Vector Databases: a quick intro

The Backbone of AI-Powered Search & Recommendations

ProtoComet

Why do we need a new kind of Database?

Traditional databases are great for structured data.

◆ But they struggle with images, text, audio, and unstructured data.

◆ Can't efficiently find "similar" items (e.g., "Find images like this").

💡 Vector Databases store & search data based on meaning, not just exact matches. *Making them **suitable for AI.***

ProtoComet

What is a Vector Database?

◆ Instead of storing raw text/images, vector databases store "vectors" (numeric representations of data).

◆ These vectors capture meaning & relationships between data points.

◆ When data is stored, it is indexed, to keep related data close.

◆ This allows it to find data that's similar or close in meaning or characteristics.

ProtoComet

→

Where does it excel?

Vector DBs are ideal for apps that require semantic understanding, such as:

◆ AI-powered search engines

◆ Image and face recognition

◆ Voice and audio processing

◆ Recommendation systems

◆ Fraud detection etc.

But what is a Vector?

◆A vector is simply an **array of numbers** representing a point in space.

◆A feature or characteristic of the data is a dimension in the array.

◆ For example, a 512 dimension vector (512D) is an array to store 512 data characteristics.

◆Higher number of dimensions allows a more detailed representation of the data, and greater accuracy.

◆However, higher dimensionality also increases the computational cost and storage requirements.⚠️

→

ProtoComet

Movie Recommendation Engine: A ⚠️ simplified example

For instance, your favorite movie can be described based on traits like:

- Action-packed: (8/10)
- Funny: (4/10)
- Inspiring: (2/10)

That list, say, **[8, 4, 2]** is its vector, a numeric fingerprint capturing its essence.

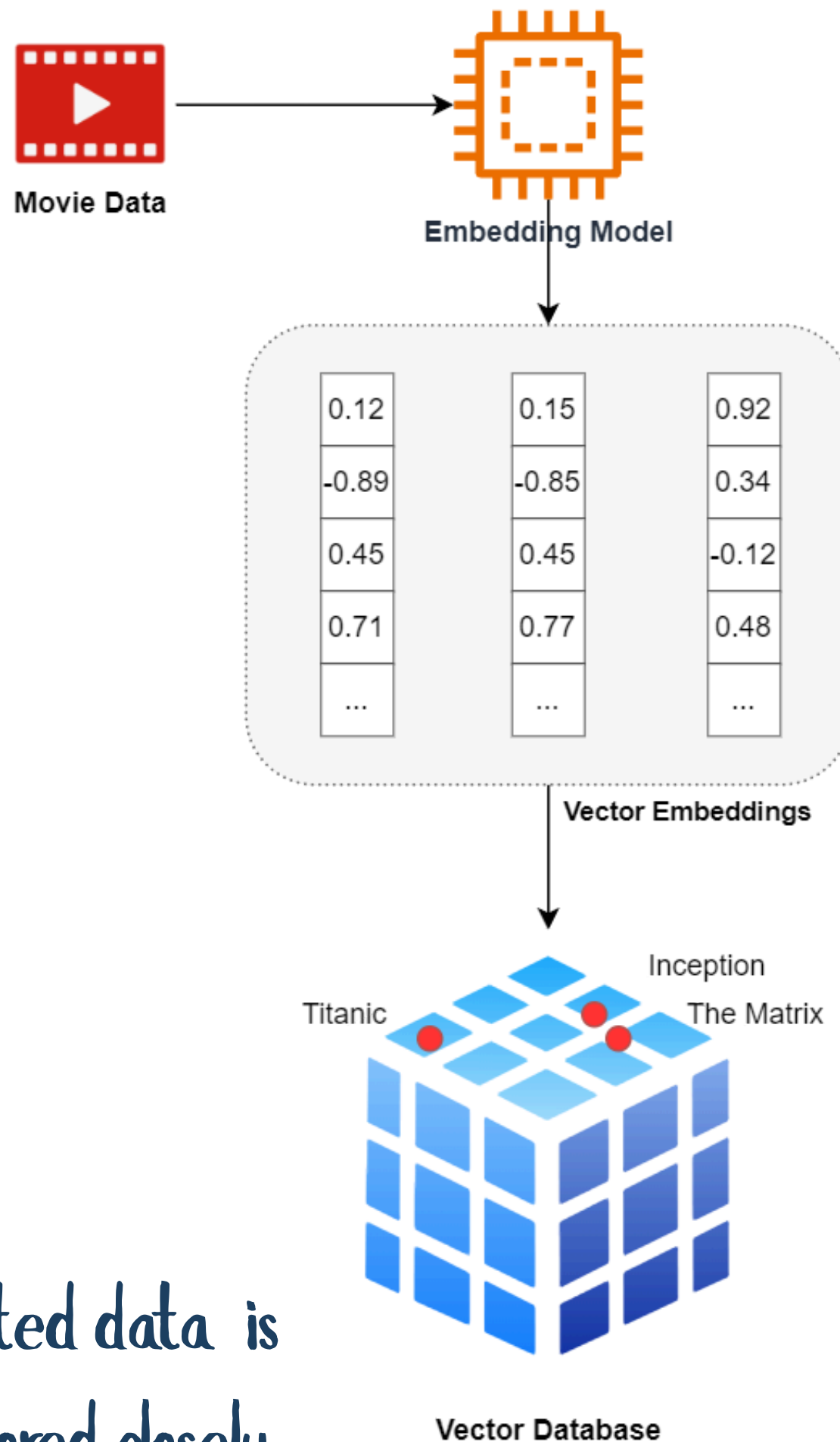💡 **In real world vector values are usually between -1 and 1. (*normalized for accuracy*)**

ProtoComet

→

... so movies could be stored something like:

MOVIE	Vector Embedding
Inception	[0.12, -0.89, 0.45, ...]
The Matrix	[0.15, -0.85, 0.40, ...]
Titanic	[0.92, 0.34, -0.12, ...]

ProtoComet

→

Simplified Visual Flow

Movie Data

Embedding Model

0.12	0.15	0.92
-0.89	-0.85	0.34
0.45	0.45	-0.12
0.71	0.77	0.48
...	...	...

Vector Embeddings

Titanic   Inception   The Matrix

Vector Database

notice how related data is indexed and stored closely

ProtoComet

... and when it is searched

Imagine you're looking for movies like "Inception".

AI converts the query (Inception movie) to a vector like:
 → [0.12, -0.89, 0.45, 0.71, -0.23, ...]

and see what the close vectors are in the database.

🟢 Matrix, Interstellar → Close vectors (similar)

🔴 Titanic → Distant vector (not similar)

🚀 This is how AI understands similarity beyond words!

Some Popular Vector DBs

- Pinecone

- Chroma

- Qdrant

- Weaviate

- Milvus

- FAISS (Facebook AI Similarity Search) - a library

- Annoy (Approximate Nearest Neighbors Oh Yeah)

→

ProtoComet

Recap

ML models convert data (text, images, etc.) into **vector embeddings**

These embeddings are stored in a **vector database**

Searching happens using **similarity measures** (*Cosine, Euclidean, Dot Product*)

Fast.
Scalable.
AI-Native 🚀

ProtoComet

Ready to dive into the future of data?

Vector databases are changing the game for AI.

What's your next big idea?

let's talk it out ☆

ProtoComet