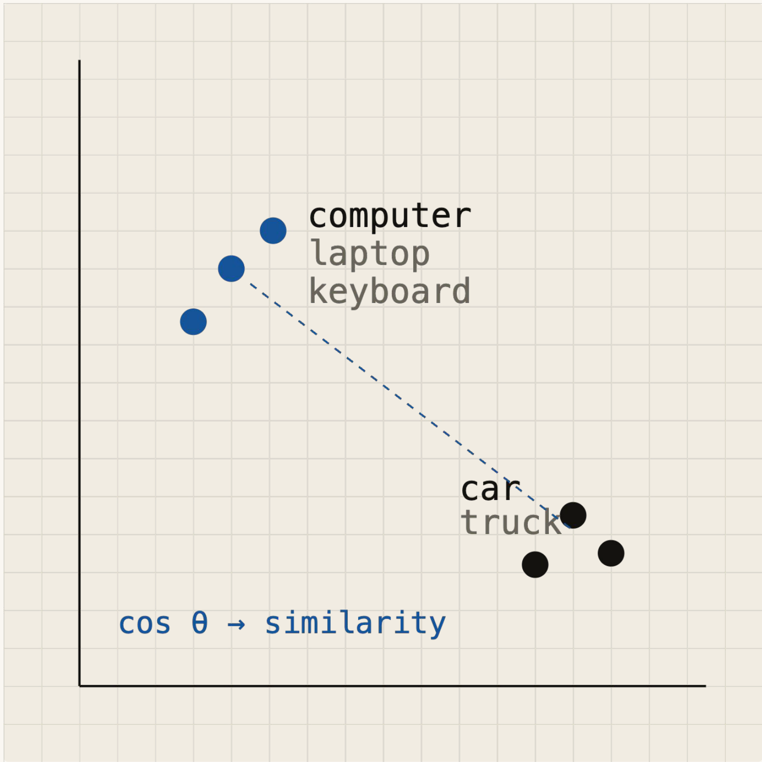
A PRIMER IN SIX SLIDES

Embeddings & Vector Stores.

Turning words, pixels & graphs into geometry a machine can search.



CORE IDEA

Latitude & longitude, but for meaning.

An embedding maps a piece of real-world data — a sentence, an image, a user — to a low-dimensional vector. Items with similar meaning land near each other; geometry becomes a proxy for semantics.

01 / compression

Lossy projection from high-dimensional raw data to a compact numeric vector — often 768 dims for BERT.

02 / similarity

Distance between vectors approximates semantic relatedness — measured by L2, cosine, or dot product.

03 / shared space

Joint embeddings put text, image and audio in one space so a text query can retrieve a video.

IN PRACTICE

Eight strings, ranked by meaning.

Embed eight short texts with one API call. Cosine similarity against the reference pangram reveals what the model considers “the same idea.”

```
# embed eight short texts
texts = ['The quick brown fox…', …]
resp = client.models.embed_content(
model='text-embedding-004',
contents=texts,
config=types.EmbedContentConfig(
task_type='semantic_similarity'))
```

reference · “The quick brown fox jumps over the lazy dog.”

COSINE SIMILARITY → REFERENCE 0.00 — 1.00

1	The quick brown fox jumps over the lazy dog.	1.00
2	The quick rbown fox jumps over the lazy dog.	0.98
3	a quick brown fox jmps over lazy dog.	0.93
4	brown fox jumping over dog	0.86
5	teh fast fox jumps over the slow woofer.	0.84
6	The five boxing wizards jump quickly. — alt. pangram	0.71
7	fox > dog — shorthand	0.57
8	Lorem ipsum dolor sit amet, consectetur adipiscing… — unrelated	0.38

Typos & word-order: still close. Different pangram, same letters: drifts. Lorem ipsum: far.

A TAXONOMY

Four families, one geometry.

01 · text

Text & document

Word vectors (GloVe, Word2Vec), then BoW & LSA, now BERT / Gecko-class LLM encoders.

BERT → 768d Gecko → 256-768d

02 · vision

Image & multimodal

CLIP-style dual encoders project pixels and captions into a single shared vector space.

text ⇄ image text ⇄ video

03 · tabular

Structured data

User/item factorization for recsys; PCA & autoencoders for general tabular features.

users × items user ≈ Σ item

04 · graph

Graph embeddings

DeepWalk, node2vec, GNNs — encode nodes so structural neighbours sit near each other.

social / KG fraud rings

Pick the family that matches the data — but remember: embeddings from different models live in different spaces, so versioning is non-negotiable in production.

WHAT IT ENABLES

Search that understands what you mean , not just what you typed.

Once your data lives in a shared vector space, finding the closest match takes milliseconds — even across a billion items.

the old way · keyword match

Looks for the exact words.

query

“affordable laptop”

results

✓ Affordable laptop sleeve
✗ Budget-friendly notebook
✗ Cheap computer for students

the new way · meaning match

Looks for the same idea.

query

“affordable laptop”

results

✓ Budget-friendly notebook
✓ Cheap computer for students
✓ Entry-level Chromebook

IN THE WILD

You use this every day — without knowing it.

01

Smarter search.

The search bar gets your intent — even when you fumble the words.

02

Personal recommendations.

“You might also like…” — picked because it sits near things you already loved.

03

AI that cites its sources.

Chatbots pull in the right documents before answering — fewer made-up facts.

Same technique. Three very different products.