O'REILLY

Data Querying

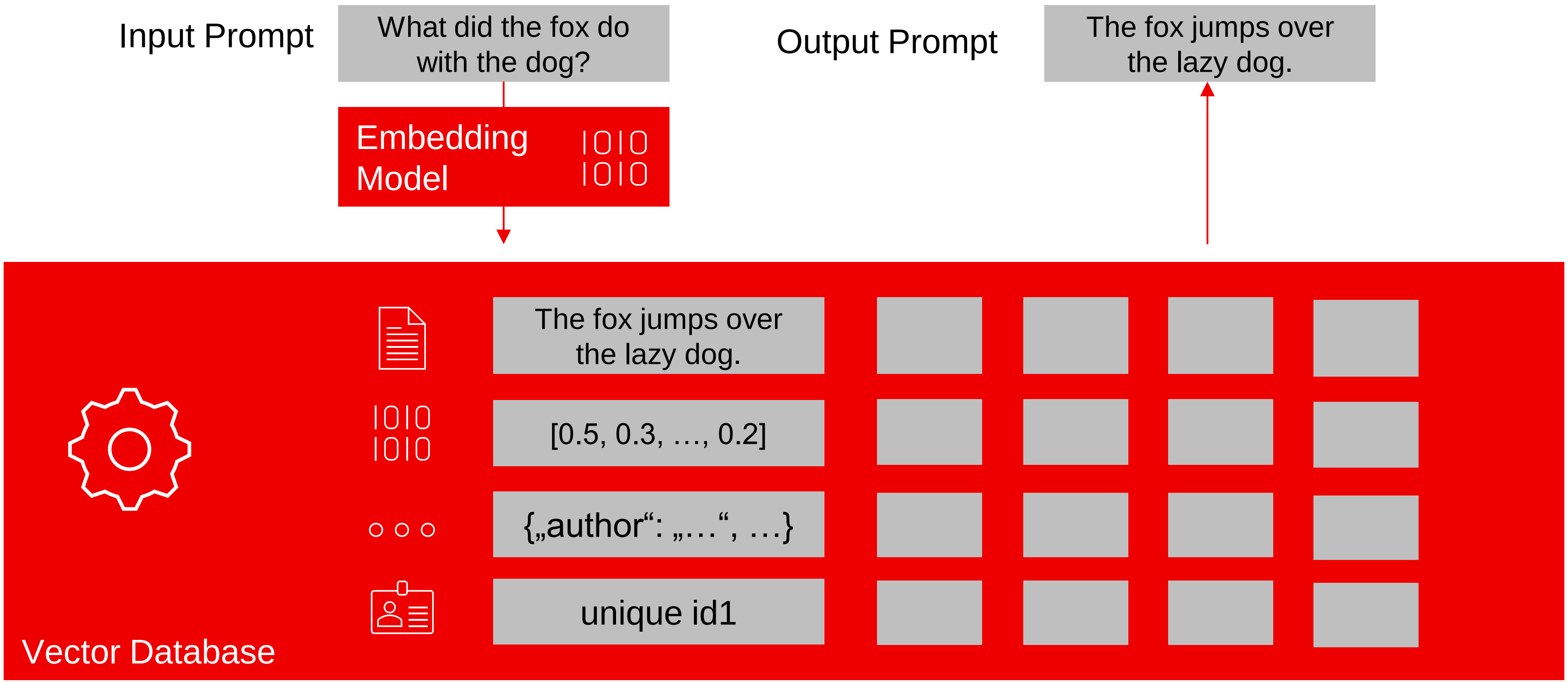
Bert Gollnick

Learning Objectives

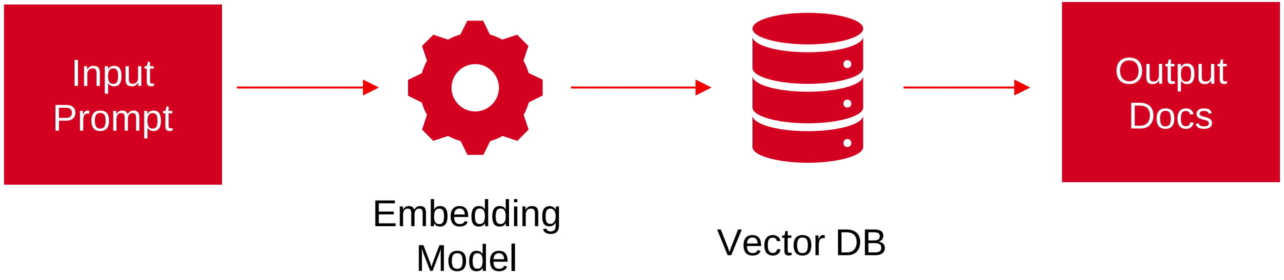
By the end of this module, you will:

- Know how to pull information from the database

Querying Introduction



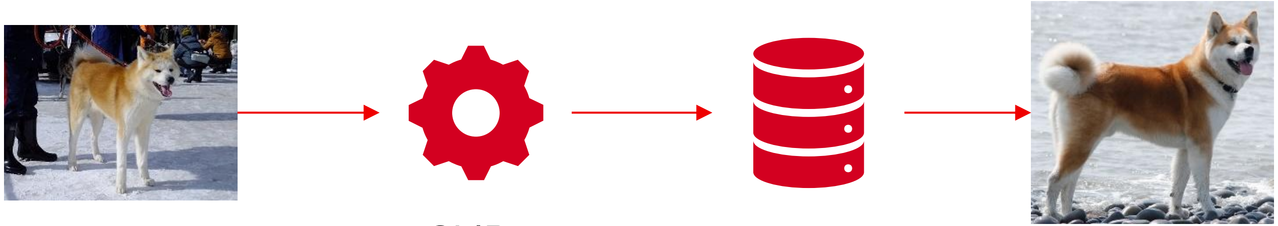
Text Querying



Practical Implementation

```
collection.query(query_texts=["This is my input text"])
```

Image Querying




```
query_list = ["../data/dogs/akita_1.jpg"]
query_result = chroma_collection.query(
        query_images = query_list,
        n_results=3,
        include=['documents’,
                 'distances’,
                 'metadatas', 'data’,
                 'uris'],)
```

```
Result 1: ../data/dogs/akita_3.jpg
          with distance: 0.17
```

Image Querying

„dog in grassland“

Query

CLIP Embeddings

Vector DB

Result

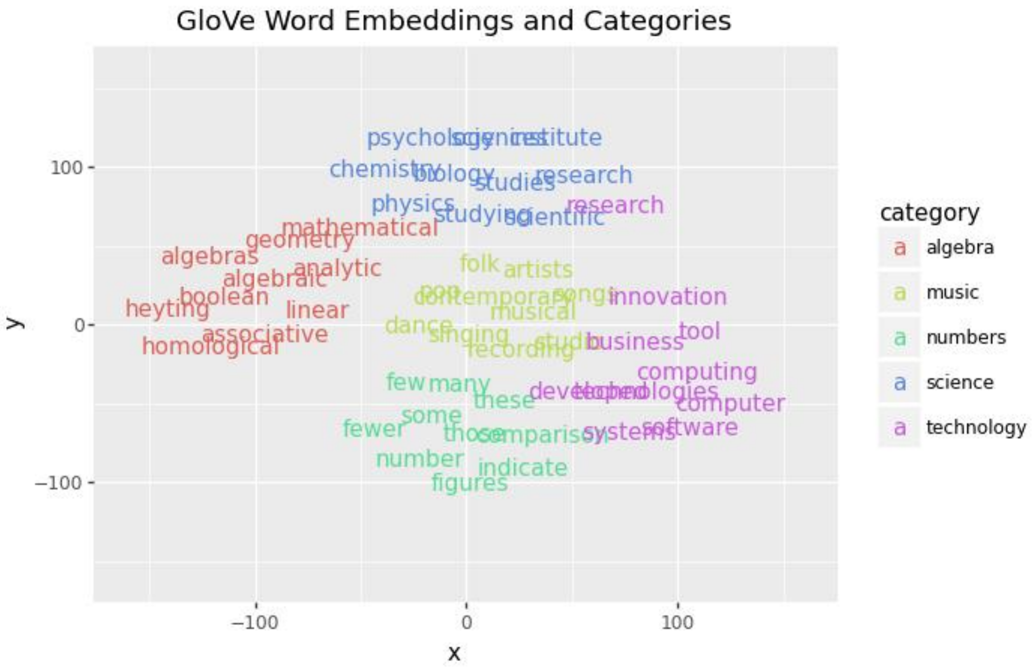
```
query_list = [“dog in grassland"]
query_result = chroma_collection.query(
        query_texts = query_list,
        n_results=3,
        include=['documents’,
                 'distances’,
                 'metadatas', 'data’,
                 'uris'],)
```

```
Query: dog in grassland Result 0:
../data/dogs/mastiff_1.jpg
with distance: 0.85
```

Similarity

- Vector DB needs to analyze similarity of query-embedding compared to document embeddings.
- Approaches:
  - Cosine Similarity
  - Maximum Margin Relevance

Similarity Search



$$dist = \sqrt{(x_1 - y_1)^2 + (x_n - y_n)^2}$$

For an embedding vector of 768 embeddings, there are 768 distance terms

Example: word embeddings reduced to 2 dimensions

Similarity Search

Imagename	Embedding					
dog1	0.3	0.02	0.8	0.6	…	0.4
dog2	0.1	0.52	0.7	0.6	…	0.4
…						
dogN	0.3	0.62	0.9	0.2	…	0.3

Vector Database

dogTest

0.3	0.02	0.8	0.6	…	0.4

$$dist = \sqrt{\sum (x_i - y_i)^2}$$

For small data, go with np.array().

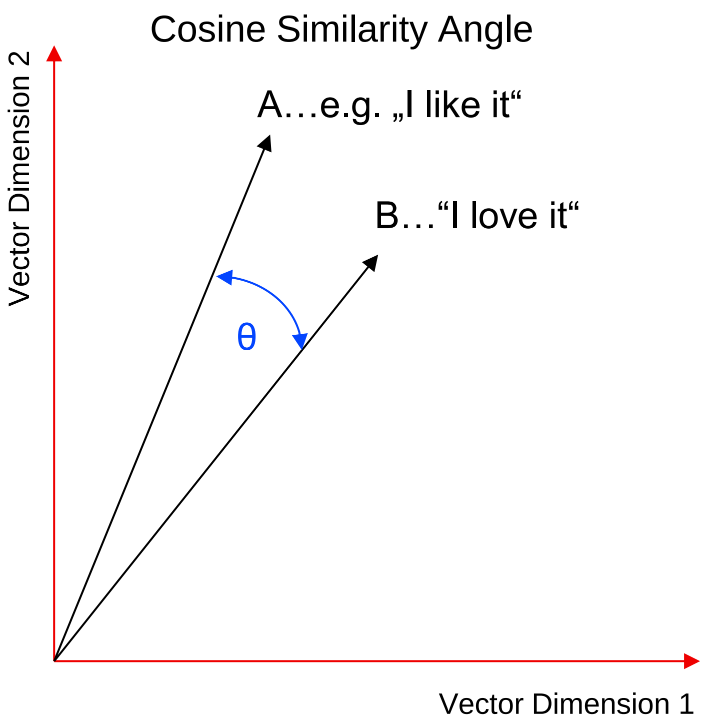
For large dataset not feasible!

Cosine Similarity

- Measures similarity between Embedding-Vectors based on angle θ.
  - Vectors maximally dissimilar

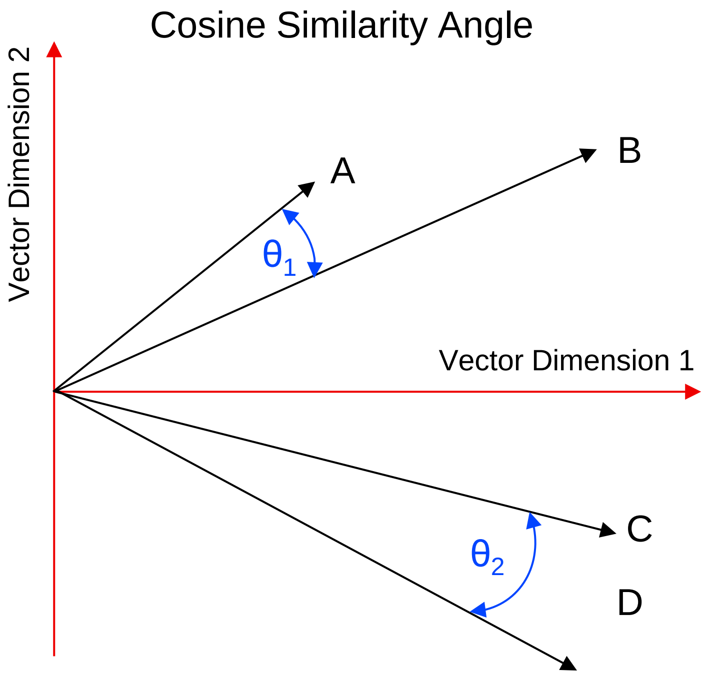
    → vectors perpendicular (θ = 90°)
  - Vectors completely similar

    → vectors parallel (θ = 0°)



Cosine Similarity

- Only the angle defines the similarity
- NOT the euclidean distance or magnitude of a vector
- Example
  - A: "The cat sleeps."
  - B: "The feline slumbers peacefully on the soft cushion."
  - C: "Trees grow leaves in spring.“
  - D: "Fish swim in the ocean."



Maximum Margin Relevance

- Aproach: reduce redundance while maintaining relevance and diversity
- Redundancy…similar vectors
- Relevance…how closely do query and documents match
- Avoid clustering effect

Topic: Renewable Energies

What are the main types of renewable energy sources and how do they work?