O'REILLY®

Embeddings

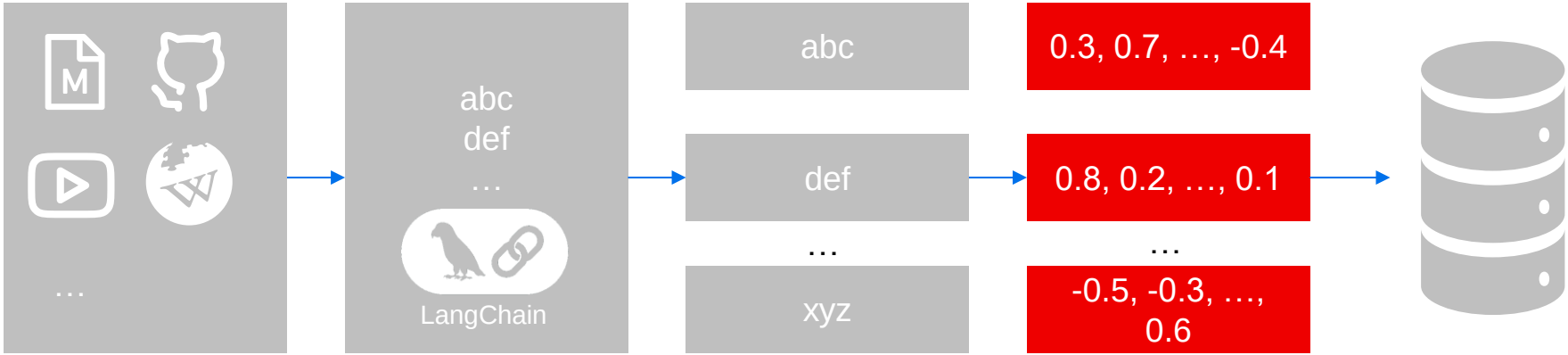
Bert Gollnick

Learning Objectives

By the end of this module, you will:

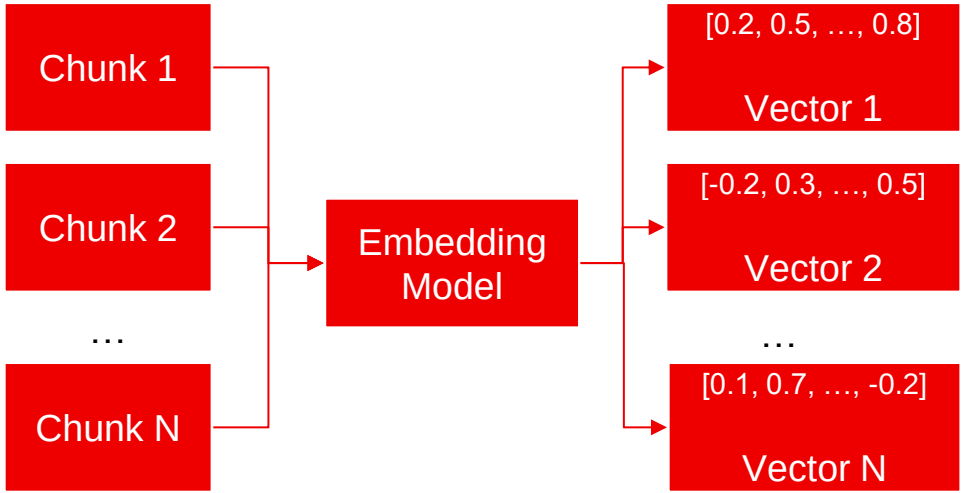
- What embeddings are
- Why they are important
- How they are created
- Which types are available

Data Ingestion Process



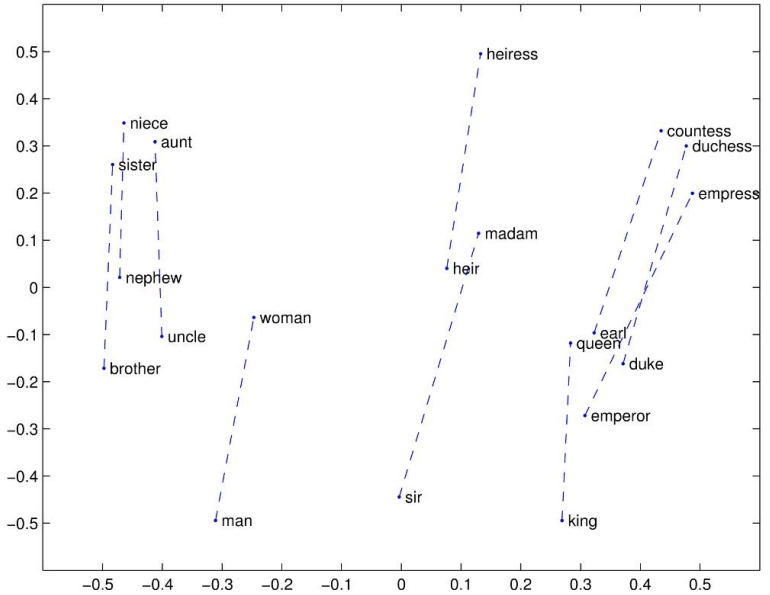
What are embeddings?

- Conversion of text data into numeric vectors
- Each word / sentence is represented as vectors
- Vector has „low“ number of dimensions



Why use embeddings?

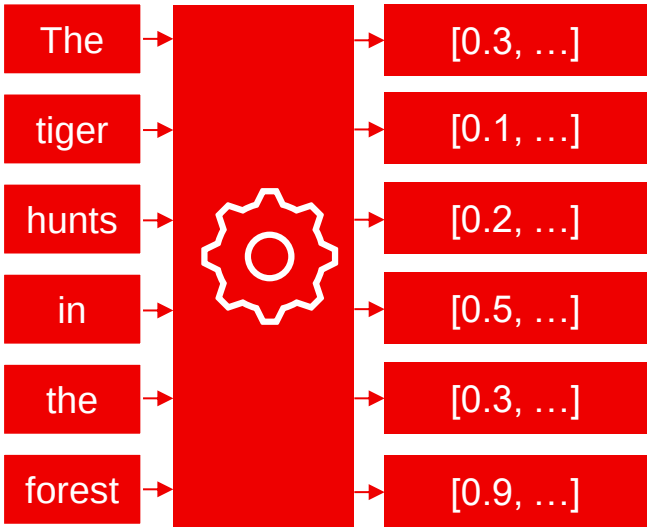
- Semantic representation
  - capture meaning of data
  - enable comparison and analysis
- Lower dimensionality
  - computational complexity is reduced
  - high-dimensional data can be represented in lower dimensions
- Reusability
  - usable across different applications



Source: https://nlp.stanford.edu/projects/glove/

How are embeddings created?

Word Embeddings



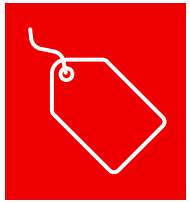
Sentence Embeddings

Which types are available?

Type	Model	Provider	Price	Vector Size
Online	text-embedding-3-small	OpenAI	0.02$ / 1M tokens	1536
Online	text-embedding-3-large	OpenAI	0.13$ / 1M tokens	3072
Online	mistral-embed	MistralAI	0.10$ /1M tokens	1024
Offline	all-MiniLM-L6-v2	Open Source	---	384
…				

Benchmark: https://huggingface.co/spaces/mteb/leaderboard

Factors to consider



Price

Speed

Off-/Online

Benchmark
Performance


O'REILLY®