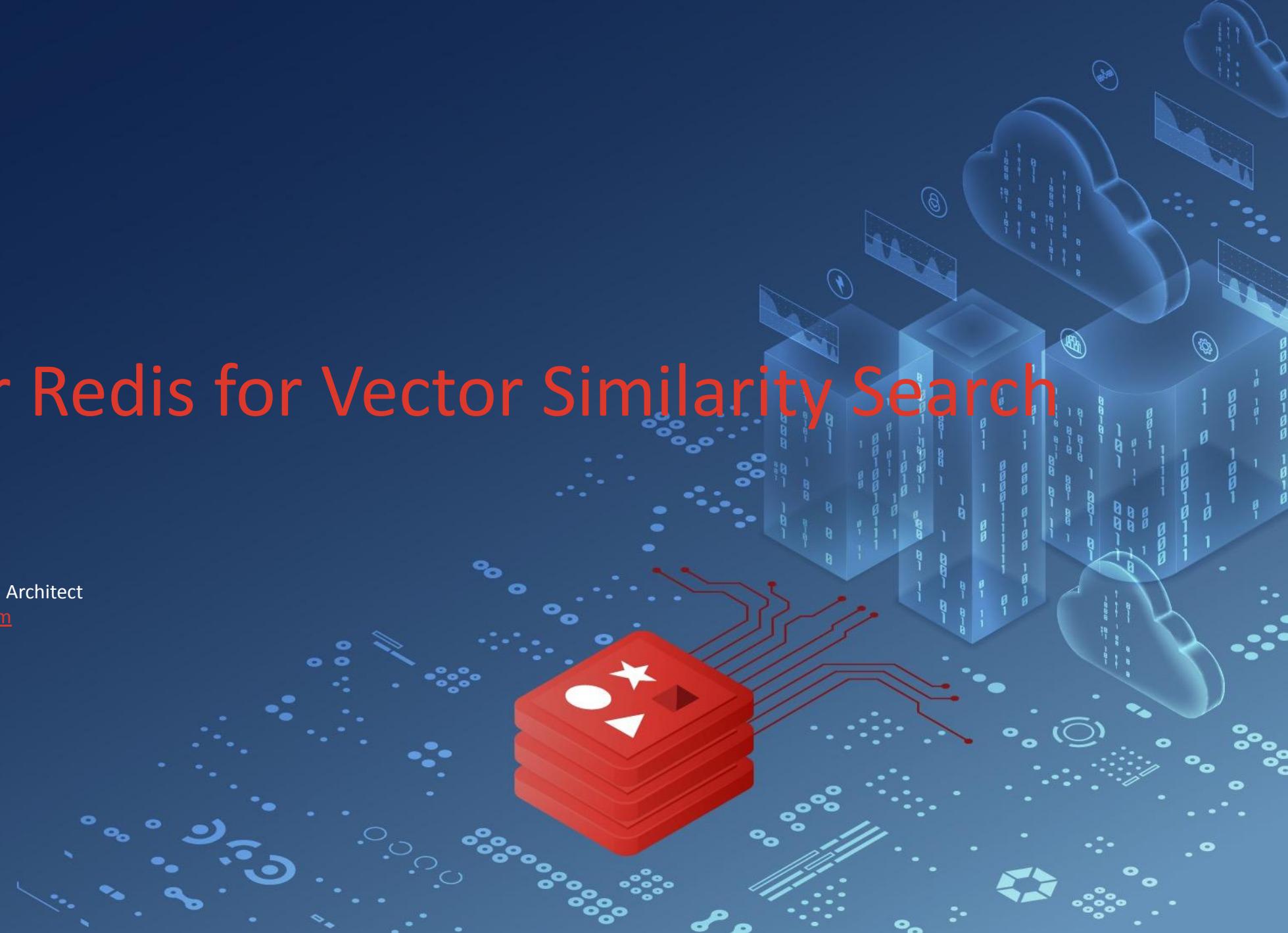




Rediscover Redis for Vector Similarity Search

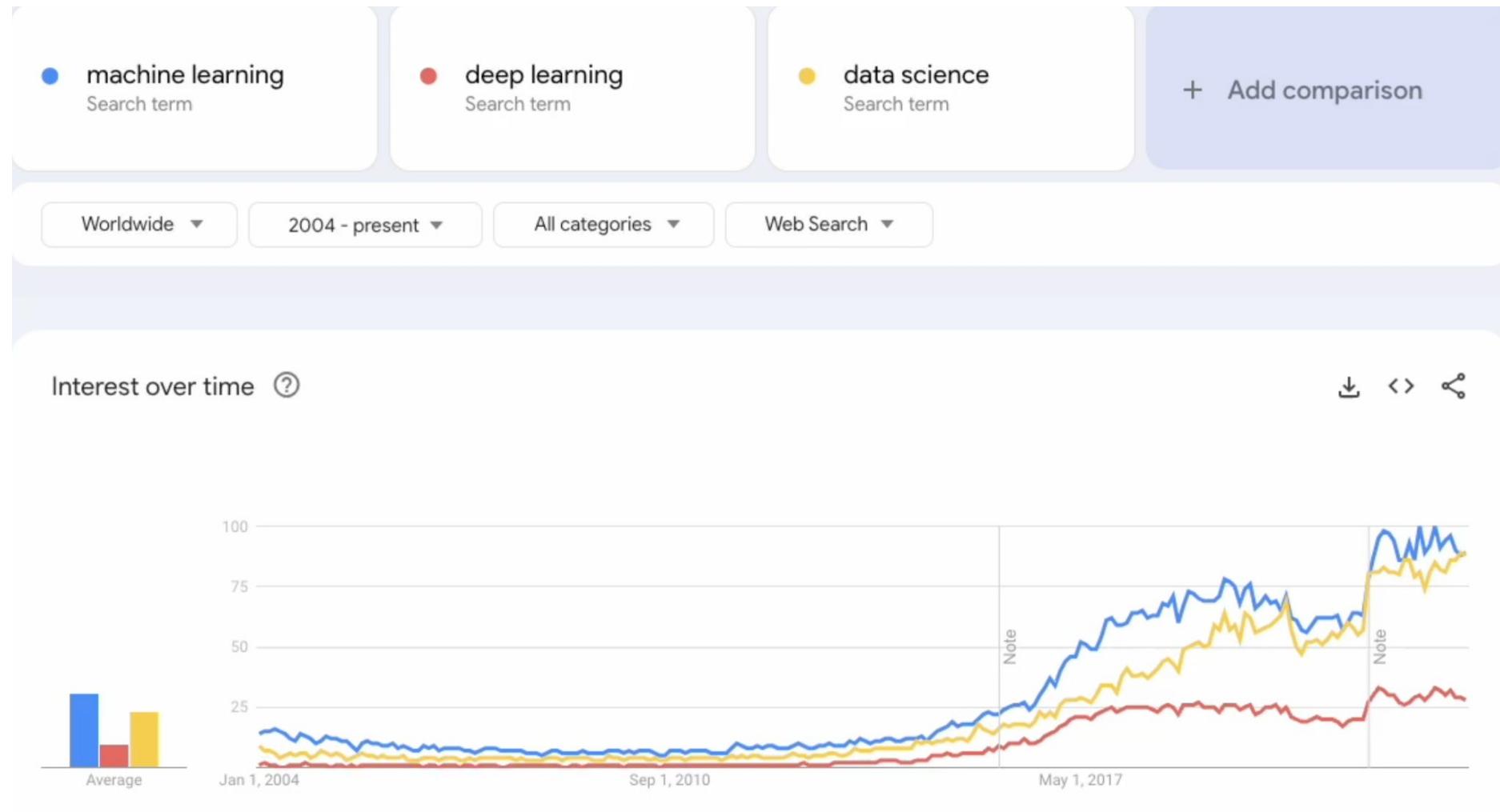


Amine El Kouhen, Solution Architect
amine.elkouhen@redis.com



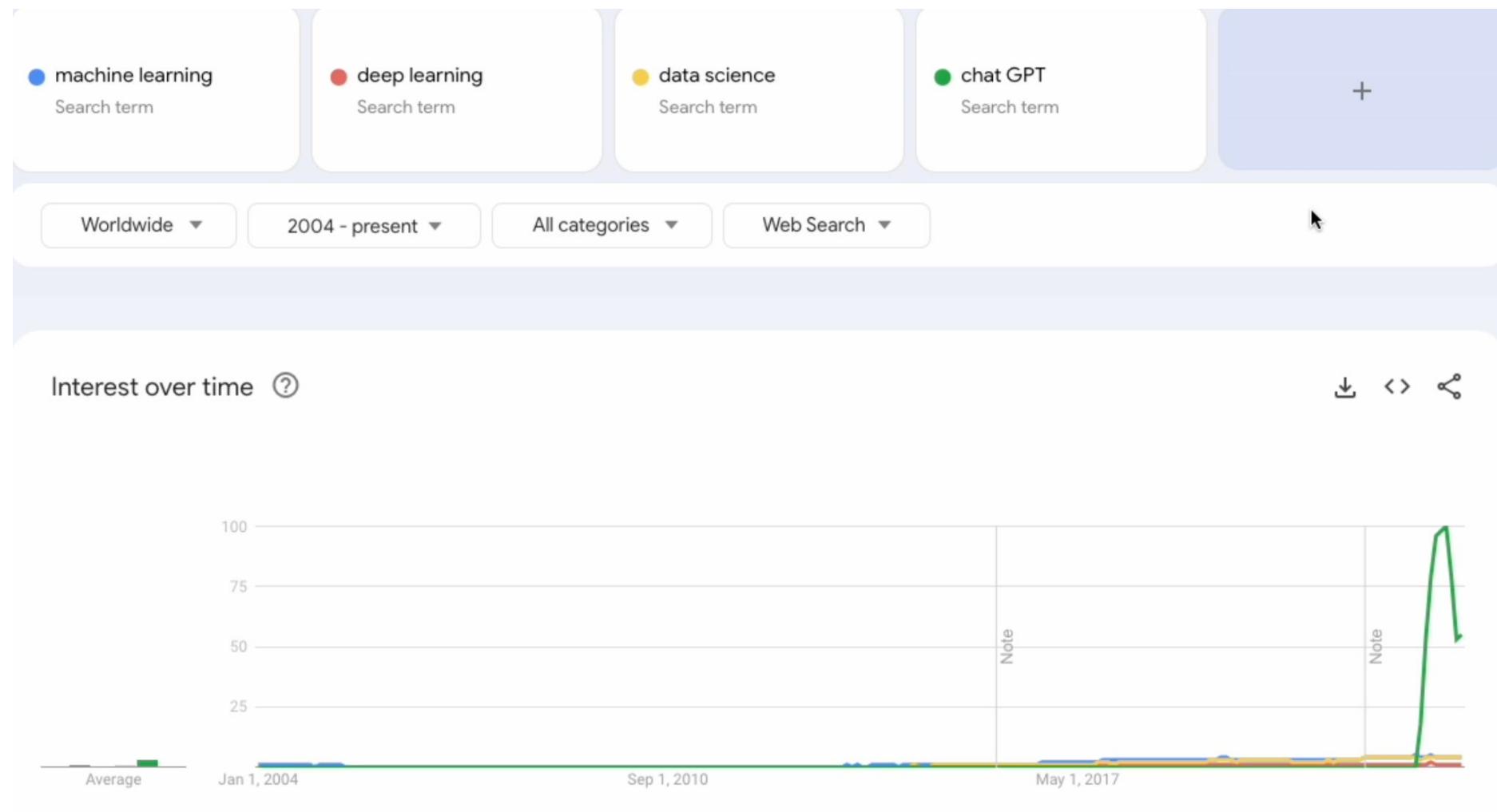
The Changing AI Landscape

2004 → 2022



The Changing AI Landscape

2004 → Now



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

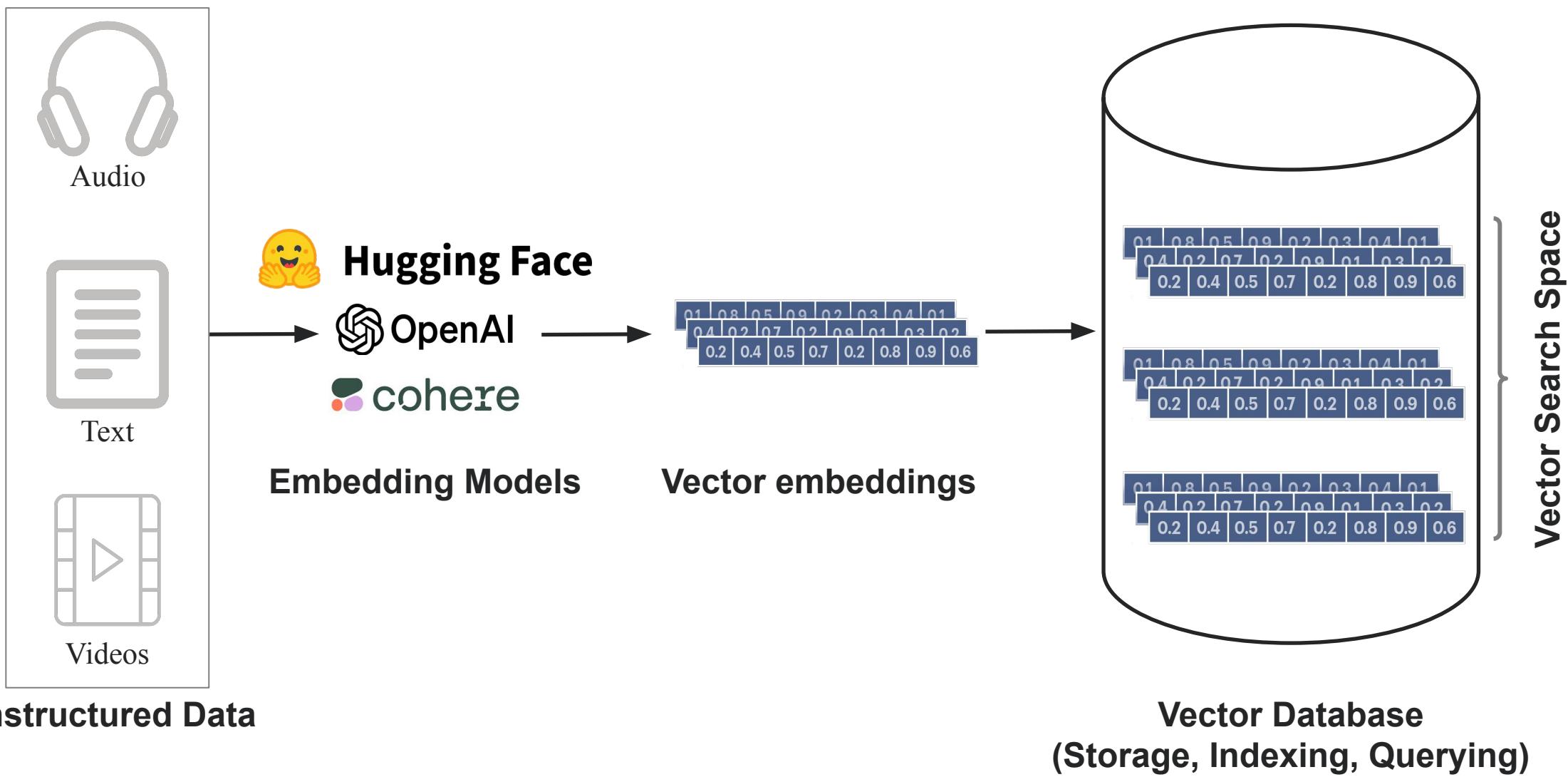
What is a Vector Embedding?

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

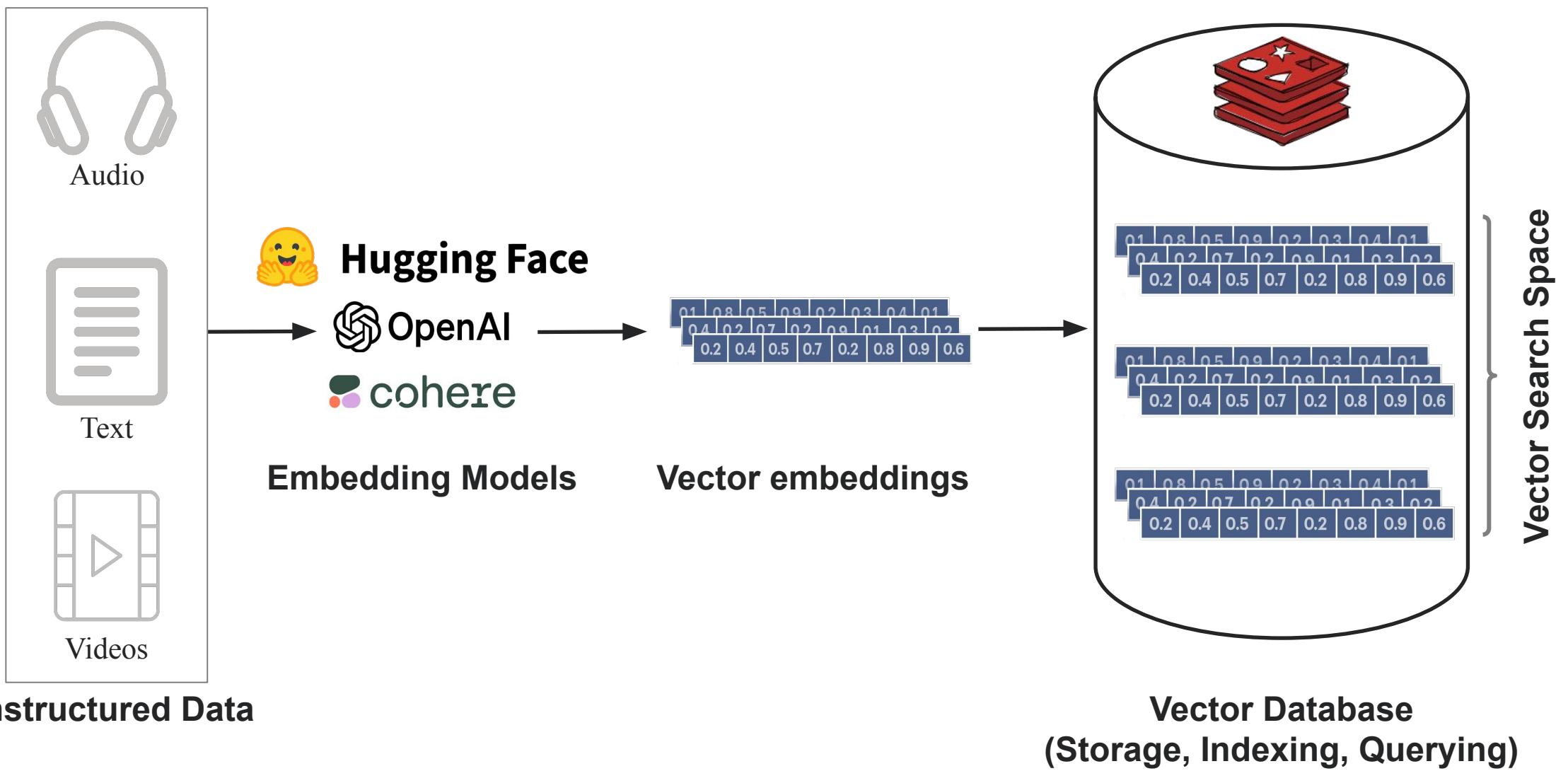
- Compact & dense data representation: A fixed-sized **list of floating-point numbers**,
- More often, used to represent **Unstructured Data** (Audio, Images, Text...),
- Produced by feature engineering or by pre-trained models (e.g., **Model providers**),
- Translate Perceived **Semantic Similarity**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



What is a Vector Database?



Redis as a Vector Database



Redis?! isn't it just a cache?



Redis Enterprise



Enterprise Grade Capabilities



Linear Scalability



High Availability



Durability



Backup & Restore



Geo-Distribution



Tiered-memory Access

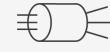


Multi-tenant



Security

Data Processing Engines



Event Streaming &
Data Integration



Query & Search



Triggered functions,
write-through,
write-behind

Developer Experience



RedisInsight



Data Structures



Strings



BitFields



Sorted Sets



Geospatial



Lists



JSON



Bitmaps



Hashes



Sets



HyperLog



Streams



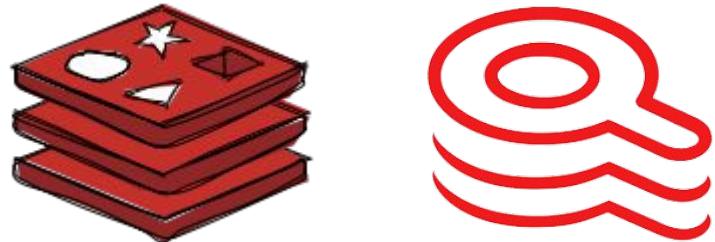
Probabilistic

Open Source Core



Redis - Vector Similarity Search

Redis + RedisSearch



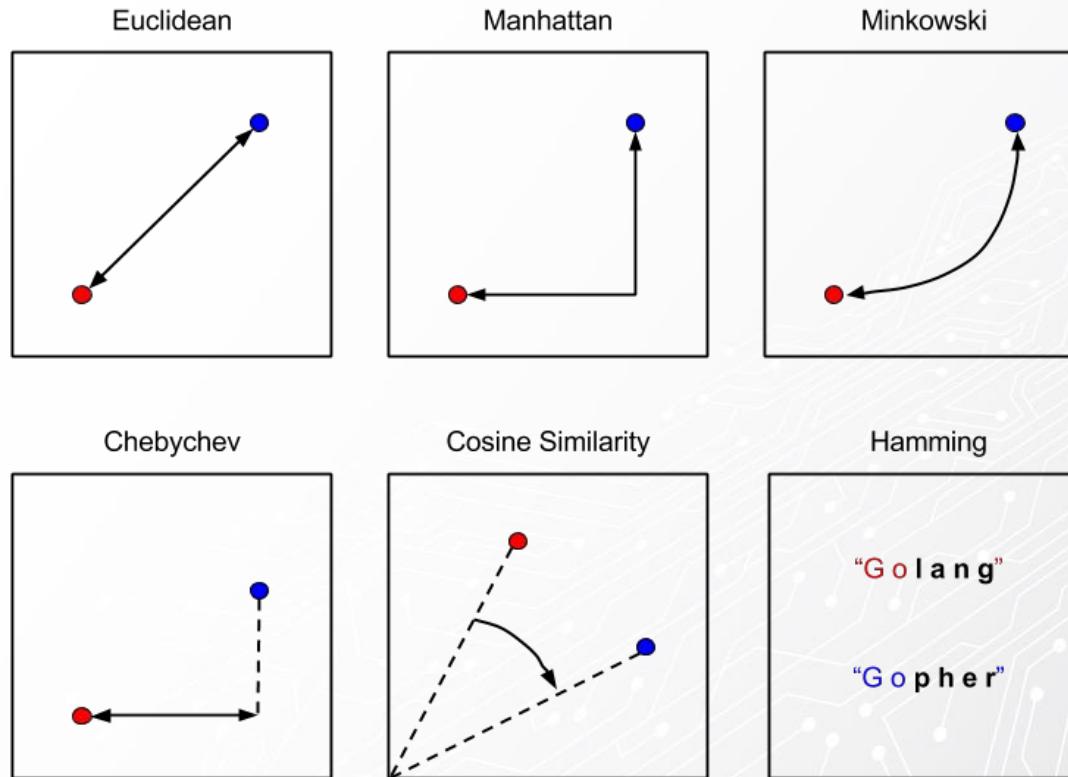
Integrations



- **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**:

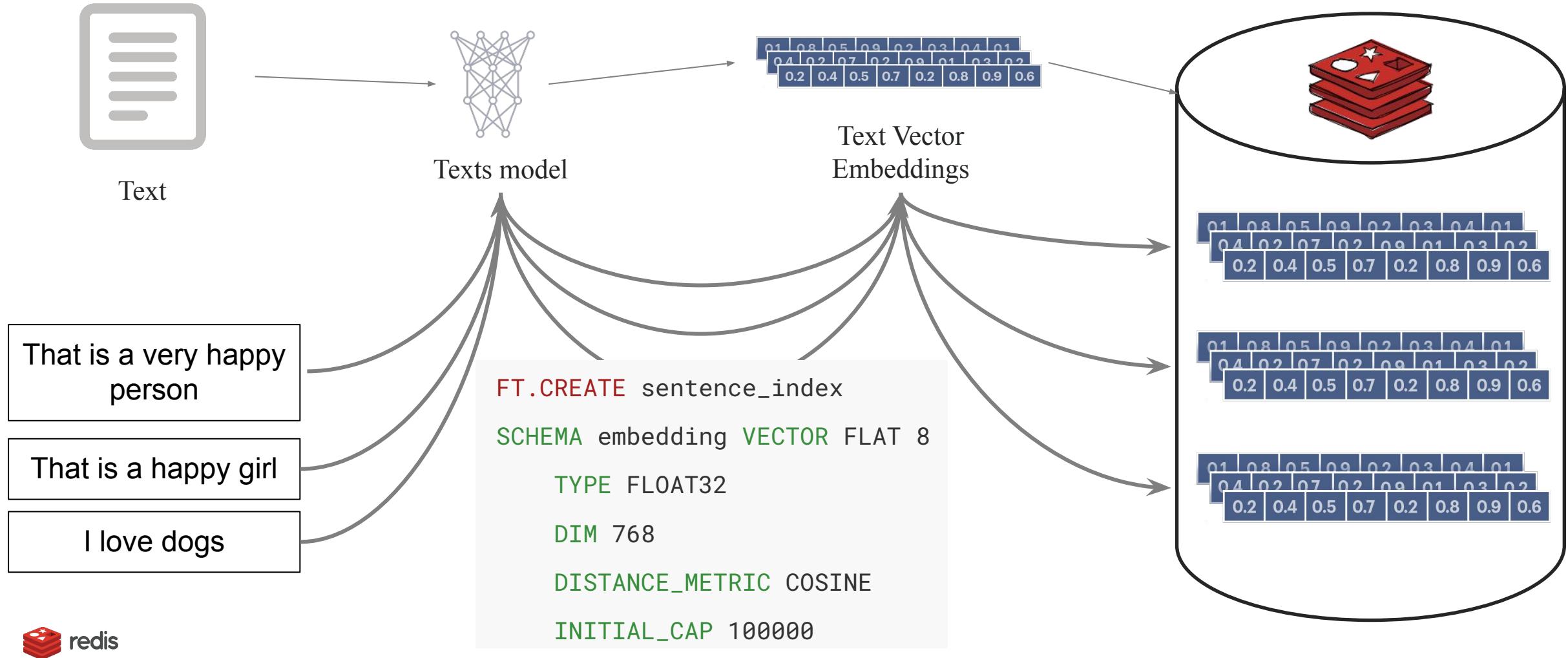




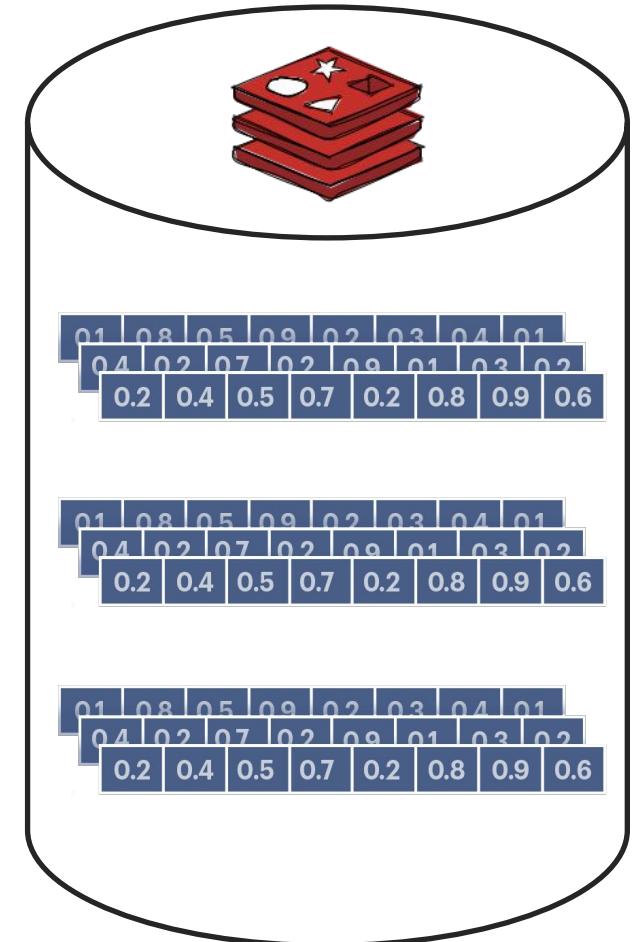
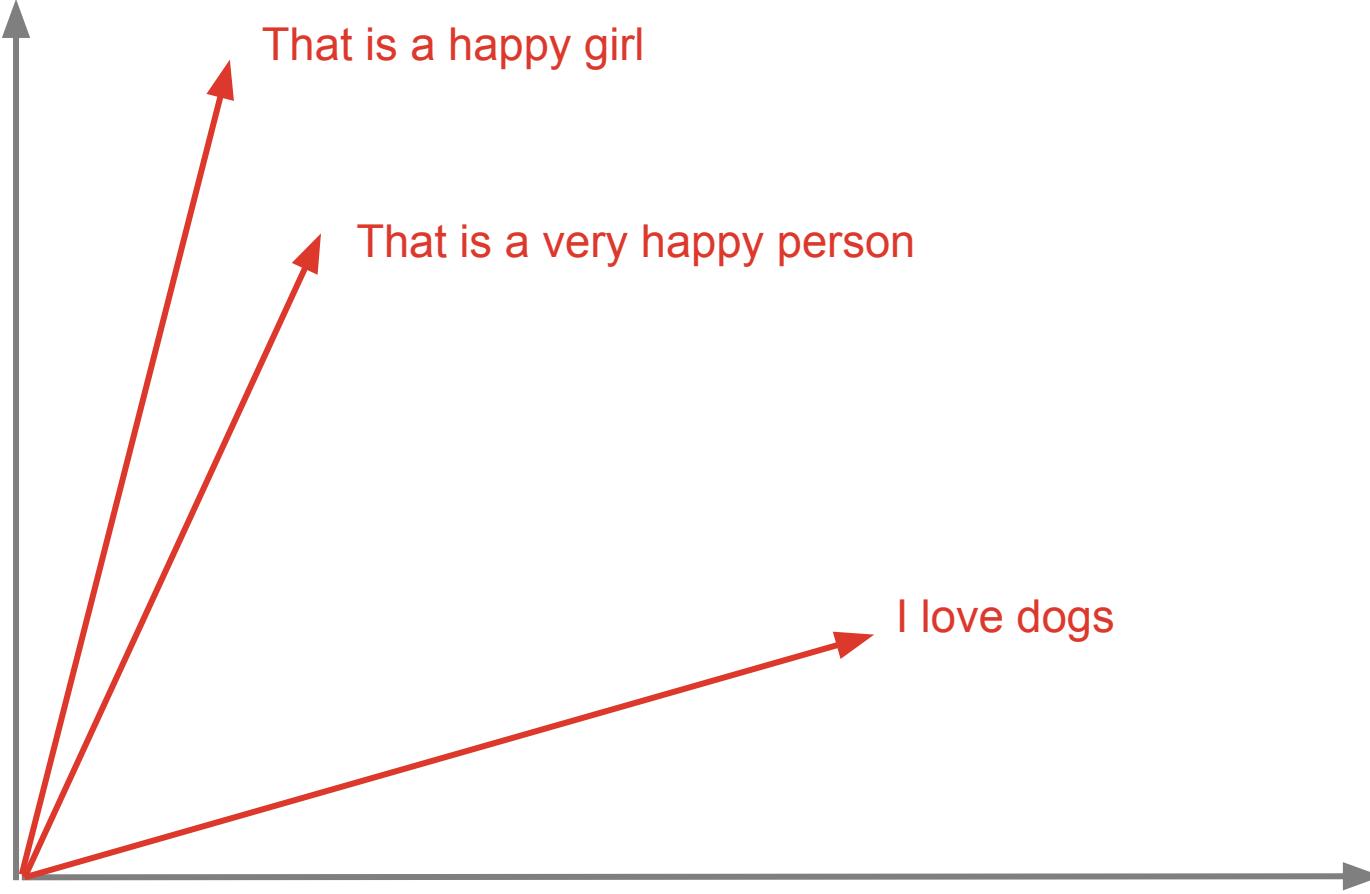
Demo 1: Text Vector Search



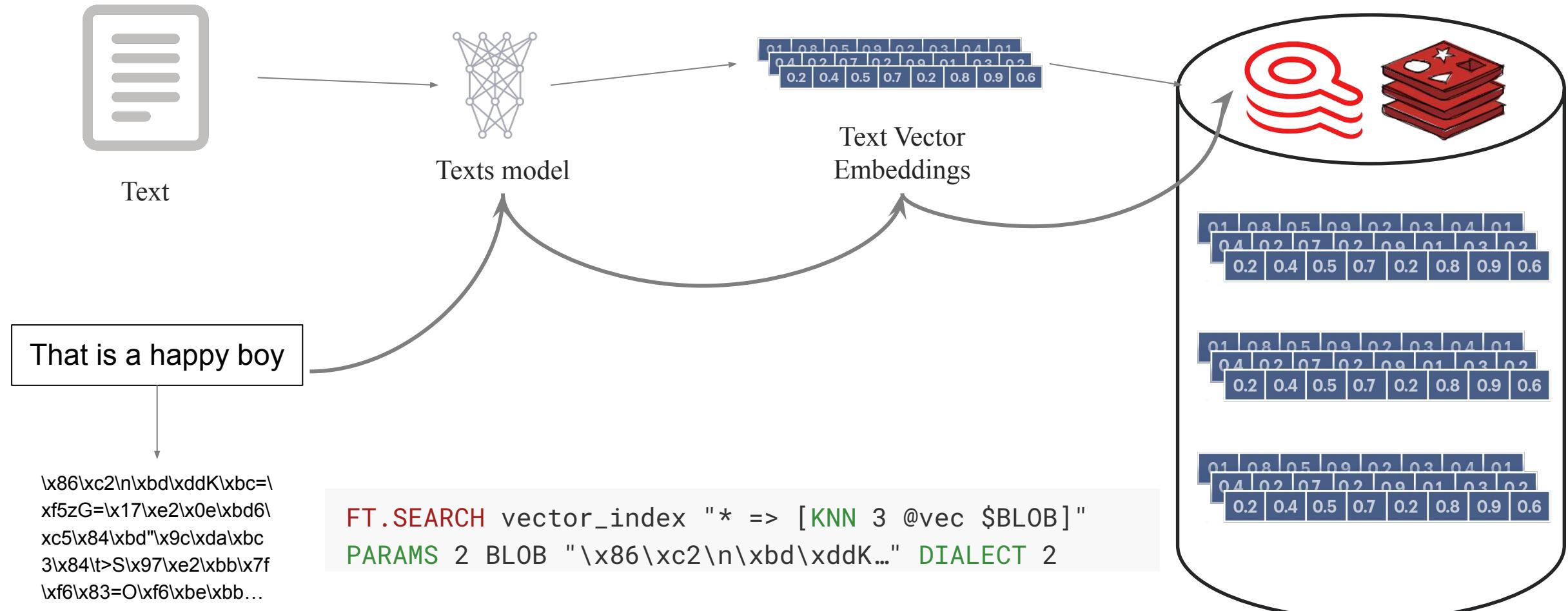
Text Vector Search - Vector Storage & Indexing



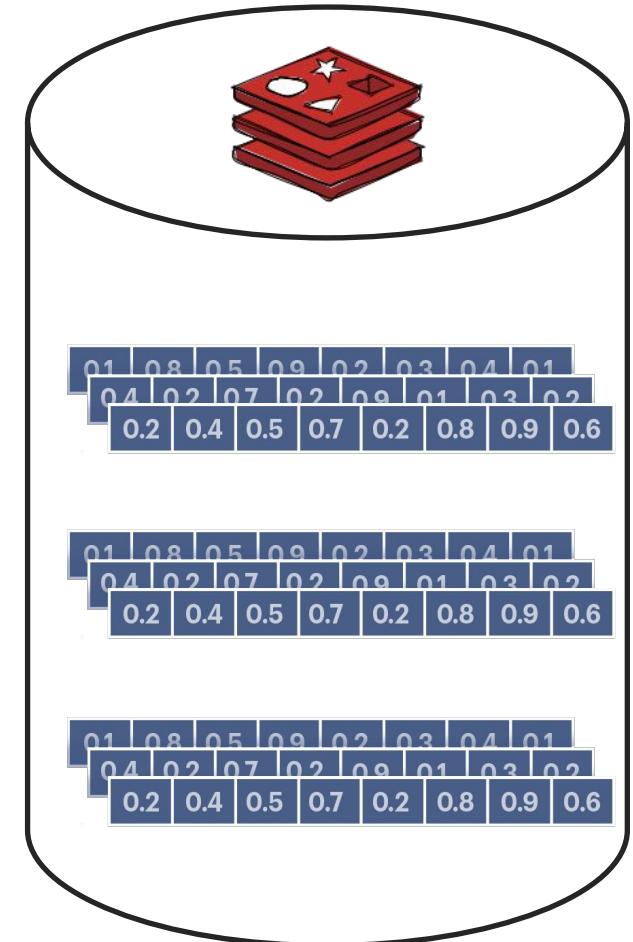
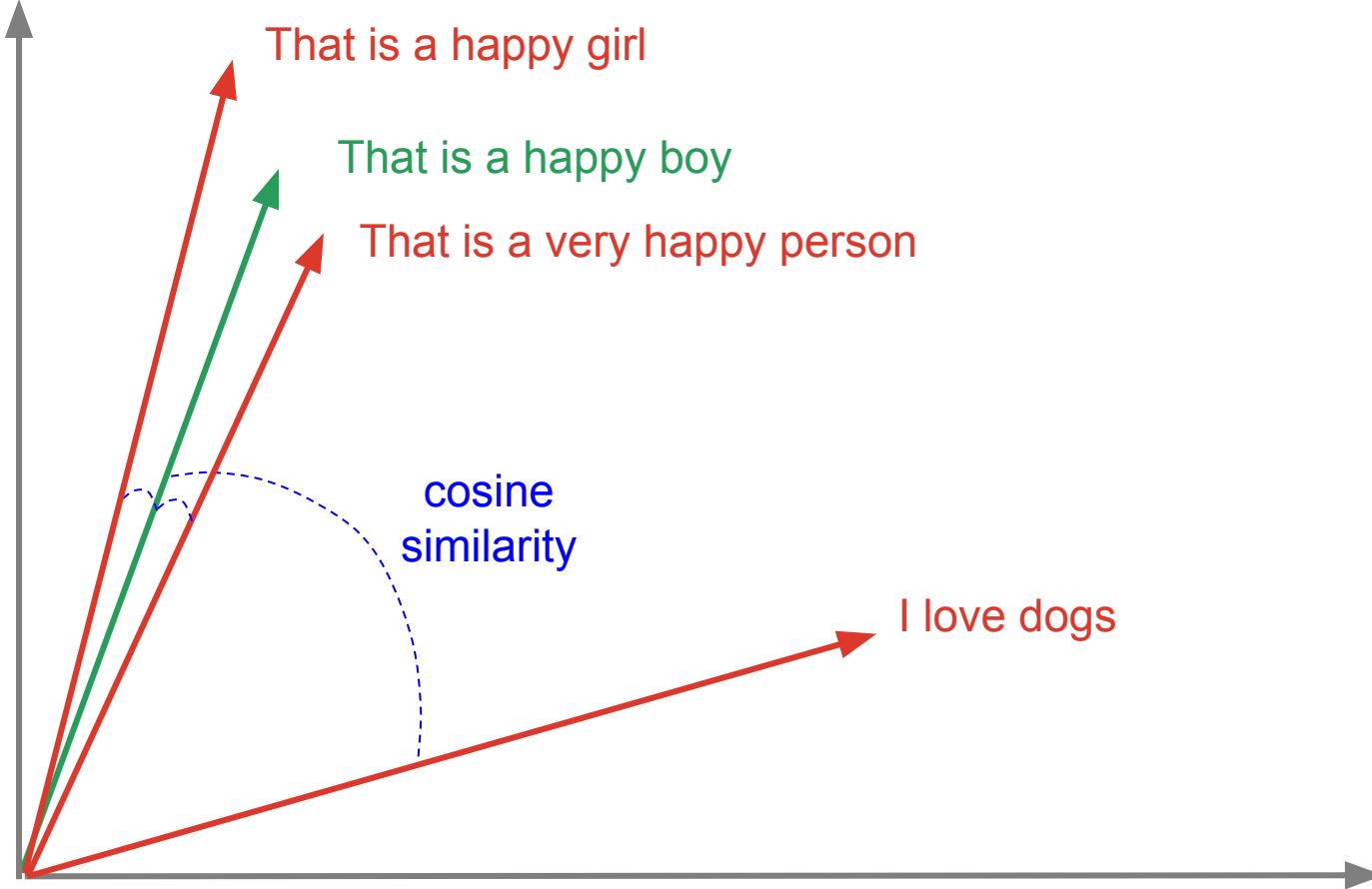
Text Vector Search - Vector Space



Text Vector Search - Vector Querying



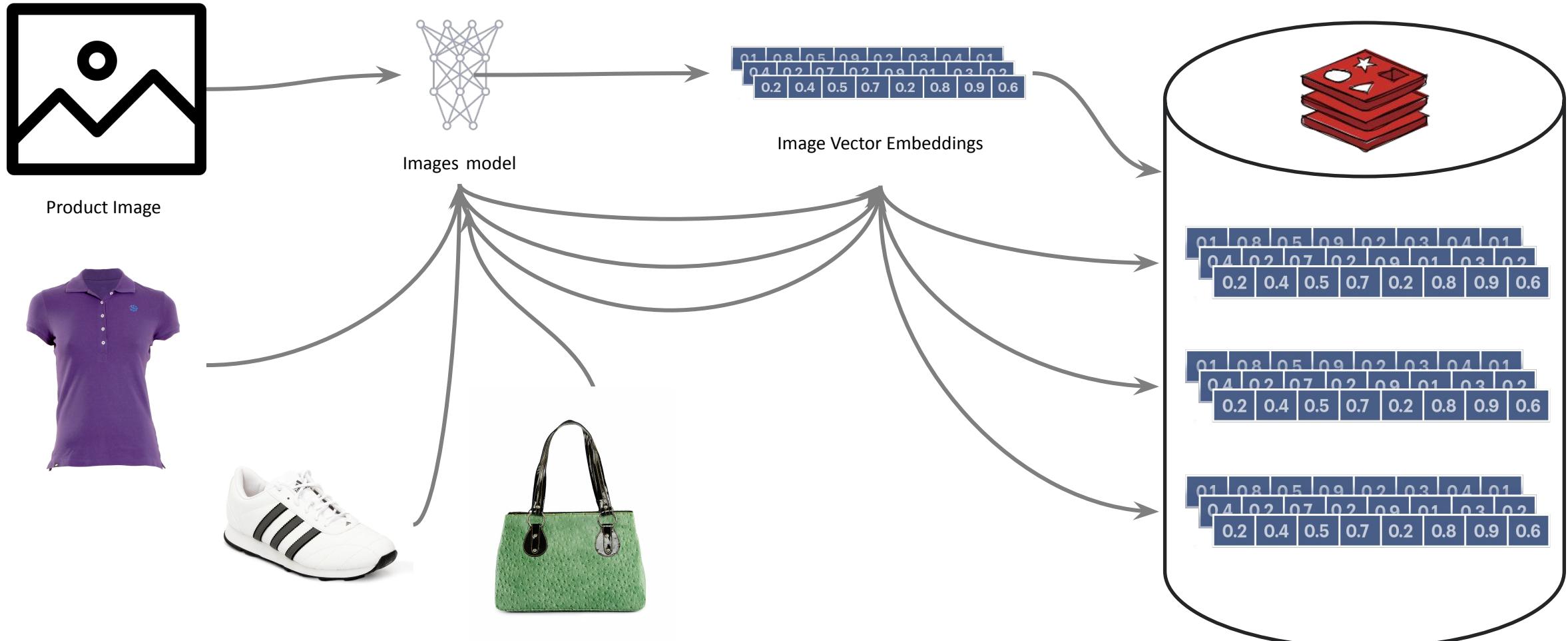
Text Vector Search - Vector Search Results



Demo 2: Visual + Hybrid Search



Visual Vector Search - Vector Storage



Redis VSS Demo - Visual Vector Search

The screenshot shows a web browser window for the "Redis Vector Search Demo" at ecommerce.redisventures.com. The title bar says "Redis VSS Demo". The main content is titled "Fashion Product Finder" and includes a subtext: "This demo uses the built in Vector Search capabilities of Redis Enterprise to show how unstructured data, such as images and text, can be used to create powerful search engines." Below this are two buttons: "Apply Filters" and "Load More Products". A message "40000 searchable products" is displayed above a grid of 10 product cards. Each card contains a product image, its name, and a "View Similar:" button with "By Text" and "By Image" options. The first product card, featuring a white ADIDAS sneaker, has its "By Image" button circled in red. A red arrow from the left points to this circled button, labeled "Search by Image".

40000 searchable products

ADIDAS Men White Sports Shoes

View Similar: [By Text](#) [By Image](#)

ADIDAS Originals Women Rekapi Purple Casual Shoes

View Similar: [By Text](#) [By Image](#)

Raymond Men Beige Socks

View Similar: [By Text](#) [By Image](#)

Giordano Men White Dial Watch

View Similar: [By Text](#) [By Image](#)

Levi's Men Printed Blue Tshirts

View Similar: [By Text](#) [By Image](#)

Puma Men's Faas Blue White Silver Red Shoe

View Similar: [By Text](#) [By Image](#)

Kraus Jeans Women Beige Shorts

View Similar: [By Text](#) [By Image](#)

Roxy Women Curves Blue Belt

View Similar: [By Text](#) [By Image](#)

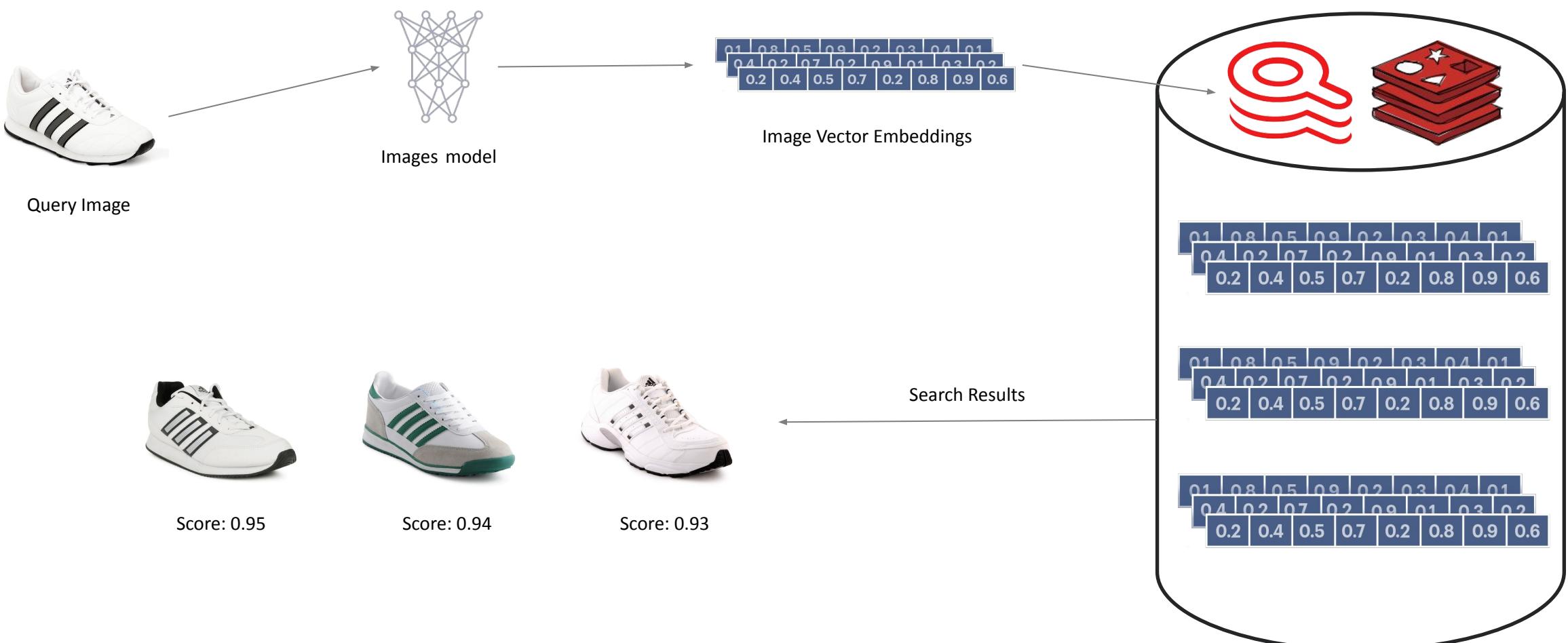
Indigo Nation Men Alnoite Black Belts

View Similar: [By Text](#) [By Image](#)

Boss Men Pure Perfume

View Similar: [By Text](#) [By Image](#)

Visual Vector Search - Vector Querying



Redis VSS Demo - Visual Vector Search



Query Image

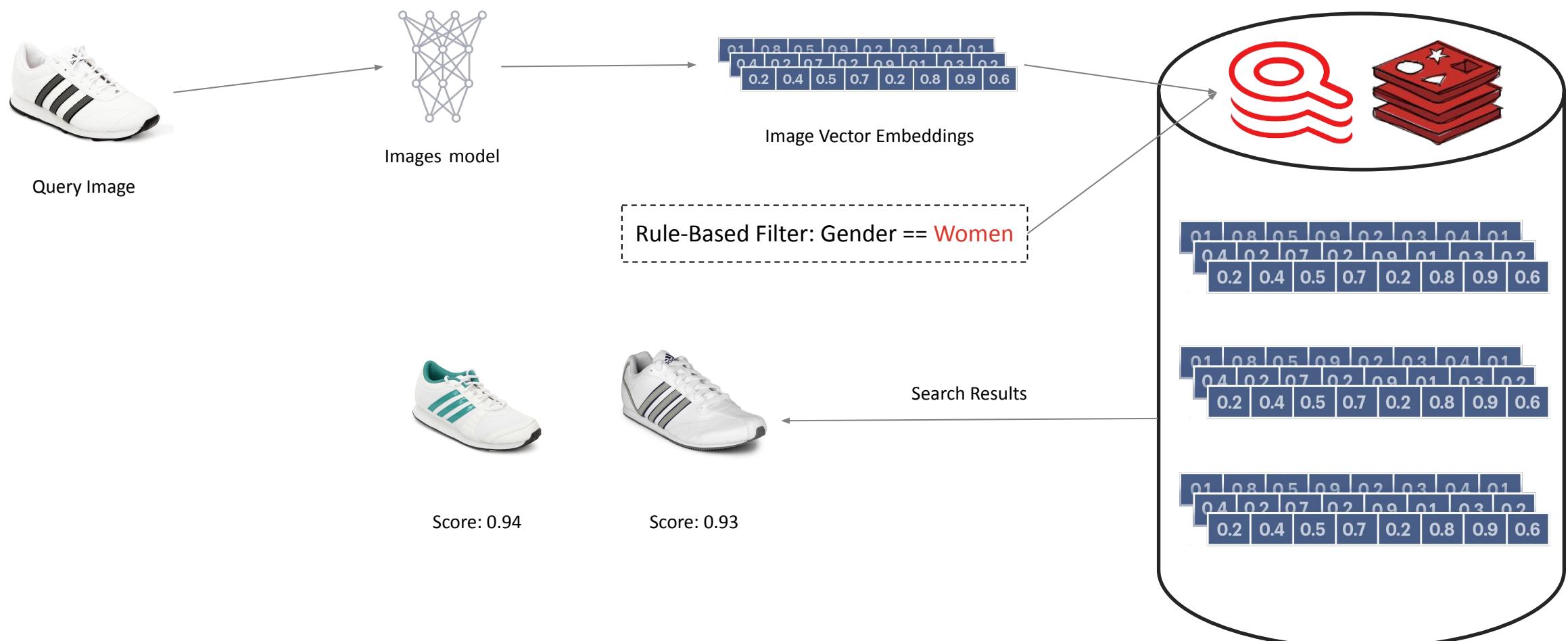
The screenshot shows a web browser window titled "Redis VSS Demo" at "ecommerce.redisventures.com". The page is titled "Fashion Product Finder" and explains the use of Redis Vector Search for unstructured data like images and text. It features a grid of 10 product cards, each with a thumbnail, name, and a "View Similar" section. The third product in the top row, "Numero Uno Men White Casual Shoes", has its "View Similar" section highlighted with an orange circle around the similarity score "0.94". A red arrow points from this score to the text "Product (vector) Similarity Score" located on the right side of the page.

Product Name	Similarity Score
ADIDAS Men White Sports Shoes	1.00
ADIDAS Men White Pluto Sports Shoes	0.95
Numero Uno Men White Casual Shoes	0.94
ADIDAS Men White Desma Sports Shoes	0.91
Numero Uno Men White Casual Shoes	0.91
Fila Men Paramount Plus White Shoes	0.94
ADIDAS Women White Bolt Sports Shoes	0.94
ADIDAS Men Sports White Sports Shoes	0.93
Numero Uno Men White Casual Shoes	0.93
ADIDAS Men Black Shoes	0.93

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

Product (vector)
Similarity Score

Hybrid Search - Filtered Vector Querying



Redis VSS Demo - Hybrid Search

Gender used as
pre-filter for
candidate selection



Query Image

The screenshot shows a web browser displaying the "Redis Vector Search Demo" website at ecommerce.redisventures.com. The page title is "Fashion Product Finder". A text overlay on the left states "40000 searchable products". A dropdown menu titled "Gender: Women" is highlighted with an orange circle. Another box highlights the "Category" section with radio buttons for "Apparel", "Accessories", and "Footwear", with "Footwear" selected. Below this, a grid of product cards shows various women's shoes. Each card includes a thumbnail, the product name, and a "View Similar" section with "By Text" and "By Image" buttons and a similarity score. An orange arrow points from the "Query Image" on the left to the "Gender: Women" filter on the demo page.

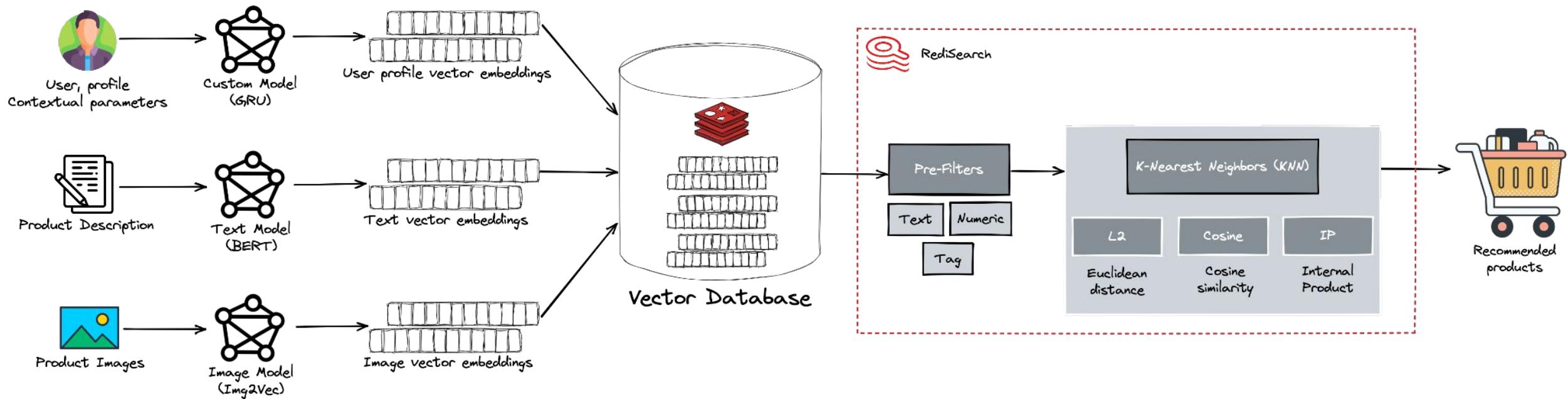
Hybrid Search: The power of vector search enhanced with rule-based filters.

Supports all RedisSearch Types:

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

Examples of Vector Similarity Search use-cases

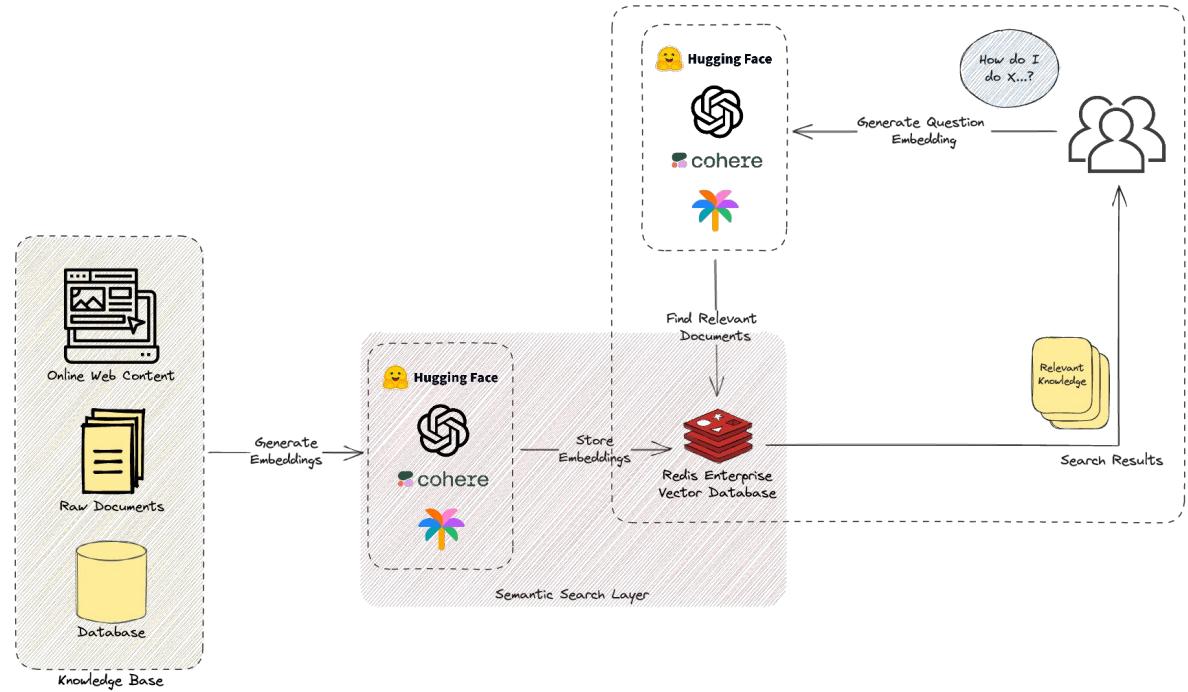
Hybrid Search for Recommendation Systems



Demo 3: Semantic Search



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

The screenshot shows a web browser window titled "Redis VSS Demo" at "docsearch.redisventures.com". The page is titled "arXiv Paper Search" and describes Redis as a real-time data platform. A search bar contains the query "machine learning model that can track trends in the fashion industry". Below the search bar are filters for "Embedding Model" (set to HuggingFace), "Year" (dropdown), and "Category" (dropdown). The results section displays two papers:

- A Novel Approach to Analyze Fashion Digital Archive from Humanities**
Authors: Satoshi Takahashi and Keiko Yamaguchi and Asuka Watanabe
Categories: cs.DL
Year: 2021
Similarity Score: 0.71
- Style in the Age of Instagram: Predicting Success within the Fashion Industry using Social Media**
Authors: Jaehyuk Park, Giovanni Luca Ciampaglia, Emilio Ferrara
Categories: cs.CY, cs.SI, physics.soc-ph
Year: 2016
Similarity Score: 0.70

Annotations include a red arrow pointing from the "Query" label to the search bar, and an orange arrow pointing from the "Paper (vector) Similarity Score" label to the numerical scores (0.71 and 0.70).

Semantic Vector Search vs Traditional Search

Semantic (Vector) Search Results

The screenshot shows a web browser window titled "Redis VSS Demo" with the URL "docsearch.redisventures.com". The search bar contains the query "machine learning model that can track trends in the fashion industry". Below the search bar, there are filters for "Embedding Model" (set to HuggingFace), "Year" (dropdown menu), and "Category" (dropdown menu). The main results area displays three arXiv papers:

- A Novel Approach to Analyze Fashion Digital Archive from Humanities**
Authors: Satoshi Takahashi and Keiko Yamaguchi and Asuka Watanabe
Categories: cs.DL
Year: 2021
Score: 0.71
Buttons: More Like This, Download
- Style in the Age of Instagram: Predicting Success within the Fashion Industry using Social Media**
Authors: Jaehyuk Park, Giovanni Luca Ciampaglia, Emilio Ferrara
Categories: cs.CY, cs.SI, physics.soc-ph
Year: 2016
Score: 0.70
Buttons: More Like This, Download
- Leveraging Multiple Relations for Fashion Trend Forecasting Based on Social Media**
Authors: Yujuan Ding, Yunshan Ma, Lizi Liao, Wai Keung Wong, Tat-Seng Chua
Categories: cs.LG, cs.IR, cs.MM
Year: 2021
Score: 0.69
Buttons: More Like This, Download

Traditional (Syntactic) Search Results

The screenshot shows a web browser window titled "Search | arXiv e-print repository" with the URL "arxiv.org/search/?query=machine+learning+model+that+can+track+trends+in+the+fashion+industry&searchtype=all&a...". The page header includes the Cornell University logo, the arXiv logo, and a message about Simons Foundation support. The search bar also contains the same query as the Redis demo.

Search
Search v0.5.6 released 2020-02-24 | Feedback | Advanced Search | Login

machine learning model that can track trends in the fashion industry | All fields | Search

(Show abstracts) (Hide abstracts)

Sorry, your query for all: machine learning model that can track trends in the fashion industry produced no results.

Searching by Author Name

- Using the **Author(s)** field produces best results for author name searches.
- For the most precise name search, follow **surname(s), forename(s)** or **surname(s), initial(s)** pattern: example Hawking, S or Hawking, Stephen
- For best results on multiple author names, separate individuals with a ; (semicolon). Example: Jin, D S; Ye, J
- Author names enclosed in quotes will return only **exact matches**. For example, "Stephen Hawking" will not return matches for Stephen W. Hawking.
- Diacritic character variants are automatically searched in the Author(s) field.
- Queries with no punctuation will treat each term independently.

Searching by subcategory

- To search within a subcategory select **All fields**.
- A subcategory search can be combined with an author or keyword search by clicking on **add another term** in advanced search.

Tips

Wildcards:

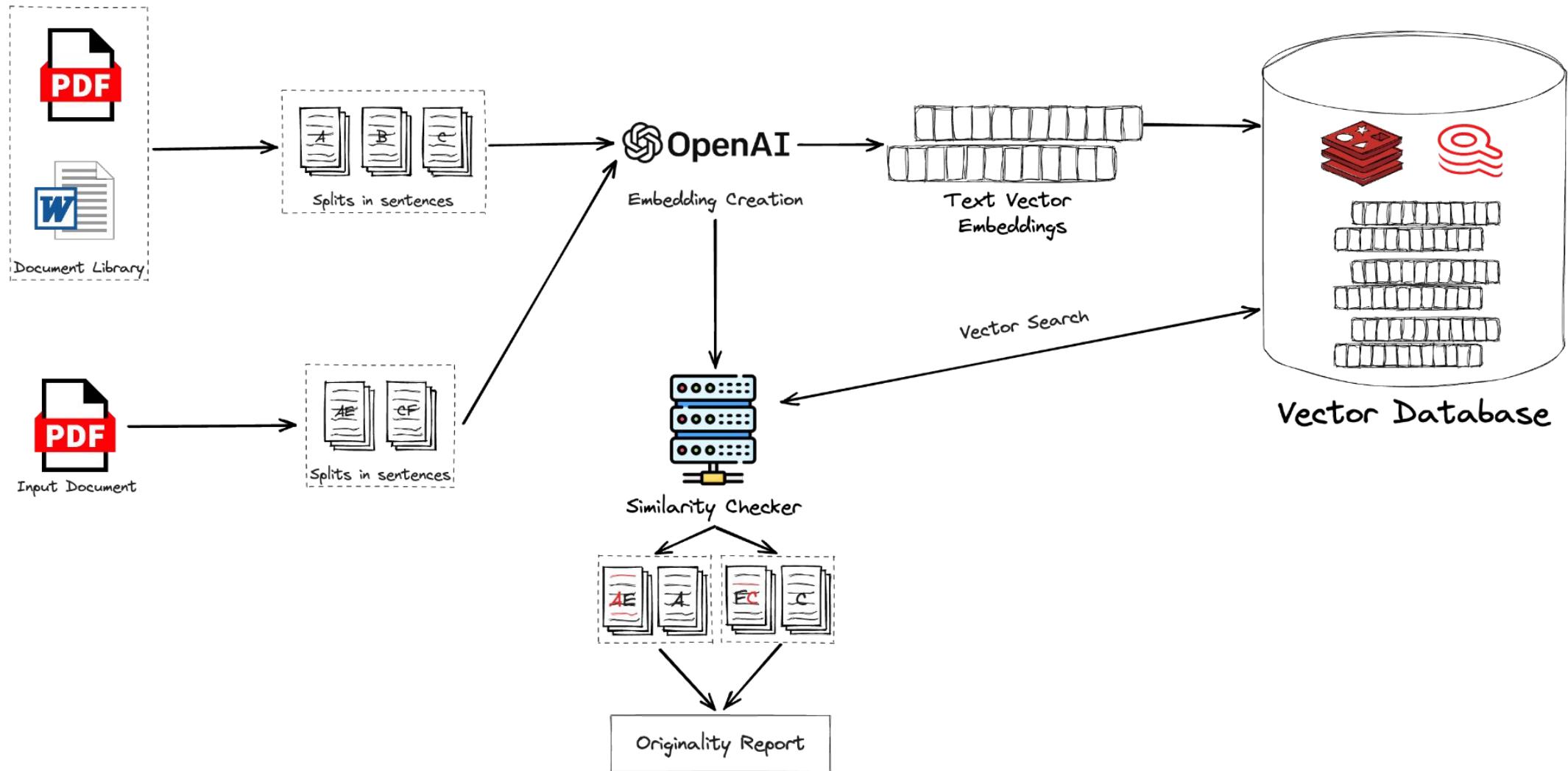
- Use ? to replace a single character or * to replace any number of characters.
- Can be used in any field, but not in the first character position. See Journal References tips for exceptions.

Expressions:

- TeX expressions can be searched, enclosed in single \$ characters.

Examples of Vector Similarity Search use-cases

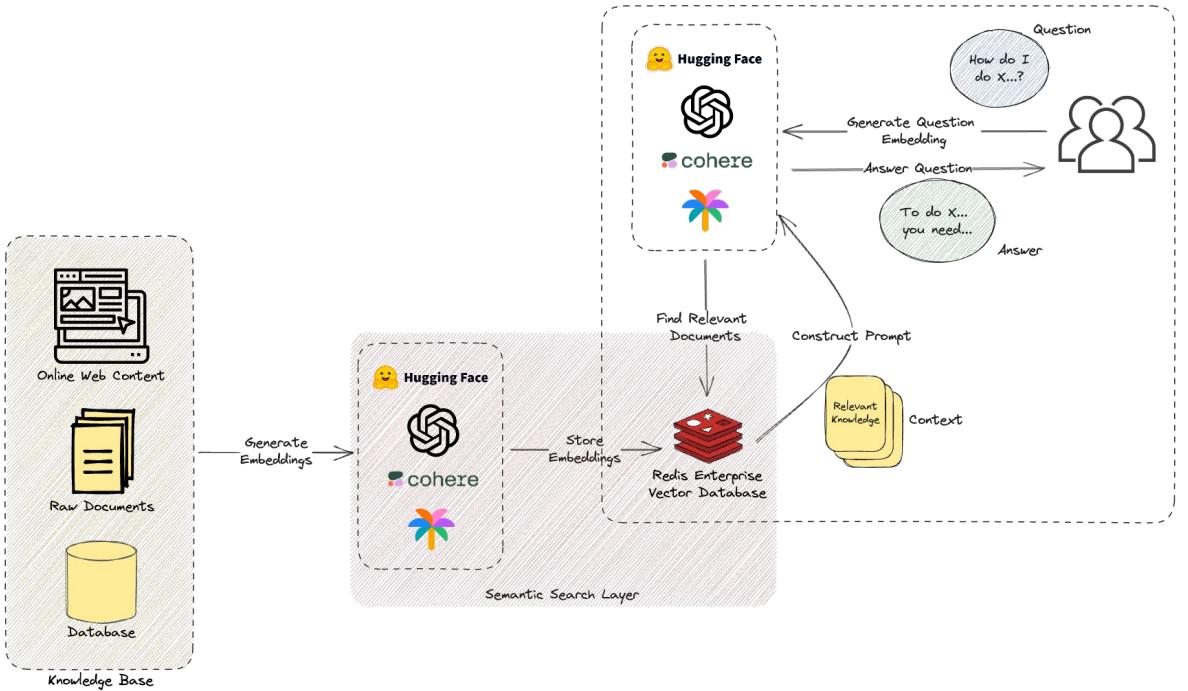
Semantic Search for Plagiarism Detection



Demo 4: Retrieval-Augmented Generation



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.

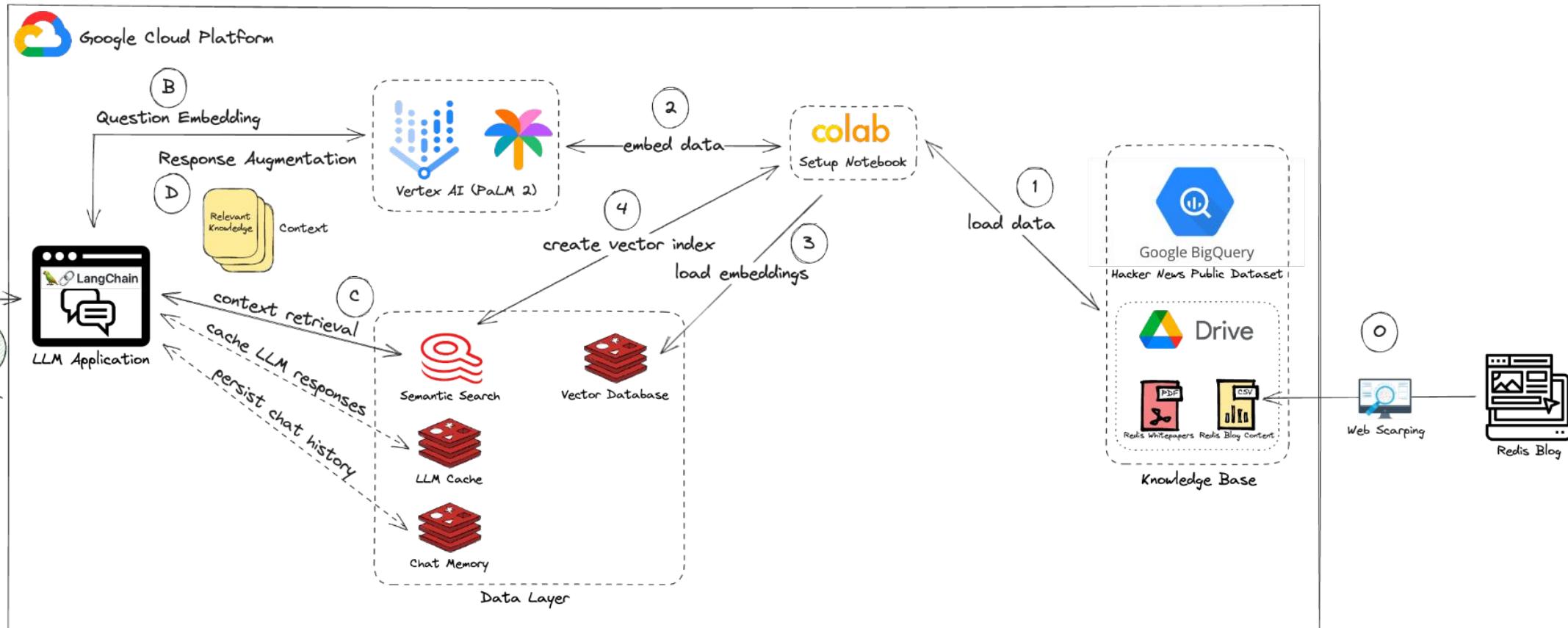
Benefits:

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

Use Cases:

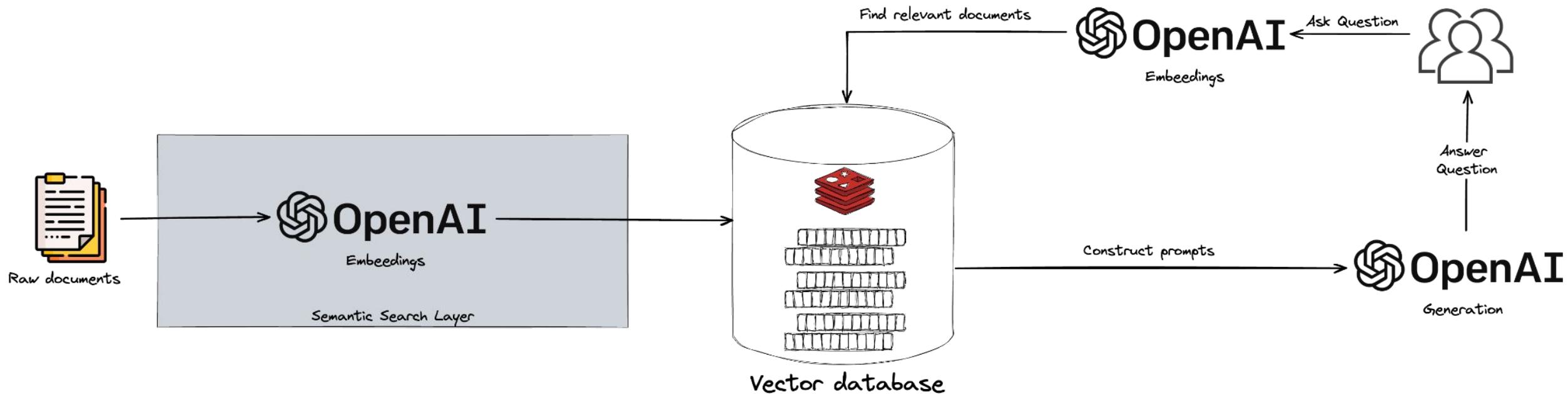
- Documents Summarization
- Shopping Assistants

Redis VSS Demo - Semantic Vector Search



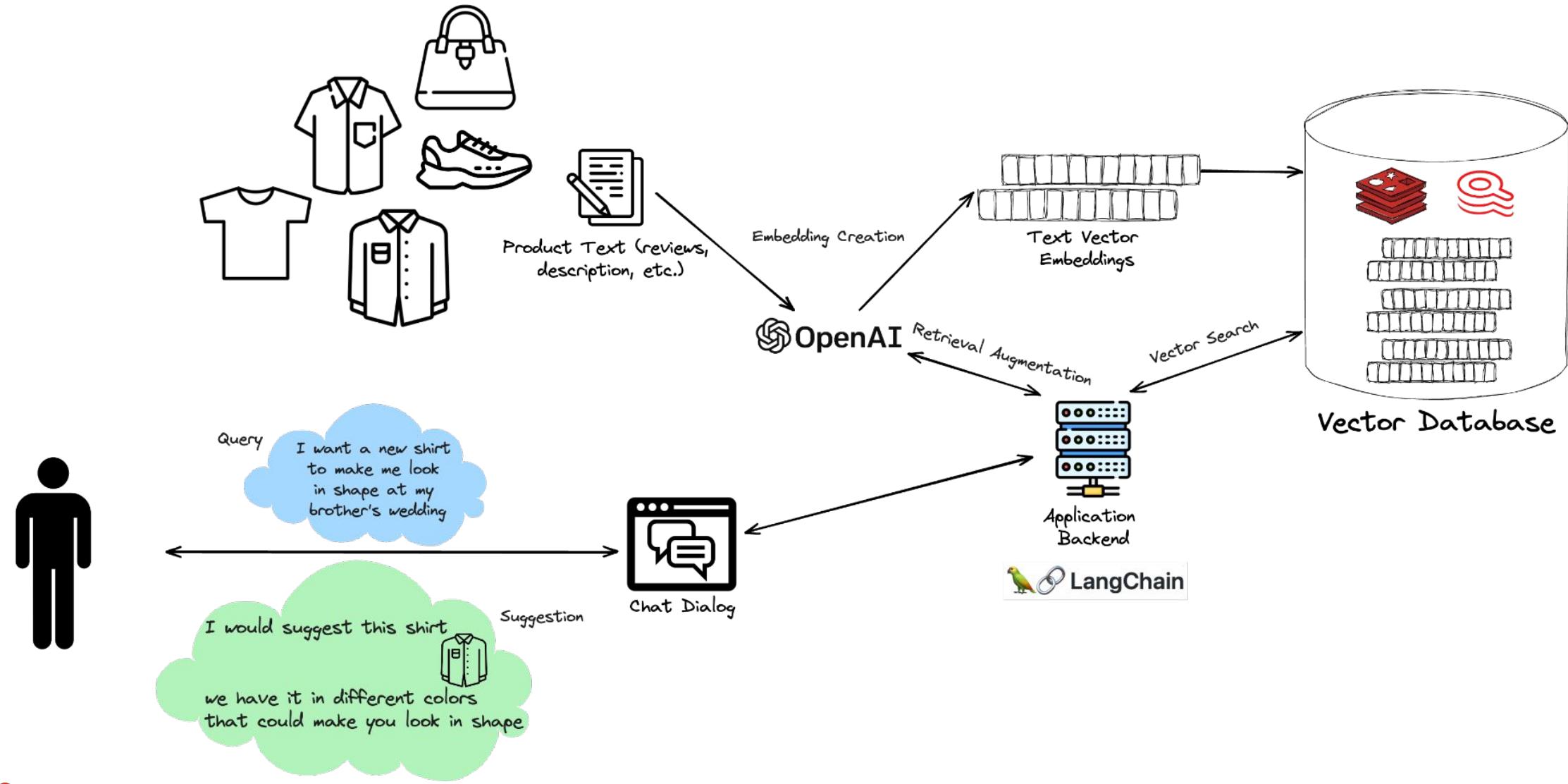
Examples of Vector Similarity Search use-cases

RAG for Domain-Specific Conversational ChatBots



Examples of Vector Similarity Search use-cases

RAG for E-Commerce ChatBots



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



RU101

Introduction to Redis Data Structures

RU101 is an introductory course, perfect for developers new to Redis. In this course, you'll learn about the data structures in Redis, and you'll see how...

[Learn More →](#)

RU102J

Redis for Java Developers

Redis for Java Developers teaches you how to build robust Redis client applications in Java using the jedis client library. The course focuses on writing...

[Learn More →](#)

RU102JS

Redis for JavaScript Developers

RU102JS is a deep dive into Redis for Node.js applications. You can expect to learn how to make connections to Redis, store and retrieve data, and levera...

[Learn More →](#)

RU102PY

Redis for Python Developers

RU102PY provides a deep dive into Python application development with Redis. You can expect to learn how to make connections to Redis, store and retrieve...

[Learn More →](#)

RU201

Redisearch

This advanced course covers Redisearch, the in-memory search engine built as a Redis Module. The course begins with a deep dive into the fundamentals of ...

[Learn More →](#)

RU202

Redis Streams

Redis Streams is a new feature for Redis 5.0. In this course, we'll cover the basic concepts of streaming, and then provide a broad overview of Redis Str...

[Learn More →](#)

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

