Vector Embeddings

The Emerging Language of AI

Pawel Zimoch
CTO, Featrix
pawel@featrix.ai

Who am I?

CTO @ Featrix

On a mission to make accurate, high-quality data modeling available to everyone, on any data.

10+ years of experience in AI/ML

Research in physics and thermodynamics, then worked on probabilistic models for language and tabular data at a DARPA-funded startup.

Software Engineer and Developer

Built apps and systems for distributed inference

What is Featrix?

AI company focused on embeddings

We bring our own embeddings system to AI problems.

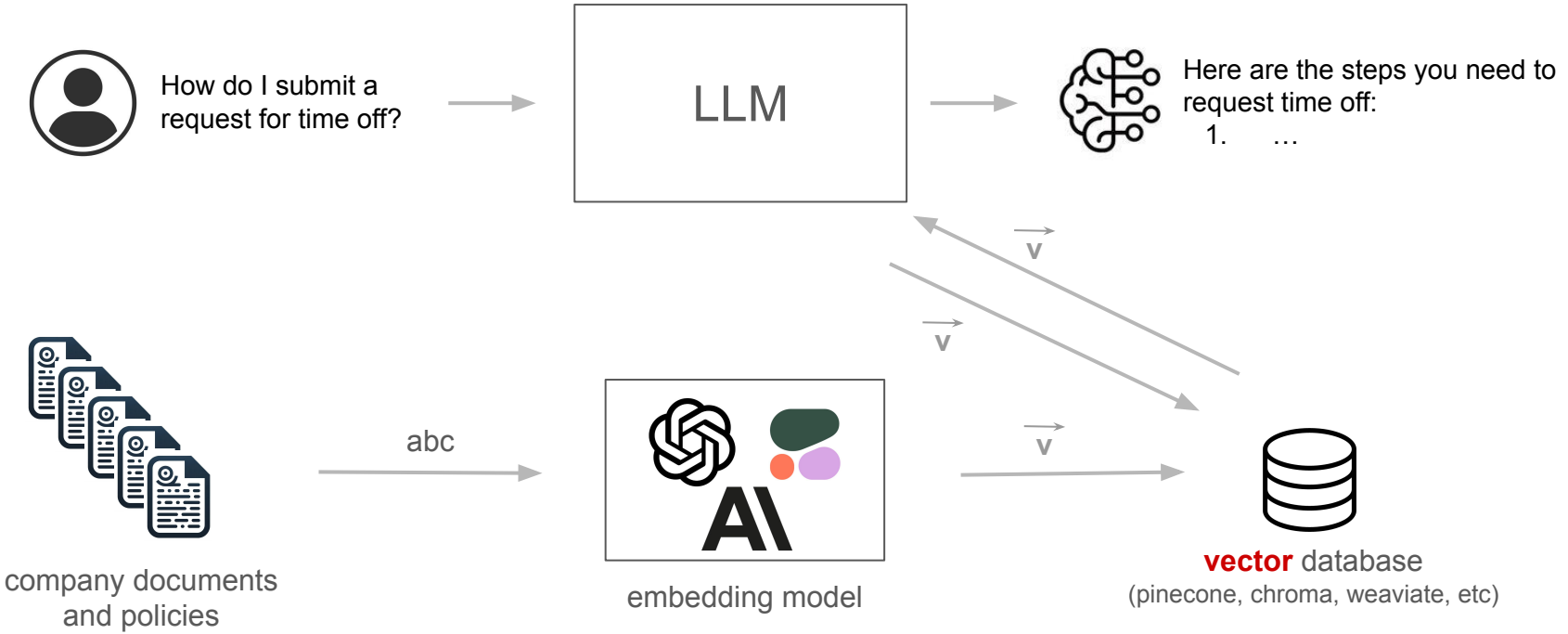
The embeddings approach lets us drastically simplify AI problems that used to be very difficult.

Contact us for:

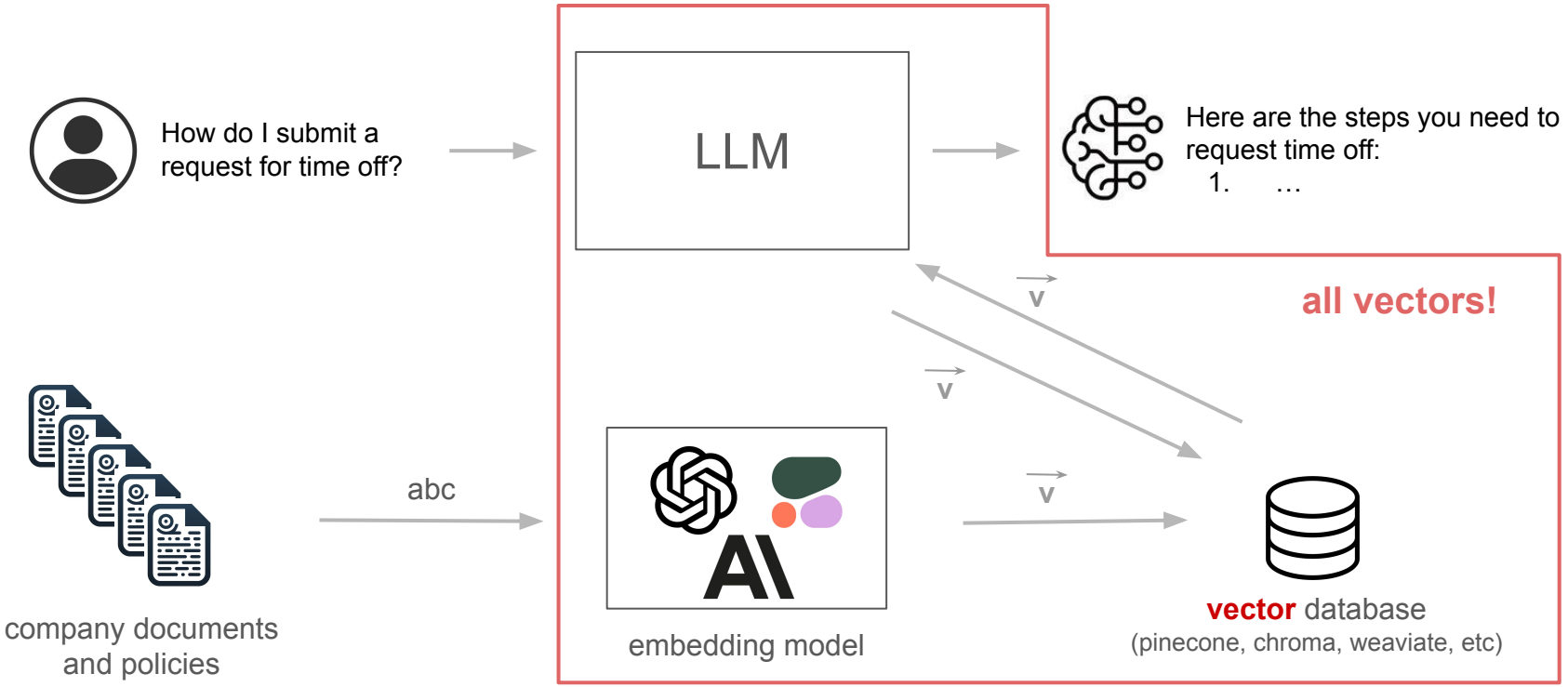
- “I have some data; can I do AI on it?”
- “I have a bunch of sales contacts; which are most like deals I have closed before?”
- “I need to do cutting edge AI but don’t have the budget for a giant AI team and infrastructure.”

pawel@featrix.ai

Retrieval-augmented generation (RAG)



AI systems communicate using vectors



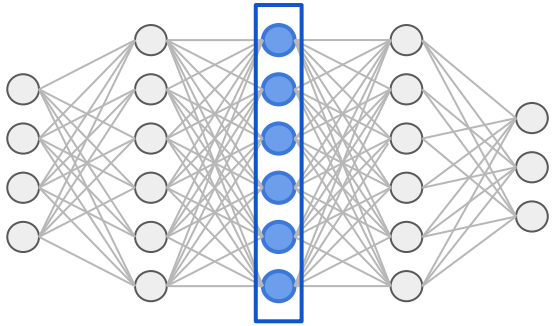
Vector embeddings are so much more than just RAG

list of numbers

[0.12, 0.52, 0.91, -0.23, …, 0.05, 0.87]

arrows

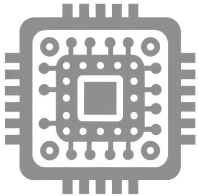
layers of a neural network



very fast

parallelized

hardware accelerated

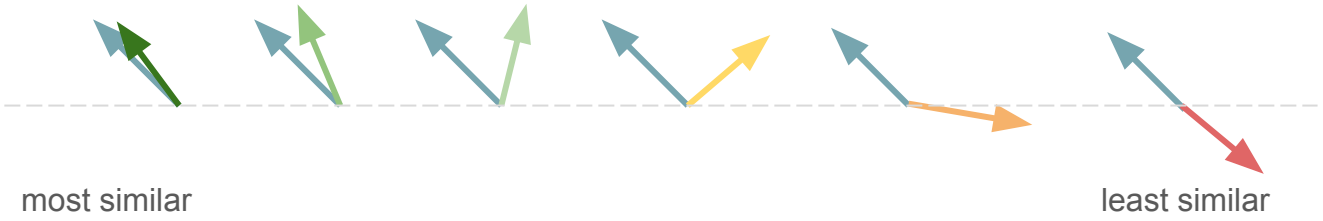


GPU

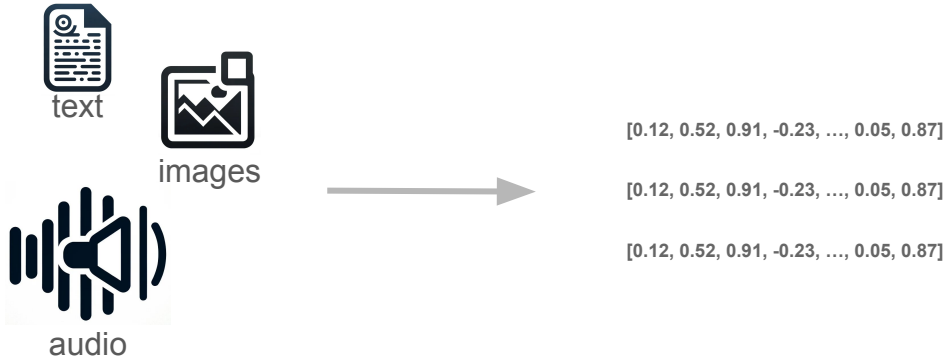
and more…

dot product

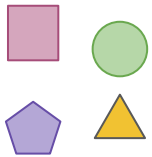
a spectrum of similarity



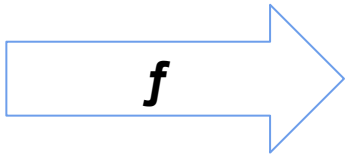
universal encoding



Embedding: a function from any set to vectors



a set of objects

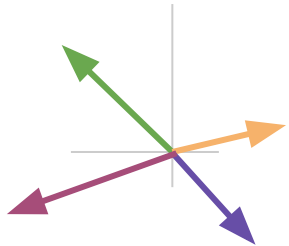


neural network

[0.12, 0.42, 0.34, …, 0.13]
[0.52, 0.38, 0.82, …, 0.61]
…

a set of vectors

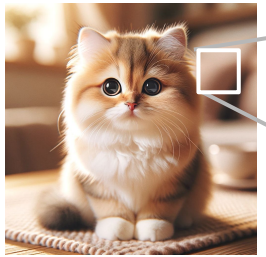
- images
- words, sentences, paragraphs
- user profiles, ecommerce catalogs
- movies, songs
- integers, floats, lists, dictionaries
- ***anything!***



“embedding space”

So what?

How data is represented affects which computations are easy



[145, 245, 1] [98, 198, 57] [67, 122, 178]
[123, 233, 15] [101, 189, 50] [34, 167, 155]
[113, 256, 8] [85, 167, 45] [23, 135, 189]

This is what an RGB image “looks like” to a computer

make a color histogram

individual values

EASY

adjust contrast

linear combinations

MEDIUM

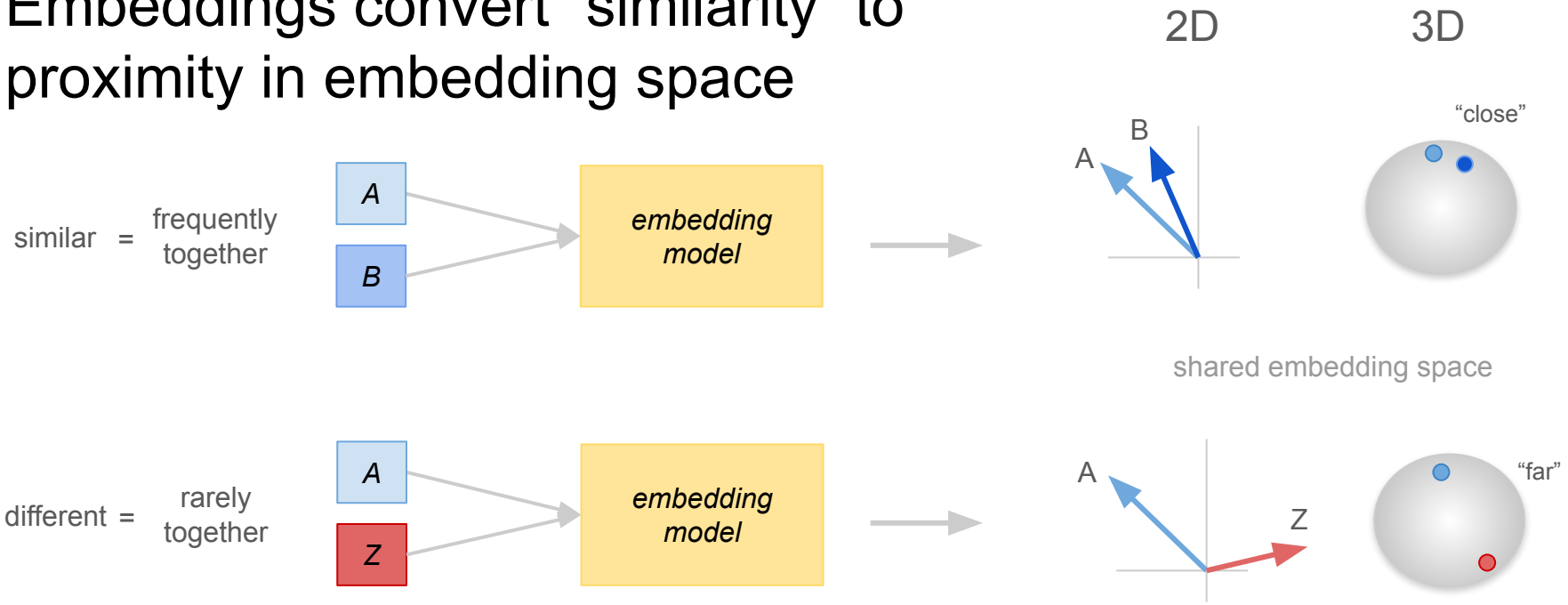
move the cat left

non-linear combinations

HARD

computational barrier

Embeddings convert “similarity” to proximity in embedding space

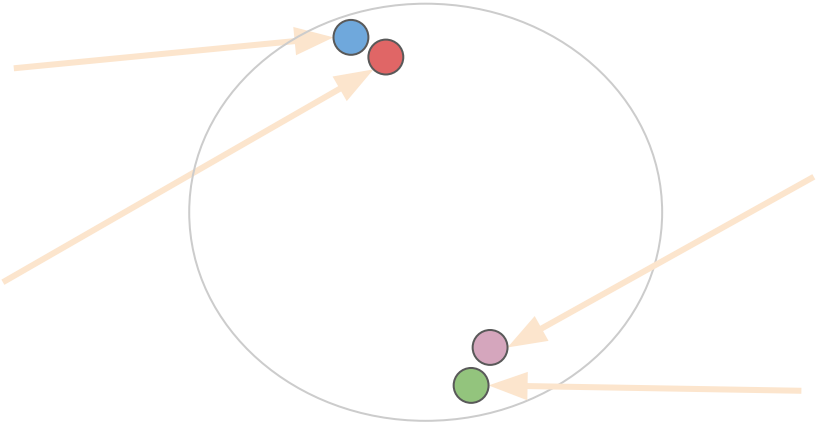


A and B are similar ⟹ their embeddings are similar (close)

A and B are different ⟹ their embeddings are different (far)

Embeddings convert “similarity” to proximity in embedding space

“A picture of a cat sleeping in bed”



shared embedding space

“A cute dog reading a magazine and wearing glasses”

Embeddings: Hard things get easier

Embeddings let us bust through the barrier with low effort.

EASY

make a color histogram

individual values

EASY

adjust contrast

linear combinations

MEDIUM

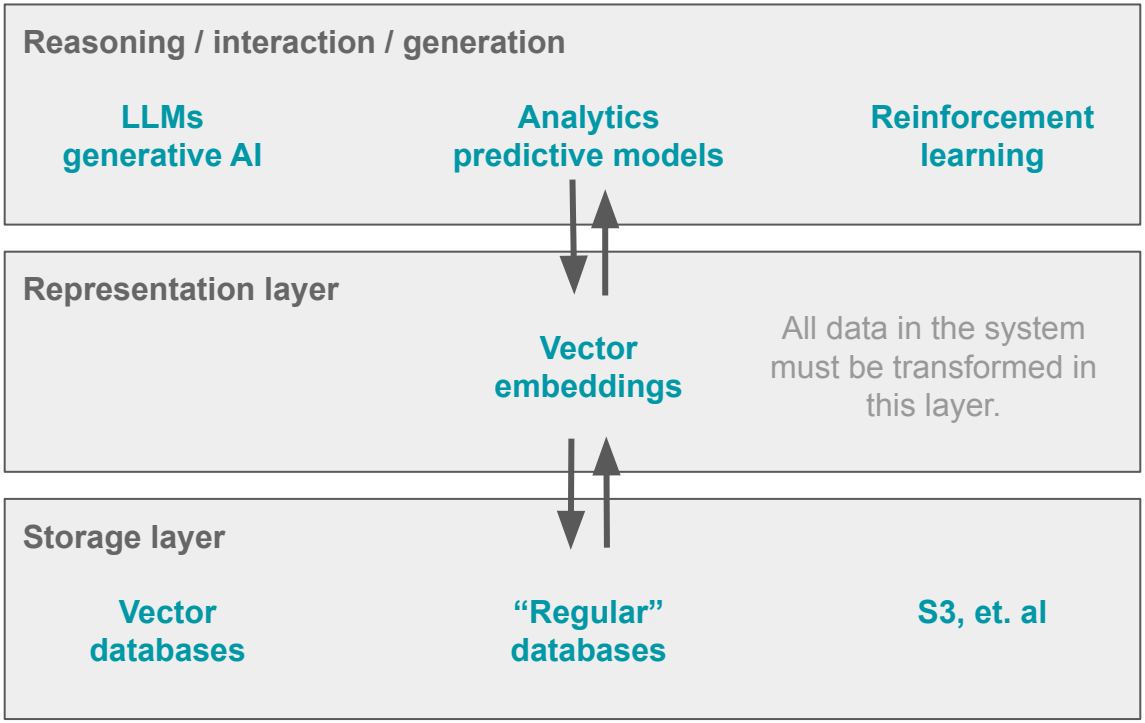
move the cat left

non-linear combinations

~~HARD~~

old computational barrier

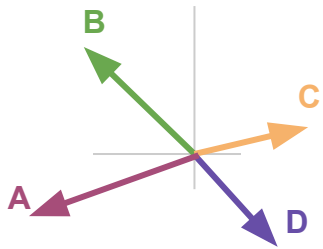
The emerging vector-based computation stack for AI



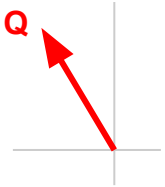
The best mental models for embeddings

(Pawel’s version)

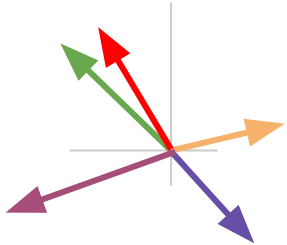
1/4 A trainable search index for a database



items in the database



query

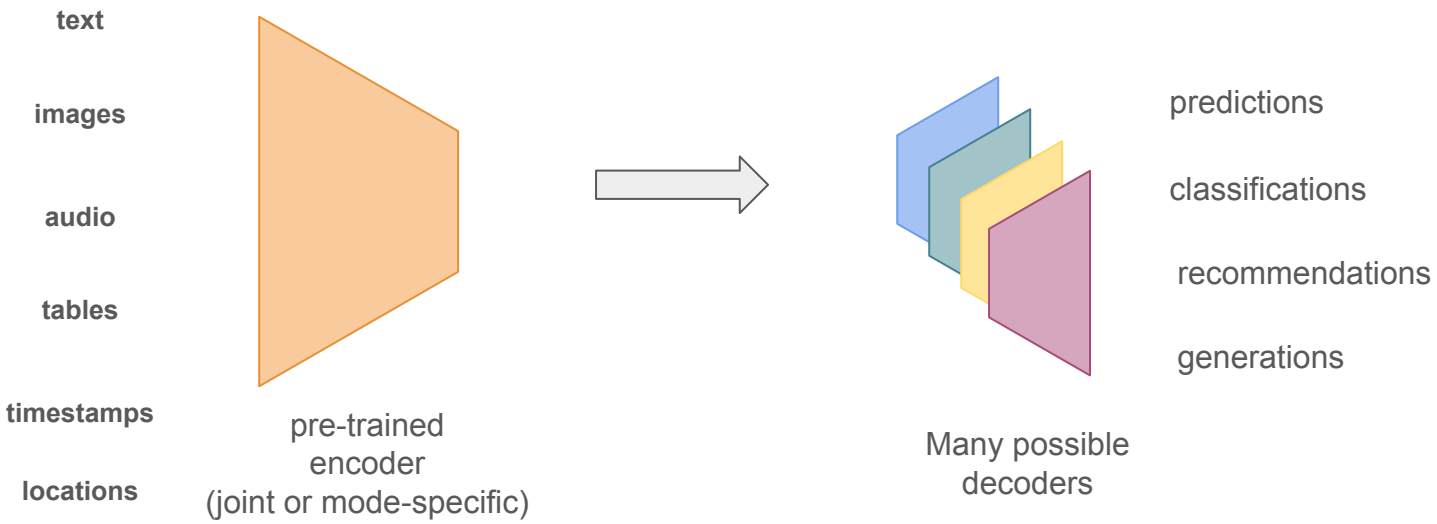


results

	similarity to Q
1. record **B**	0.95
2. record **A**	0.6
3. record **C**	0.5
4. record **D**	0.05

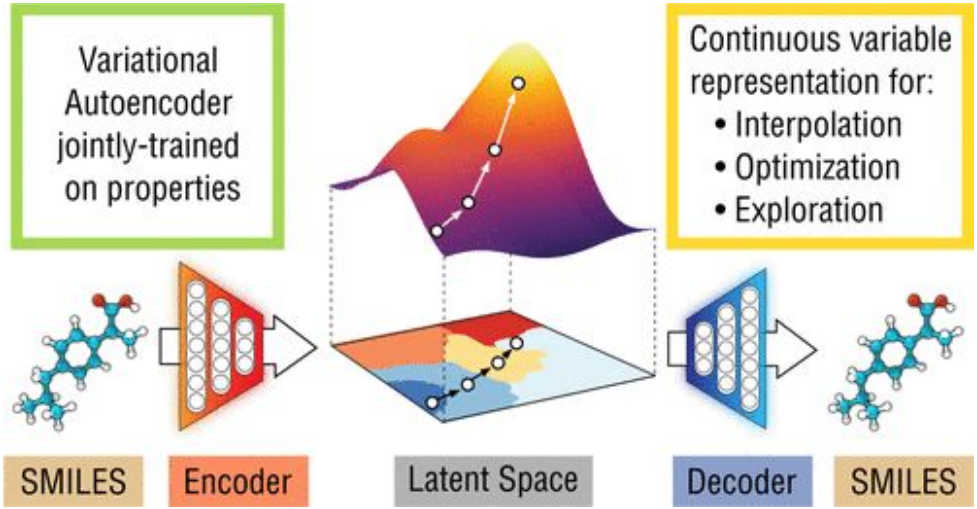
document snippets	user prompt	most relevant documents	**RAG**
movies, songs, images	user profile	recommended content	**recommendations**
store catalog	shopping cart	most likely to also buy…	**sales/marketing**

2/4 An interface for composing neural networks



3/4 Dimensionality reduction

what's in here? ***what can I do with this?*** ***what else is possible?***



Gómez-Bombarelli et. al. *Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules.*
ACS Cent. Sci. 2018, 4, 2, 268–276

4/4 A trainable abstraction - multimodal embeddings

“A picture of a cat sleeping in bed”

embedding space

abstraction barrier

4/4 A trainable abstraction - train / query asymmetry

reconstruction

Image embedding model

Diffusion model

image embedding

training

query

“a cute dog reading a magazine and wearing glasses”

text embedding model

text embedding

Computational barriers are not unique to unstructured data

timestamp	user_id	transaction_id	product_id	price
1699907568088	0000123124423	a3df-a4d2-123a	0000000351	$158.32
1699907592034	0000123124672	a340-dfsd-1203	0000000103	$35.99

sum up order value by customer

EASY

compute revenue for Q3 2023

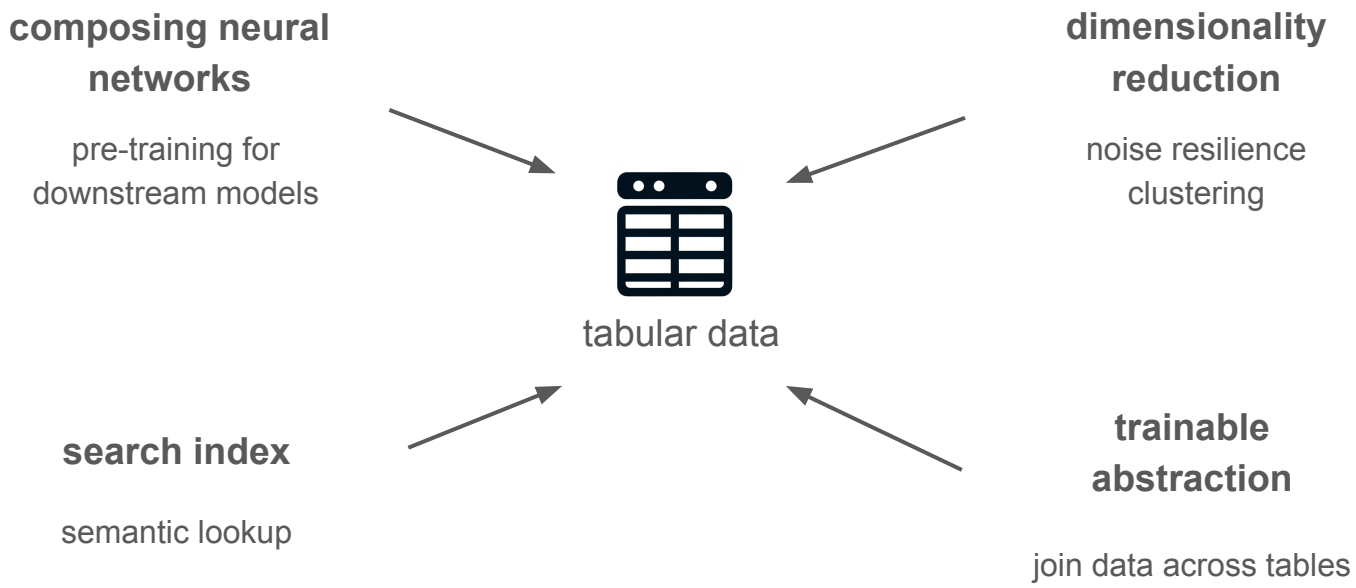
MEDIUM

computational barrier

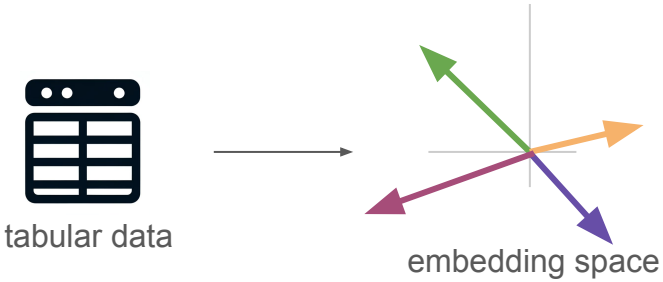
Find customers who are about to churn

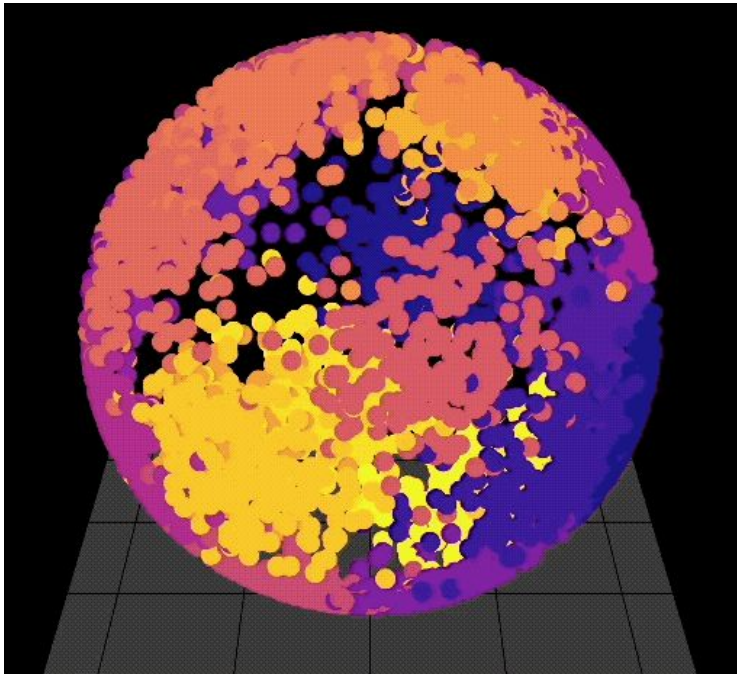
HARD

Embeddings are perfect for tabular data



Featrix: Custom embeddings for tabular data





3D visualization of the Featrix training process

Embeddings

the emerging language for AI

Vector embeddings are the *lingua franca* of AI

Embeddings encode the notion of “similarity”

Off-the-shelf embeddings are available for most common data types

The scope for computation has increased by A LOT because of lowering computational barriers

AI Primer poster giveaway!

Get up to speed on the most common AI terms and buzzwords

Get a PDF or a free 22” x 28” hard copy

https://www.featrix.ai/poster