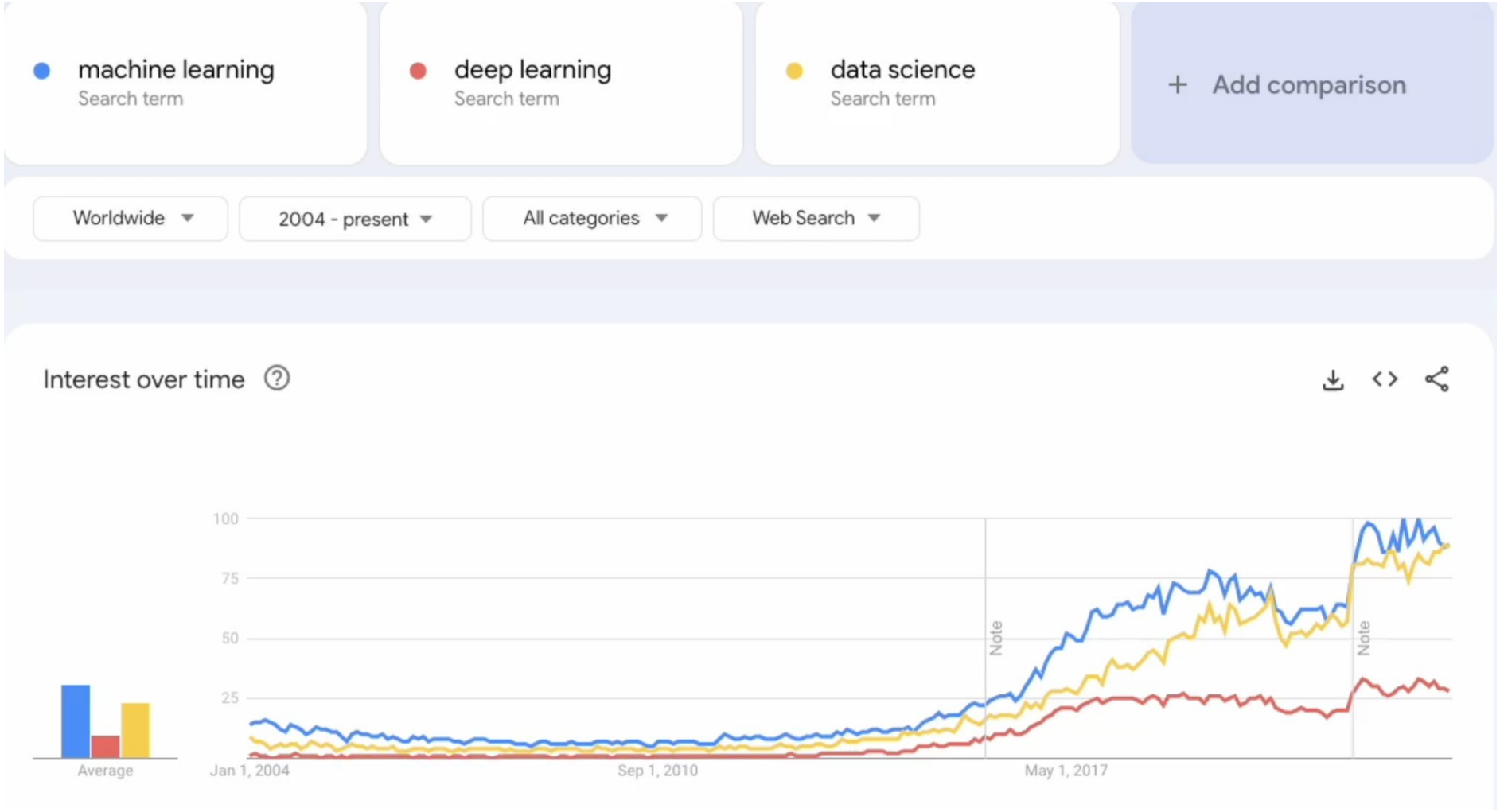
Rediscover Redis for Vector Similarity Search

Amine El Kouhen, Solution Architect
amine.elkouhen@redis.com

The Changing AI Landscape

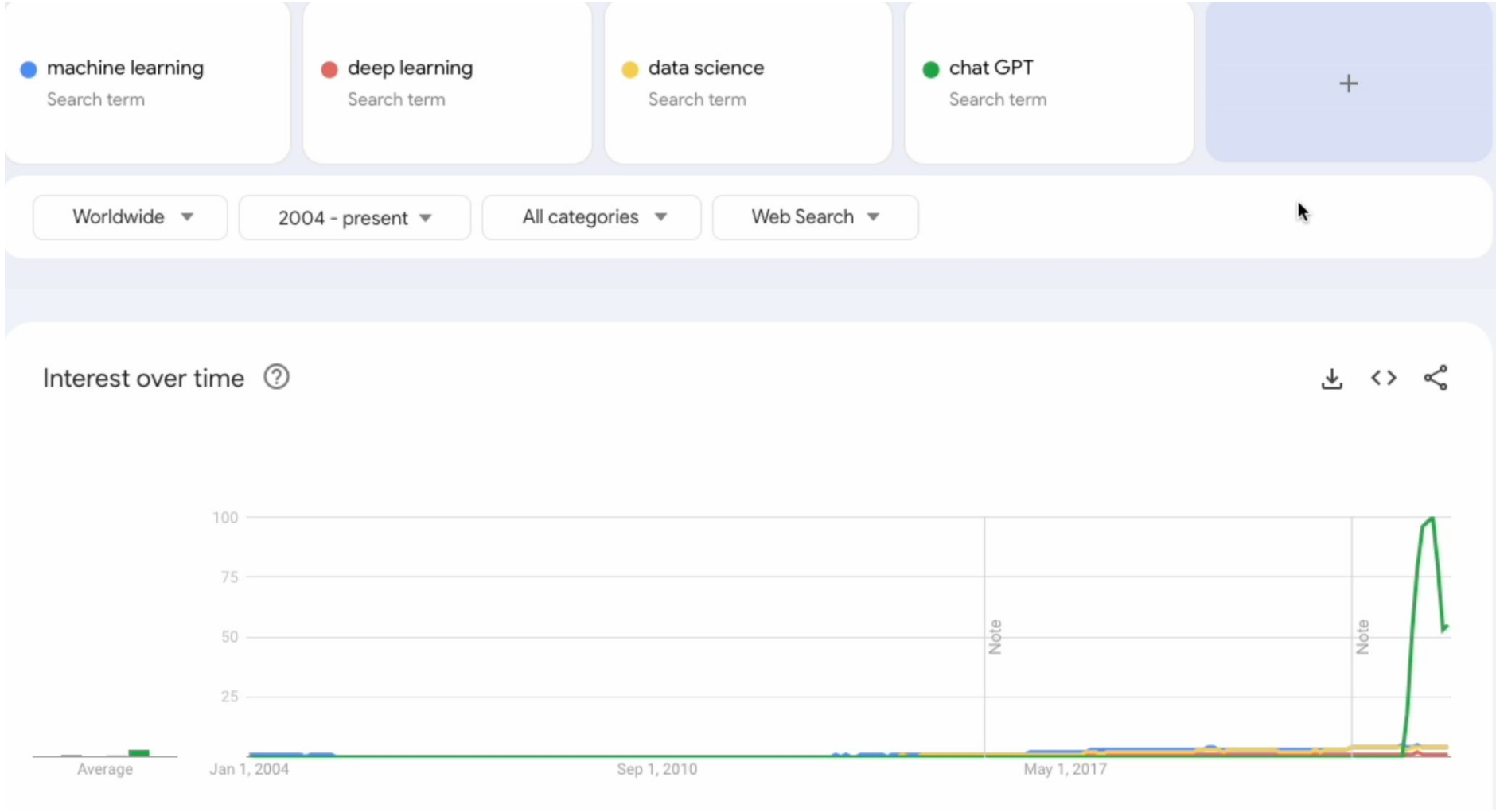
2004 → 2022



Google Trends

The Changing AI Landscape

2004 → Now



Google Trends

Generative AI = Large Language Models + Vector Similarity Search

Application Development

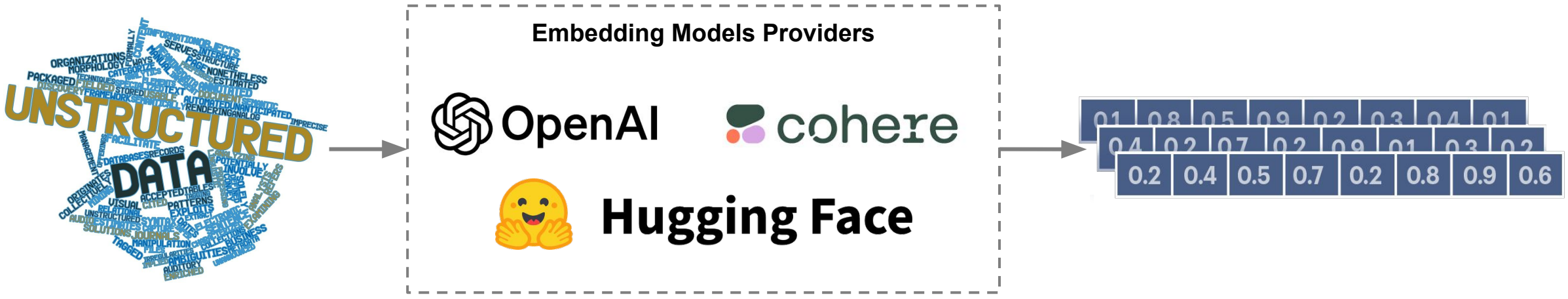
LLM Providers

- LLMs are massive general purpose neural networks pre-trained on large amount of text.
- Specifically focused on language **understanding** and **generation** (GPT, BERT, LLaMA, PaLM)
- Utilize **Vector Similarity Search** to retrieve information from external **databases**
- Use cases:
  - Translation
  - Sentiment Analysis
  - Content Generation/Summarization
  - Question Answering

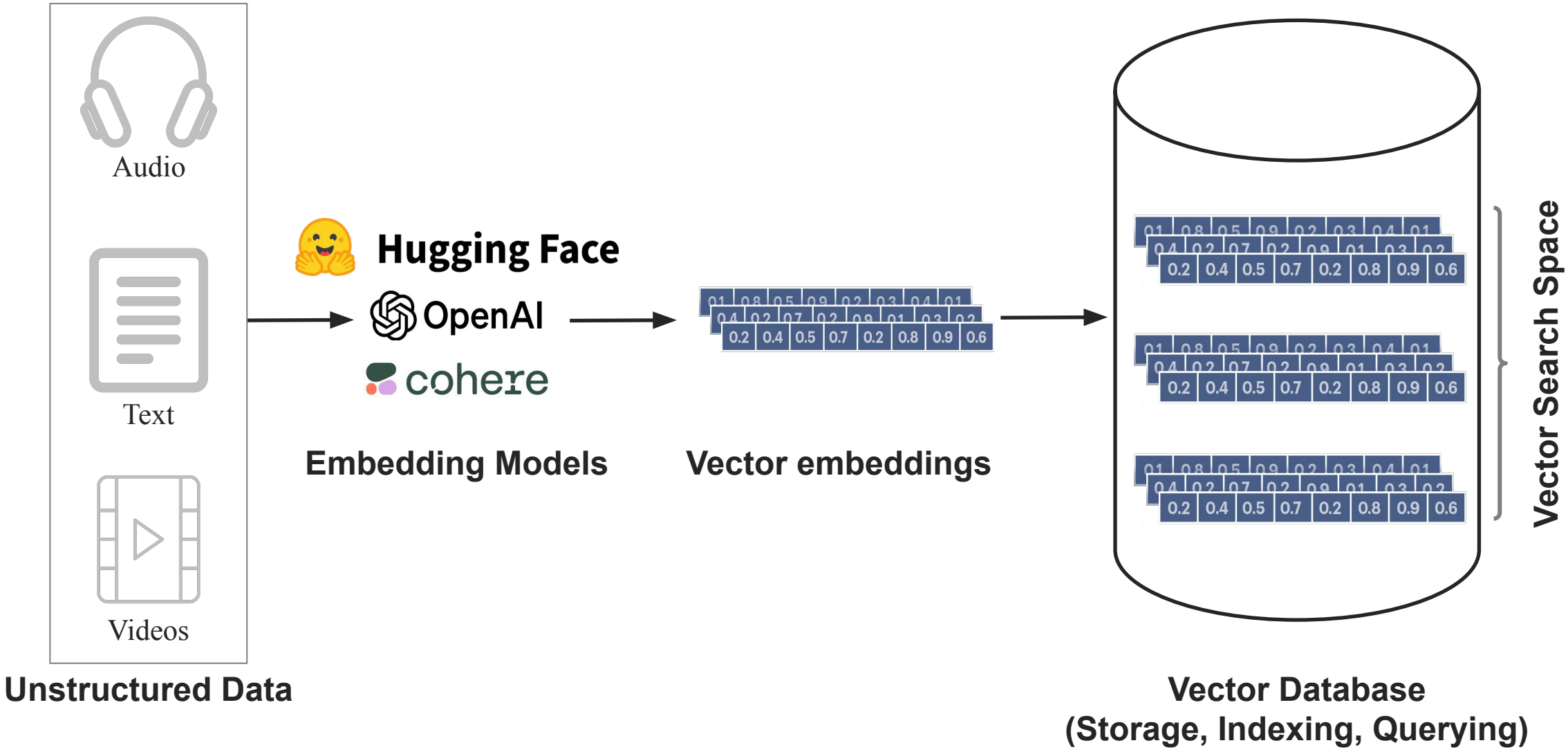
What is a Vector Embedding?

Vectors embeddings are mathematical representations of data points where each vector dimension corresponds to a specific feature or attribute of the data:

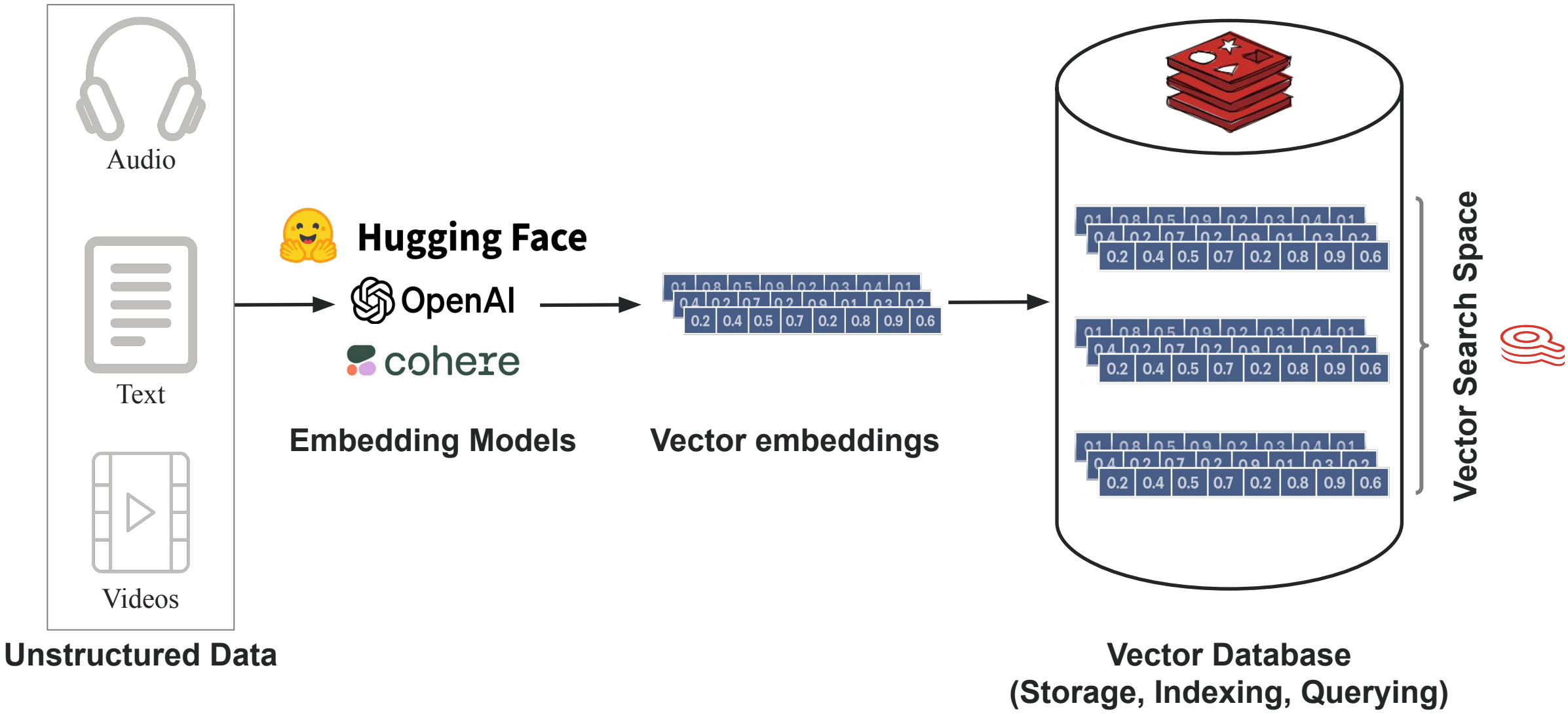
- Compact & dense data representation: A fixed-sized **list of floating-point** numbers,
- More often, used to represent **Unstructured Data** (Audio, Images, Text…),
- Produced by feature engineering or by pre-trained models (e.g., **Model providers**),
- Translate Perceived **Semantic Similarity** to the **Vector Search Space**



What is a Vector Database?

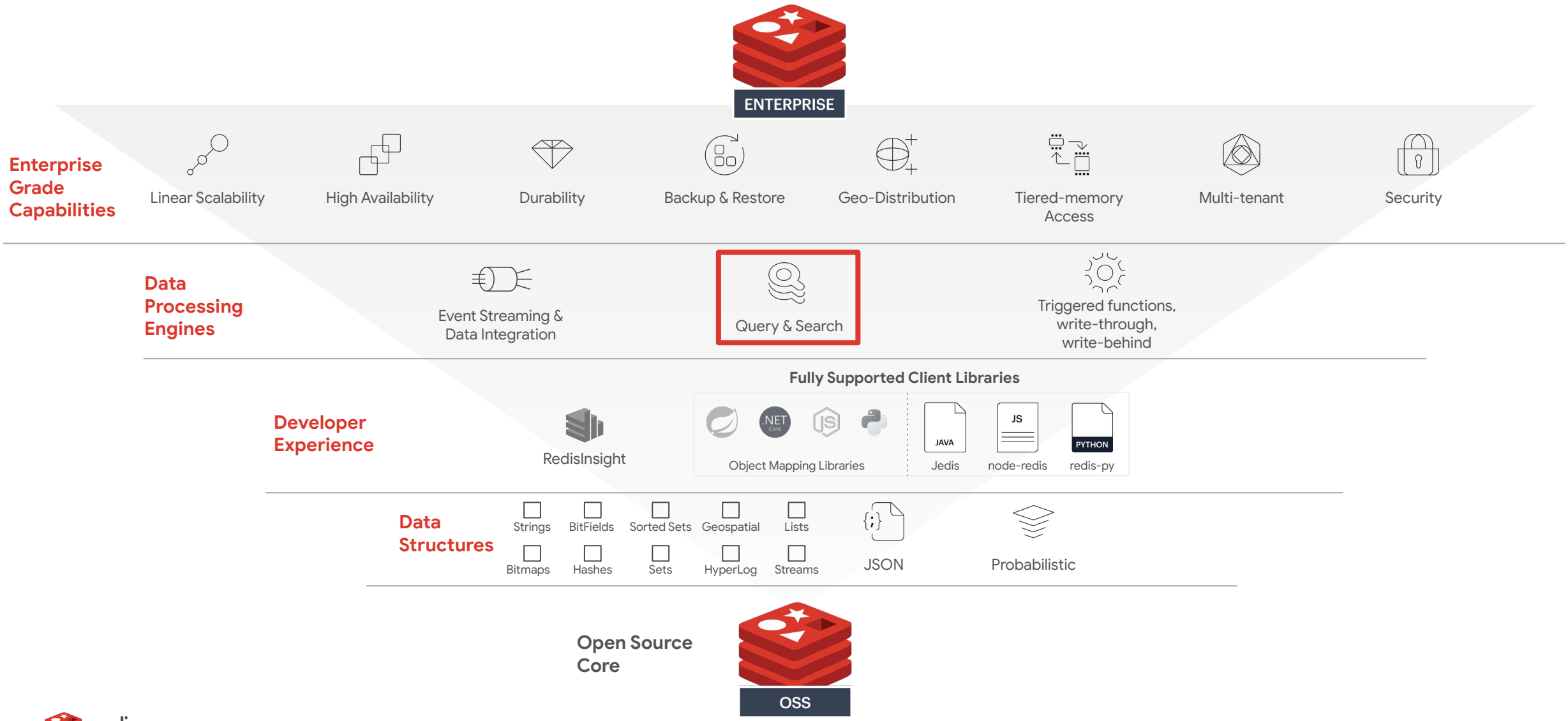


Redis as a Vector Database



Redis?! isn't it just a cache?

Redis Enterprise



redis

Redis - Vector Similarity Search

Redis + RediSearch

Integrations

LangChain

jina

metal.

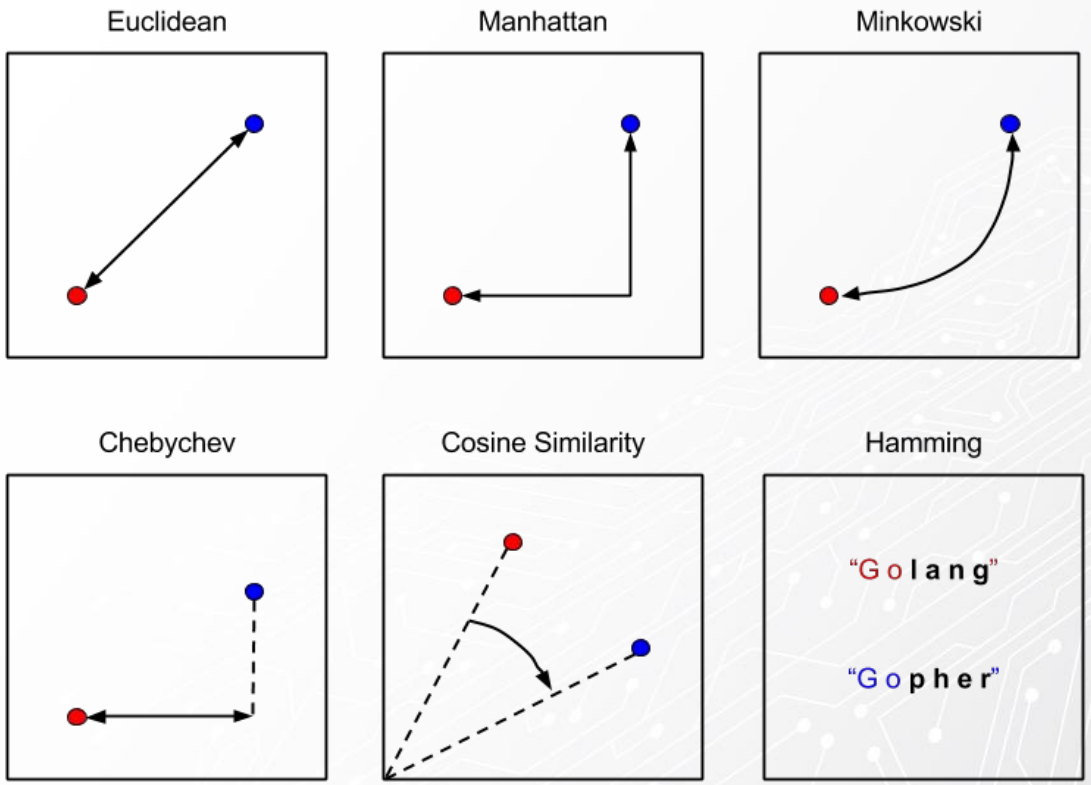
Relevance AI

NVIDIA

- **Redis**: Low-latency, scalable, in-memory database
- Real-time updates
- Indexing methods:
  - HSNW (ANN)
  - Flat (KNN)
- Distance metrics:
  - Euclidean (L2)
  - Internal Product (IP)
  - COSINE
- Support Hybrid Search:
  - Vector Search + Filtering by text, geo…
- Stores Vectors in JSON & Hash

Redis - Vector Similarity Search

Vector Similarity Search focuses on finding out how alike or different two vectors are. To achieve this in a reliable and measurable way, we need a specific type of score that can be calculated and compared objectively. These scores are known as **distance metrics**:



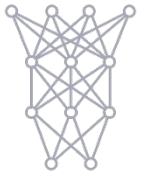

DEMO TIME!

Demo 1: Text Vector Search

redis

Text Vector Search - Vector Storage & Indexing

Text



Texts model

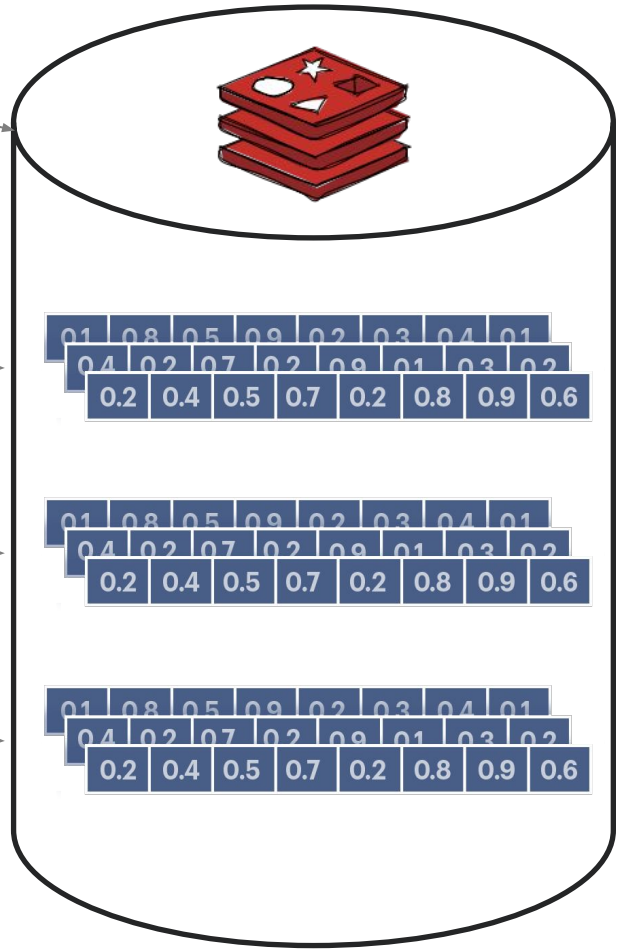
Text Vector Embeddings

That is a very happy person

That is a happy girl

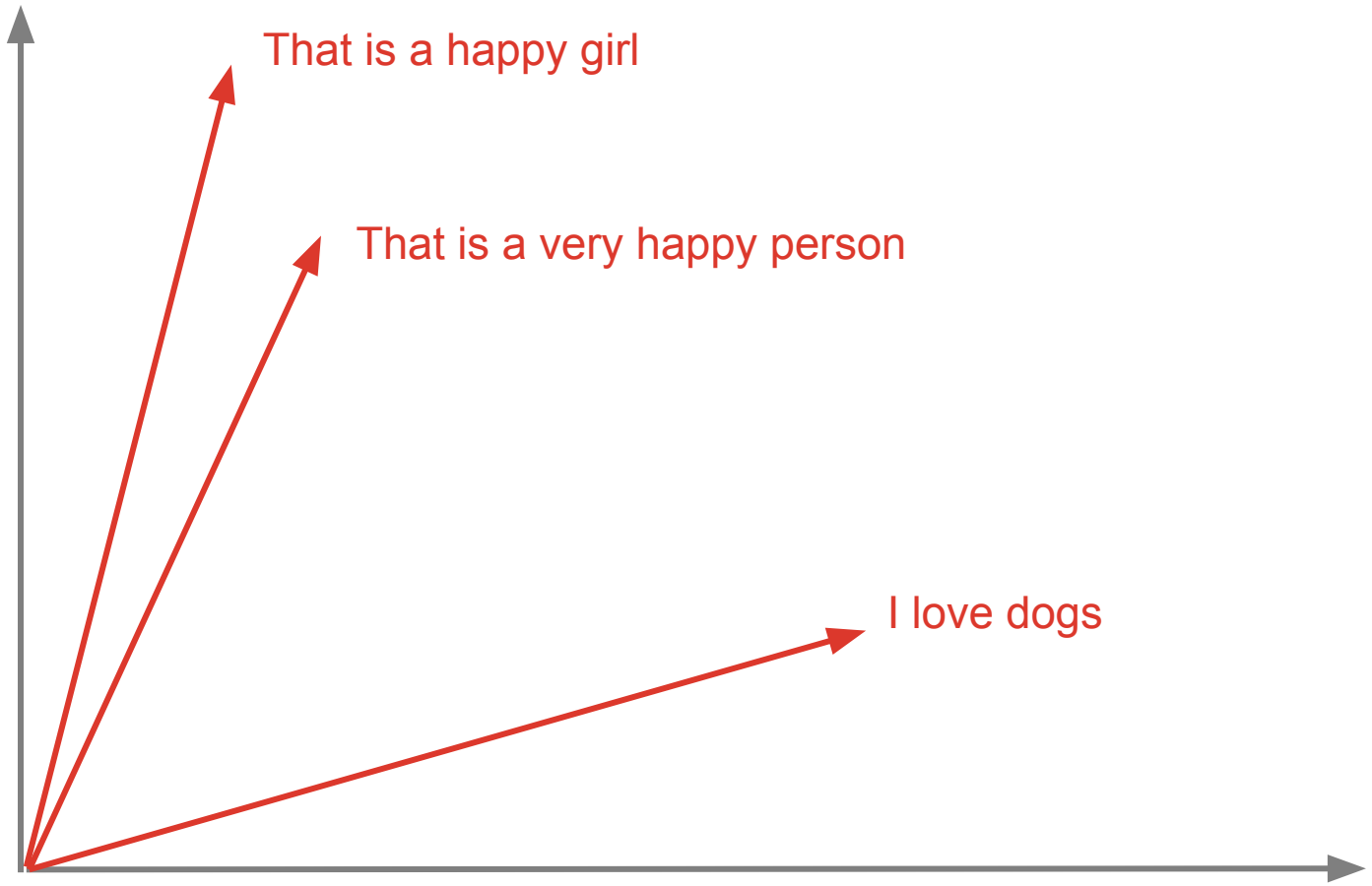
I love dogs

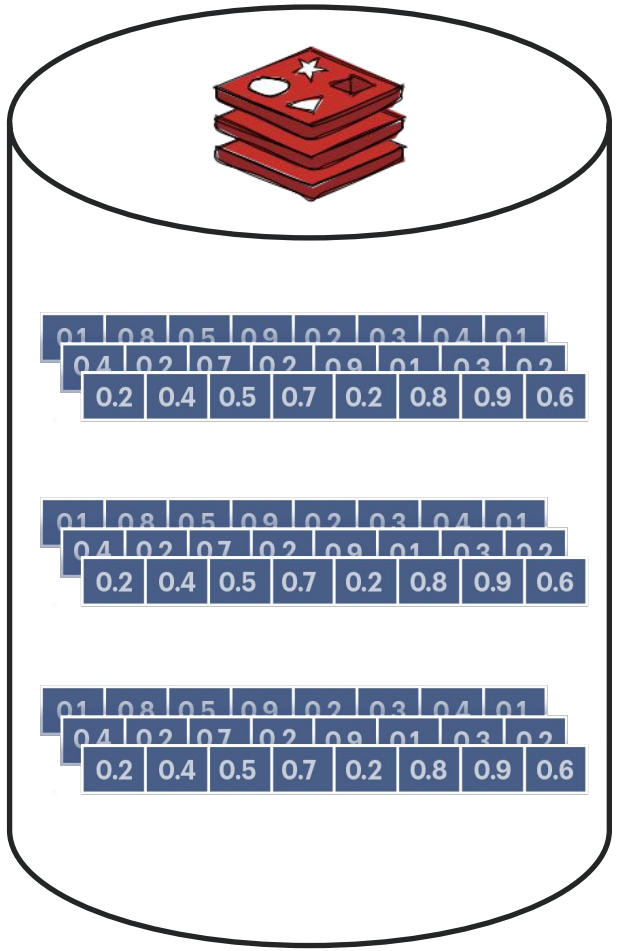
```
FT.CREATE sentence_index
SCHEMA embedding VECTOR FLAT 8
    TYPE FLOAT32
    DIM 768
    DISTANCE_METRIC COSINE
    INITIAL_CAP 100000
```



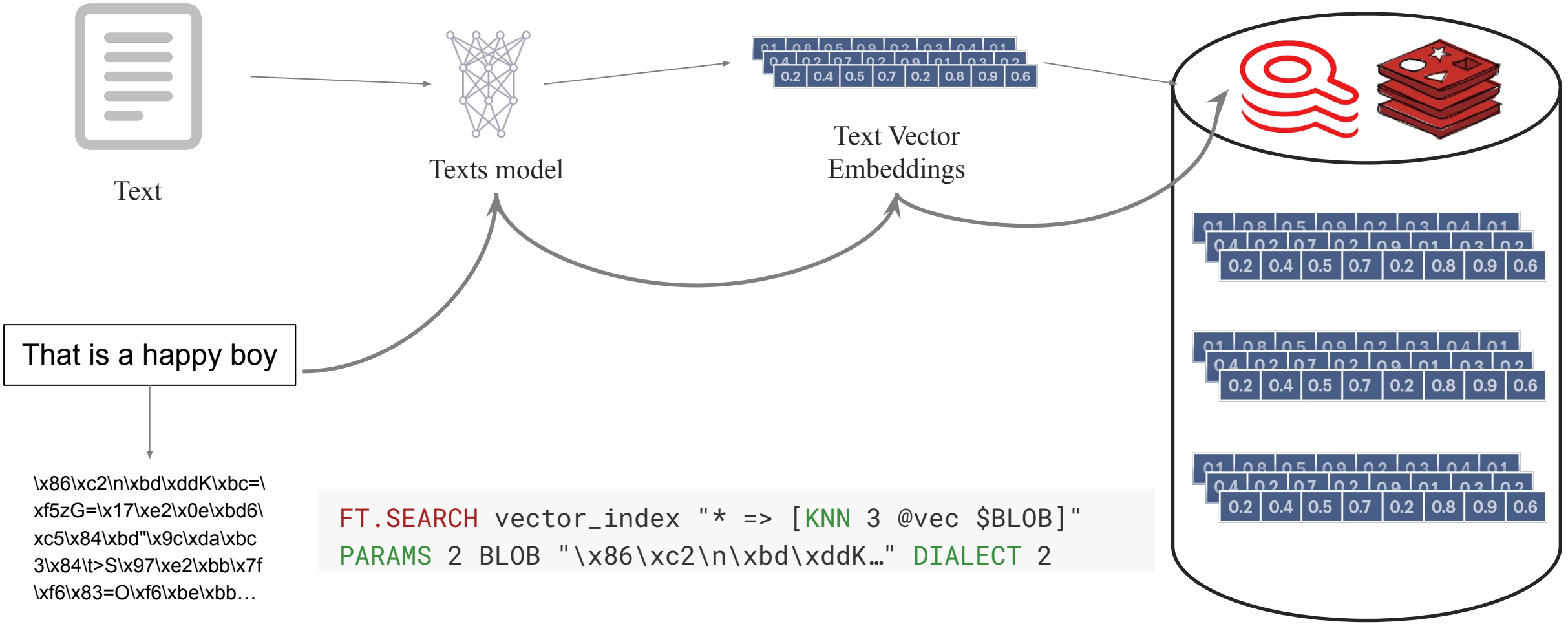
redis

Text Vector Search - Vector Space



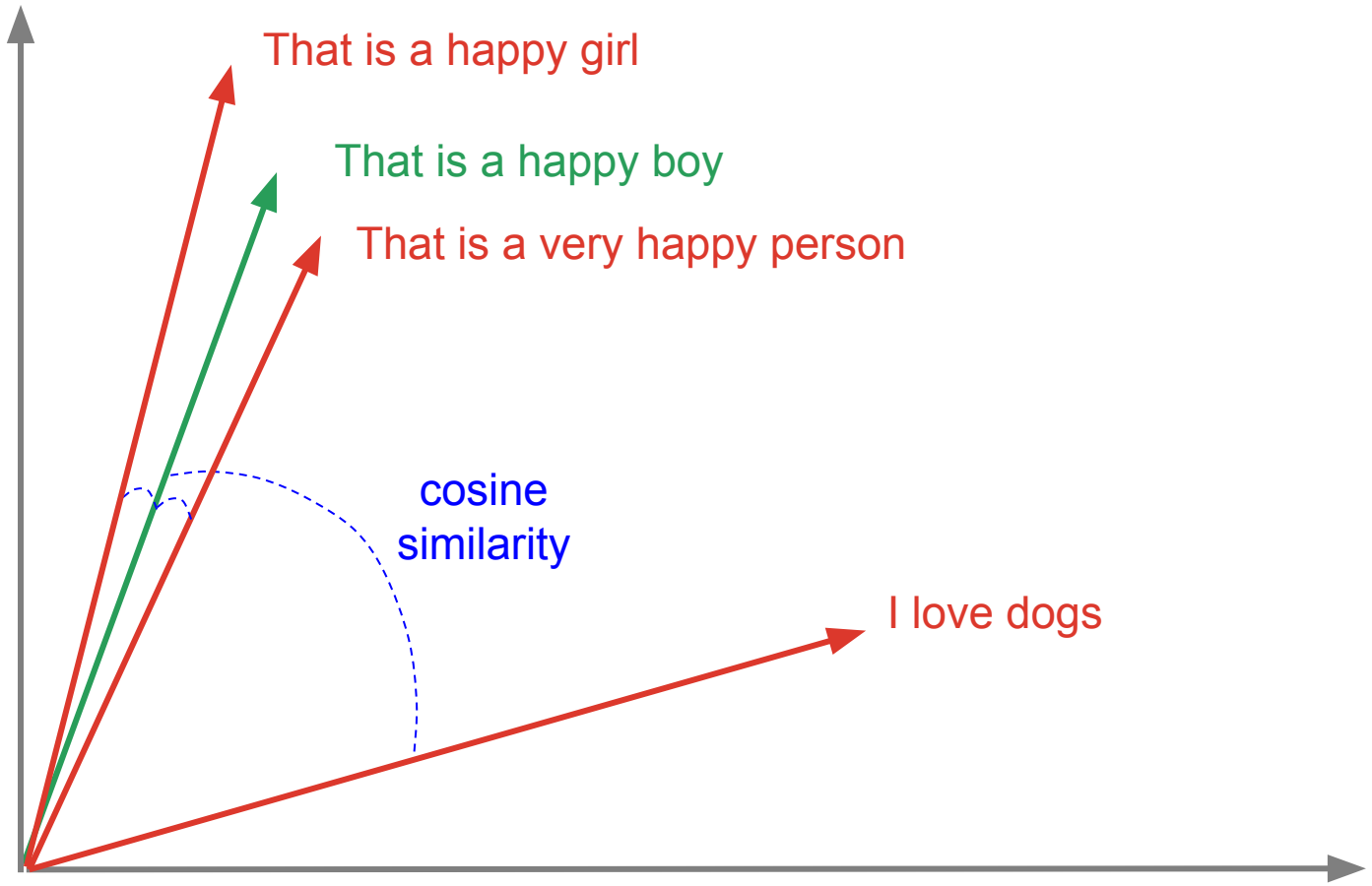


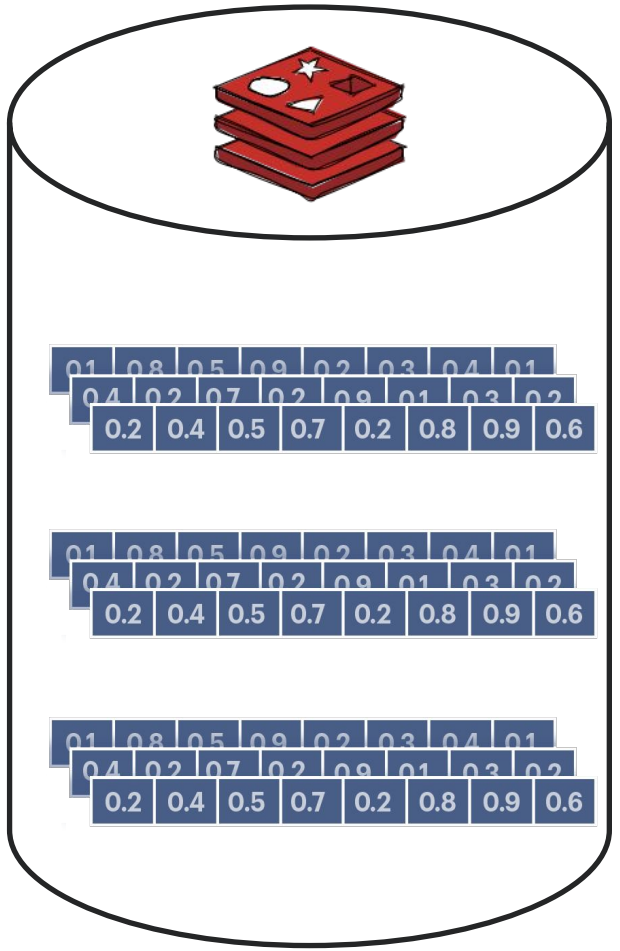
Text Vector Search - Vector Querying



redis

Text Vector Search - Vector Search Results

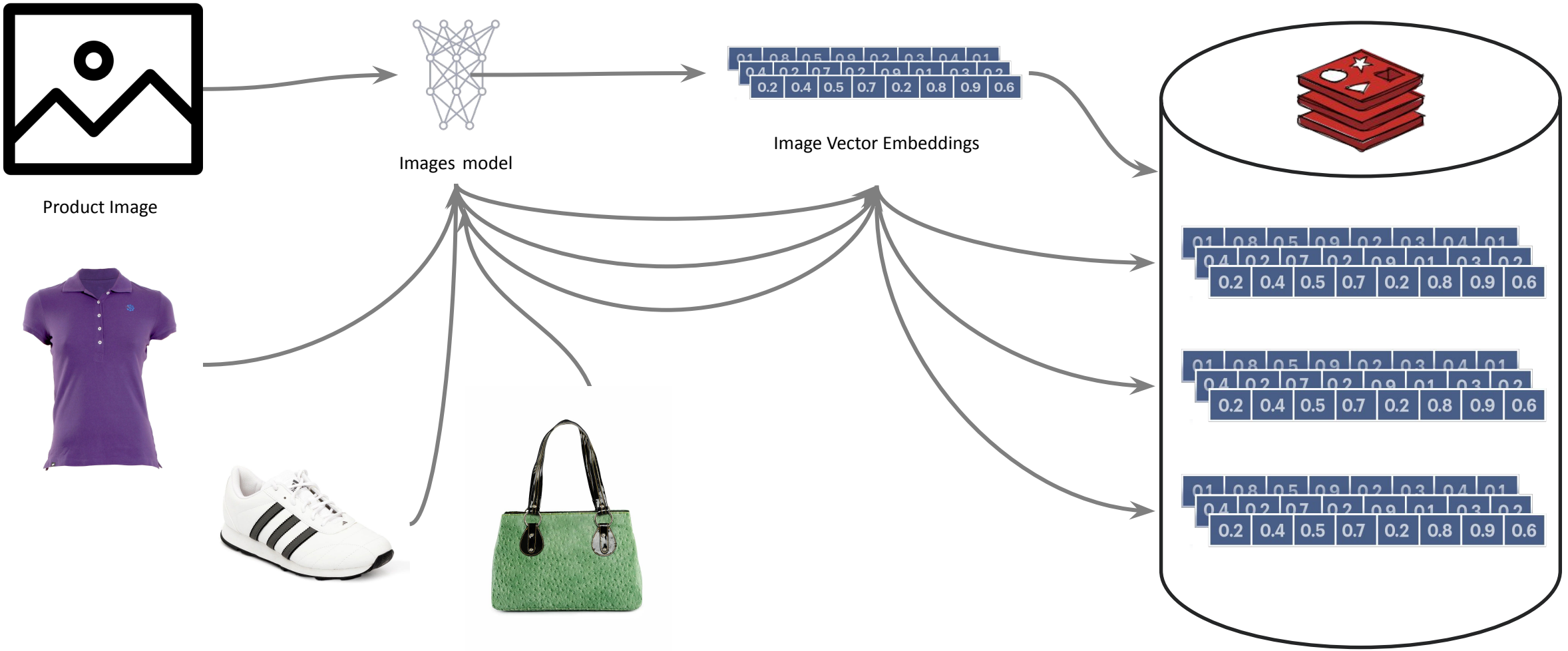




Demo 2: Visual + Hybrid Search

redis

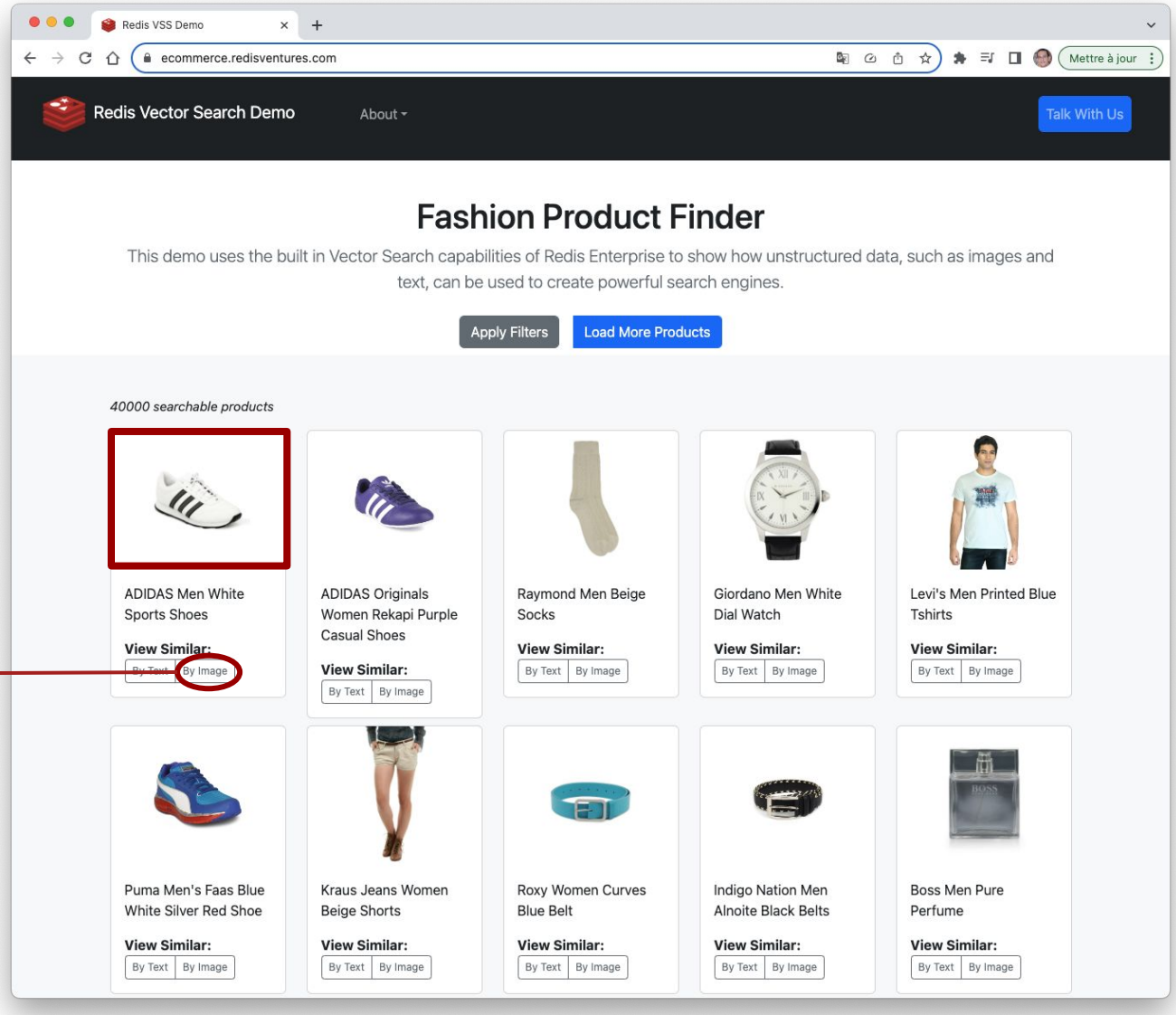
Visual Vector Search - Vector Storage



redis

Redis VSS Demo - Visual Vector Search

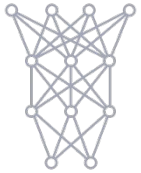
Search by Image



redis

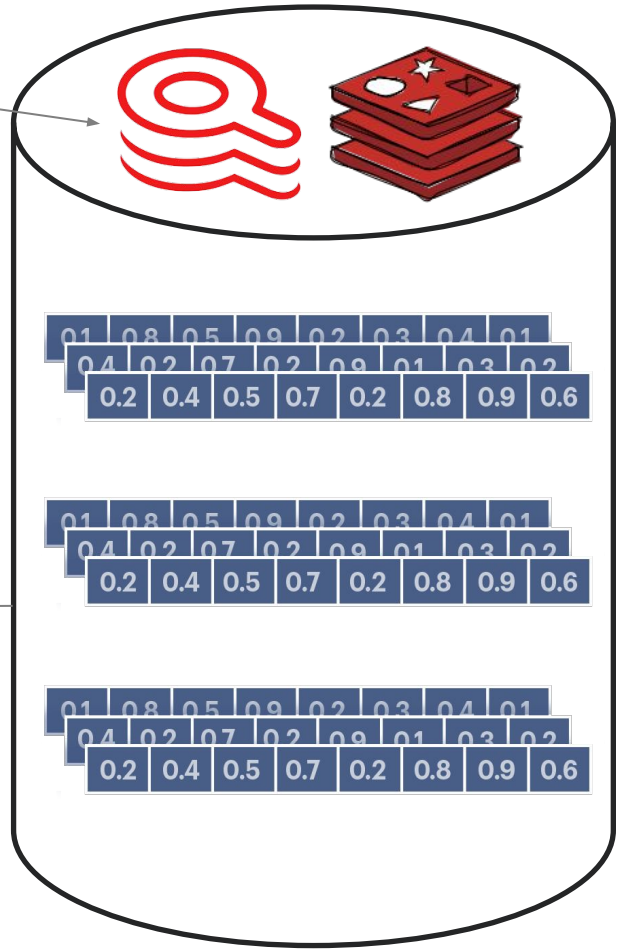
Visual Vector Search - Vector Querying

Query Image



Images model

Image Vector Embeddings



Search Results

Score: 0.95

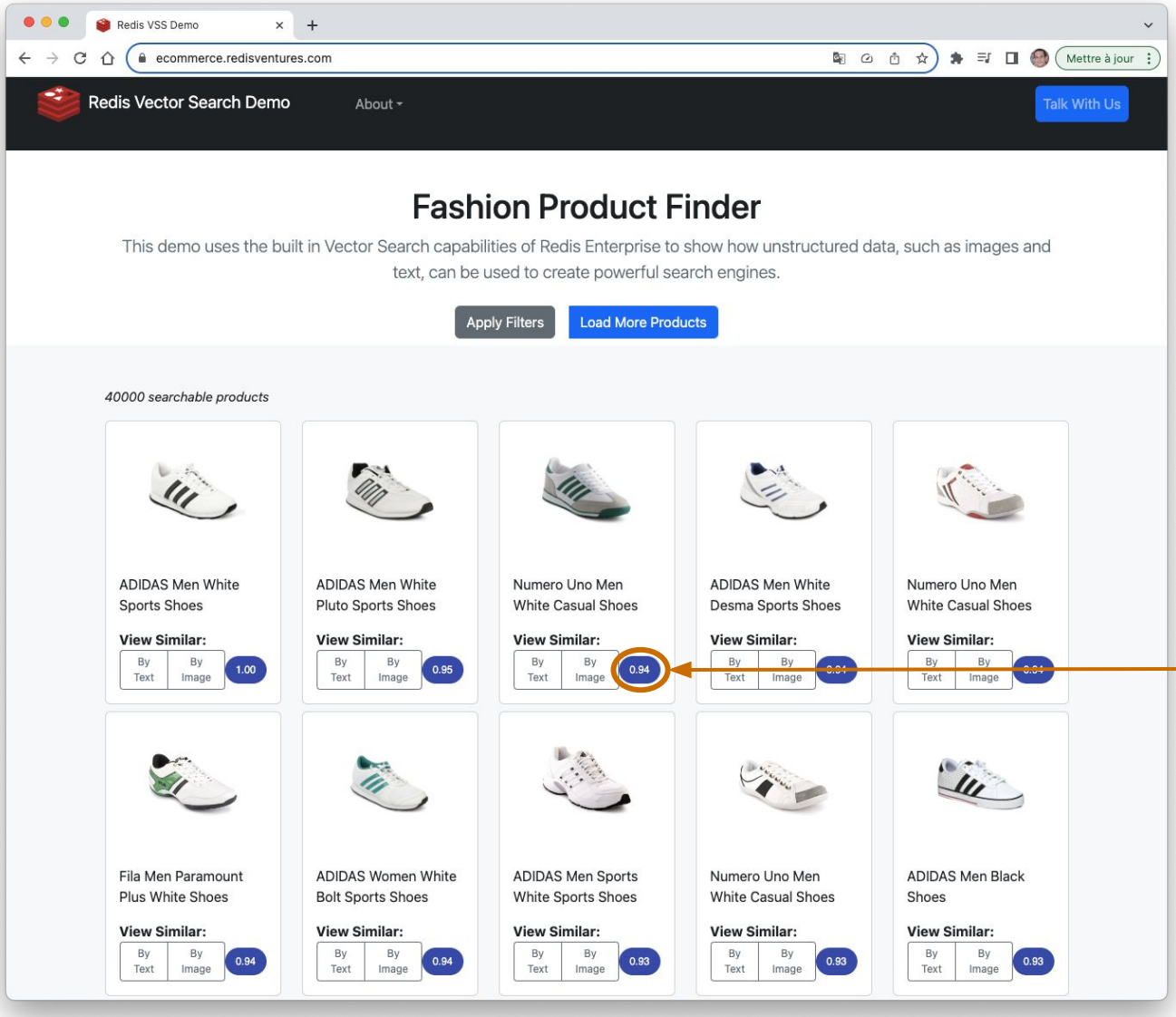
Score: 0.94

Score: 0.93

redis

Redis VSS Demo - Visual Vector Search

Query Image



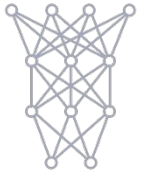
What if I want a women's shoe that looks like this men's shoe?

Product (vector) "Similarity Score"

redis

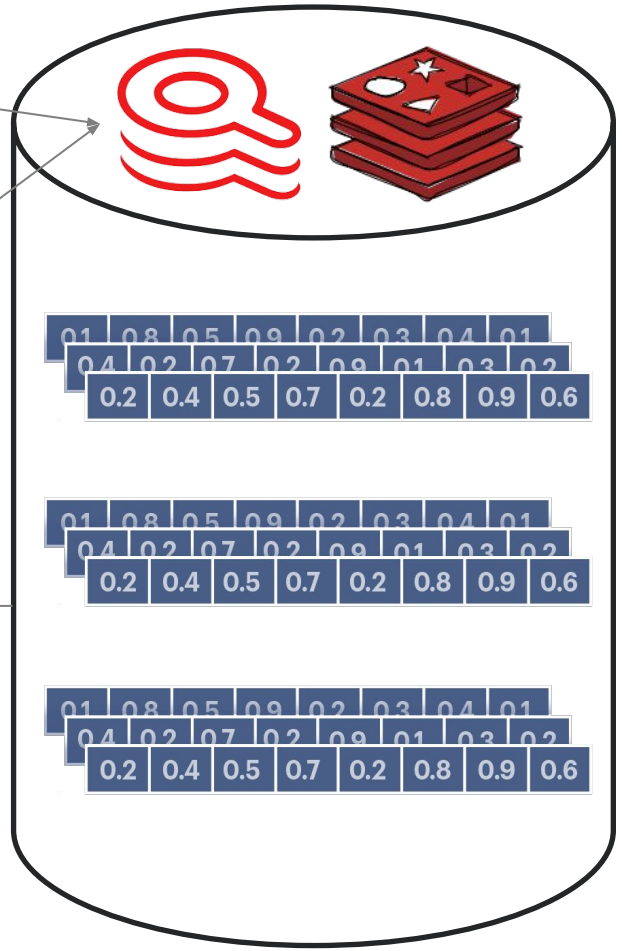
Hybrid Search - Filtered Vector Querying

Query Image



Images model

Image Vector Embeddings



Rule-Based Filter: Gender == Women

Score: 0.94

Score: 0.93

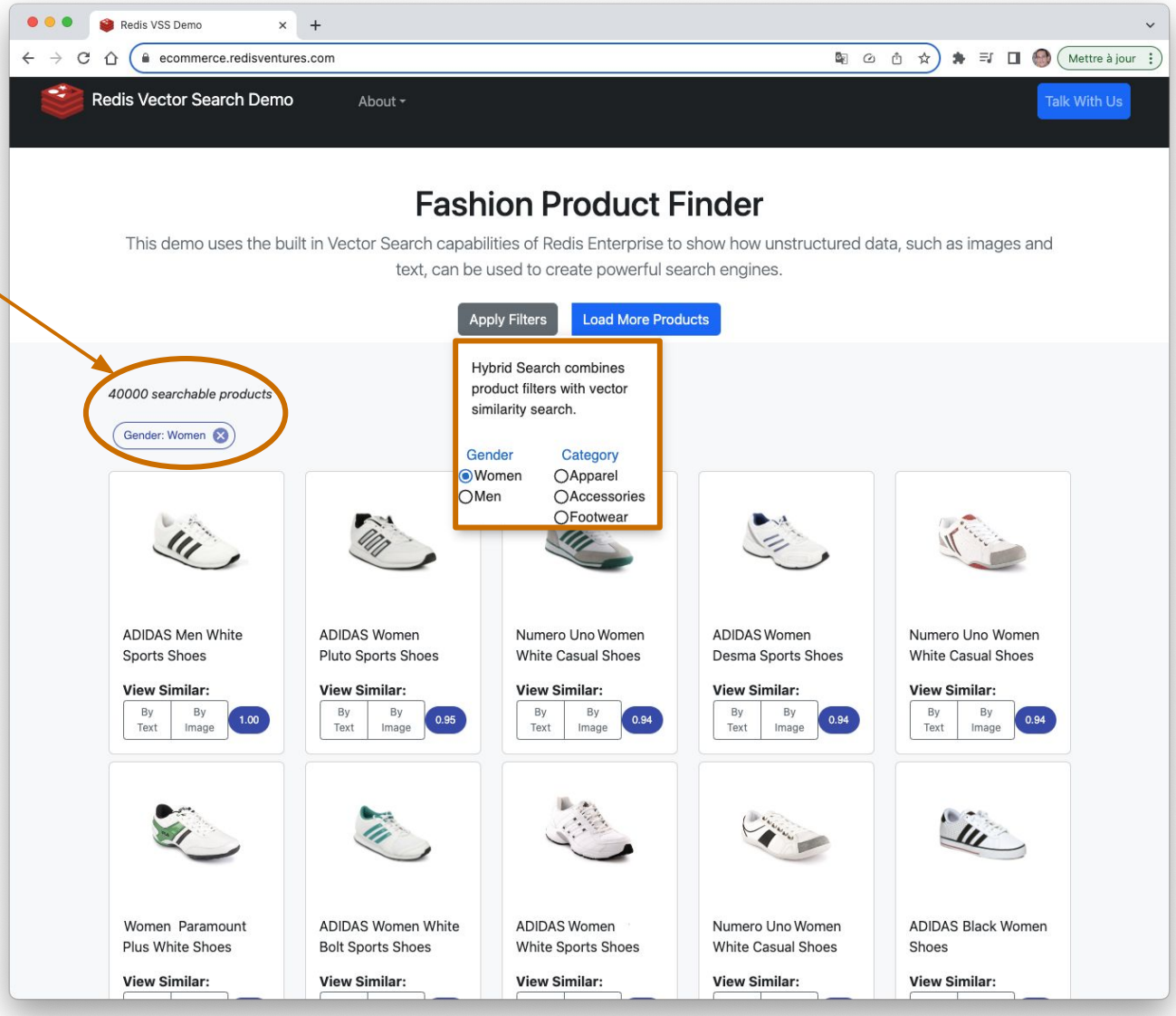
Search Results

redis

Redis VSS Demo - Hybrid Search

Gender used as pre-filter for candidate selection

Query Image



Hybrid Search: The power of vector search enhanced with rule-based filters.

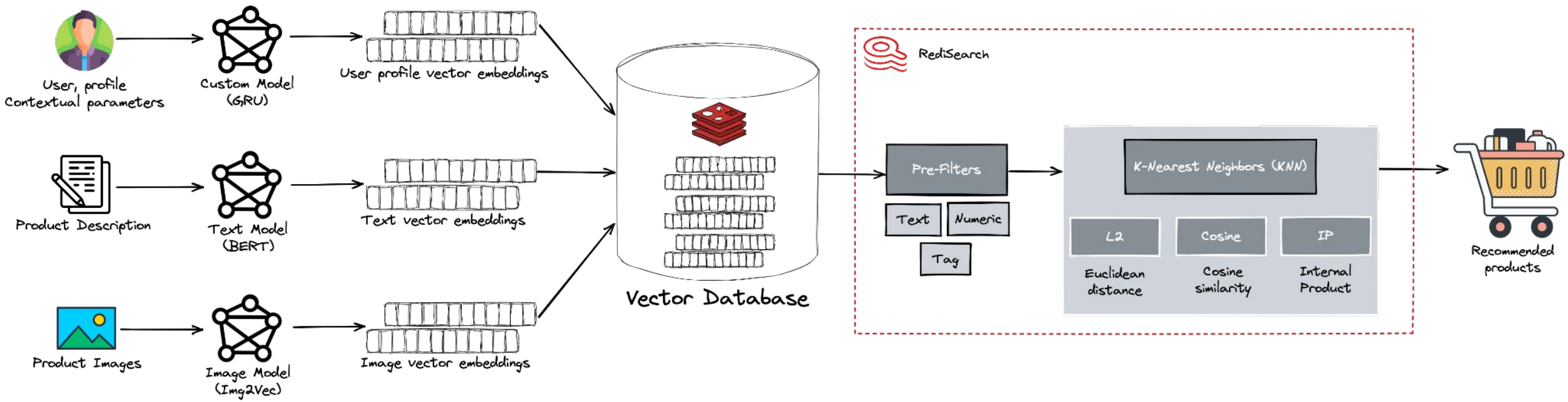
Supports all RediSearch Types:

- Tag
- Numeric (Range)
- Geographic
- Text
- Conditional

redis

Examples of Vector Similarity Search use-cases

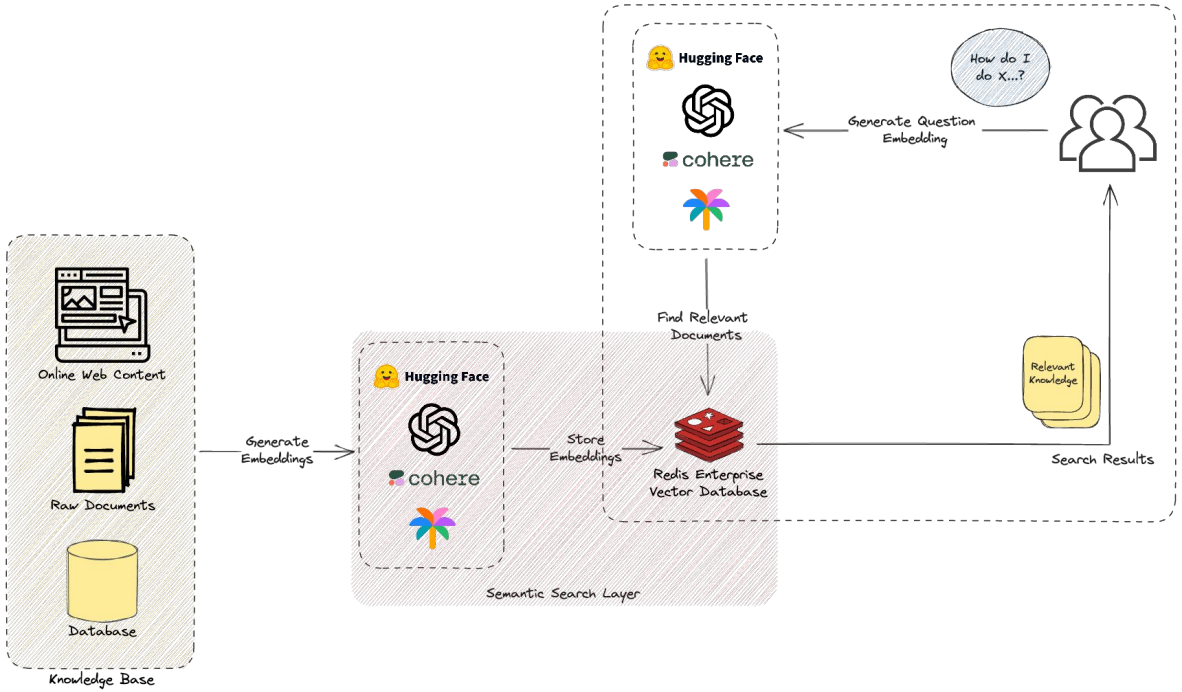
Hybrid Search for Recommendation Systems



Demo 3: Semantic Search

redis

Semantic Vector Search - Concepts



Description:

- Redis is used as an external knowledge base for the large language model.
- Queries are used to find relevant information within the knowledge base.

Benefits:

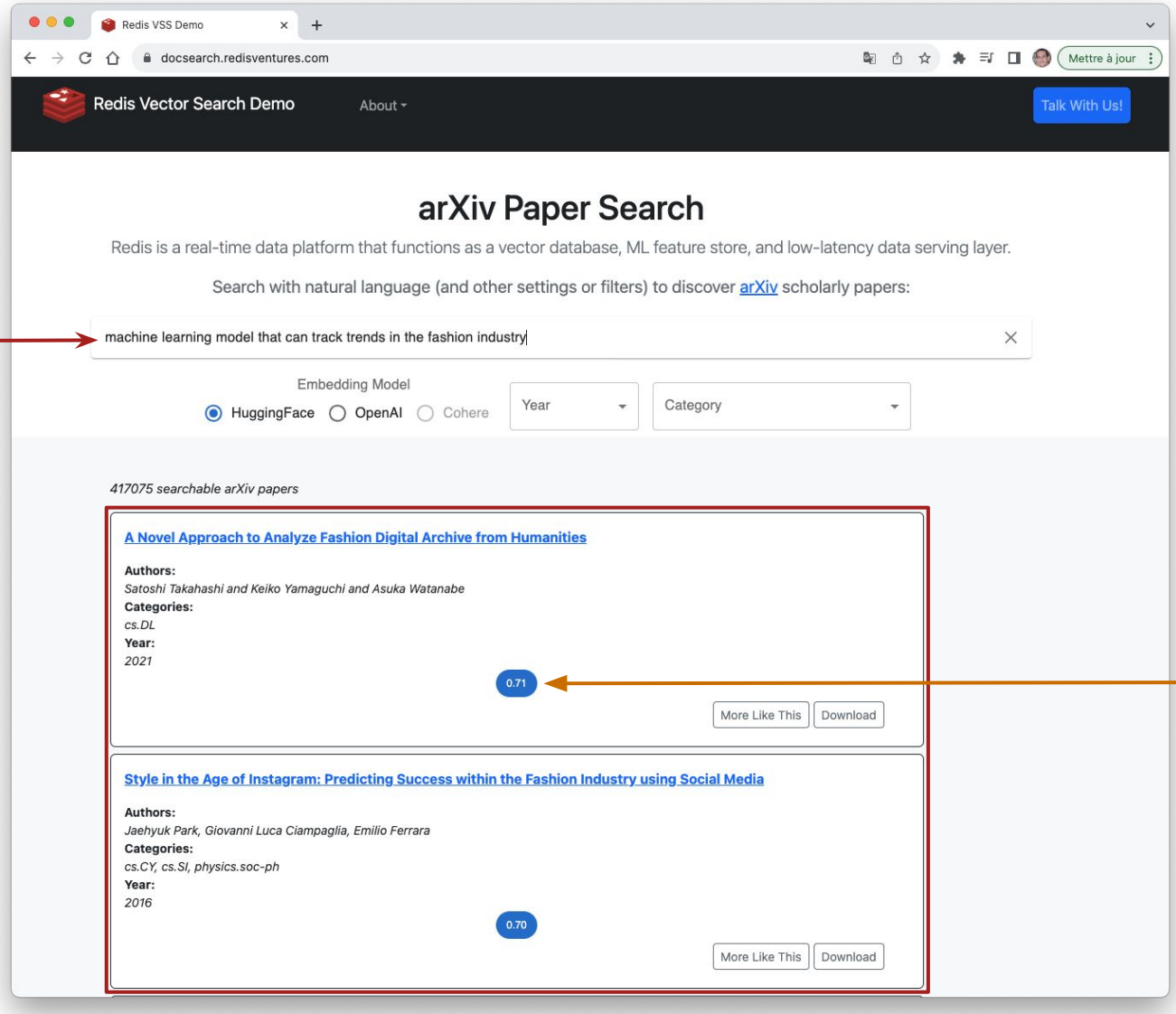
- **Cheaper and faster** than fine-tuning.
- **Real-time updates** to the knowledge base.
- **Sensitive data** doesn't need to be used in model training or fine-tuning.

Use Cases:

- Documents Discovery & Analysis
- ChatBots

Redis VSS Demo - Semantic Vector Search

Query

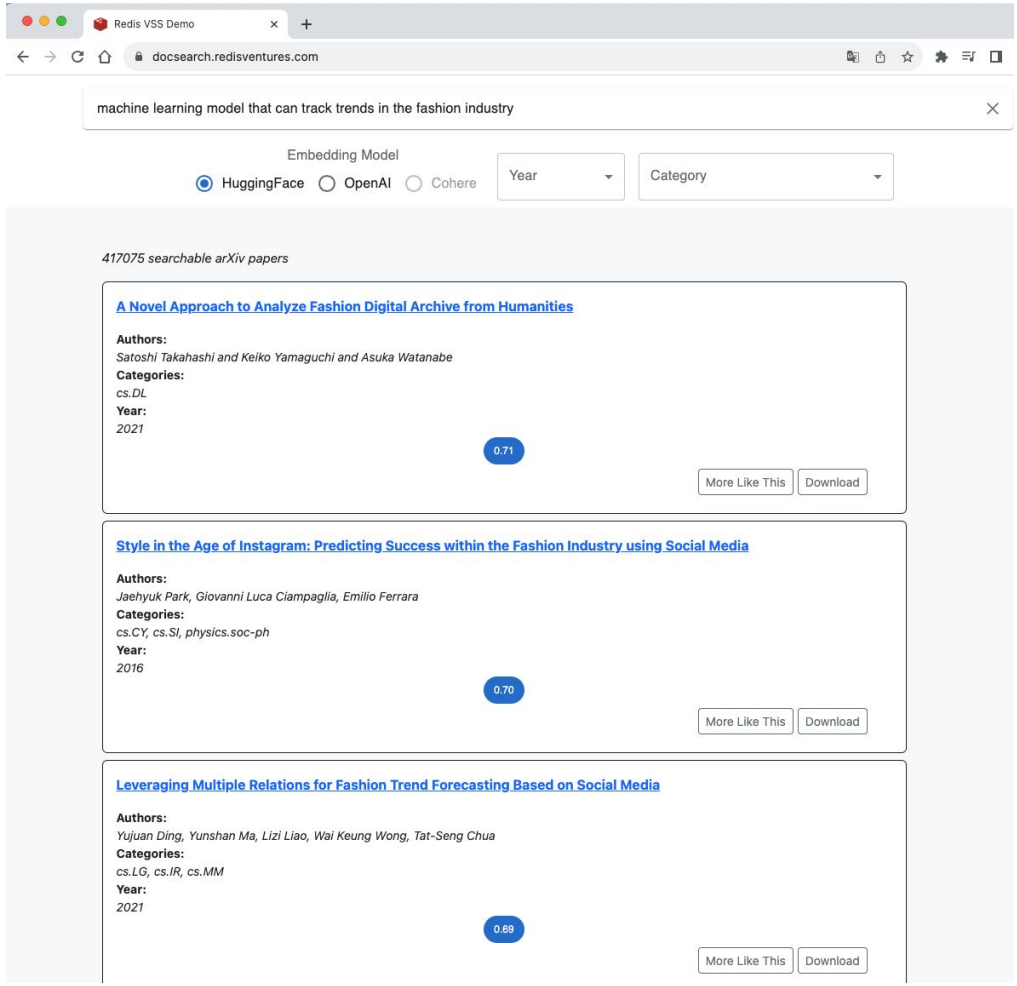


Paper (vector)
"Similarity Score"

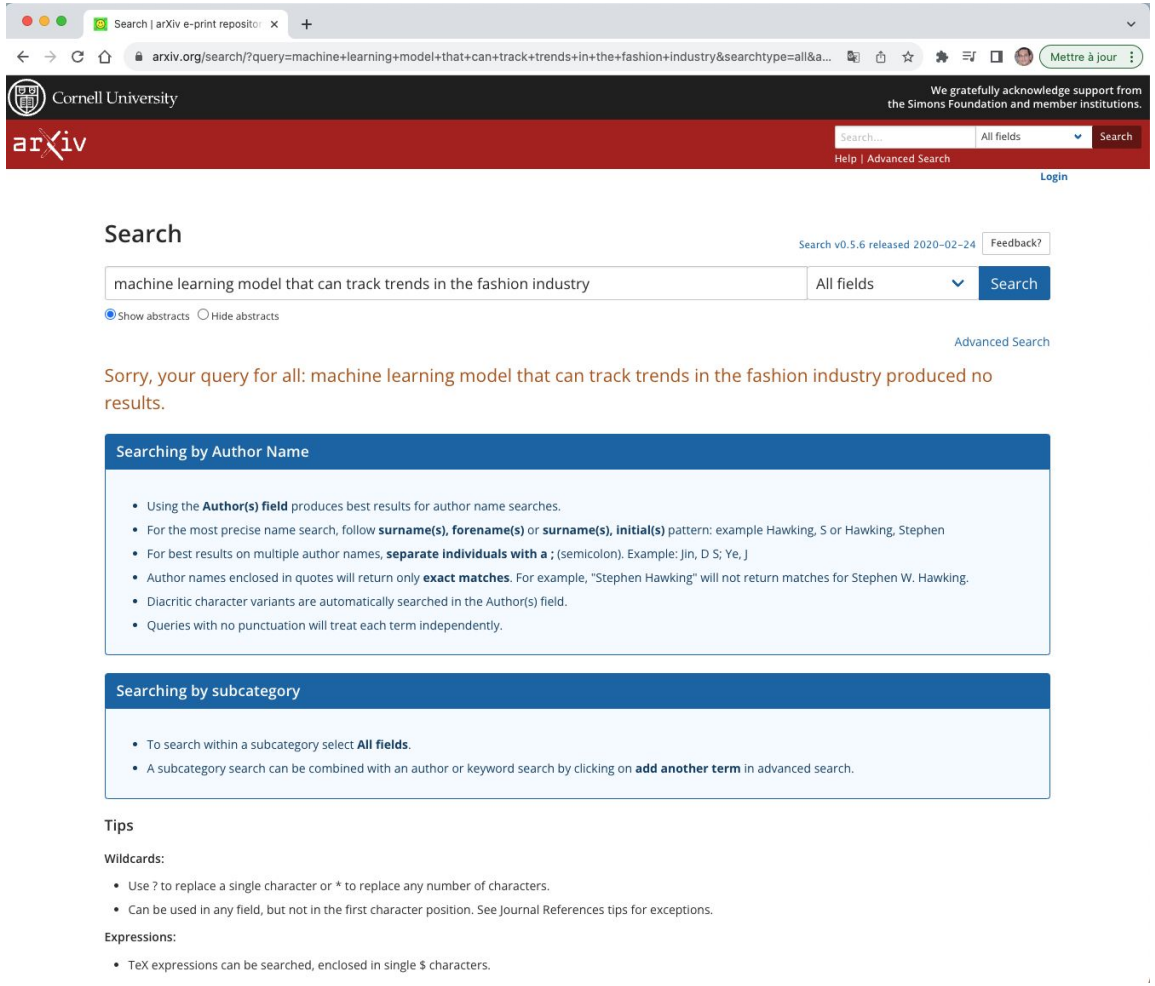
redis

Semantic Vector Search vs Traditional Search

Semantic (Vector) Search Results



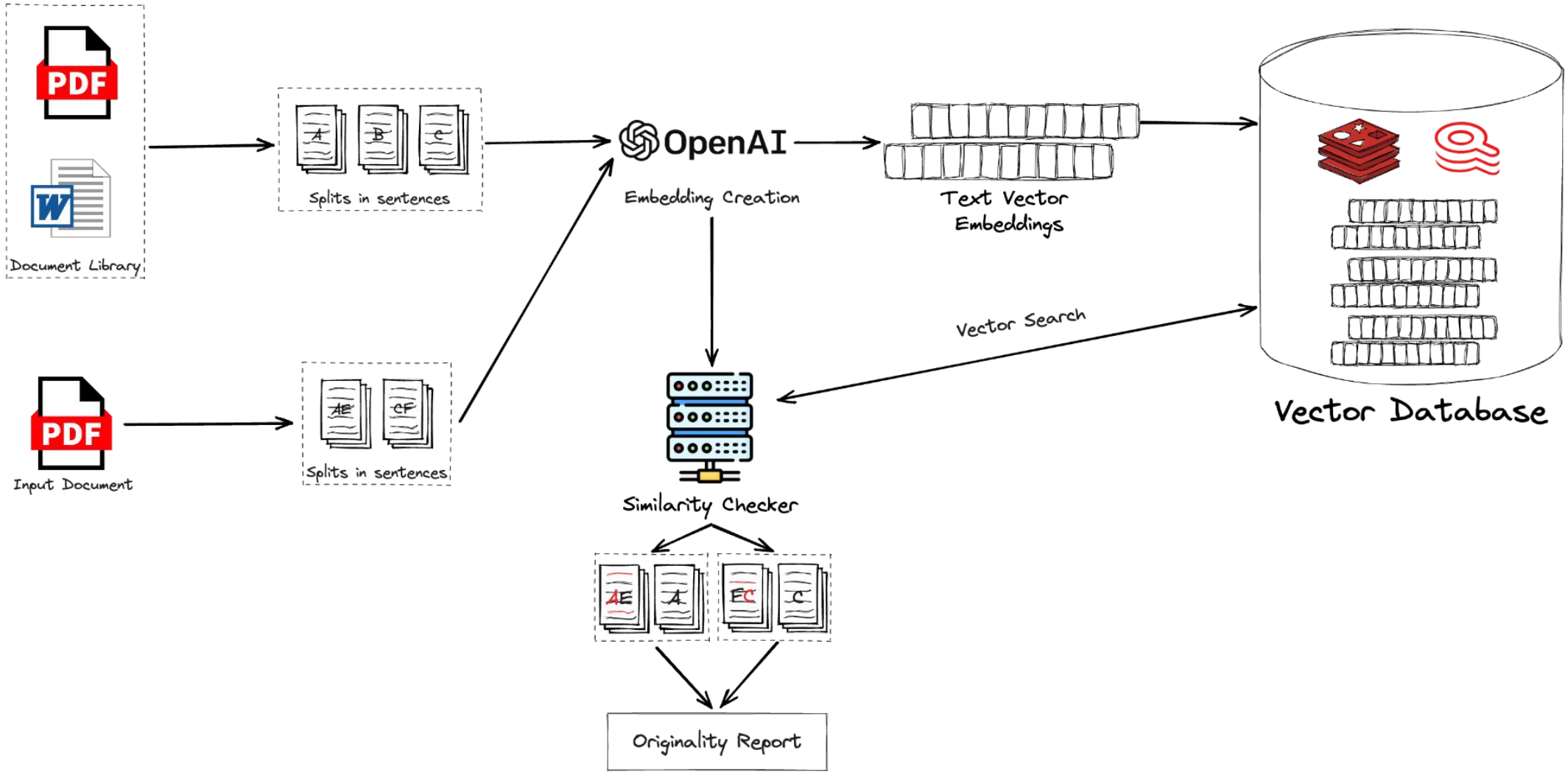
Traditional (Syntaxic) Search Results



redis

Examples of Vector Similarity Search use-cases

Semantic Search for Plagiarism Detection

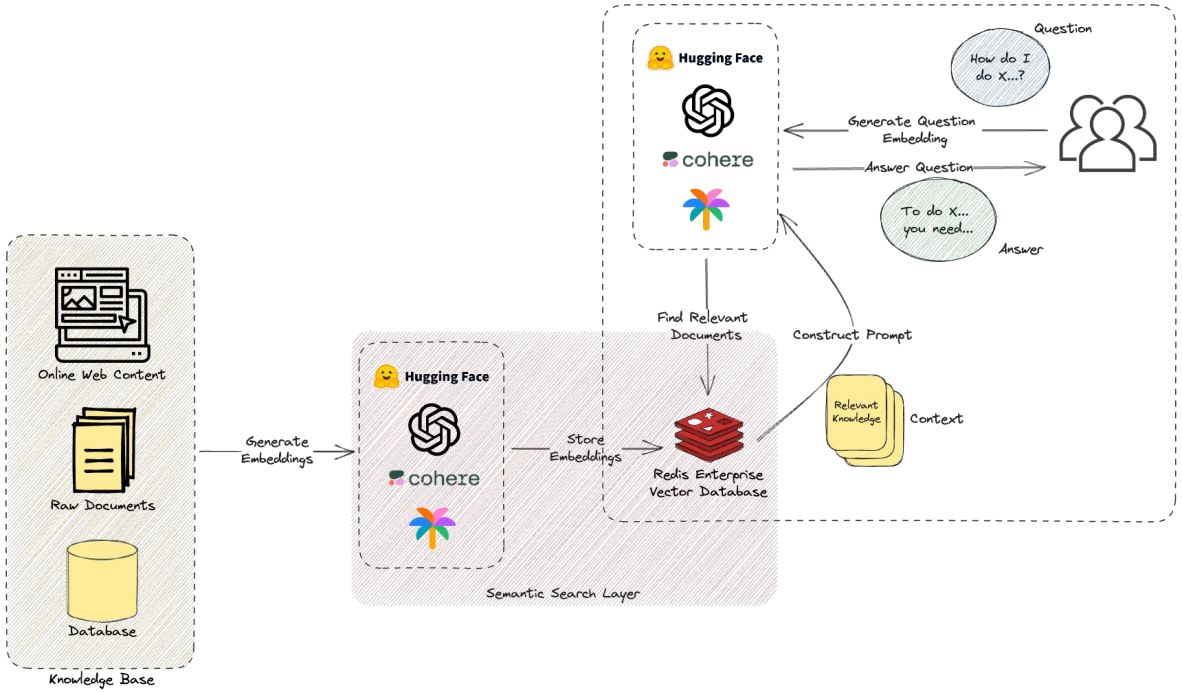


redis

Demo 4: Retrieval-Augmented Generation

redis

Retrieval-Augmented Generation - Concepts



Description:

- Redis is used as an external knowledge base for the large language model.
- Queries are used to detect similar information (context) within the knowledge base.
- LLM are used to make answers (based on context) look more human.

Benefits:

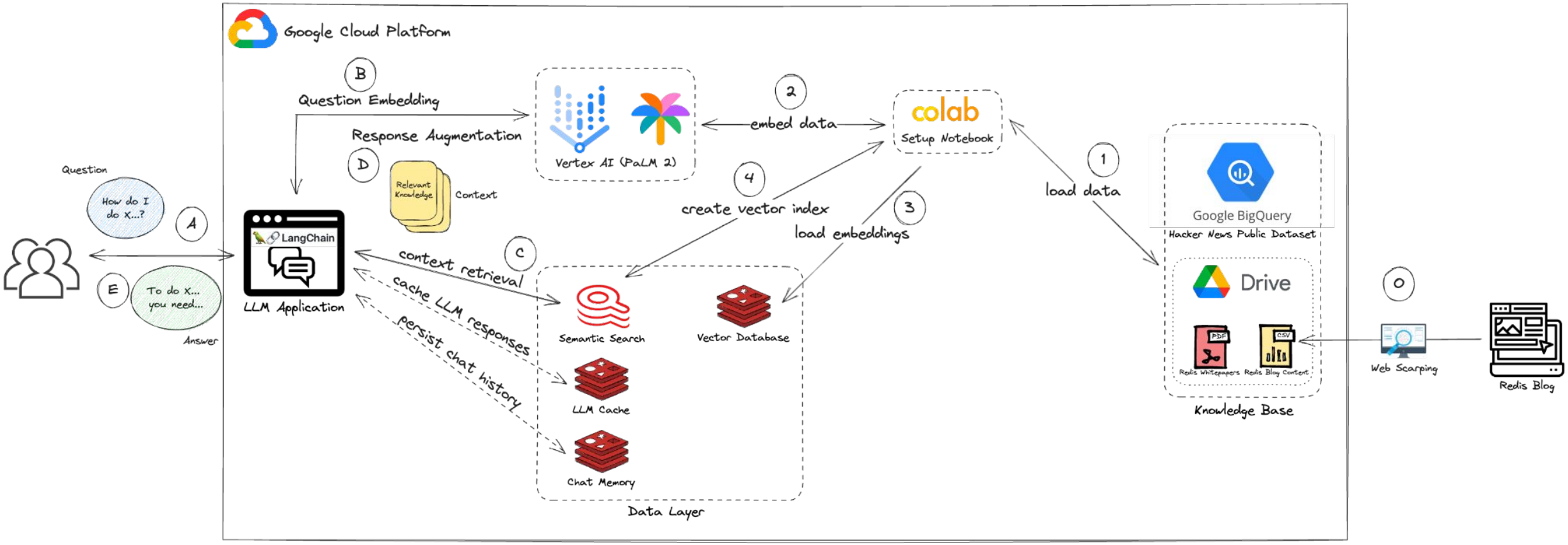
- All benefits of Semantic Search.
- **Efficacy** in handling complex conversational tasks.

Use Cases:

- Documents Summarization
- Shopping Assistants

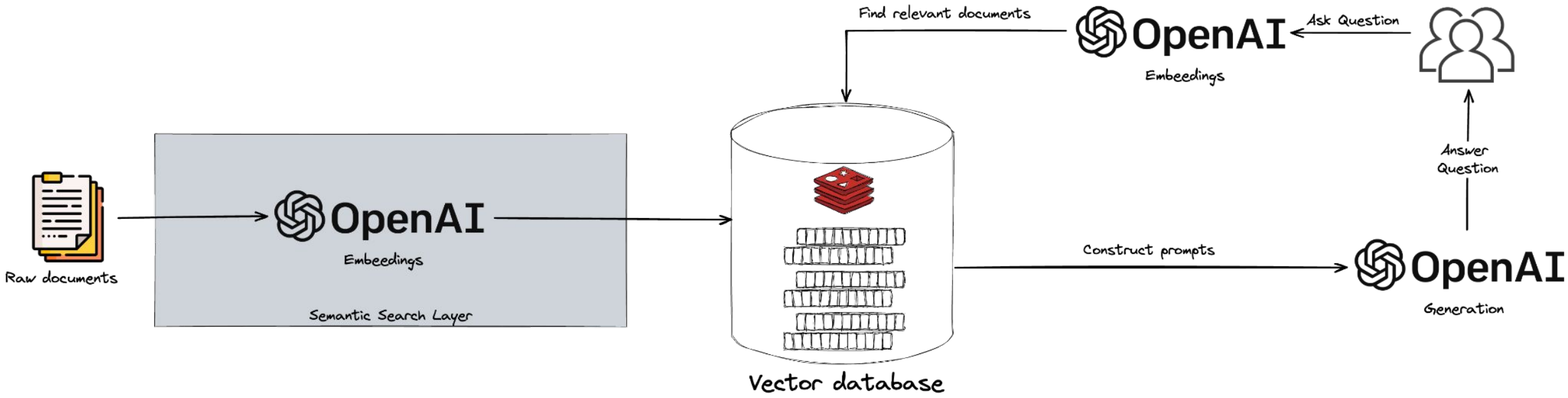
redis

Redis VSS Demo - Semantic Vector Search



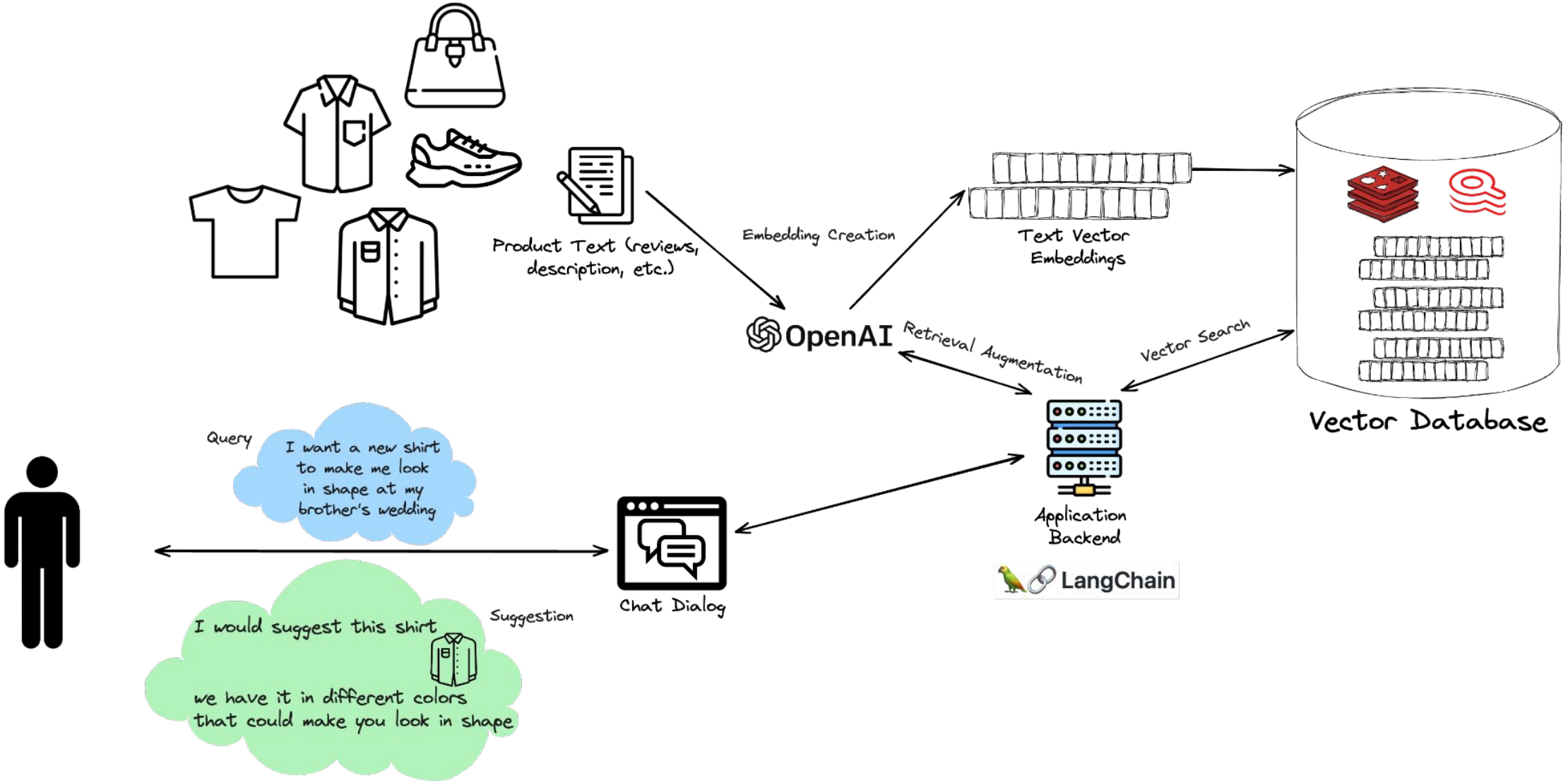
Examples of Vector Similarity Search use-cases

RAG for Domain-Specific Conversational ChatBots



Examples of Vector Similarity Search use-cases

RAG for E-Commerce ChatBots



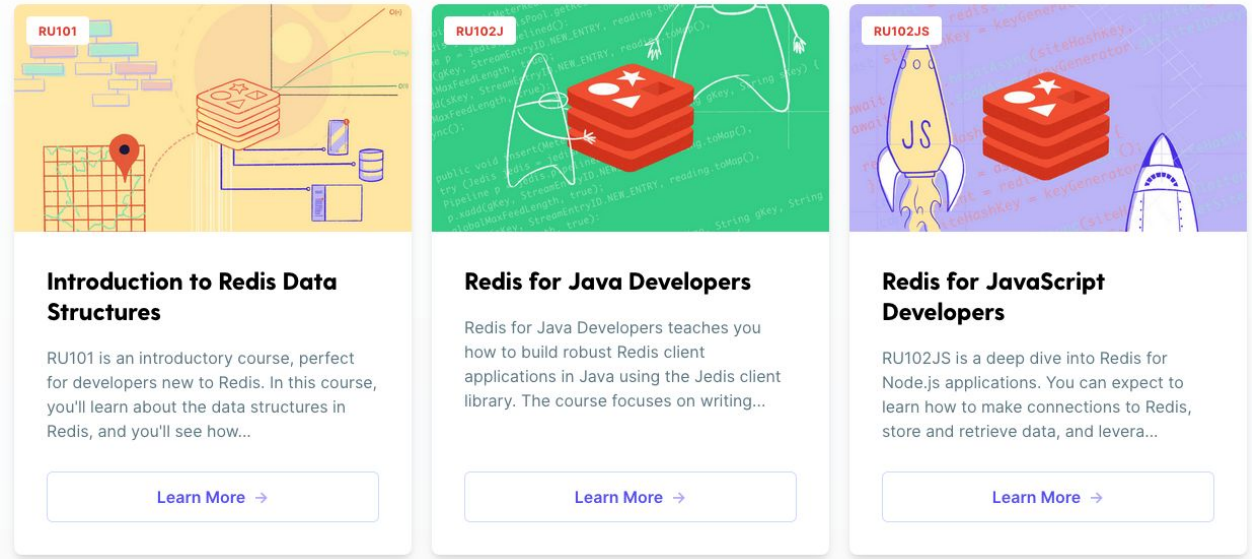
redis

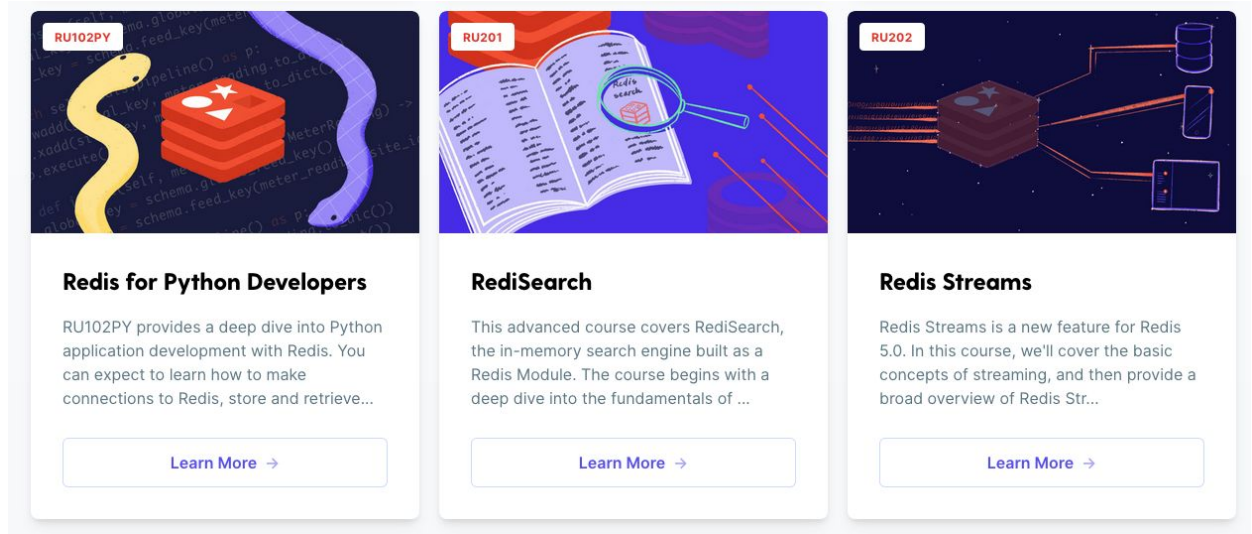
Resources

Redis University

- 7 free online courses
  - Introduction to data structures
  - Streams
  - RediSearch
  - Security
  - Development (Java, JS, Python)
- Video lectures, online lab environments, quizzes, and tests
- Take at your own pace
- Pass to earn a LinkedIn certificate

- Prepare for Redis certification





university.redislabs.com

redis

Resources about VSS

- https://developer.nvidia.com/blog/how-to-build-a-distributed-inference-cache-with-nvidia-triton-and-redis/
- https://openai.com/blog/march-20-chatgpt-outage
- https://github.com/RedisVentures/redis-openai-qna
- https://github.com/continuum-llms/chatgpt-memory
- https://platform.openai.com/docs/tutorials/web-qa-embeddings
- https://docsearch.redisventures.com
- https://ecommerce.redisventures.com
- https://www.datacrafterslab.com/2023-05-22-data-redis-part-7/

Thank You