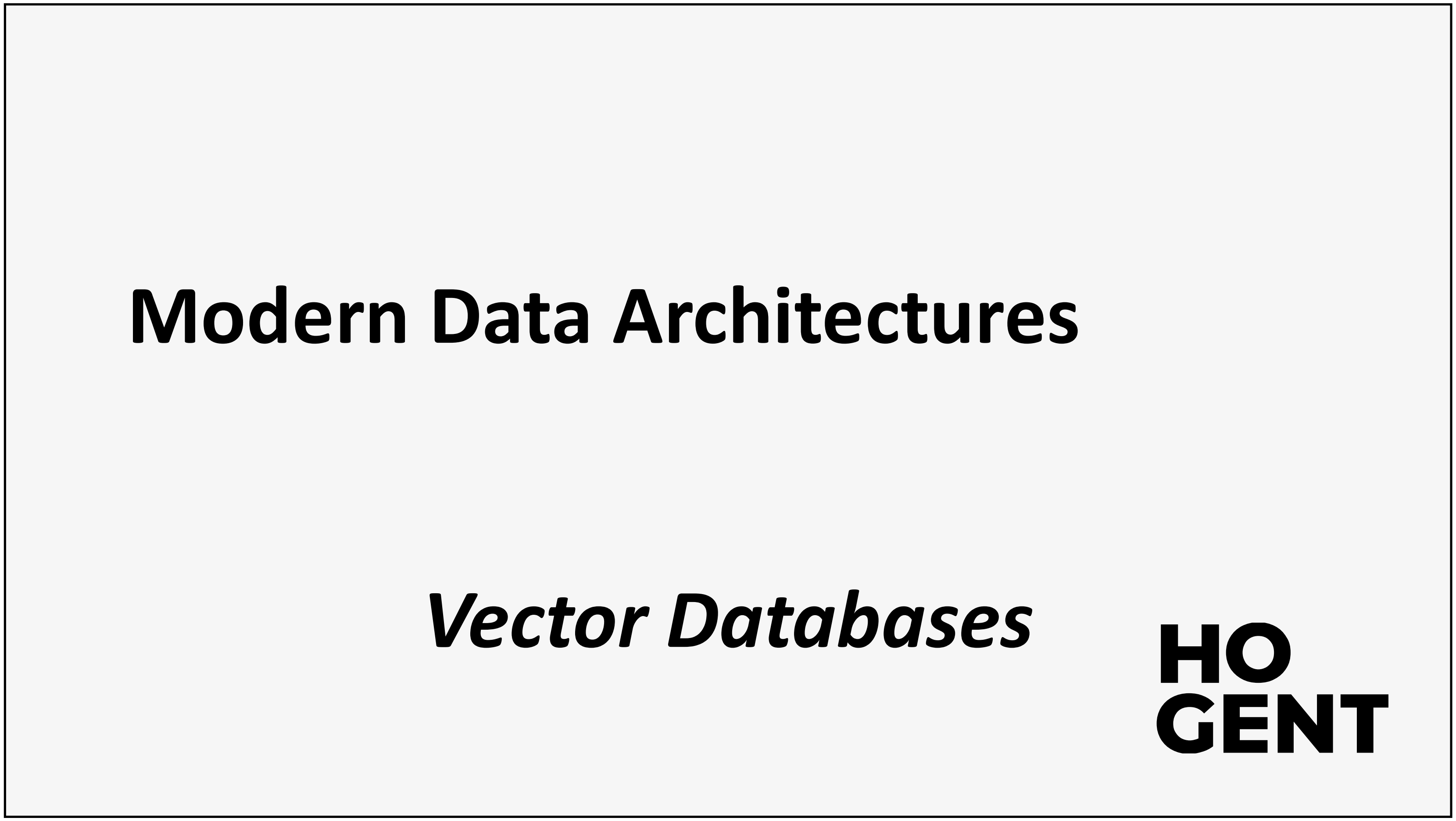
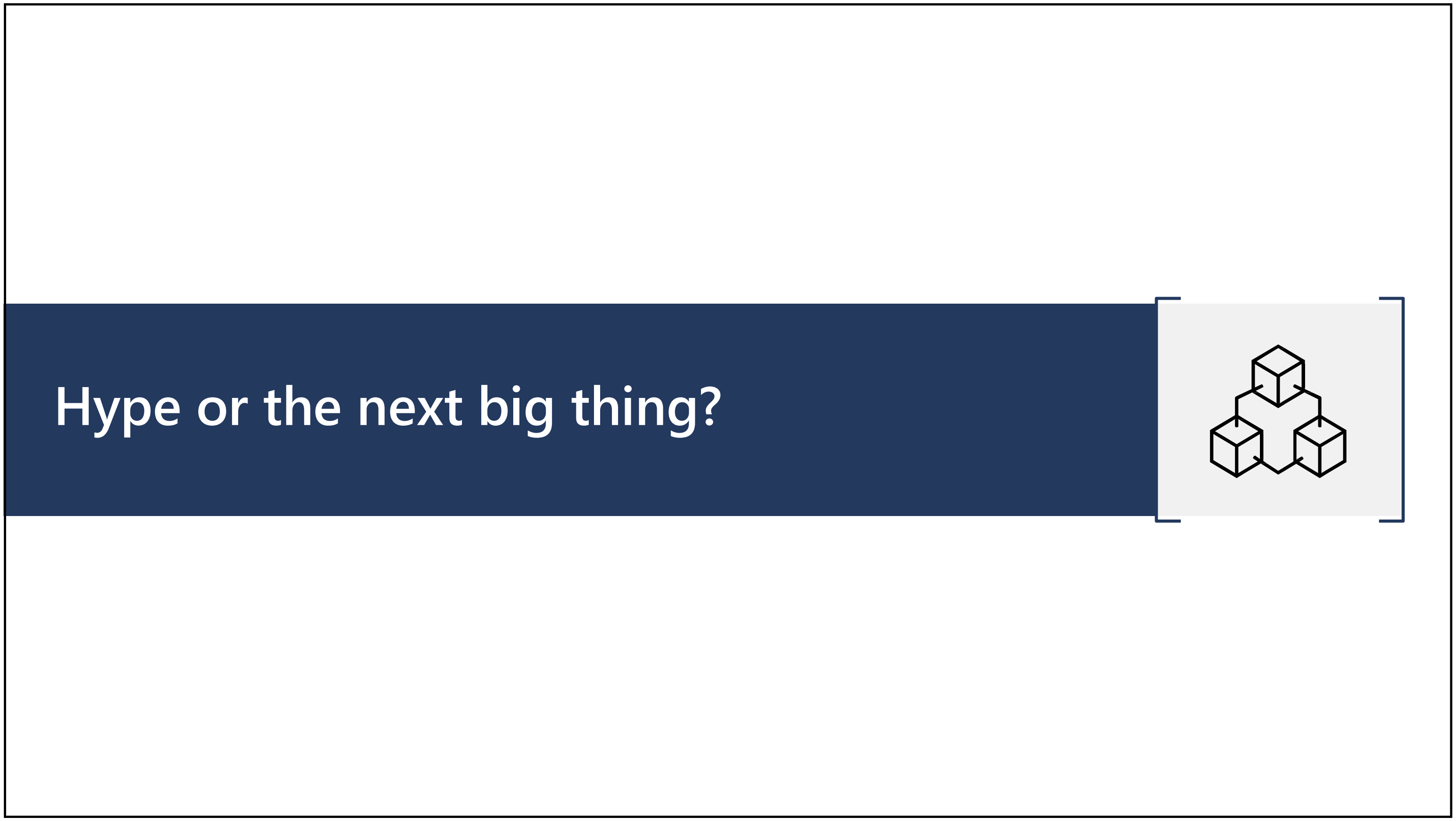


Modern Data Architectures

Vector Databases

HO
GENT

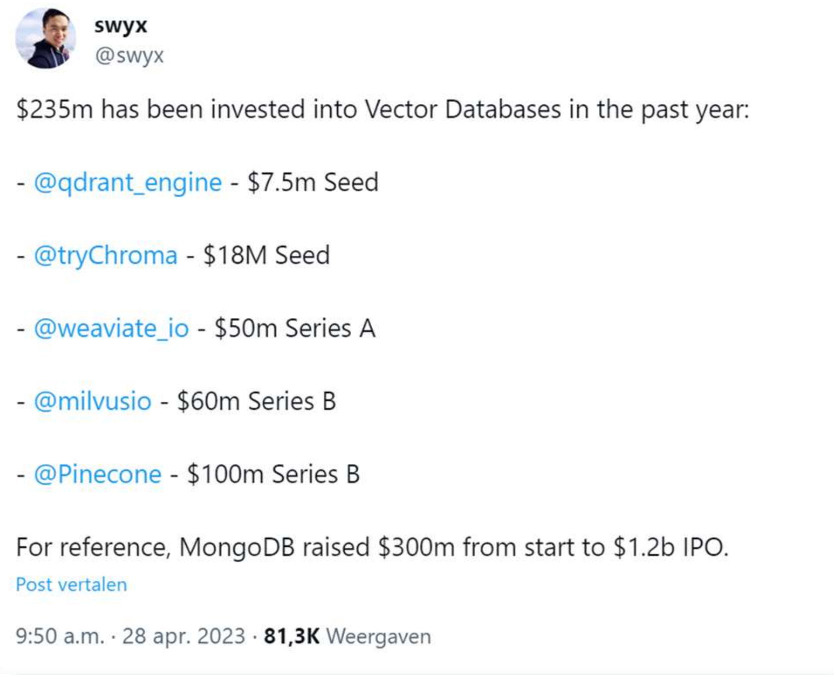


Hype or the next big thing?

Hype or the next big thing?

Hype?:

- “New kind of database for the AI era”
- For/by/focused on how AI models work
- Data → VectorDB → LLM



swyx
@swyx

$235m has been invested into Vector Databases in the past year:

- @qdrant_engine - $7.5m Seed

- @tryChroma - $18M Seed

- @weaviate_io - $50m Series A

- @milvusio - $60m Series B

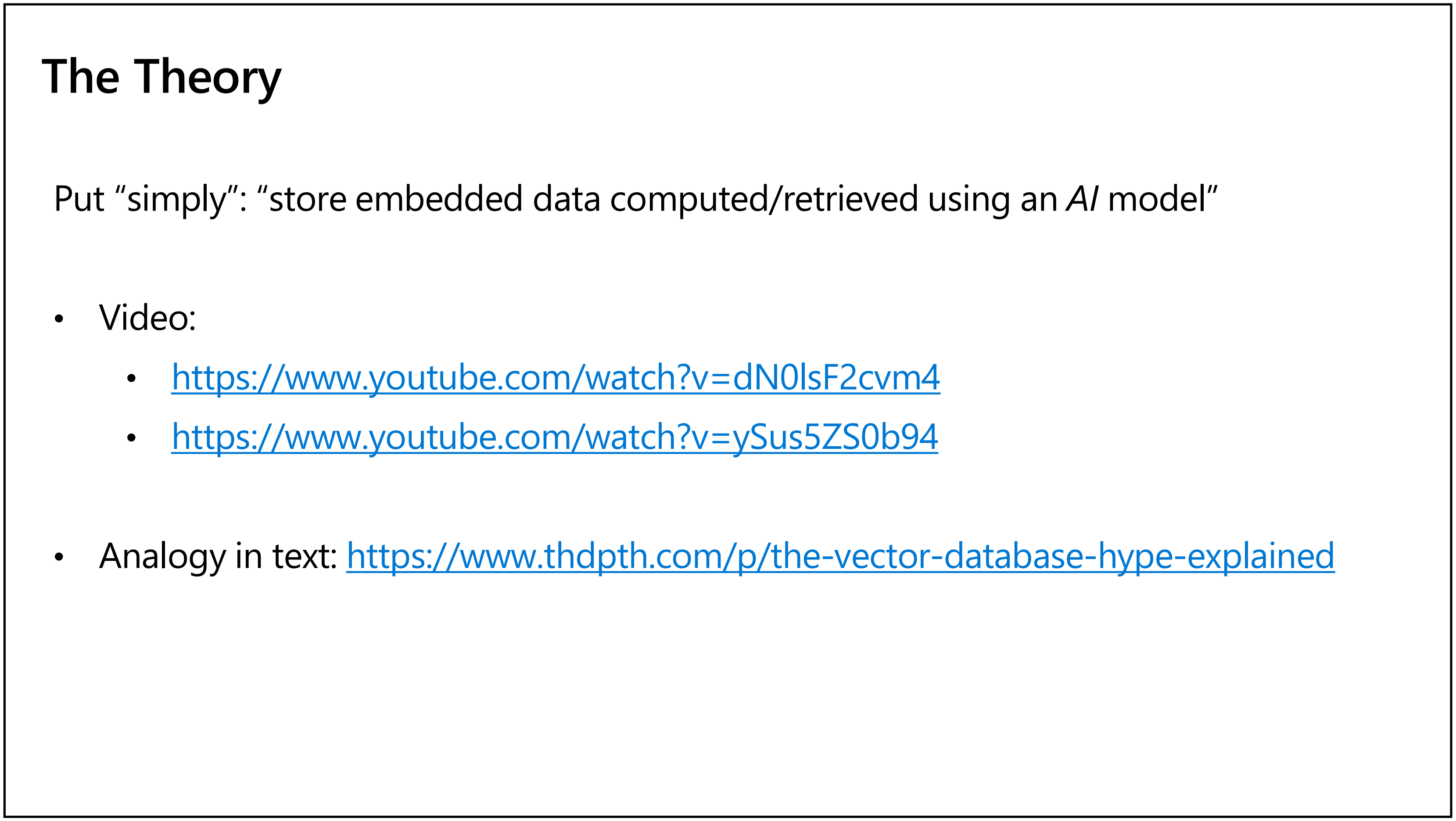
- @Pinecone - $100m Series B

For reference, MongoDB raised $300m from start to $1.2b IPO.

Post vertalen

9:50 a.m. · 28 apr. 2023 · **81,3K** Weergaven

The Theory

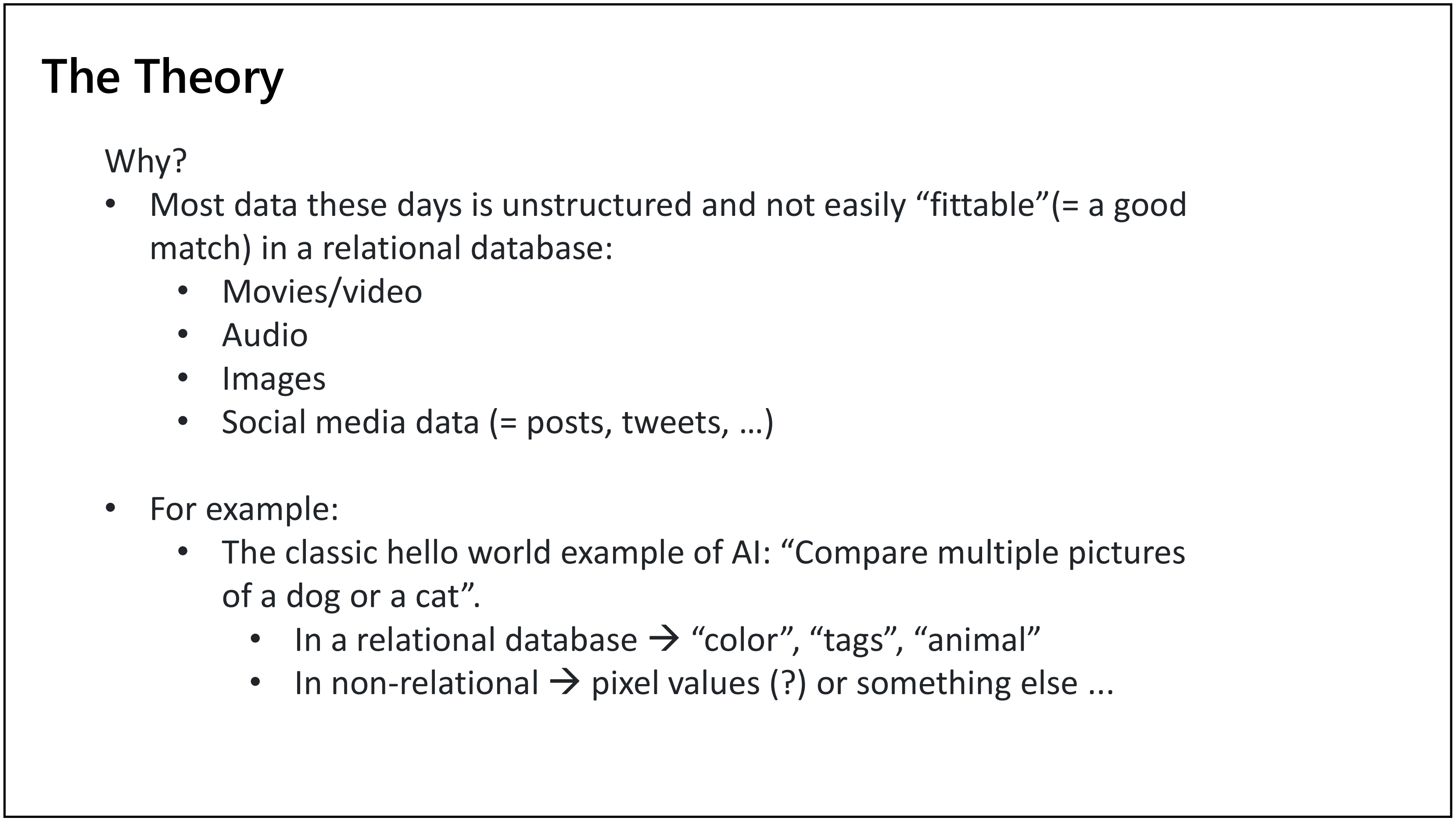


The Theory

Put “simply”: “store embedded data computed/retrieved using an *AI* model”

- Video:
  - https://www.youtube.com/watch?v=dN0lsF2cvm4
  - https://www.youtube.com/watch?v=ySus5ZS0b94

- Analogy in text: https://www.thdpth.com/p/the-vector-database-hype-explained



The Theory

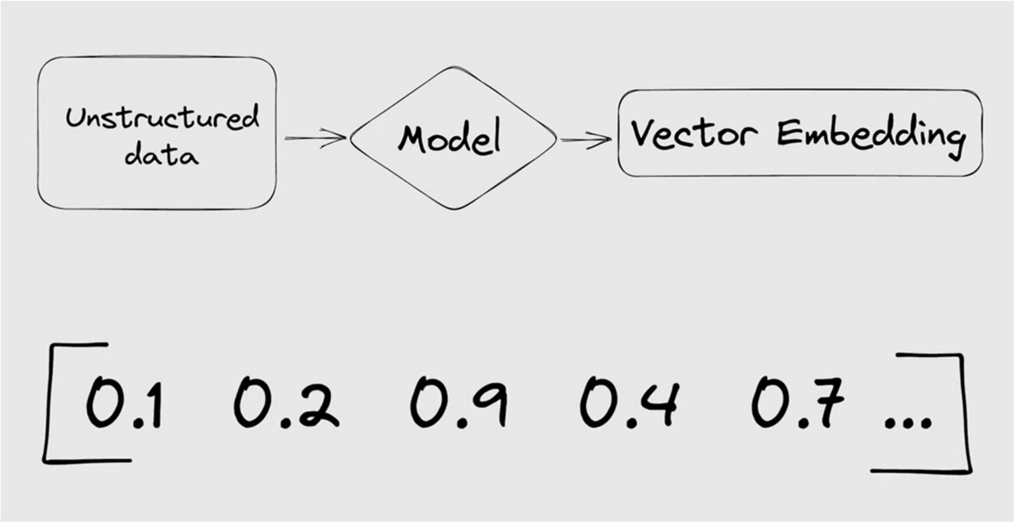
Why?

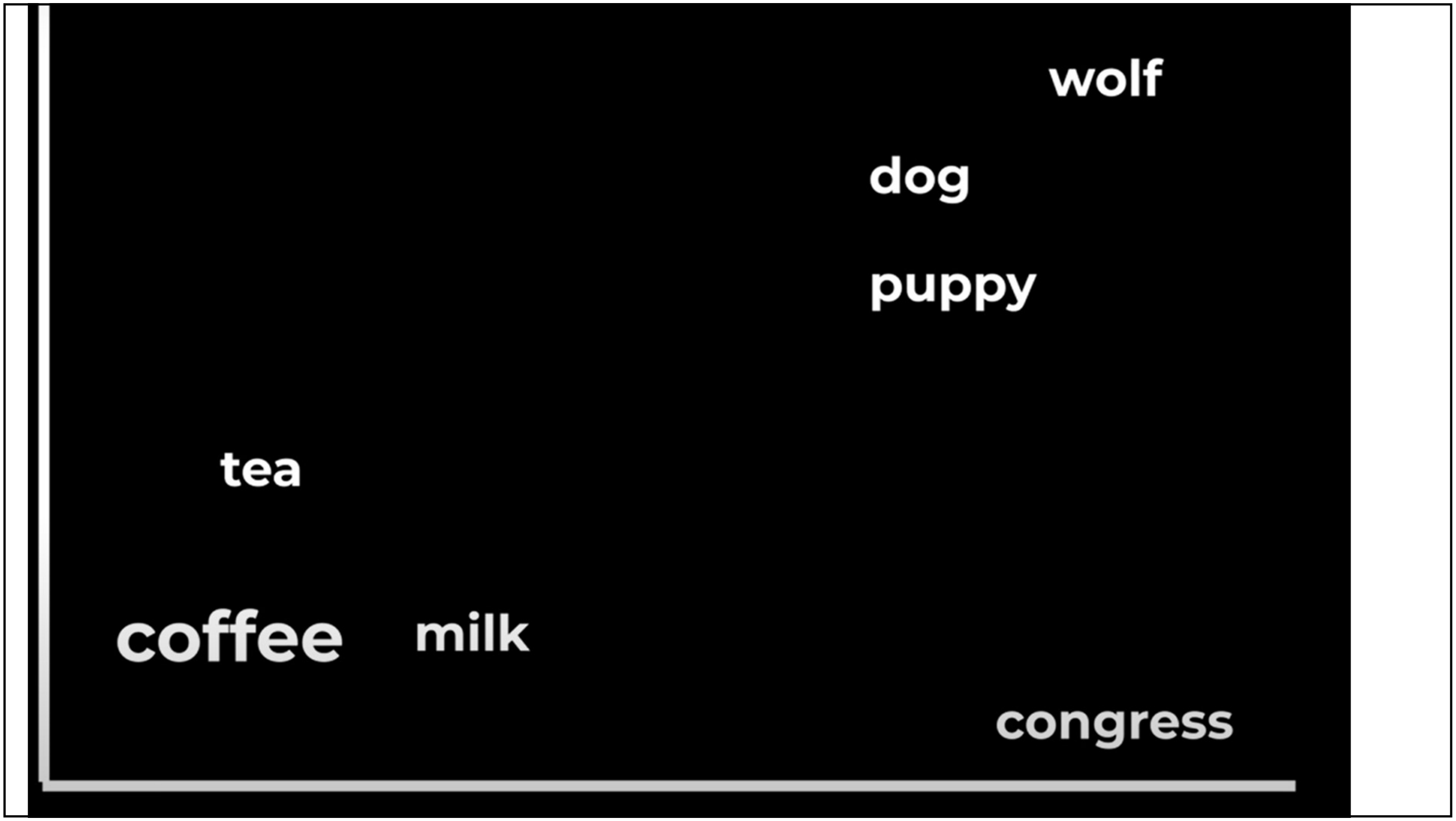
- Most data these days is unstructured and not easily “fittable”(= a good match) in a relational database:
  - Movies/video
  - Audio
  - Images
  - Social media data (= posts, tweets, …)

- For example:
  - The classic hello world example of AI: “Compare multiple pictures of a dog or a cat”.
    - In a relational database → “color”, “tags”, “animal”
    - In non-relational → pixel values (?) or something else ...

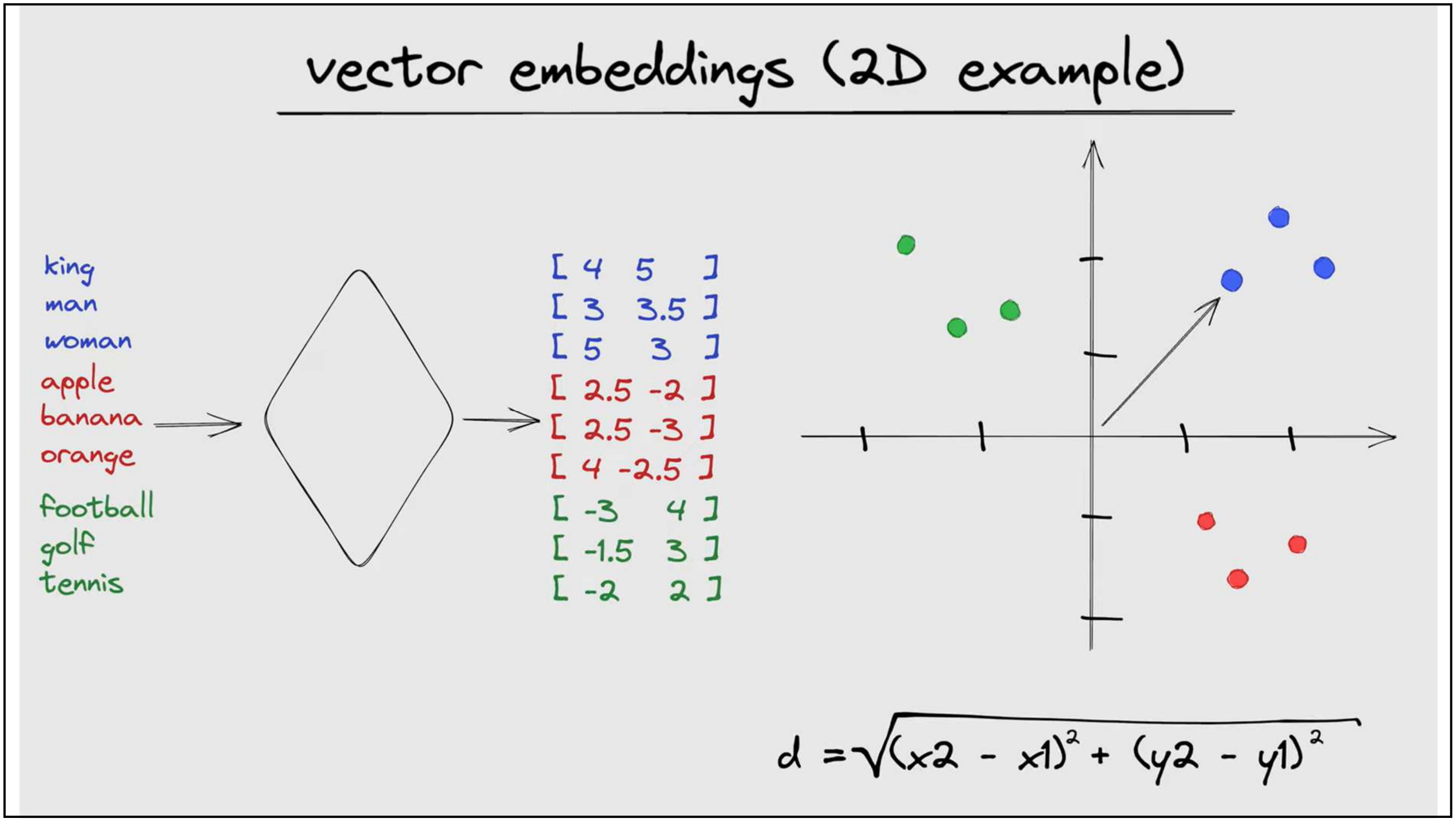
Vector embeddings ~ Vector databases

Definition:
“A vector database **indexes** and stores **vector embeddings** for **fast retrieval** and **similarity search**”





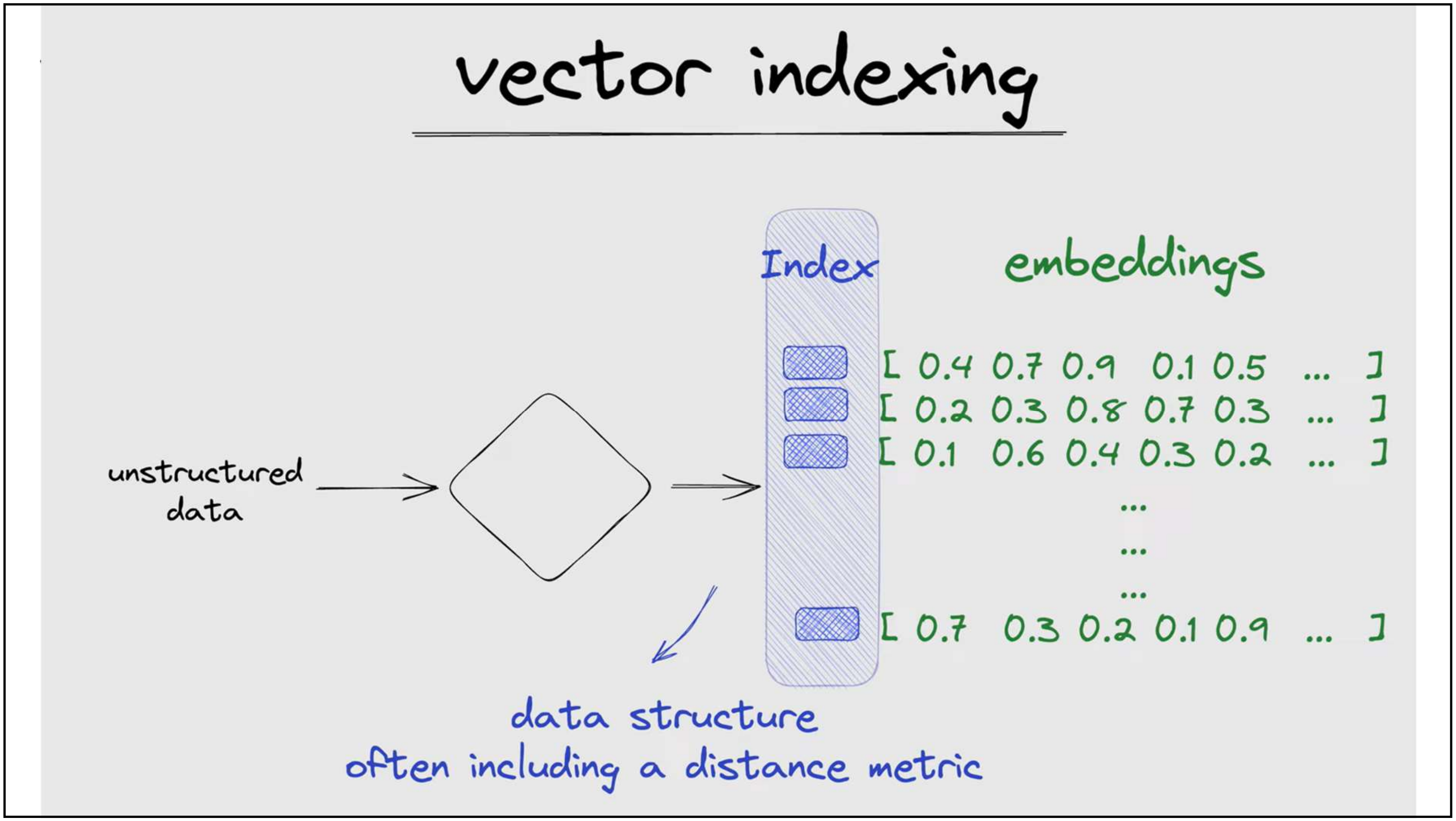
8



vector embeddings (2D example)

king	[4 5]
man	[3 3.5]
woman	[5 3]
apple	[2.5 -2]
banana	[2.5 -3]
orange	[4 -2.5]
football	[-3 4]
golf	[-1.5 3]
tennis	[-2 2]

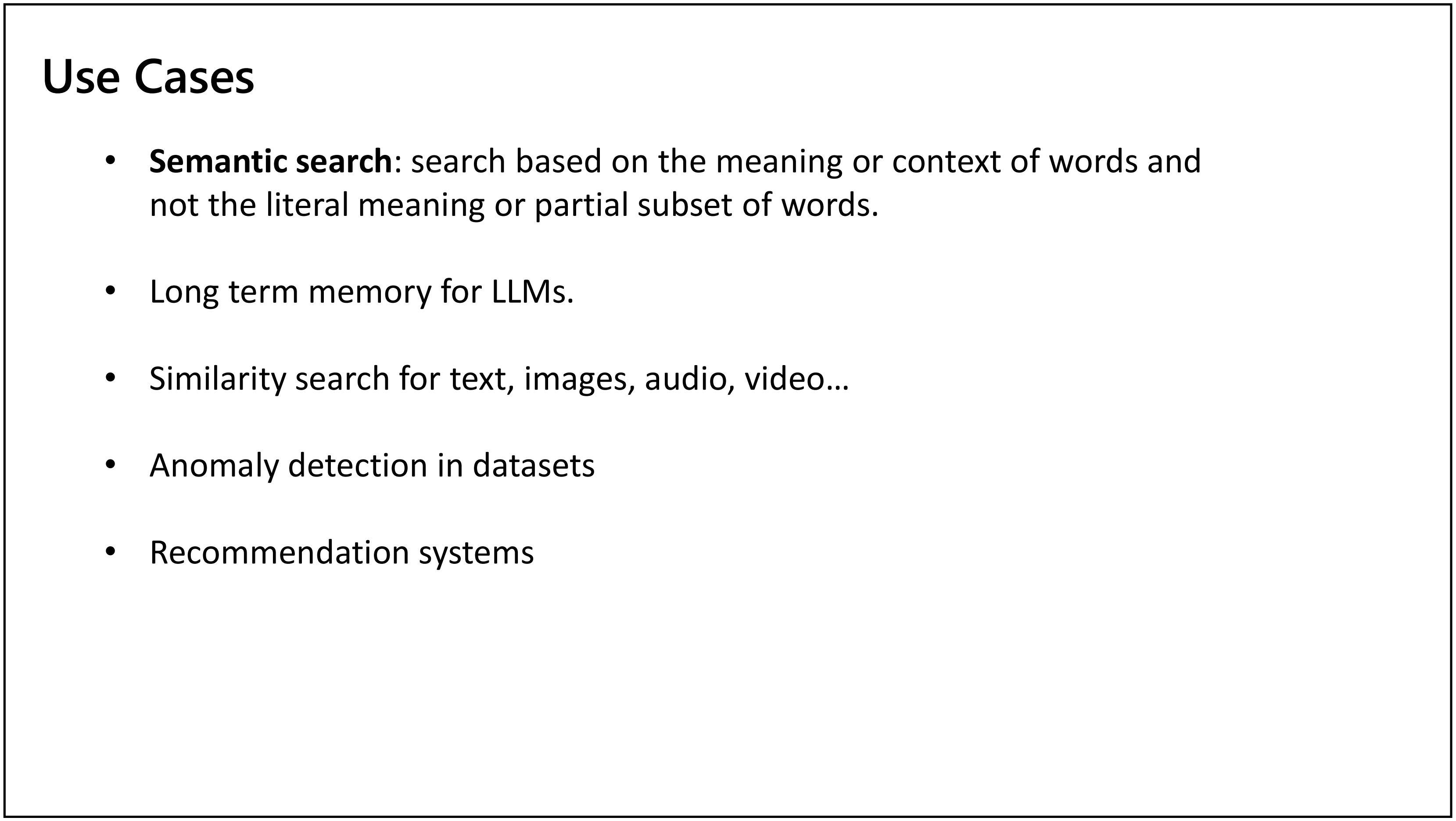
$$d = \sqrt{(x2 - x1)^2 + (y2 - y1)^2}$$



vector indexing

[0.4 0.7 0.9 0.1 0.5 ...]
[0.2 0.3 0.8 0.7 0.3 ...]
[0.1 0.6 0.4 0.3 0.2 ...]
...
...
...
[0.7 0.3 0.2 0.1 0.9 ...]

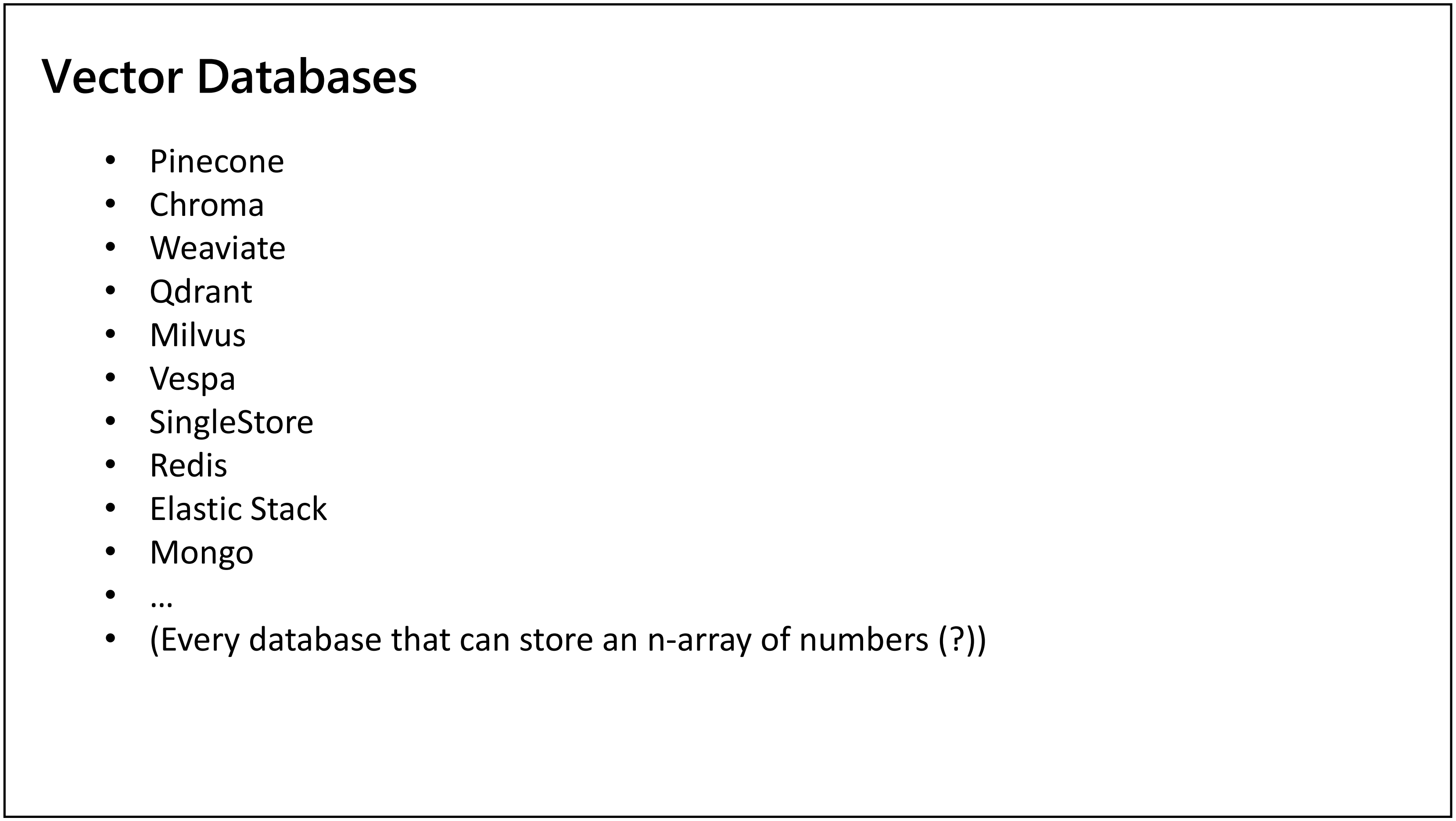
Use Cases



Use Cases

- **Semantic search**: search based on the meaning or context of words and not the literal meaning or partial subset of words.
- Long term memory for LLMs.
- Similarity search for text, images, audio, video…
- Anomaly detection in datasets
- Recommendation systems

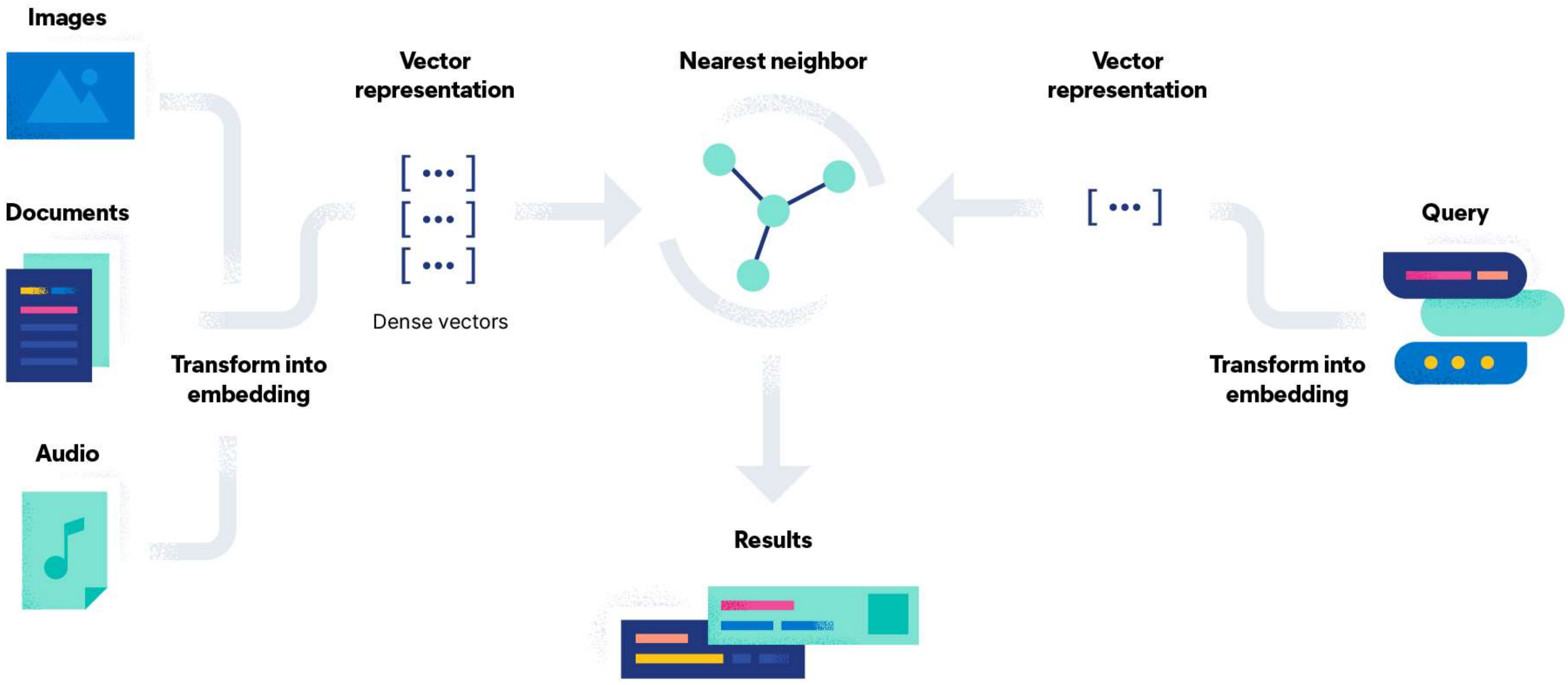
Different vector databases



Vector Databases

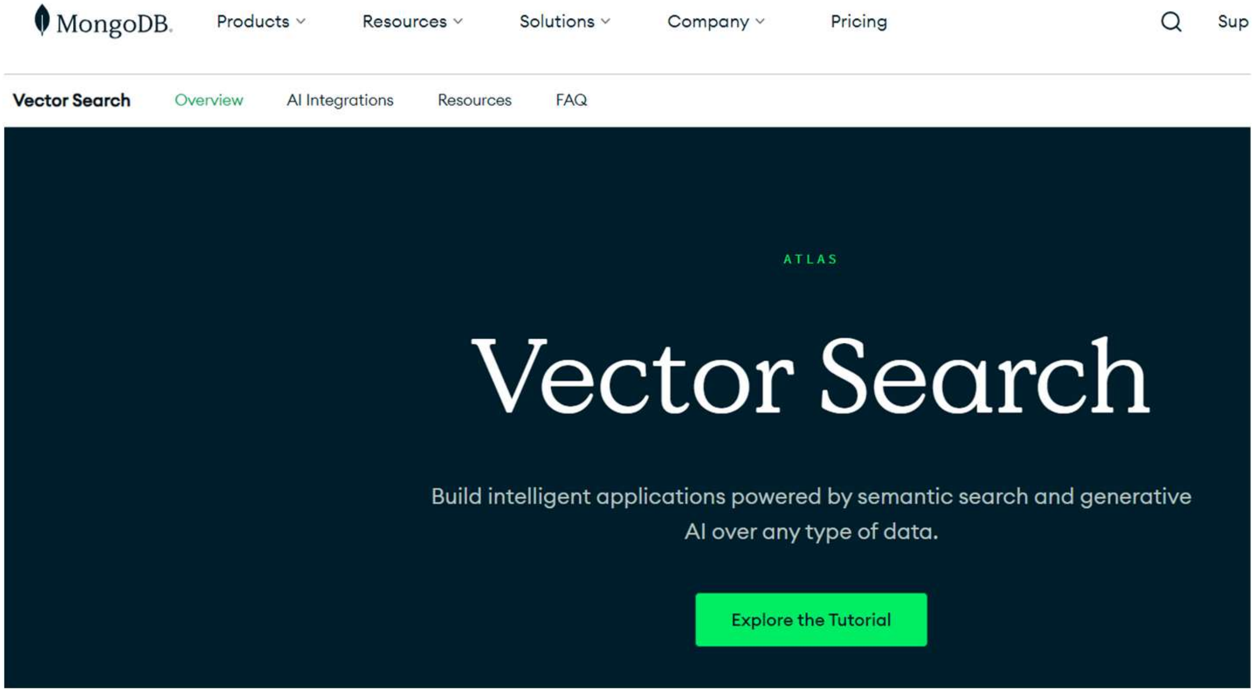
- Pinecone
- Chroma
- Weaviate
- Qdrant
- Milvus
- Vespa
- SingleStore
- Redis
- Elastic Stack
- Mongo
- …
- (Every database that can store an n-array of numbers (?))

14



https://www.elastic.co/what-is/vector-search

Vector Databases



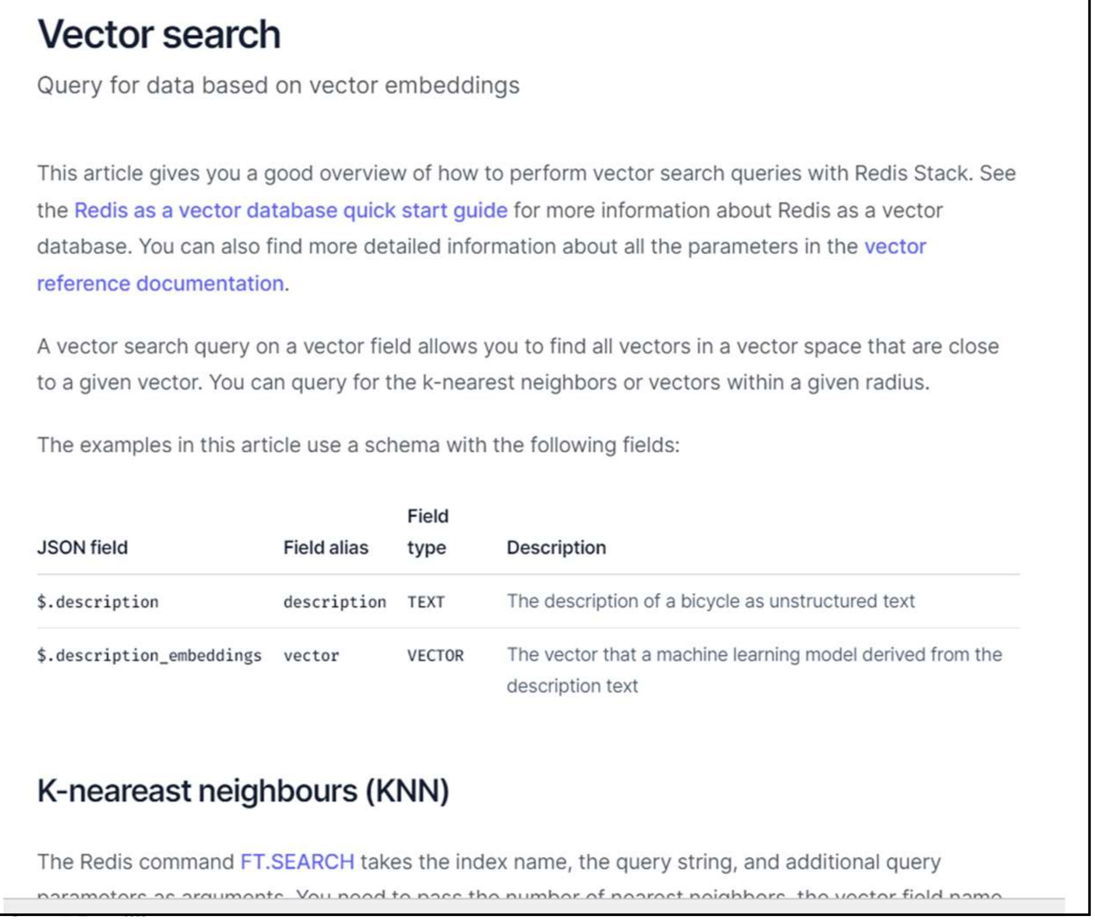
Vector Search

Build intelligent applications powered by semantic search and generative AI over any type of data.

https://www.mongodb.com/products/platform/atlas-vector-search

Vector Databases

https://redis.io/docs/interact/search-and-query/query/vector-search/



Vector search

Query for data based on vector embeddings

This article gives you a good overview of how to perform vector search queries with Redis Stack. See the Redis as a vector database quick start guide for more information about Redis as a vector database. You can also find more detailed information about all the parameters in the vector reference documentation.

A vector search query on a vector field allows you to find all vectors in a vector space that are close to a given vector. You can query for the k-nearest neighbors or vectors within a given radius.

The examples in this article use a schema with the following fields:

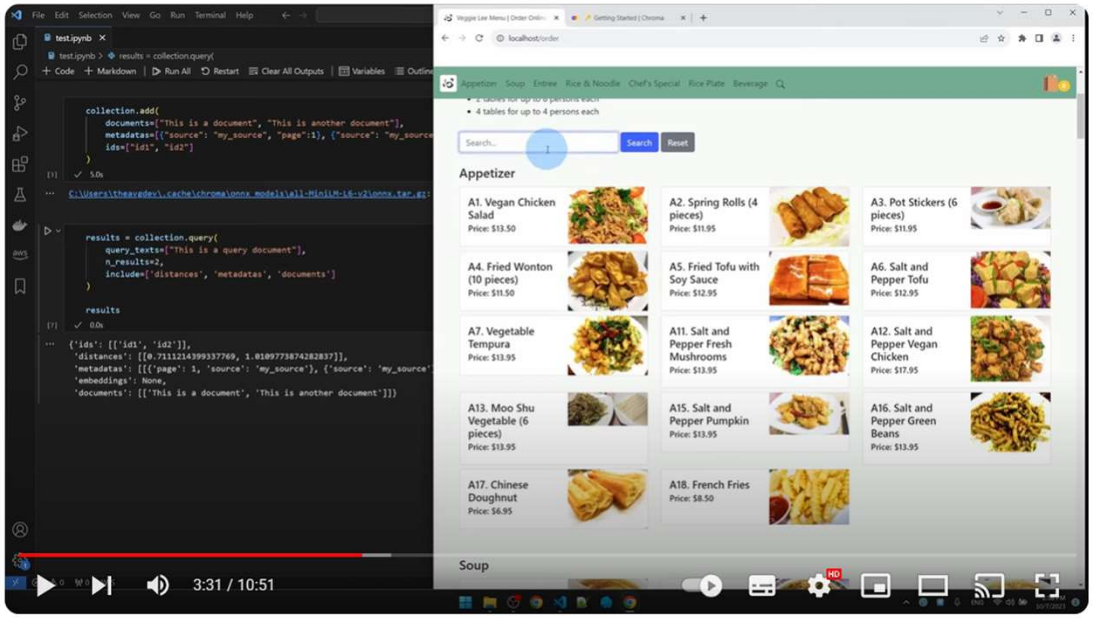
JSON field	Field alias	Field type	Description
`$.description`	`description`	`TEXT`	The description of a bicycle as unstructured text
`$.description_embeddings`	`vector`	`VECTOR`	The vector that a machine learning model derived from the description text

K-neareast neighbours (KNN)

The Redis command FT.SEARCH takes the index name, the query string, and additional query parameters as arguments. You need to pass the number of nearest neighbors, the vector field name

Demo use case

Demo: Use case


File Edit Selection View Go Run Terminal Help
test.ipynb
test.ipynb > results = collection.query(
+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline
localhost/order
Appetizer Soup Entree Rice & Noodle Chef's Special Rice Plate Beverage
Search...
Search
Reset
3:31 / 10:51
HD

```
collection.add(
    documents=["This is a document", "This is another document"],
    metadatas=[{"source": "my_source", "page":1}, {"source": "my_source
    ids=["id1", "id2"]
)
```

[3] ✓ 5.0s

C:\Users\theavgdev\.cache\chroma\onnx_models\all-MiniLM-L6-v2\onnx.tar.gz:

```
results = collection.query(
    query_texts=["This is a query document"],
    n_results=2,
    include=['distances', 'metadatas', 'documents']
)

results
```

[7] ✓ 0.0s

```
{'ids': [['id1', 'id2']],
 'distances': [[0.7111214399337769, 1.0109773874282837]],
 'metadatas': [[{'page': 1, 'source': 'my_source'}, {'source': 'my_source'
 'embeddings': None,
 'documents': [['This is a document', 'This is another document']]}
```

- 4 tables for up to 4 persons each

Appetizer

A1. Vegan Chicken Salad Price: $13.50

A2. Spring Rolls (4 pieces) Price: $11.95

A3. Pot Stickers (6 pieces) Price: $11.95

A4. Fried Wonton (10 pieces) Price: $11.50

A5. Fried Tofu with Soy Sauce Price: $12.95

A6. Salt and Pepper Tofu Price: $12.95

A7. Vegetable Tempura Price: $13.95

A11. Salt and Pepper Fresh Mushrooms Price: $13.95

A12. Salt and Pepper Vegan Chicken Price: $17.95

A13. Moo Shu Vegetable (6 pieces) Price: $13.95

A15. Salt and Pepper Pumpkin Price: $13.95

A16. Salt and Pepper Green Beans Price: $13.95

A17. Chinese Doughnut Price: $6.95

A18. French Fries Price: $8.50

Soup

Getting Started with ChromaDB - Lowest Learning Curve Vector Database & Semantic Search

https://www.youtube.com/watch?v=QSW2L8dkaZk

19