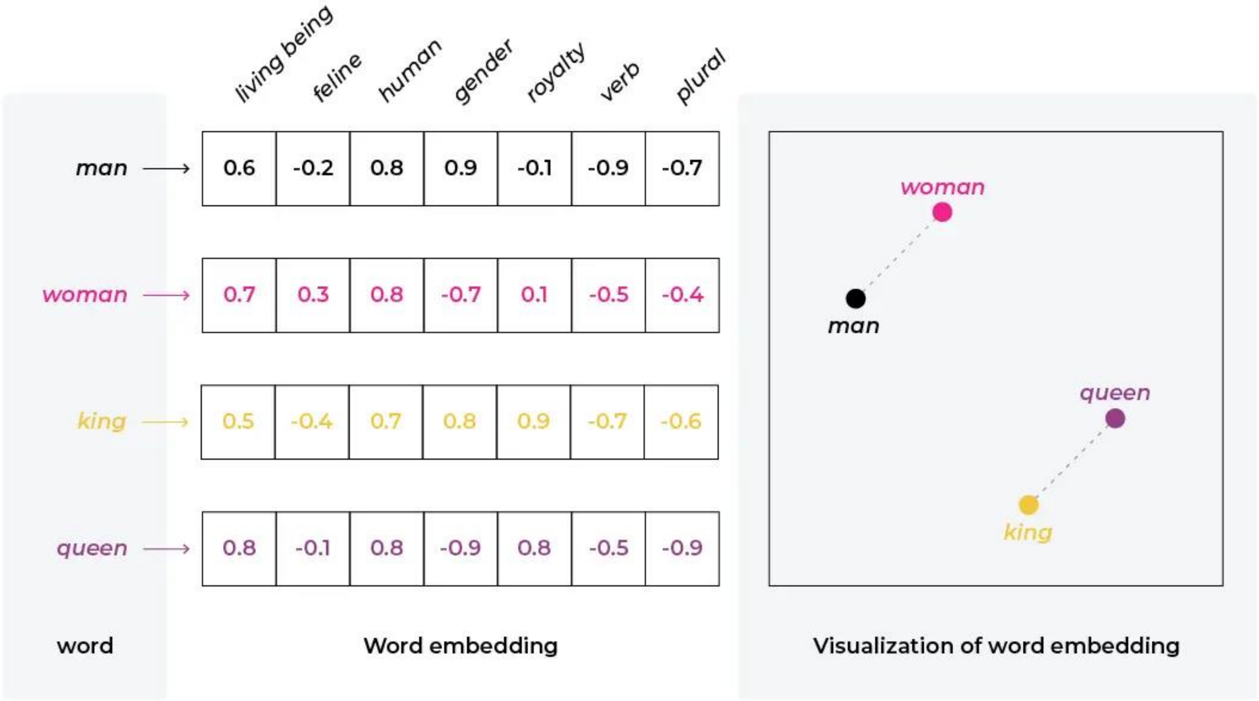
Semantically related words map to numerically similar vectors!

Geometry of word embeddings



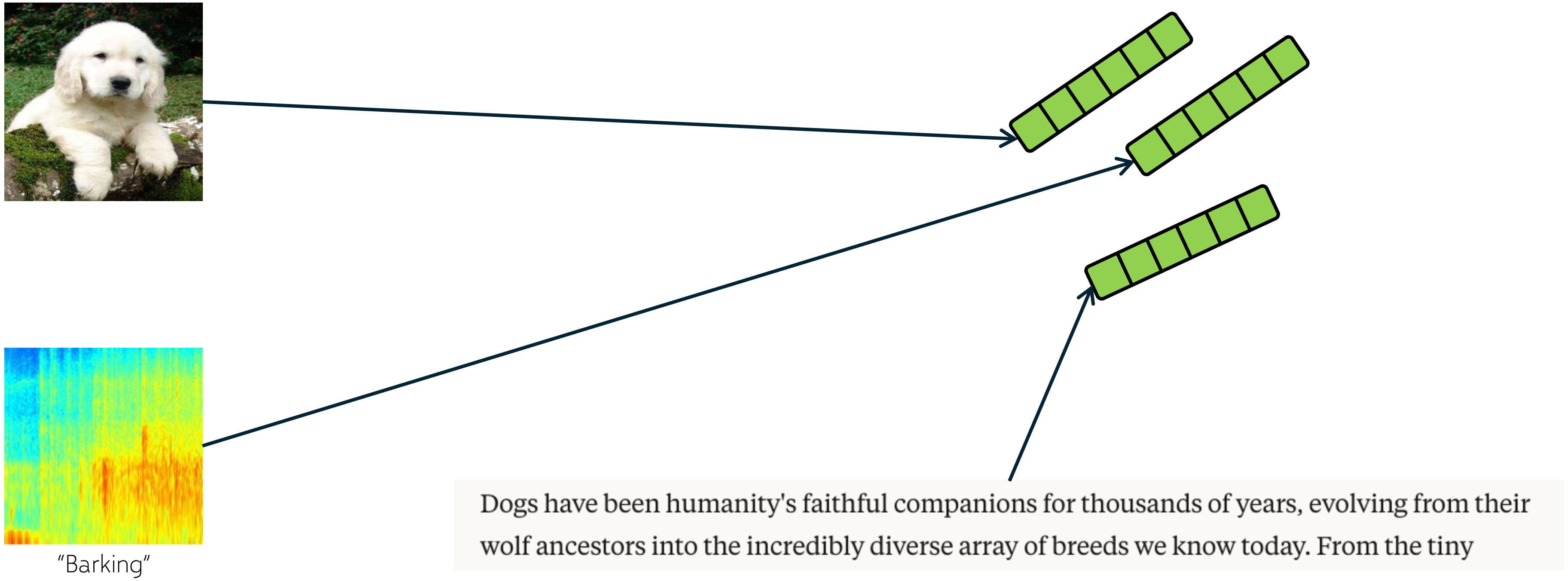
source: Google

Geometry of word embeddings



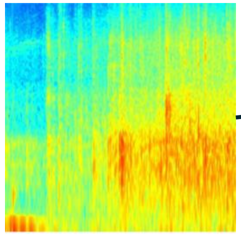
"king" – "man" + "woman" = "queen"

Semantic similarity in embeddings



Modern encoders generalize this notion to whole documents and **different modalities.**

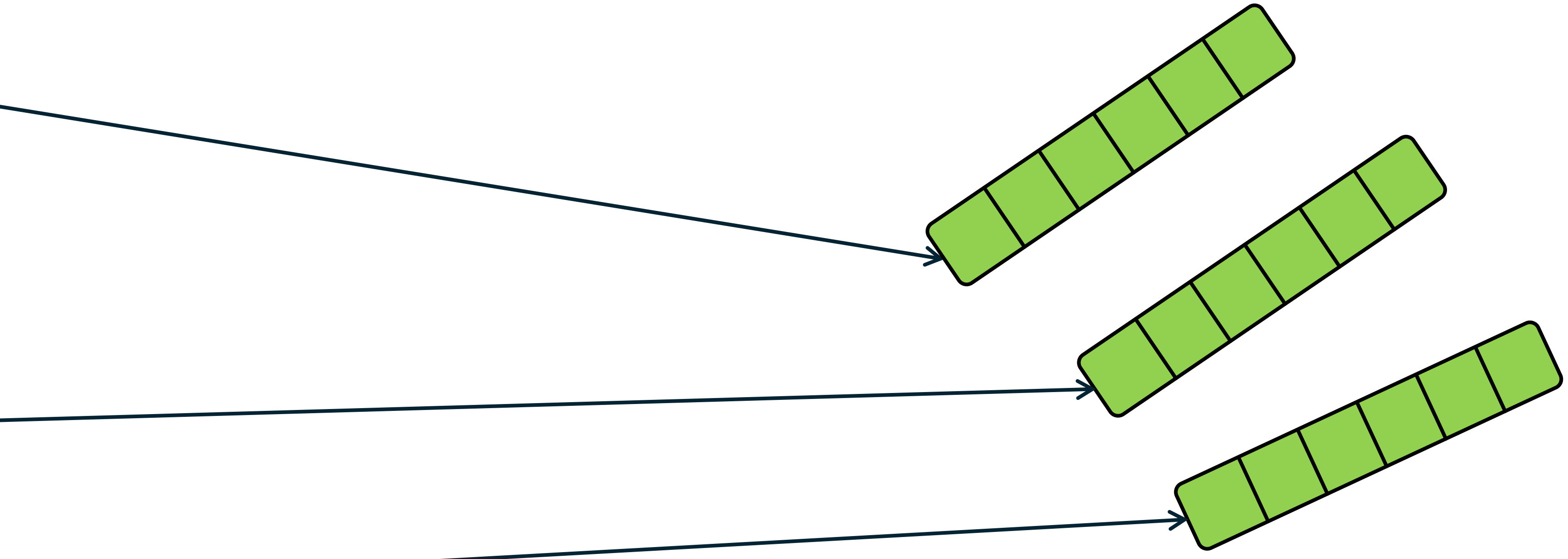
Semantic similarity in embeddings



Dogs have been humanity's faithful companions for thousands of years, evolving from their wolf ancestors into the incredibly diverse array of breeds we know today. From the tiny

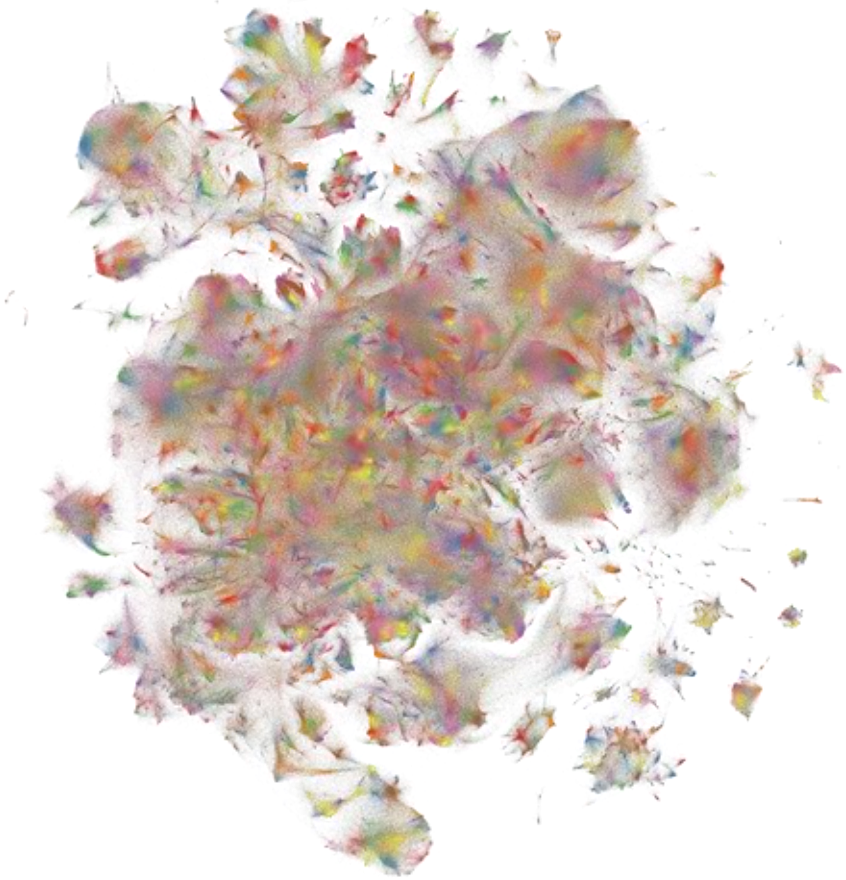
Closeness is a property of the inputs…

Semantic similarity in embeddings



Regardless of encoder or modality!

Modern embeddings



Retrieval (Semantic Search)

Summarization

Retrieval augmented generation (RAG) in LLMs

Companies training embedding models

Embeddings power memory

The Verge

Menu +

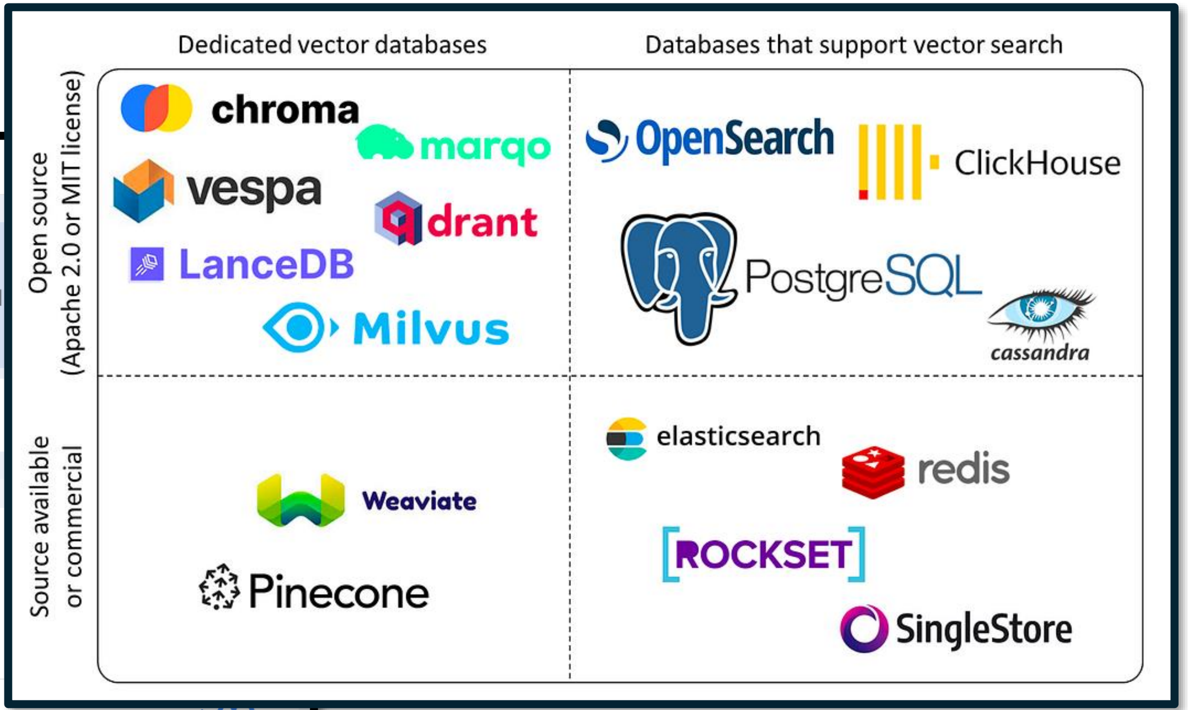
NEWS

ChatGPT will now remember your old conversations / Long-term memory allows ChatGPT to reference details you discussed, even if you didn't manually save them.

by **Jess Weatherbed**
Apr 11, 2025, 5:43 AM EDT

2 Comments (2 New)

The rise of vector databases



Startups

Pinecone drops $100M i on $750M valuation, as v database demand grows

AI

Qdrant, an open source vector databas develop data

Paul Sawers @psawers

Premium HOME > TECH

Vector database Chroma scored $18 mil million valuation. Here's why its technol generative AI startups.

Stephanie Palazzolo Apr 6, 2023, 8:00 AM EDT

Weaviate Raises $50 Million Series B F Meet Soaring Demand for AI Native Vector Database Technology

USA - English

NEWS PROVIDED BY
Weaviate →
21 Apr, 2023, 08:00 ET

Company's open source vector database and new cloud service ar

FORBES > INNOVATION > CLOUD

The Rise Of Vector Databases

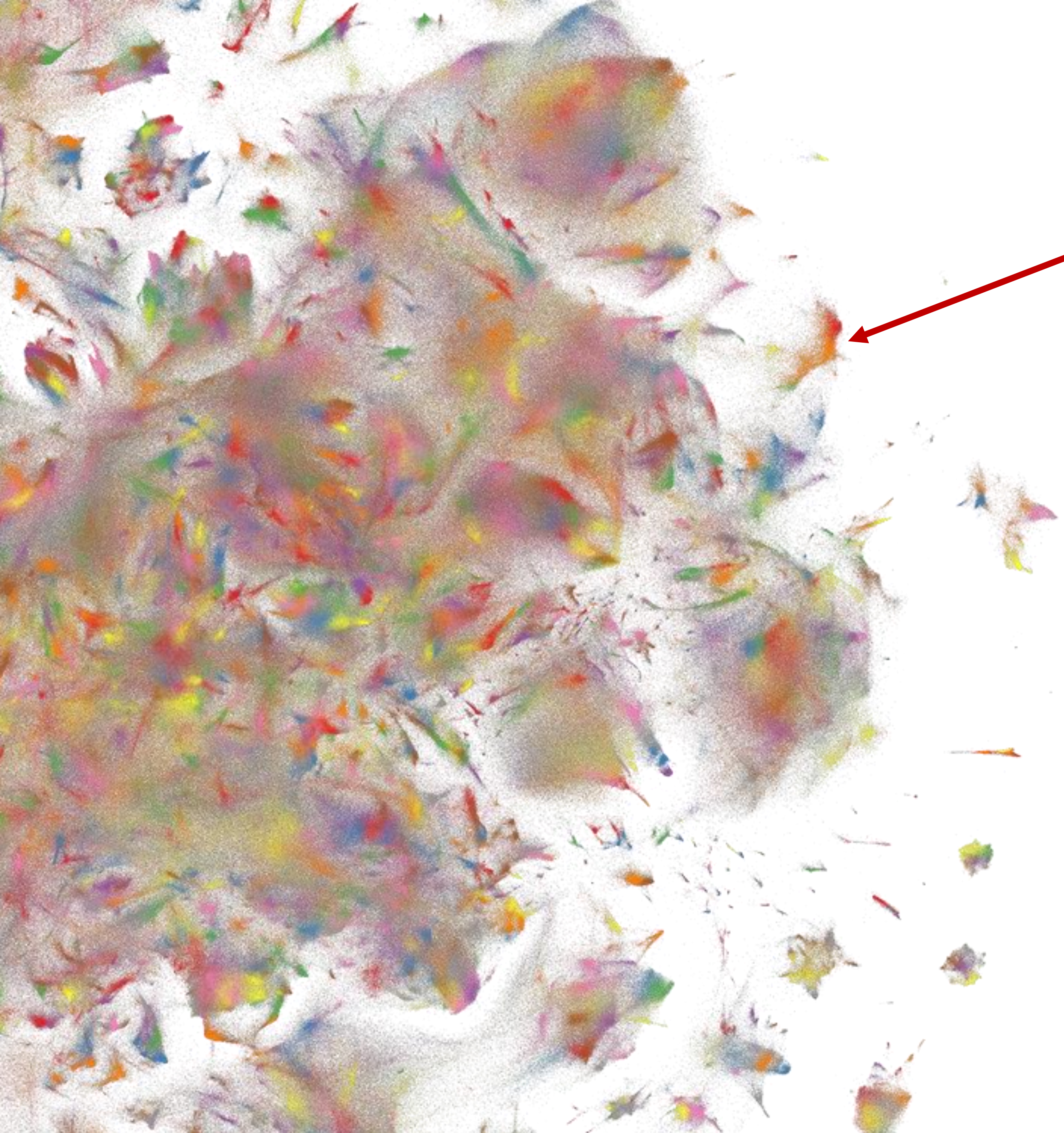
Adrian Bridgwater Senior Contributor

I track enterprise software application development & data management.

Follow

Gameplan

1. Background: what are embeddings?
2. **vec2text: How much information do embeddings leak?**
3. vec2vec: Translating embeddings with no help
4. Conclusion



Question: How much information about a document is preserved by *its vector representation*?

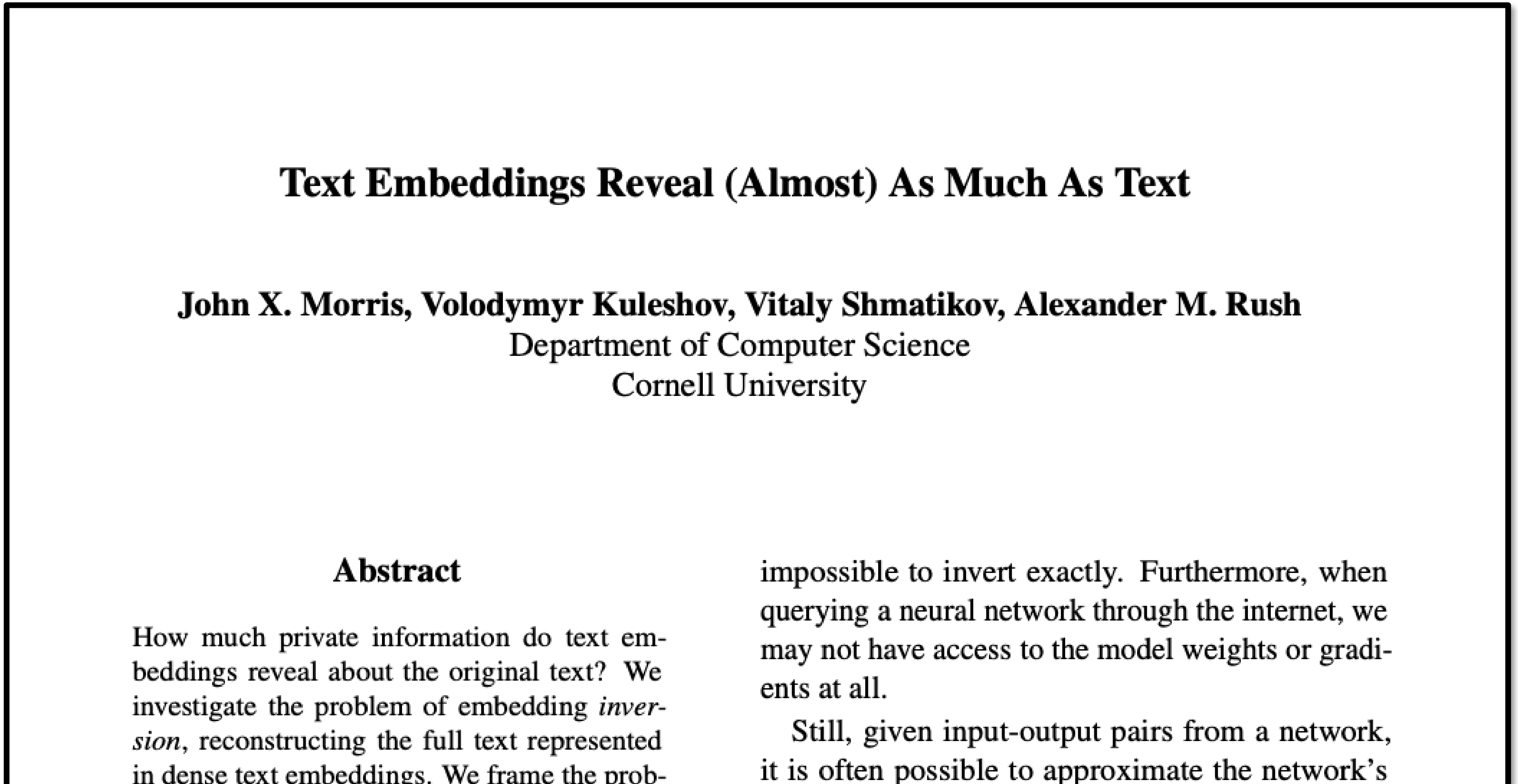
Reframed: What can a bad actor learn from just looking at embeddings of text?

The **data processing inequality** is an information theoretic concept that states that the information content of a signal cannot be increased via a local physical operation. This can be expressed concisely as 'post-processing cannot increase information'.[1]

Challenges:

1. Small changes in input text (one word!) produce different vectors
2. Data processing inequality

Answer: Embeddings leak almost everything!



Text Embeddings Reveal (Almost) As Much As Text

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, Alexander M. Rush
Department of Computer Science
Cornell University

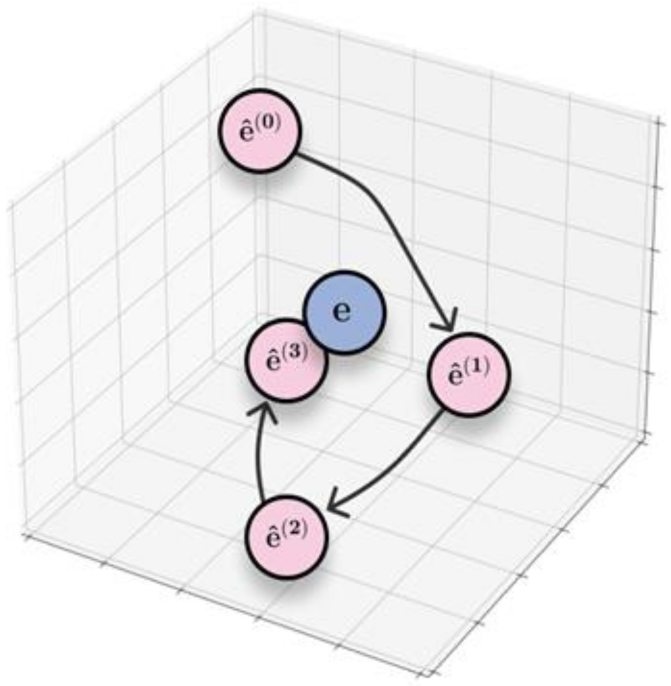
Abstract

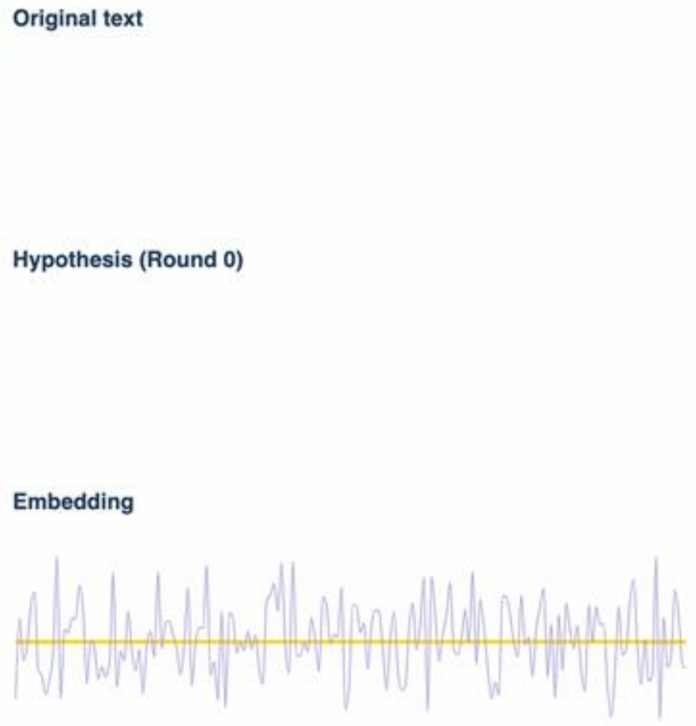
How much private information do text embeddings reveal about the original text? We investigate the problem of embedding *inversion*, reconstructing the full text represented in dense text embeddings. We frame the prob-

impossible to invert exactly. Furthermore, when querying a neural network through the internet, we may not have access to the model weights or gradients at all.

Still, given input-output pairs from a network, it is often possible to approximate the network's

vec2text





Intuition: Repeatedly **query an encoder** with candidate texts until the embeddings are close to target!

Key Idea: Encoders imbue **lots** of semantic information in their embeddings.

Key Idea: Encoders imbue **lots** of semantic information in their embeddings.

… an attacker with access to embeddings **and encoder** can reconstruct original text!

Gameplan

1. Background: what are embeddings?
2. vec2text: How much information do embeddings leak?
3. **vec2vec: Translating embeddings with no help**
4. Conclusion

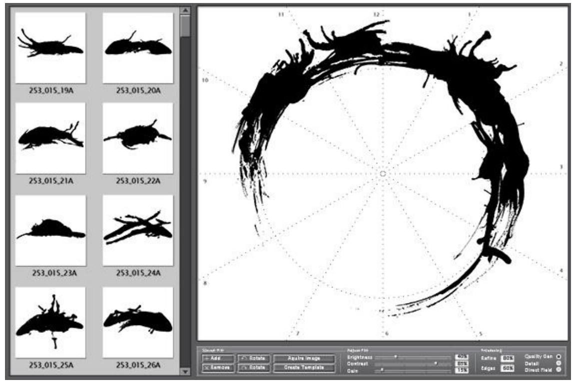
Question: What if we don’t have access to the original encoder? Do embeddings contain *enough* information still?

```
[5.62833 4.45560 7.09206 1.70772 8.53488 … 2.81810]
[8.32360 0.17597 6.37227 7.09399 4.30062 … 5.91650]
[2.05302 4.23975 6.58735 1.82040 8.01594 … 7.40739]
[3.08180 7.25108 5.14575 2.28853 1.18346 … 2.87634]
[0.82509 8.74585 4.85676 5.90278 1.30682 … 1.09638]
[3.36469 8.70506 6.34738 3.00865 8.25189 … 7.84836]
[8.75593 2.19901 1.14154 2.48679 8.53991 … 8.24471]
[5.72269 8.46621 3.27051 6.58750 8.80183 … 2.80392]
[3.34592 5.21735 2.51893 5.21443 8.57784 … 6.69609]
…
[1.52108 1.68765 3.82813 0.27698 7.82777 … 1.54355]
```

Invert this!

```
[5.62833 4.45560 7.09206 1.70772 8.53488 … 2.81810]
[8.32360 0.17597 6.37227 7.09399 4.30062 … 5.91650]
[2.05302 4.23975 6.58735 1.82040 8.01594 … 7.40739]
[3.08180 7.25108 5.14575 2.28853 1.18346 … 2.87634]
[0.82509 8.74585 4.85676 5.90278 1.30682 … 1.09638]
[3.36469 8.70506 6.34738 3.00865 8.25189 … 7.84836]
[8.75593 2.19901 1.14154 2.48679 8.53991 … 8.24471]
[5.72269 8.46621 3.27051 6.58750 8.80183 … 2.80392]
[3.34592 5.21735 2.51893 5.21443 8.57784 … 6.69609]
…
[1.52108 1.68765 3.82813 0.27698 7.82777 … 1.54355]
```

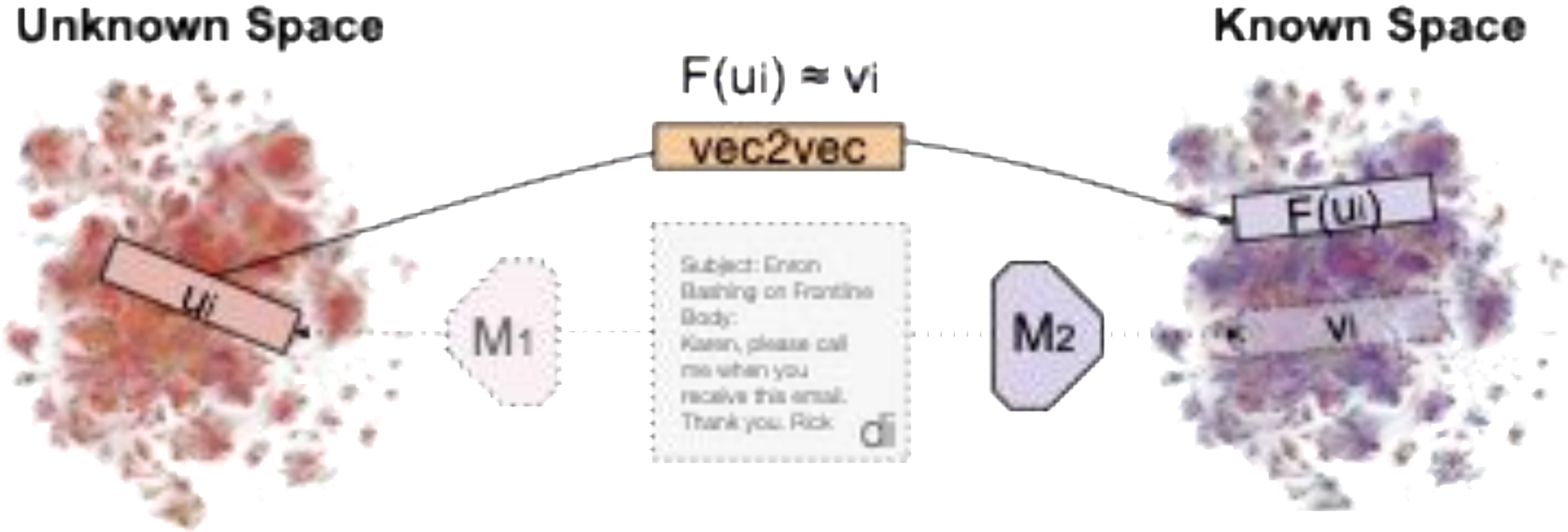
"Found some embeddings lying on the floor!"



Without knowing the encoder, the vectors are seemingly meaningless! Like trying to read alien **without any aliens!**

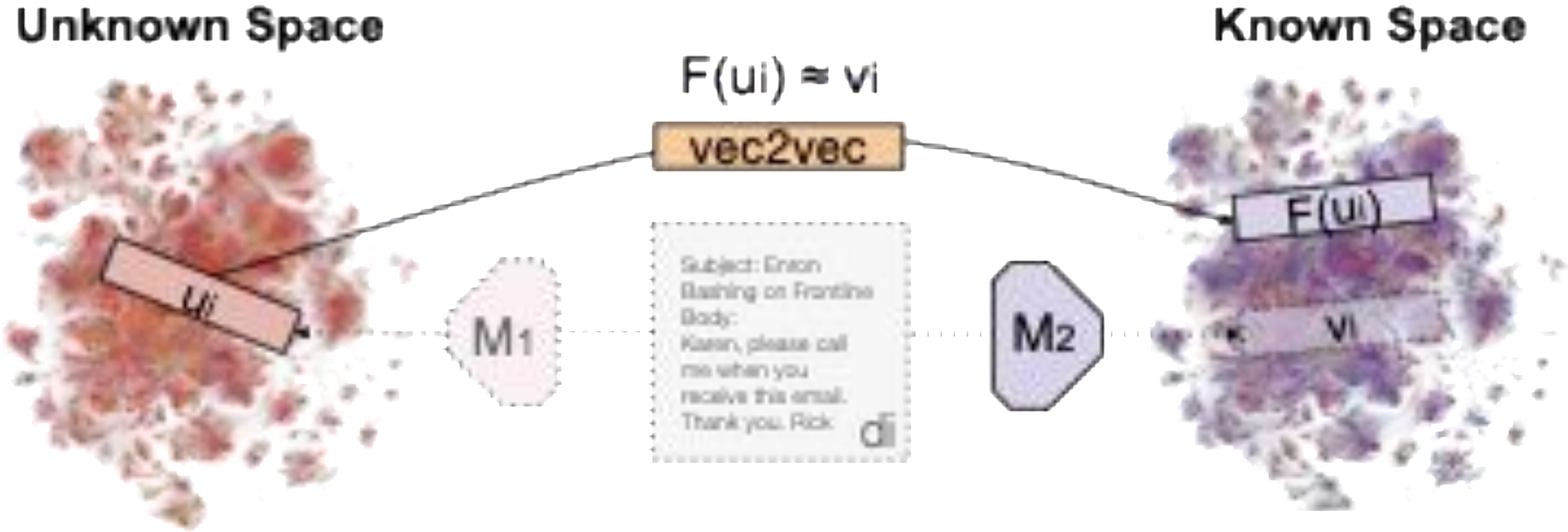
What can we do???

Unsupervised embedding translation



Idea: Translate our leaked vectors $\{u_i\}_{i=0}^{n}$ to a known space and run analysis there!

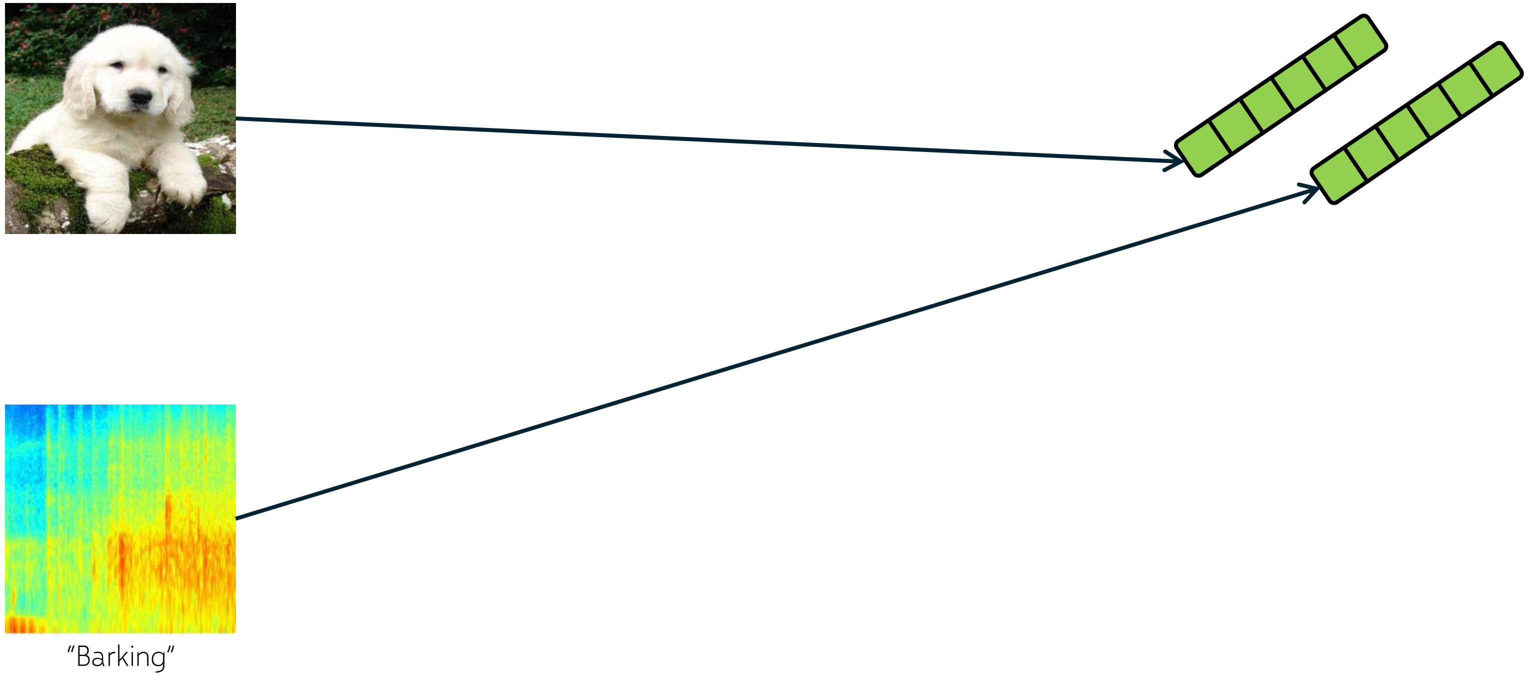
Unsupervised embedding translation



Note: We **do not** have access to the documents, M_1, or matches $\{v_i\}_{i=0}^{n}$!

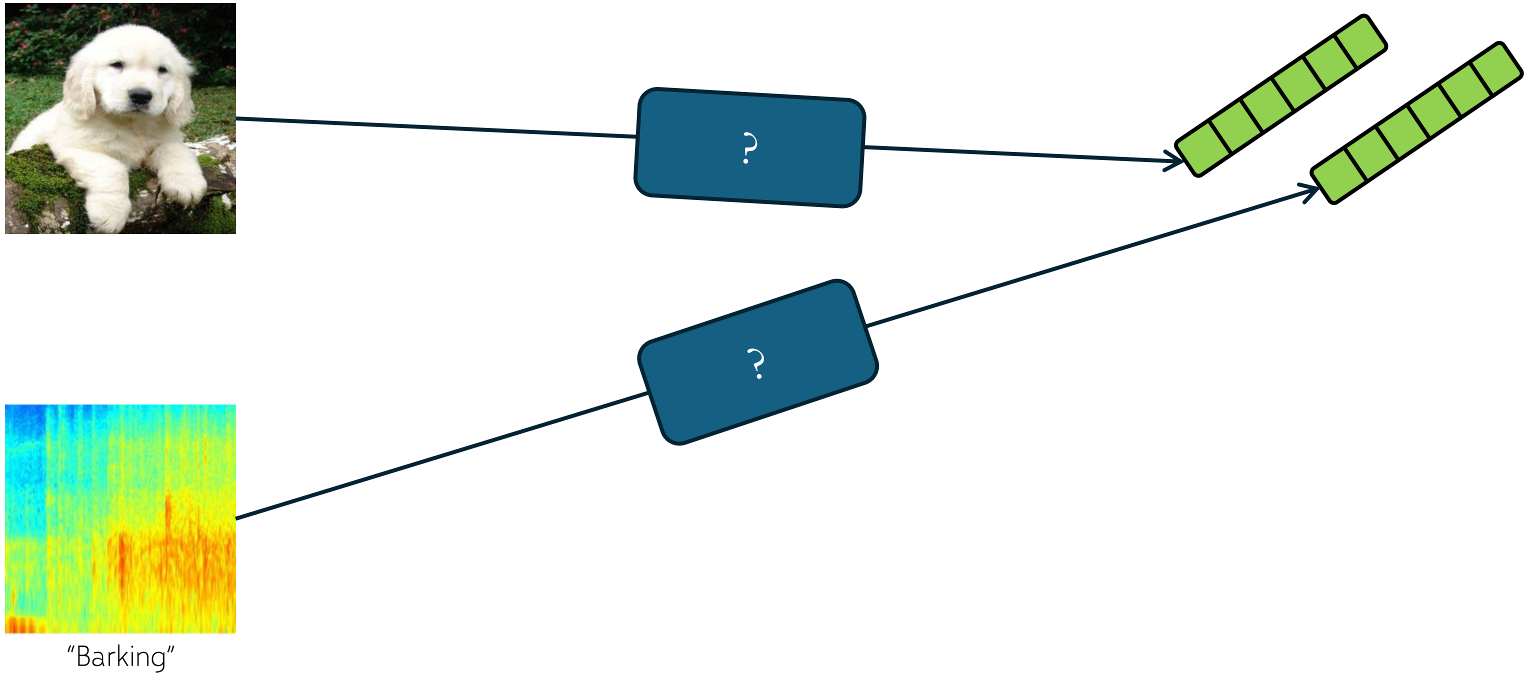
Our hope: Use the semantic structure of language as our Rosetta Stone!

Semantic similarity in embeddings



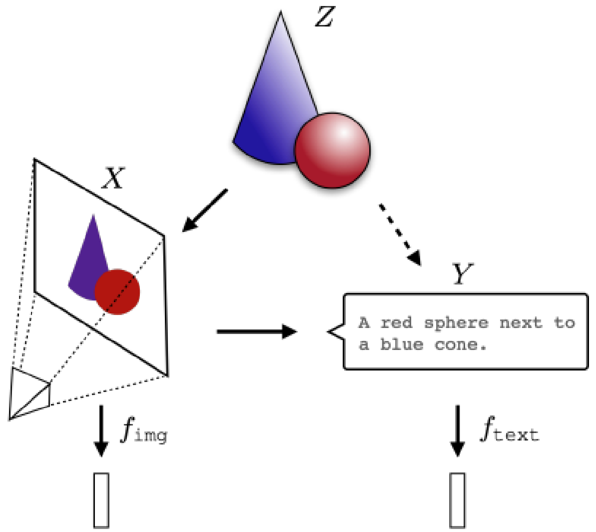
Recall: For *any* encoder to be useful, semantically related inputs must encode into similar vectors.

Semantic similarity in embeddings



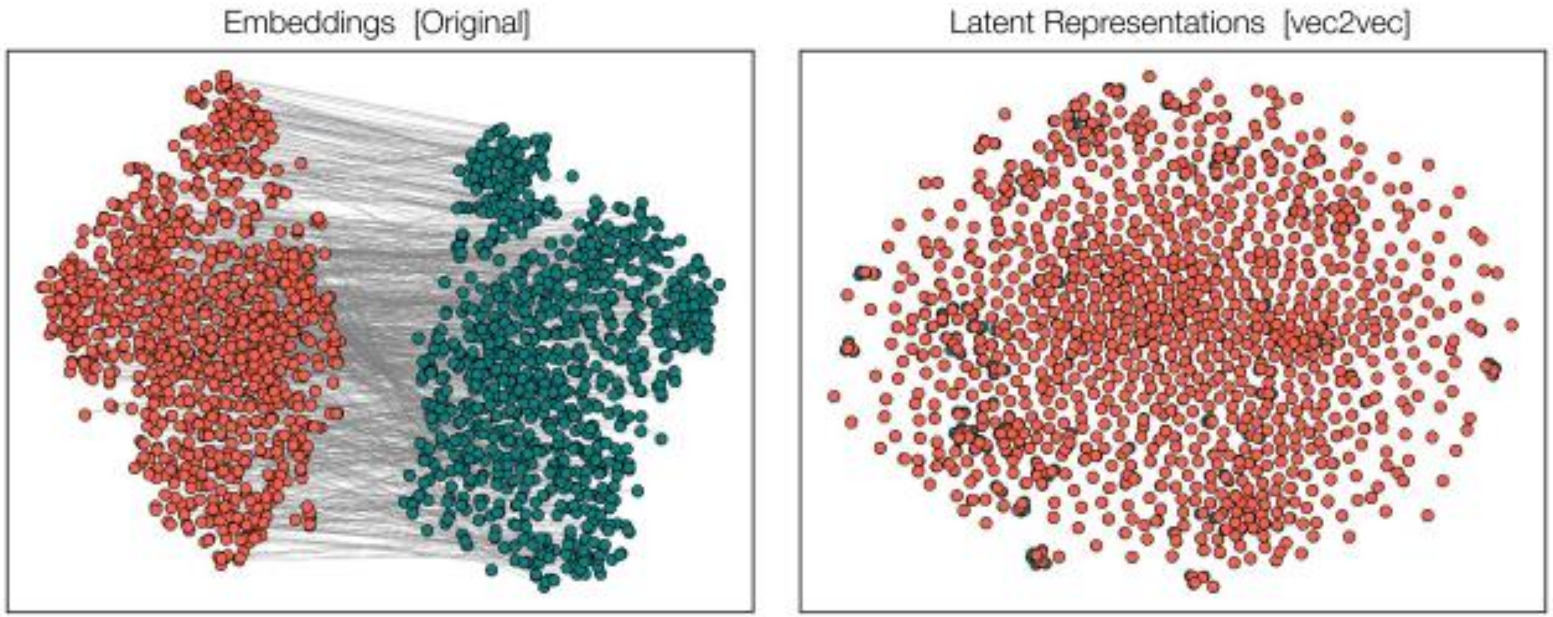
Semantic similarity is a property of **content not encoder!**

Is semantic structure universal?



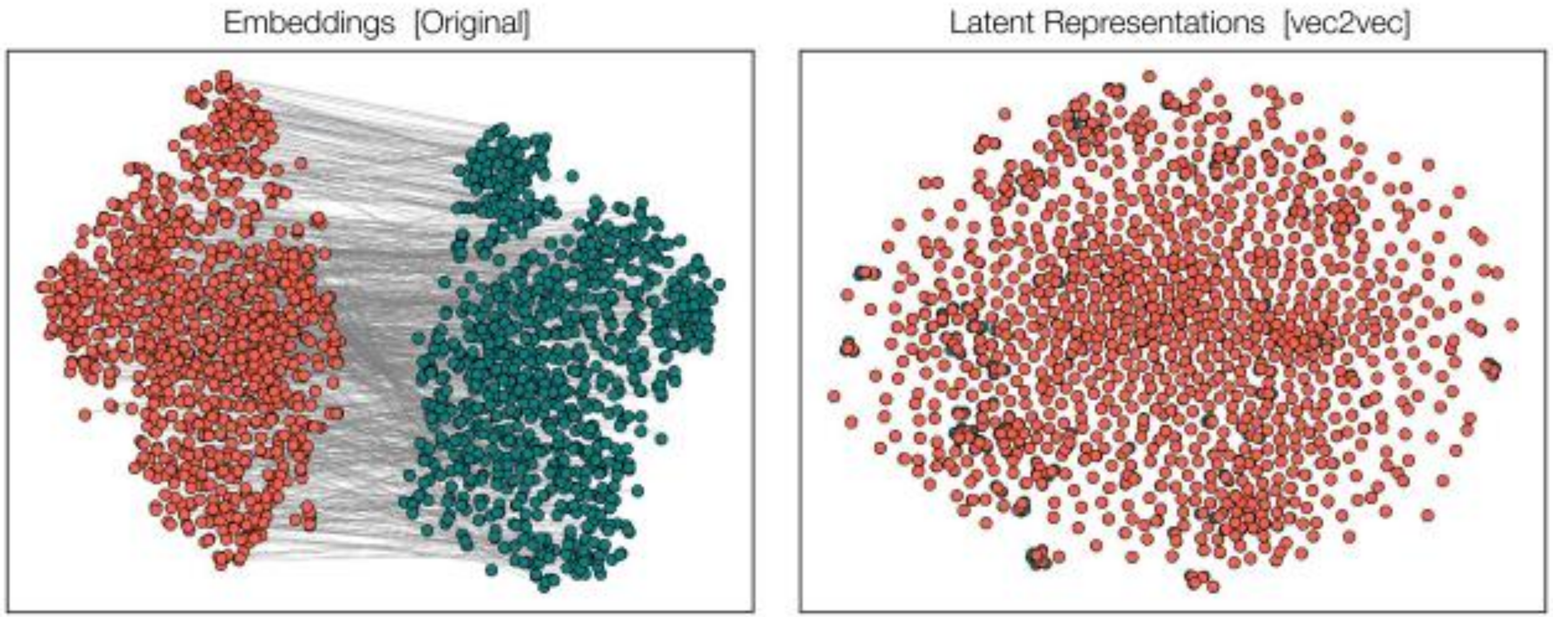
Platonic Representation Hypothesis [Huh et al., 2024]:
"Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces."

Key idea: Alignment of vector spaces



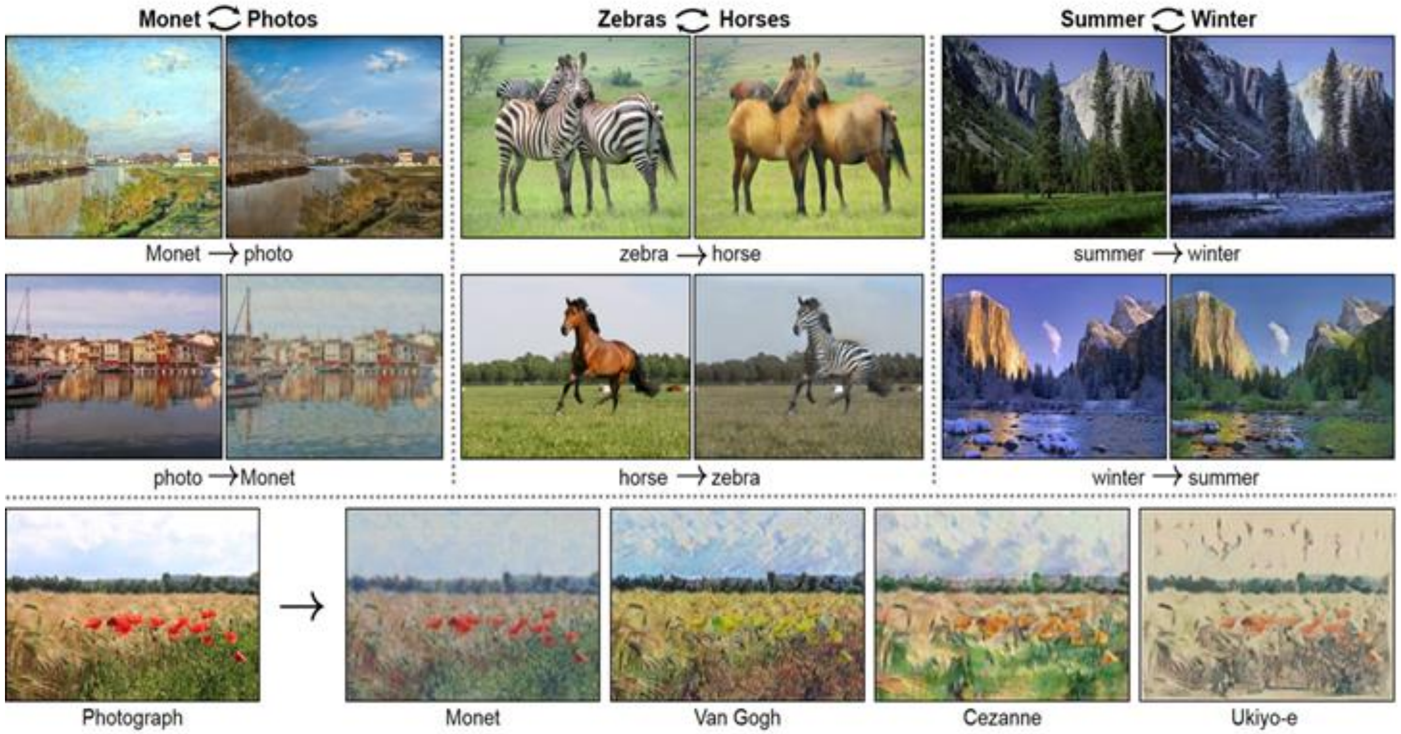
If the PRH is right, there *should be* a **shared statistical model** of the two spaces!

Key idea: Alignment of vector spaces



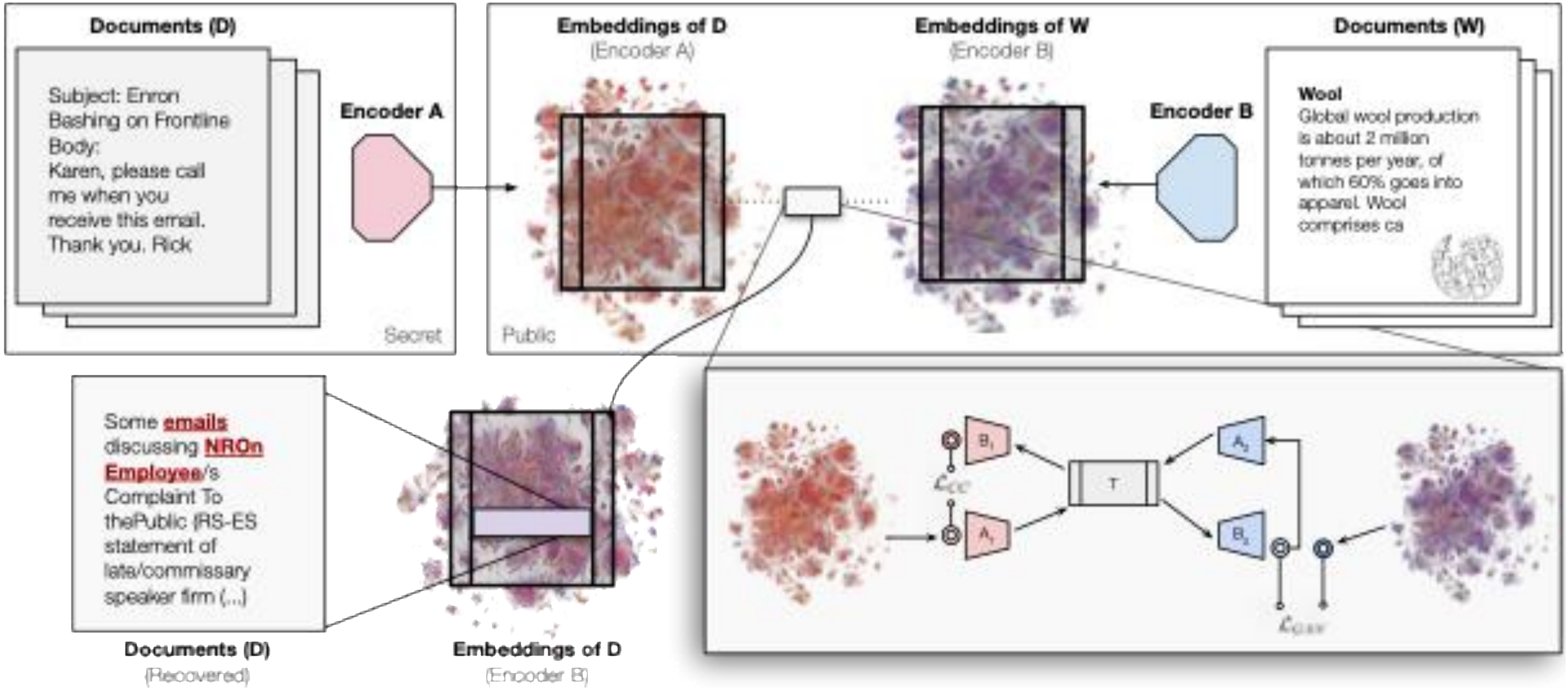
Hope: We can (1) characterize and (2) translate to and from a **shared latent representation!**

CycleGAN: A technique for image translation

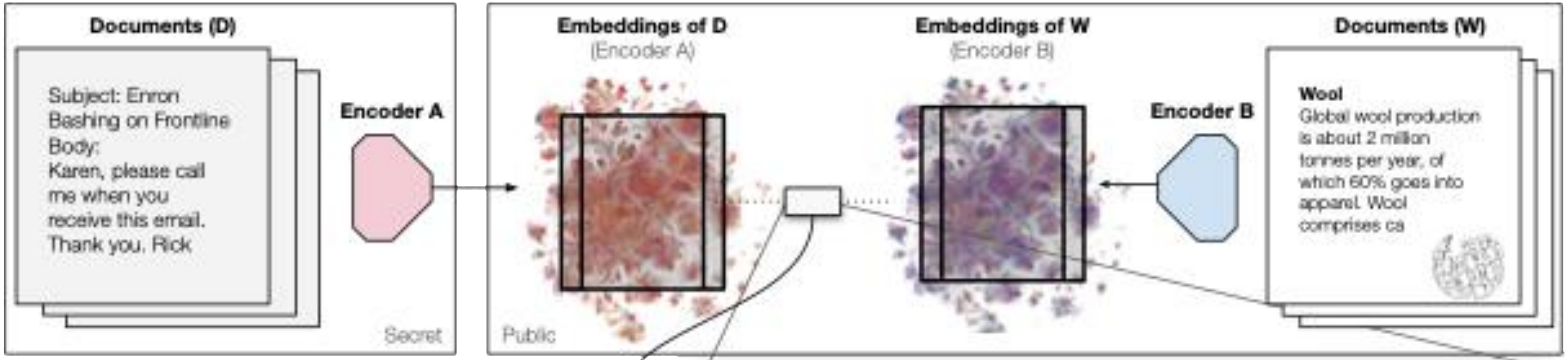


Inspiration: We adapt this method *to text* and use a different neural architecture.

Our approach: vec2vec



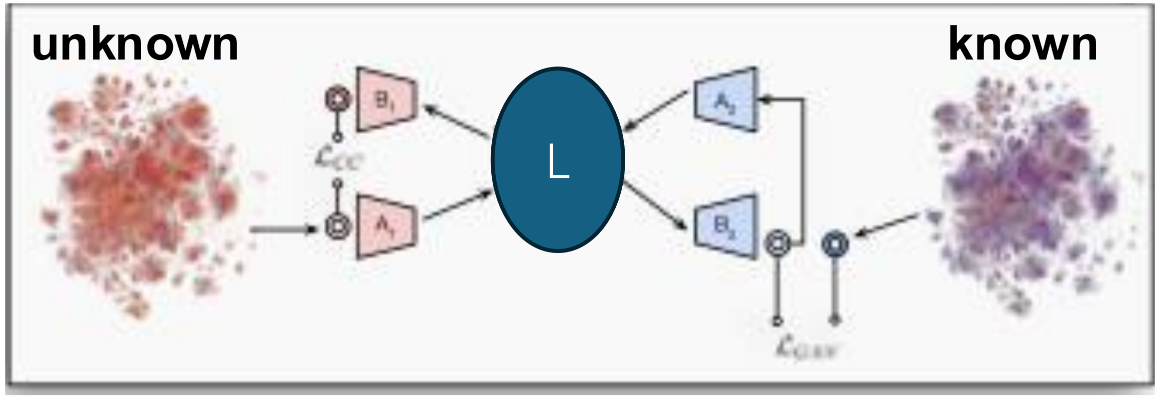
Our approach: vec2vec



Recall, we're given:

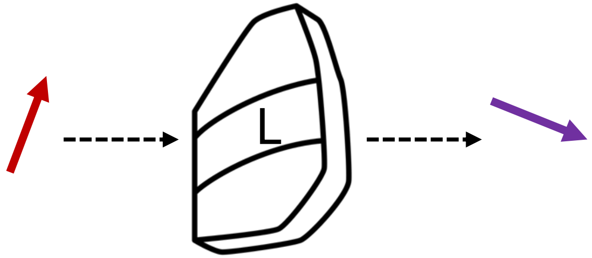
1. A set of "leaked" embeddings from an unknown encoder and unknown documents,
2. A known encoder and known (unmatched) documents.

Our approach: vec2vec



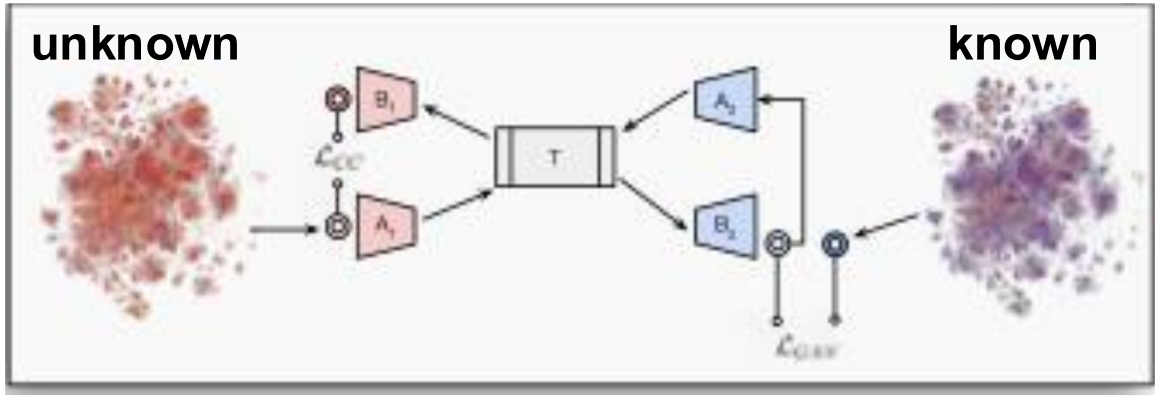
We then attempt to learn a **shared representation,** L, between the sets.

Our approach: vec2vec



L is our Rosetta Stone between embedding spaces.

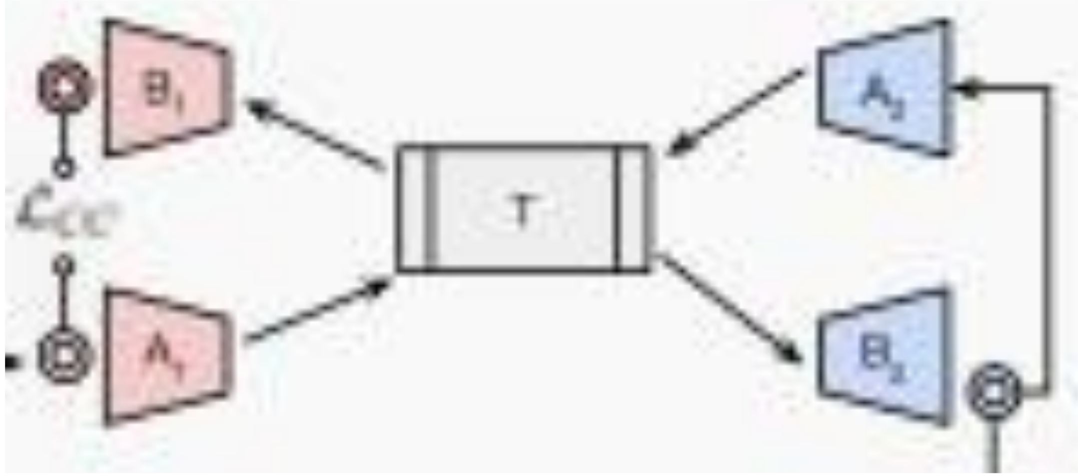
Our approach: vec2vec



Architecturally:

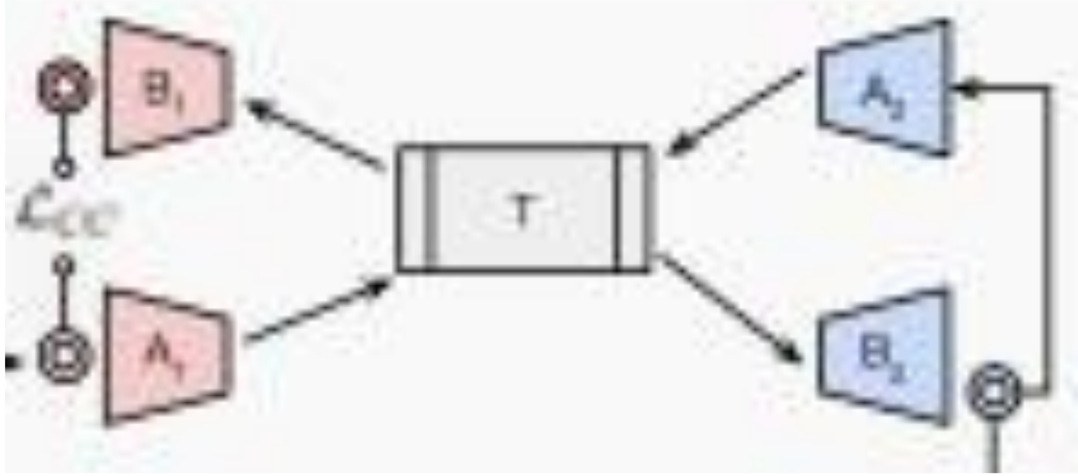
1. For each embedding space we have an input adapter A and an output B,
2. **(Important)** Input adapters share some weights T.

Our approach: vec2vec



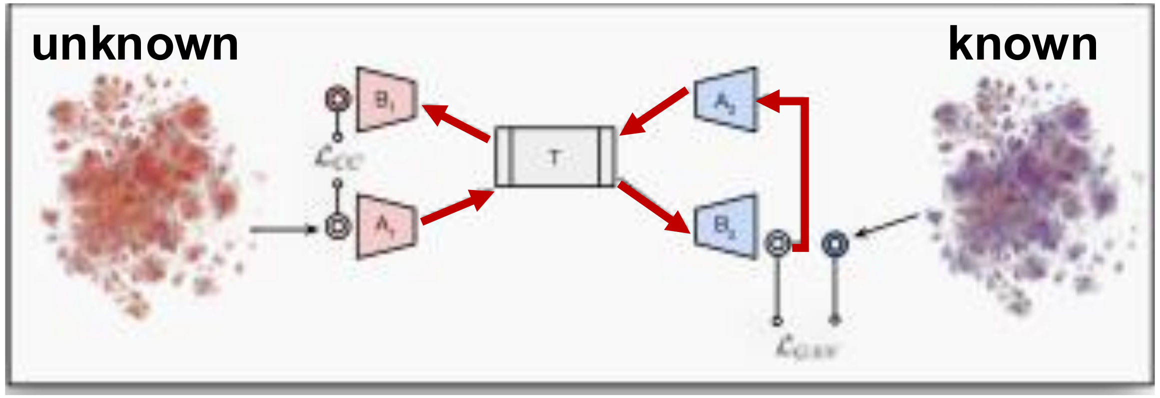
Shared weights T attempt to ensure both sets use the same parameters to encode similar semantics.

Our approach: vec2vec



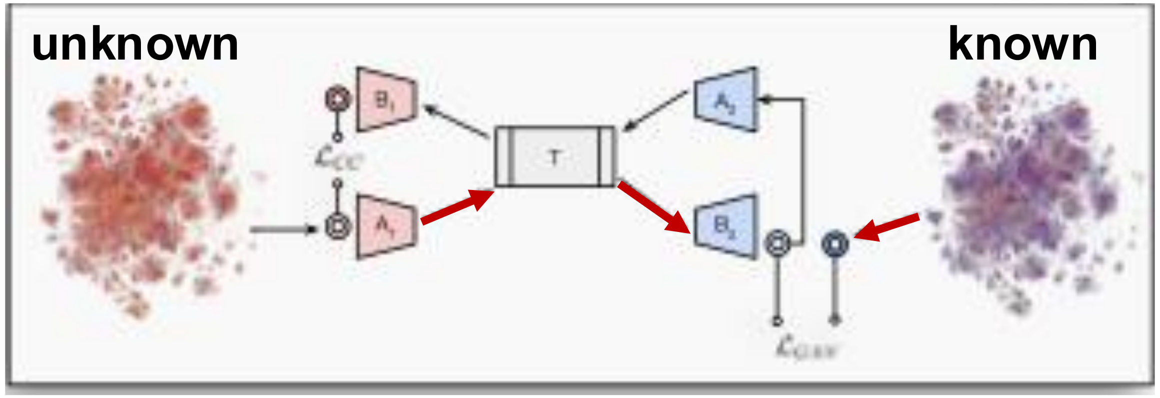
Each component is a *residual MLP* (read. standard neural network).

Our approach: vec2vec



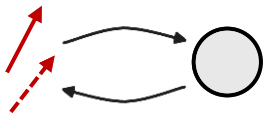
CycleGAN: compare an embedding with its out-and-back translation.

Our approach: vec2vec

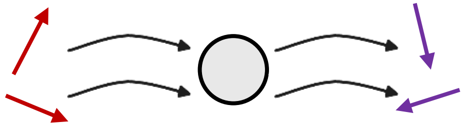


Cycle**GAN**: “Discriminator” compares *distributions* of out-translations and target embedding space.

Our approach: vec2vec



Reconstruction



Vector Space Preservation

We also include a few other structure-preserving losses.

Experiments

Model	Params (M)	Backbone	Year	Dims
[47] gtr	110	T5	2021	768
[50] clip	151	CLIP	2021	512
[58] e5	109	BERT	2022	768
[32] gte	109	BERT	2023	768
[68] stella	109	BERT	2023	768
[14] granite	278	RoBERTa	2024	768
[12] qwen3	4000	Qwen	2025	2048

We run experiments with 7 different encoders, each trained with **different algorithms and data…**

Experiments

Model	Params (M)	Backbone	Year	Dims
[47] gtr	110	T5	2021	768
[50] clip	151	CLIP	2021	512
[58] e5	109	BERT	2022	768
[32] gte	109	BERT	2023	768
[68] stella	109	BERT	2023	768
[14] granite	278	RoBERTa	2024	768
[12] qwen3	4000	Qwen	2025	2048

… vastly different parameter sizes…

Experiments

Model	Params (M)	Backbone	Year	Dims
[47] gtr	110	T5	2021	768
[50] clip	151	CLIP	2021	512
[58] e5	109	BERT	2022	768
[32] gte	109	BERT	2023	768
[68] stella	109	BERT	2023	768
[14] granite	278	RoBERTa	2024	768
[12] qwen3	4000	Qwen	2025	2048

… model architectures…

Experiments

Model	Params (M)	Backbone	Year	Dims
[47] gtr	110	T5	2021	768
[50] clip	151	CLIP	2021	512
[58] e5	109	BERT	2022	768
[32] gte	109	BERT	2023	768
[68] stella	109	BERT	2023	768
[14] granite	278	RoBERTa	2024	768
[12] qwen3	4000	Qwen	2025	2048

… vintages…

Experiments

Model	Params (M)	Backbone	Year	Dims
[47] gtr	110	T5	2021	768
[50] clip	151	CLIP	2021	512
[58] e5	109	BERT	2022	768
[32] gte	109	BERT	2023	768
[68] stella	109	BERT	2023	768
[14] granite	278	RoBERTa	2024	768
[12] qwen3	4000	Qwen	2025	2048

… and even embedding dimensionalities!

Experiments

Model	Params (M)	Backbone	Year	Dims
[47] gtr	110	T5	2021	768
[50] clip	151	CLIP	2021	512
[58] e5	109	BERT	2022	768
[32] gte	109	BERT	2023	768
[68] stella	109	BERT	2023	768
[14] granite	278	RoBERTa	2024	768
[12] qwen3	4000	Qwen	2025	2048

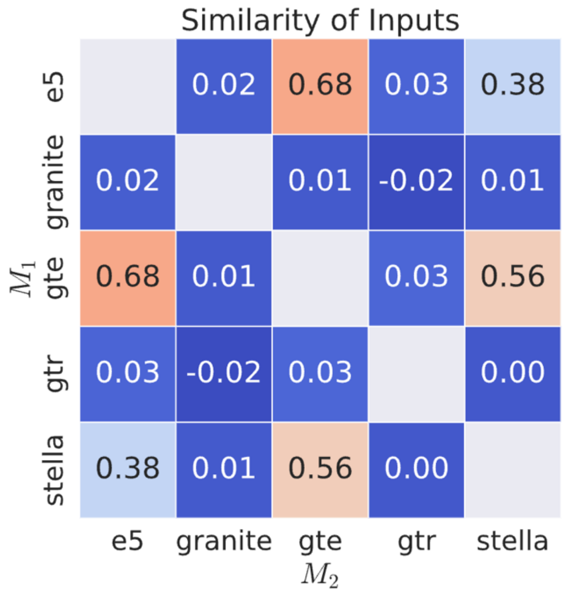
Some of the encoders are multimodal…

Experiments

Model	Params (M)	Backbone	Year	Dims
[47] gtr	110	T5	2021	768
[50] clip	151	CLIP	2021	512
[58] e5	109	BERT	2022	768
[32] gte	109	BERT	2023	768
[68] stella	109	BERT	2023	768
[14] granite	278	RoBERTa	2024	768
[12] qwen3	4000	Qwen	2025	2048

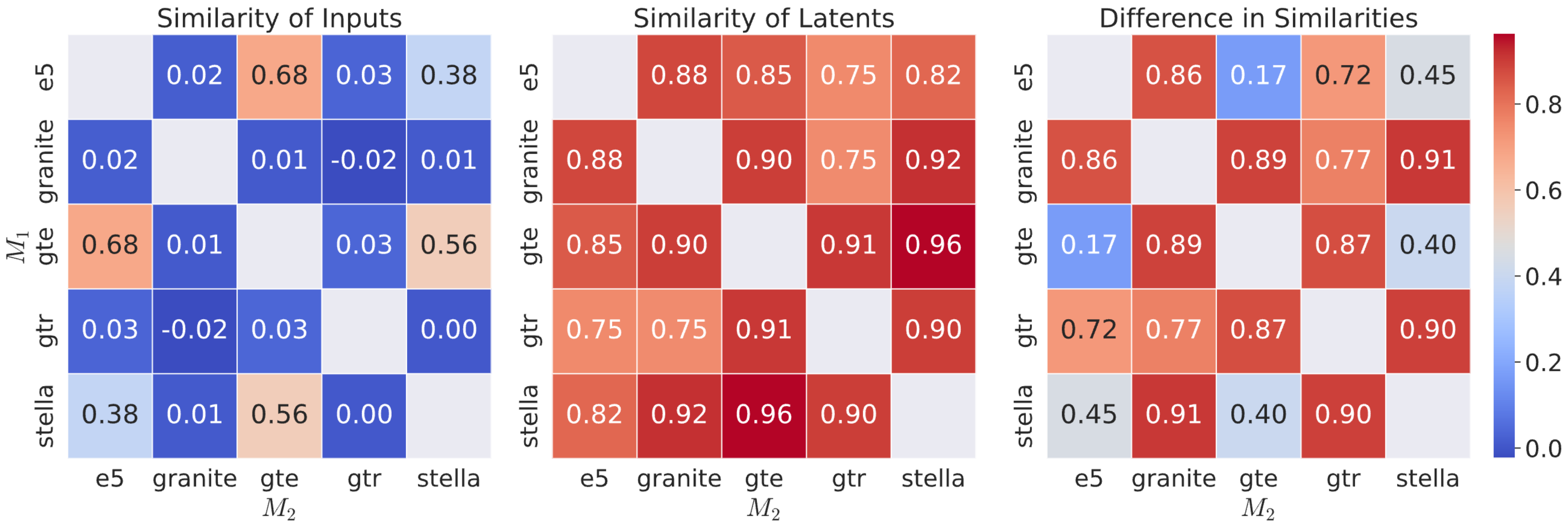
and others multilingual.

Experiments



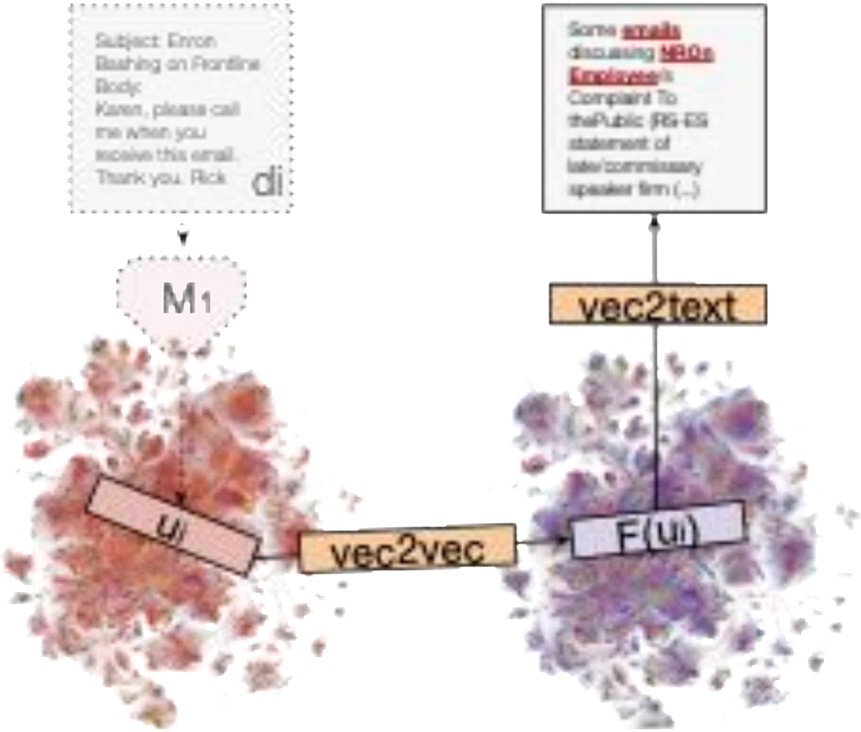
Numerically, embeddings of different architectures produce very different vectors.

Universal language of embeddings



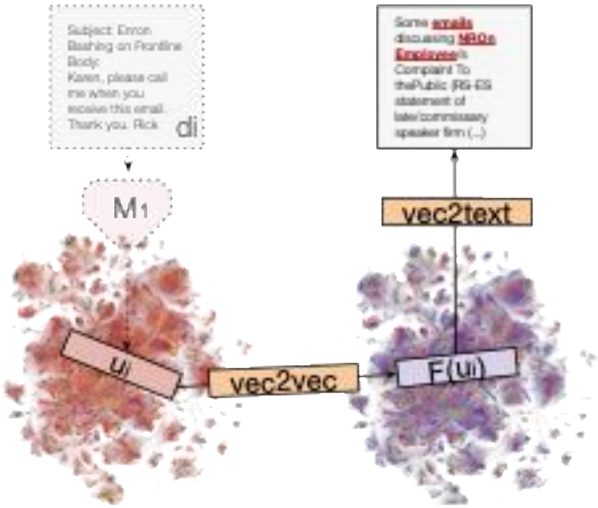
But, **in latent space,** converge to *very similar representations*!

Inverting unknown embeddings



Circling Back: The translations maintain critical semantic information.

Inverting unknown embeddings



Ground Truth: “Subject: Enron Bashing on Frontline \n Body:..."
Generation: “Some emails discussing NROn Employee/s Complaint To thePublic ..."

Ground Truth: “Subject: Trades for 3/1/02 \n Body: \n John, \n The following trades..."
Generation: “... future transactions may await John G..."

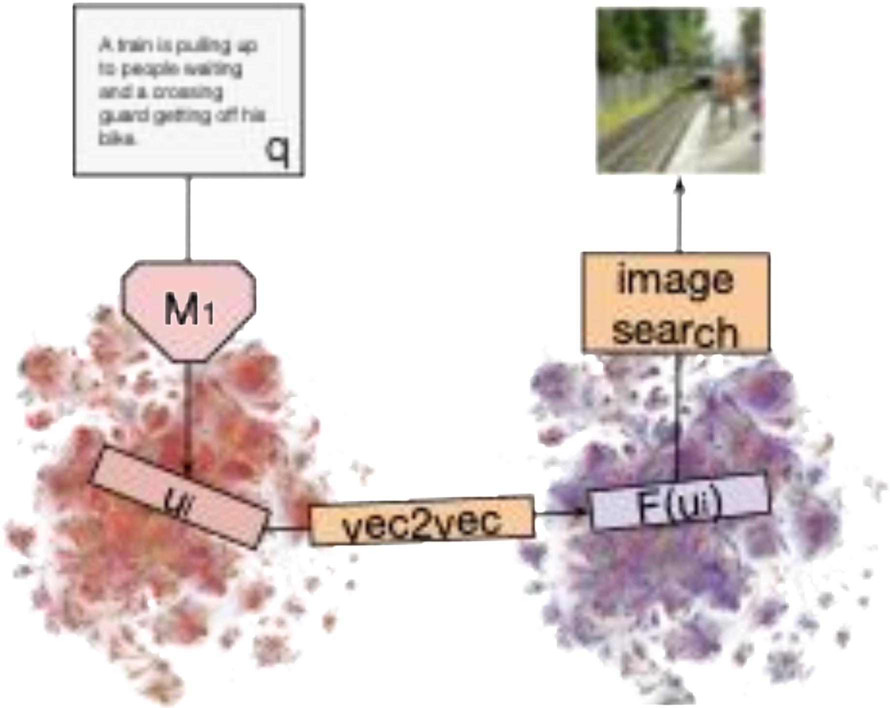
Ground Truth: “ The following expense report is ready for approval..."
Generation: “ The upcoming expense statement from YYYY MM Dec..."

And leak sensitive information when inverted!

Conjecture: If you can retrieve, we can invert!

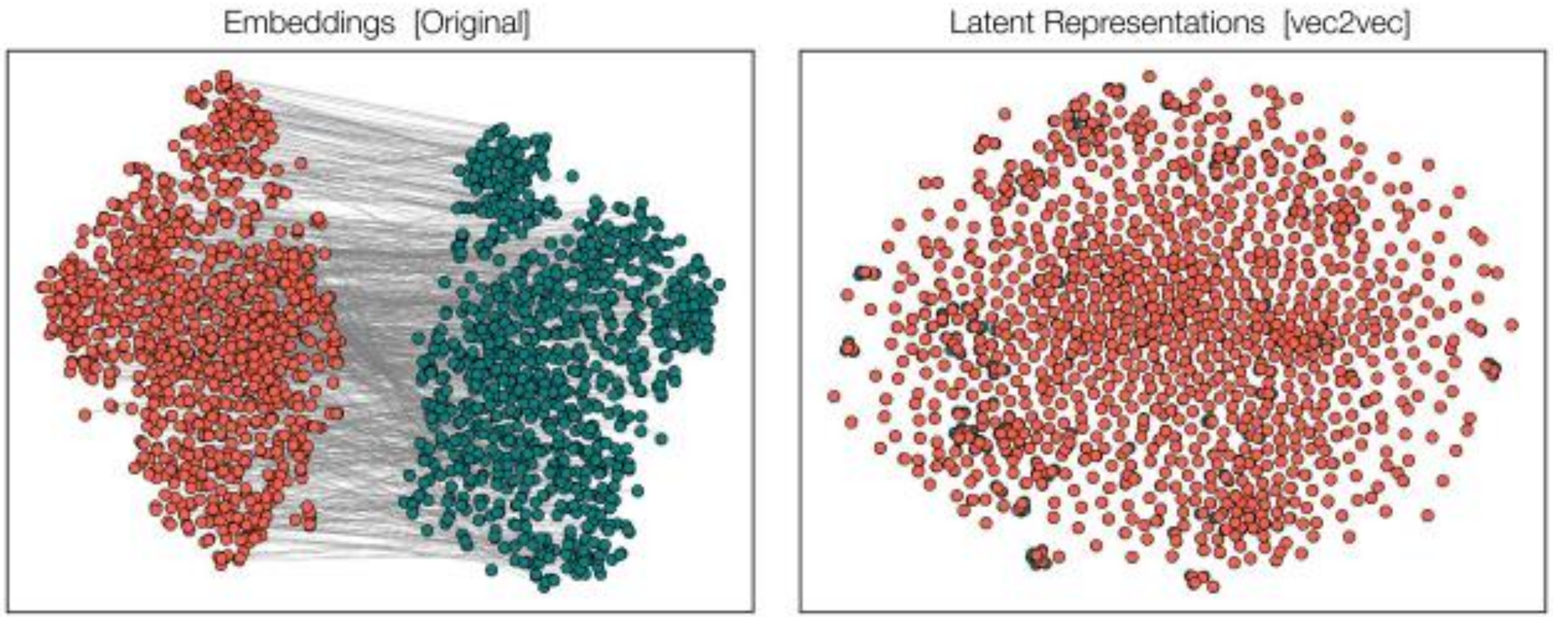
Intuition: Using homomorphic encryption, many proposals for **encrypted embeddings** look to preserve search. Search requires comparing encrypted embeddings for similarity, which is **all we need** for vec2vec!

Modality "stitching"

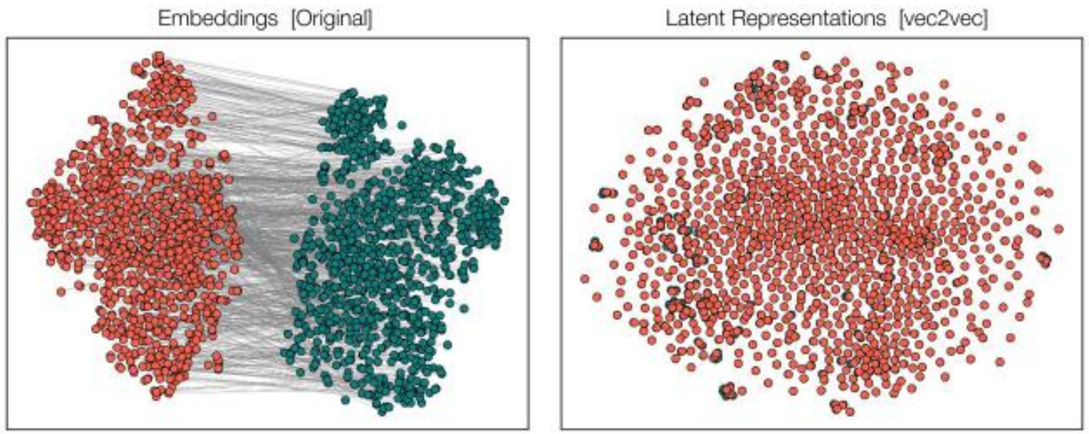


We can even "stitch" additional modalities onto unimodal models via translation!

Universal language of embeddings

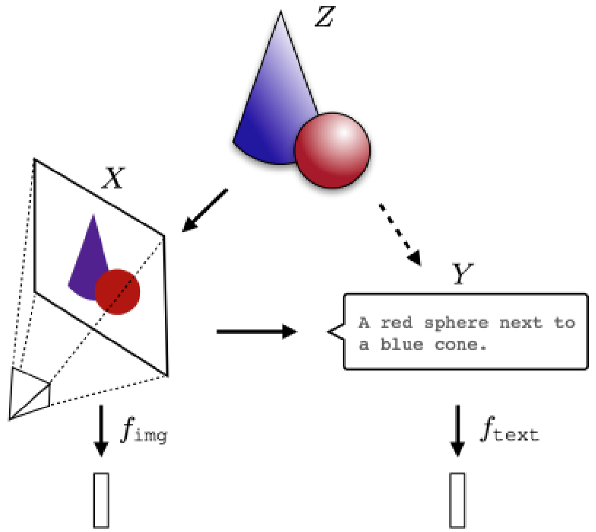


Finding: Not only do universal representations exist, but we can characterize and use them!



***Strong* Platonic representation hypothesis:** "The universal latent structure of text representations can be learned and, furthermore, harnessed to translate representations from one space to another without any paired data or encoders."

Is semantic structure universal?

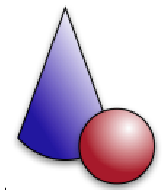


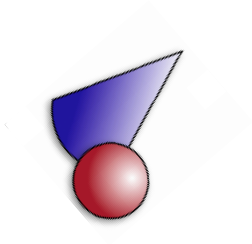
Platonic Representation Hypothesis [Huh et al., 2024]:
"Neural networks, trained with different objectives on different data and **modalities**, are converging to a shared statistical model of reality in their representation spaces."

We **conjecture** that the *Strong* Platonic Representation Hypothesis is true with embeddings of *all modalities*… we’re yet to show this!

Gameplan

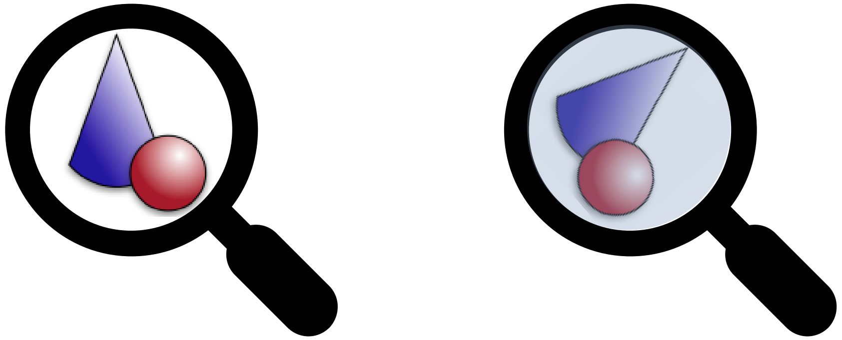
1. Background: what are embeddings?
2. vec2text: How much information do embeddings leak?
3. vec2vec: Translating embeddings with no help
4. **Conclusion**





So, are all AI models the same?

Each encoder we tested reduced to vec2vec's universal latent space: old, new, big, small, different architectures, different dimensions, and different training recipes.



Yet, each encoder has vastly different performance!

Interpretation: Each encoder is a lens onto the Platonic structure of semantics—some lenses capture the world in sharper focus, others in blurrier or more distorted form, but they seem to observe the same reality.

Ad astra per aspera

More stable translation methods

- GANs are brittle and finicky

More modalities: images, audio, …

Translate and invert encrypted embeddings

Translate internal representations of LLMs

Translate across languages

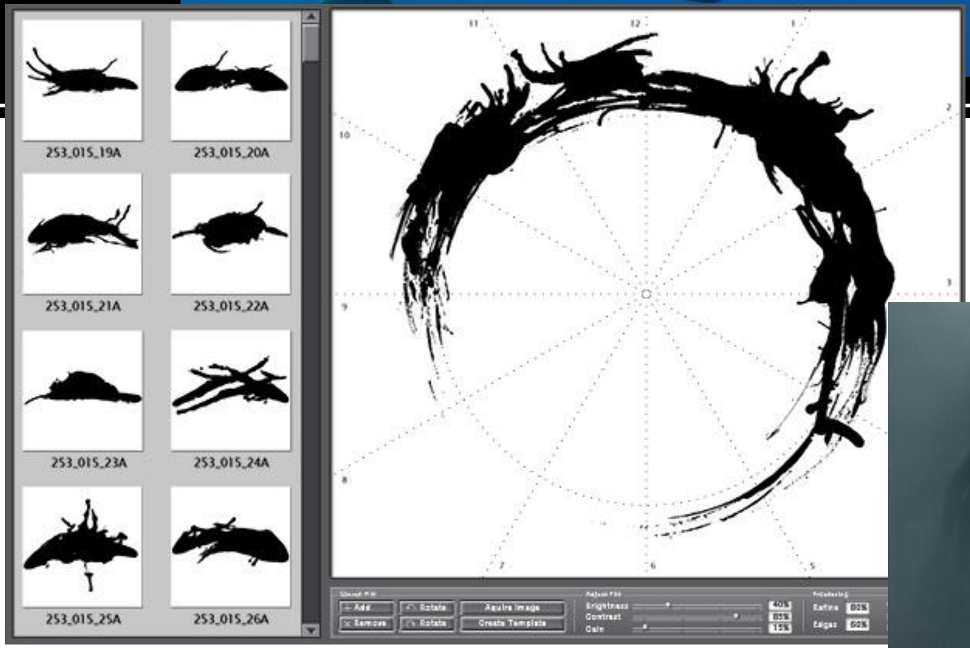
Characterize the universal geometry of meaning?


Tech

Scientists Have Reported a Breakthrough In Understanding Whale Language

By Jordan Pearson December 7, 2023, 11:19am

source: Vice



source: Wolfram

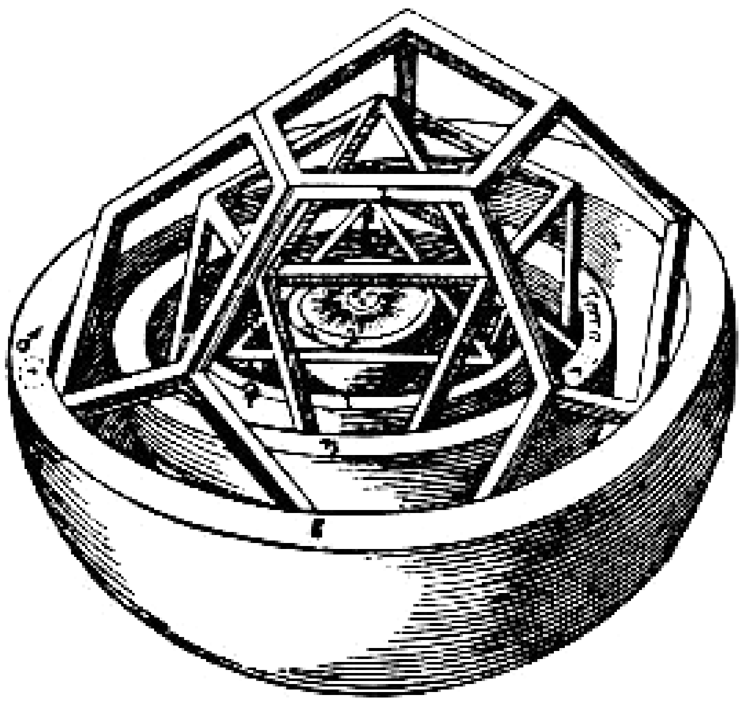
Text Embeddings Reveal (Almost) As Much As Text

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, Alexander M. Rush
Department of Computer Science
Cornell University

Harnessing the Universal Geometry of Embeddings

Rishi Jha **Collin Zhang** **Vitaly Shmatikov** **John X. Morris**
Department of Computer Science
Cornell University

QUESTIONS?



Kepler's celestial geometry of Platonic solids