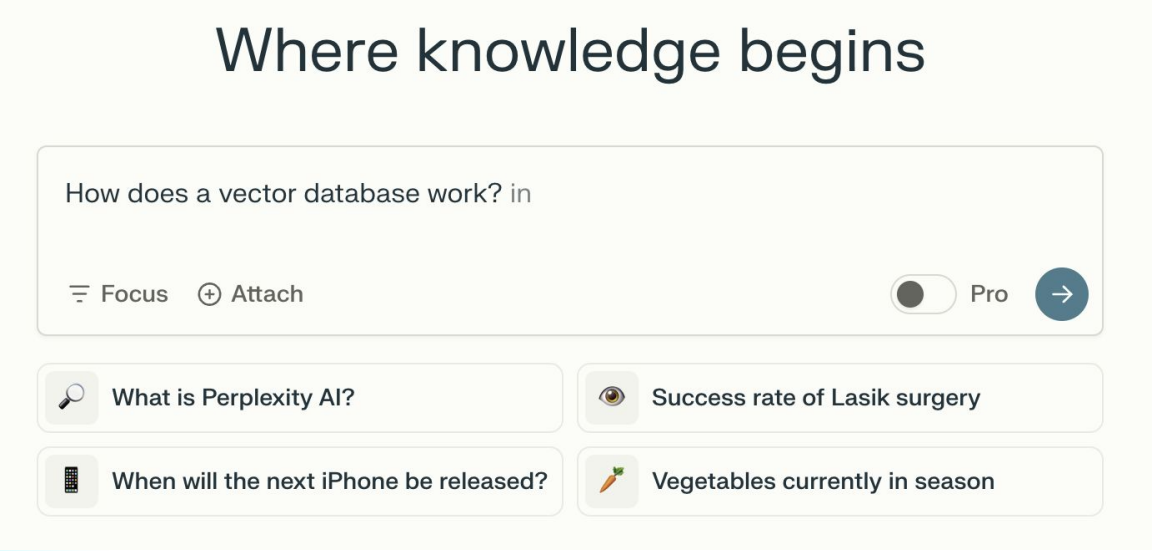
Stefan Webb

Developer Advocate, Zilliz

stefan.webb@zilliz.com
https://www.linkedin.com/in/stefan-webb
https://x.com/stefan_webb

Searching the Web with Gen AI



How does a vector database work?

Sources

What is a Vector Database? - Elastic
elastic · 1

What is a Vector Database & How Does it Work? Use Cases + ...
pinecone · 2

What are Vector Databases and How Do They Work? - Zilliz blog
zilliz · 3

View 4 more

Perplexity

A vector database works by storing, indexing, and searching data represented as high-dimensional vectors. Here's an overview of how vector databases function:

Vector Representation

Vector databases store data as vector embeddings - numerical representations of data objects like text, images, or audio [1]. These embeddings are created using machine learning models that capture semantic relationships between data points [5].

Indexing

Vector databases use specialized indexing algorithms to organize the vector data for efficient retrieval:

- Algorithms like HNSW (Hierarchical Navigable Small World), LSH (Locality-Sensitive Hashing), or PQ (Product Quantization) are used to index vectors [3].
- Indexing maps vectors to data structures that enable fast similarity searches [5].

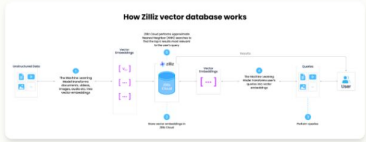
Querying

When a query is made:

1. The query is converted into a vector embedding using the same model used for indexing [1].
2. The database performs an Approximate Nearest Neighbor (ANN) search to find the

Ask follow-up Pro

Similarity Search



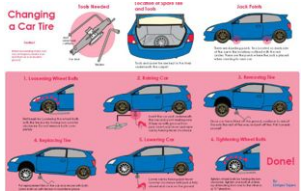
View More

Search Videos +

PRO Generate Image +

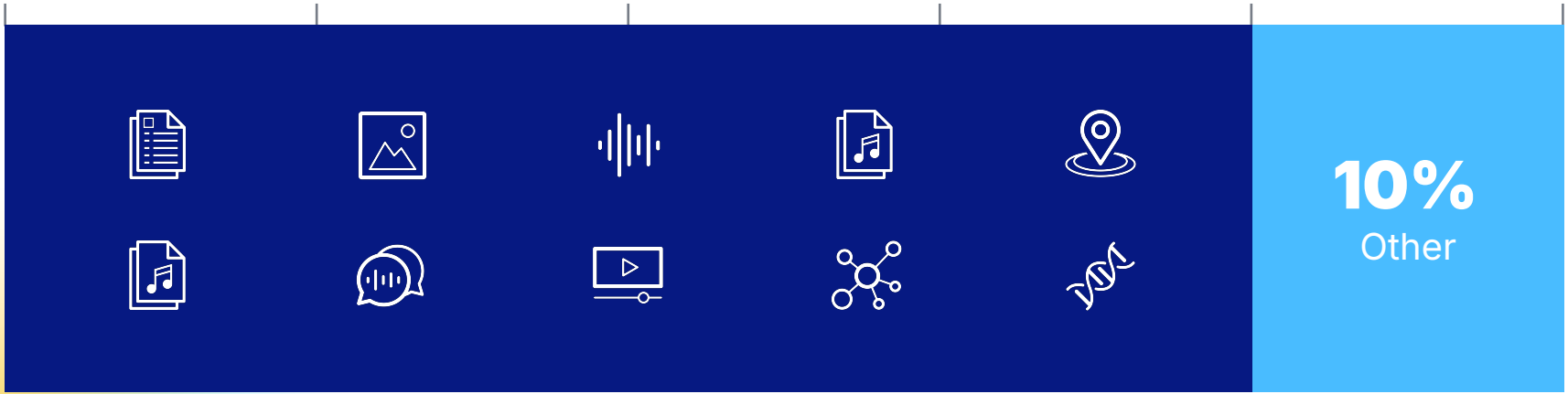
perplexity

Why is Semantic Search Difficult?

Apple	Rising dough	Change car tire
	Rising Dough ✔	
or	*or*	*or*
	Proofing Bread ❌	

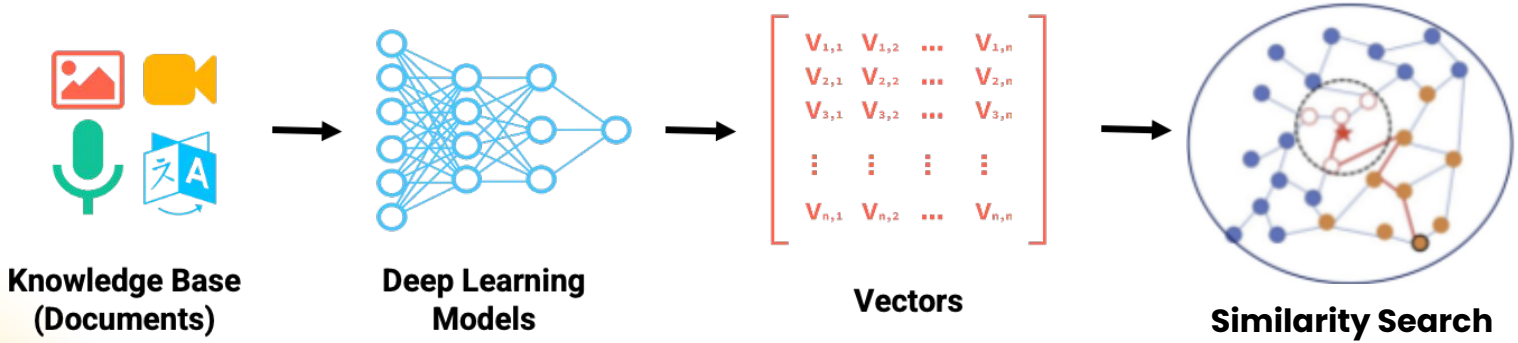
Why is Semantic Search Important?

90% newly generated data in 2025 will be unstructured data

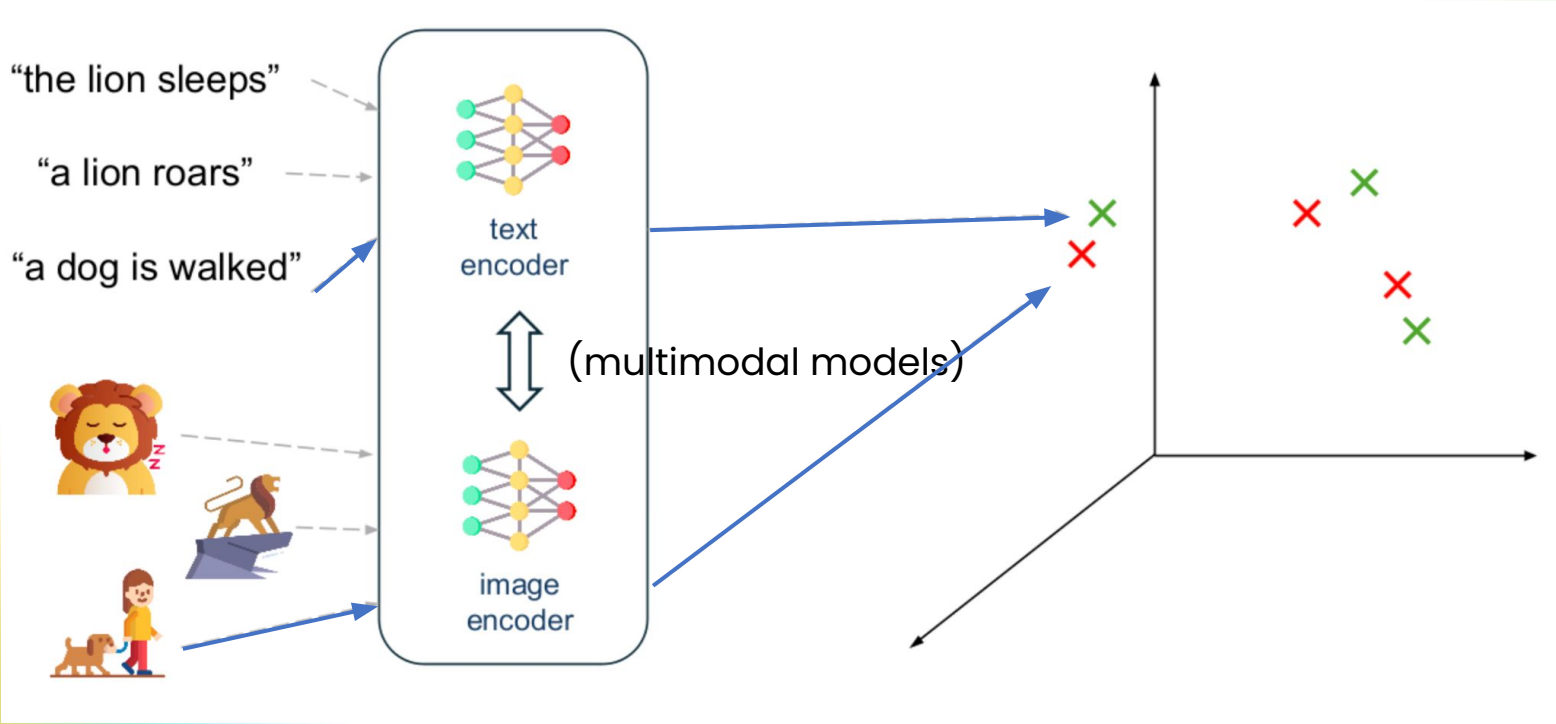


*Data Source: The Digitization of the World by *IDC*

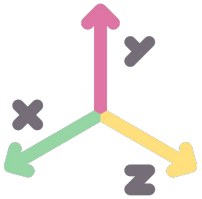
Solution: Deep Learning



Semantic Similarity?

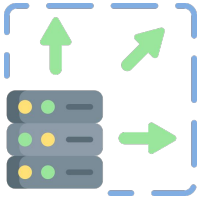


New Challenge: Search in Vector Spaces



How to Index and Search?

- High-dimensional
- > 1000 dims



How to Scale?

- 10-100 million vectors?
- Billions?
- Trillions?
- Billions of users?

Multiple Data Types?

- Text
- Images
- Audio
- Graphs
- ...

Milvus: High-performance, scalable vector database

Milvus is an **Open-Source Vector Database** to **store**, **index**, **manage,** and **use** the massive number of **embedding vectors** generated by deep neural networks and LLMs.

milvus
LF AI &DATA

400+
contributors

30K+
stars

66M+
docker pulls

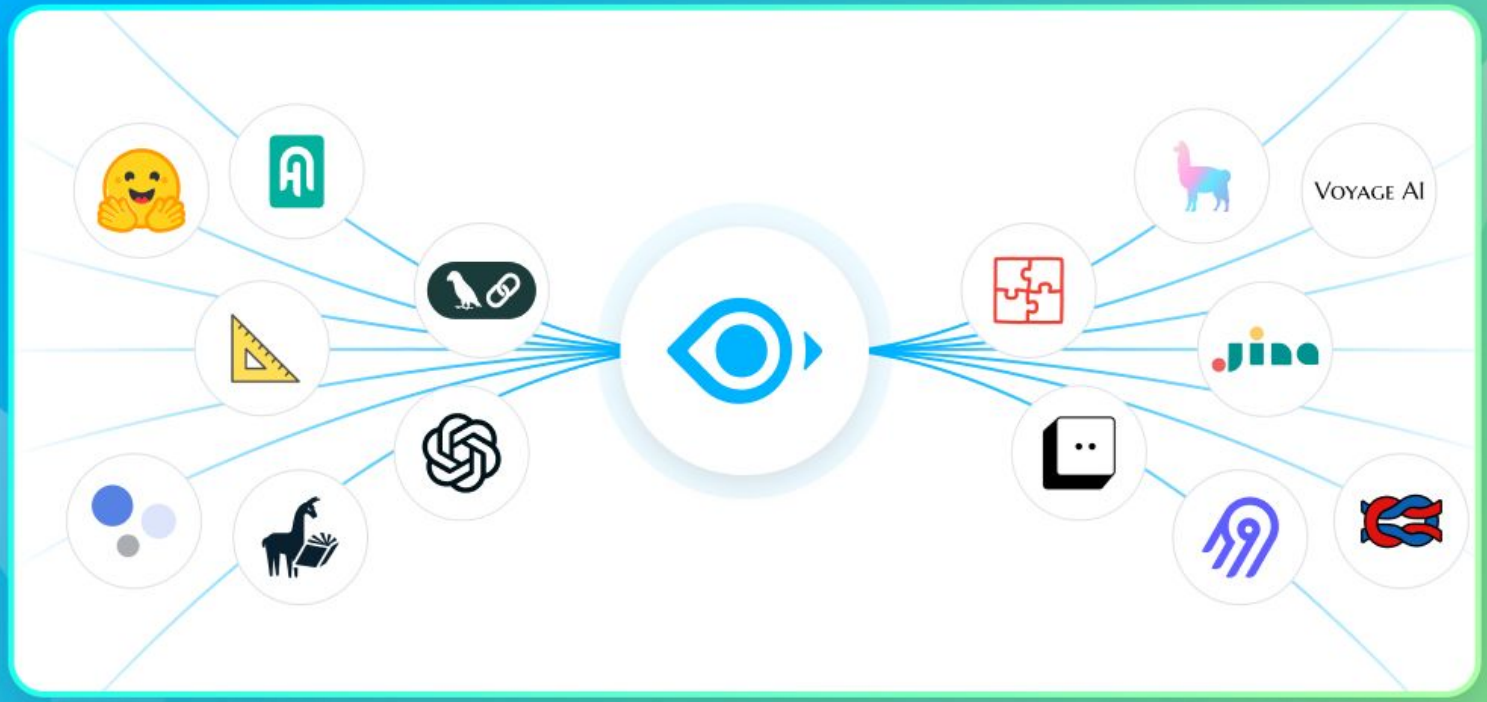
2.7K+
forks

 |

zilliz

Milvus Users

Seamless integration with all popular AI toolkits



milvus

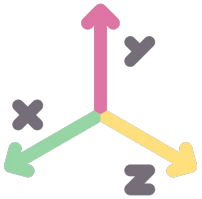
Why Open-Source?

Flexibility

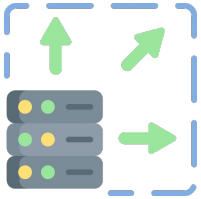
Innovation

Community

Why Not Traditional Databases?



Suboptimal Indexing / Search



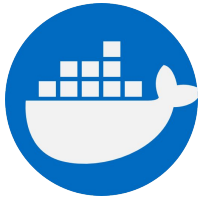
Scaling

Inadequate Query & Analytics Support

Deployment Options

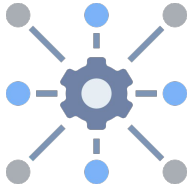
Milvus Lite

- Locally hosted
- Suitable for prototyping and demos



Milvus Standalone

- Single remote/local server
- “Medium” scale
- Simplified setup, maintenance, etc. compared to cluster



Milvus Cluster

- Distributed system
- Many different types of nodes
- Scales to 100s of billions of vectors

Getting Started

- Google Colab notebook

LET'S STAY CONNECTED!

https://milvus.io/discord

https://github.com/milvus-io/milvus

https://x.com/milvusio

https://www.linkedin.com/company/the-milvus-project