Vector Space Models and Embeddings in RAGs

Fundamentals of Embedding Vectors and Their Role in RAG Systems

Cătălin Tudose

Java Champion, PhD in Computer Science, Java and Web Technologies Expert

www.catalintudose.com | www.linkedin.com/in/catalin-tudose-847667a1

Course Introduction

Embedding vectors

Word relationships and context in a vector space

Compare techniques

- TF-IDF
- Word2Vec
- Transformer-based embedding models

Connect unstructured data to RAG retrieval pipelines

Embeddings support

- Semantic search
- Similarity matching
- Relevance ranking

Optimization techniques

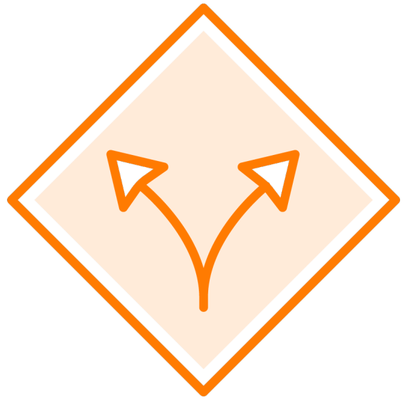
- Dimensionality reduction
- Vector compression

Bias concerns

Human Language

Rich, ambiguous,
highly contextual

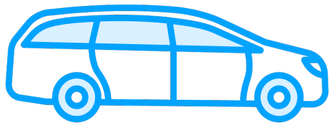
One word,
multiple meanings



Different words,
same concept

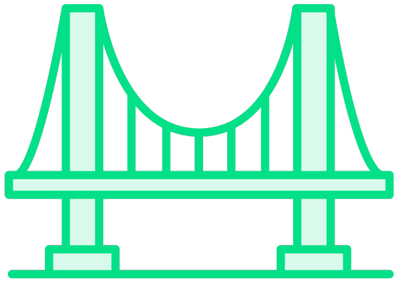
Meaning of Words

Car



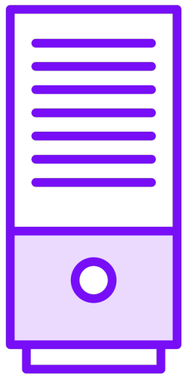
Automobile

Finance

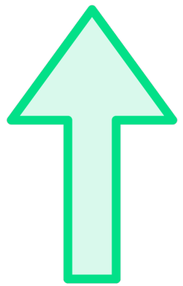


River

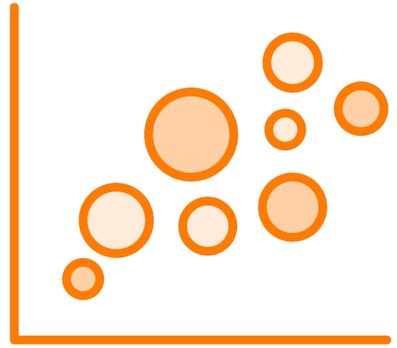
Language Representation



Machines



Embedding vectors



High-dimensional space

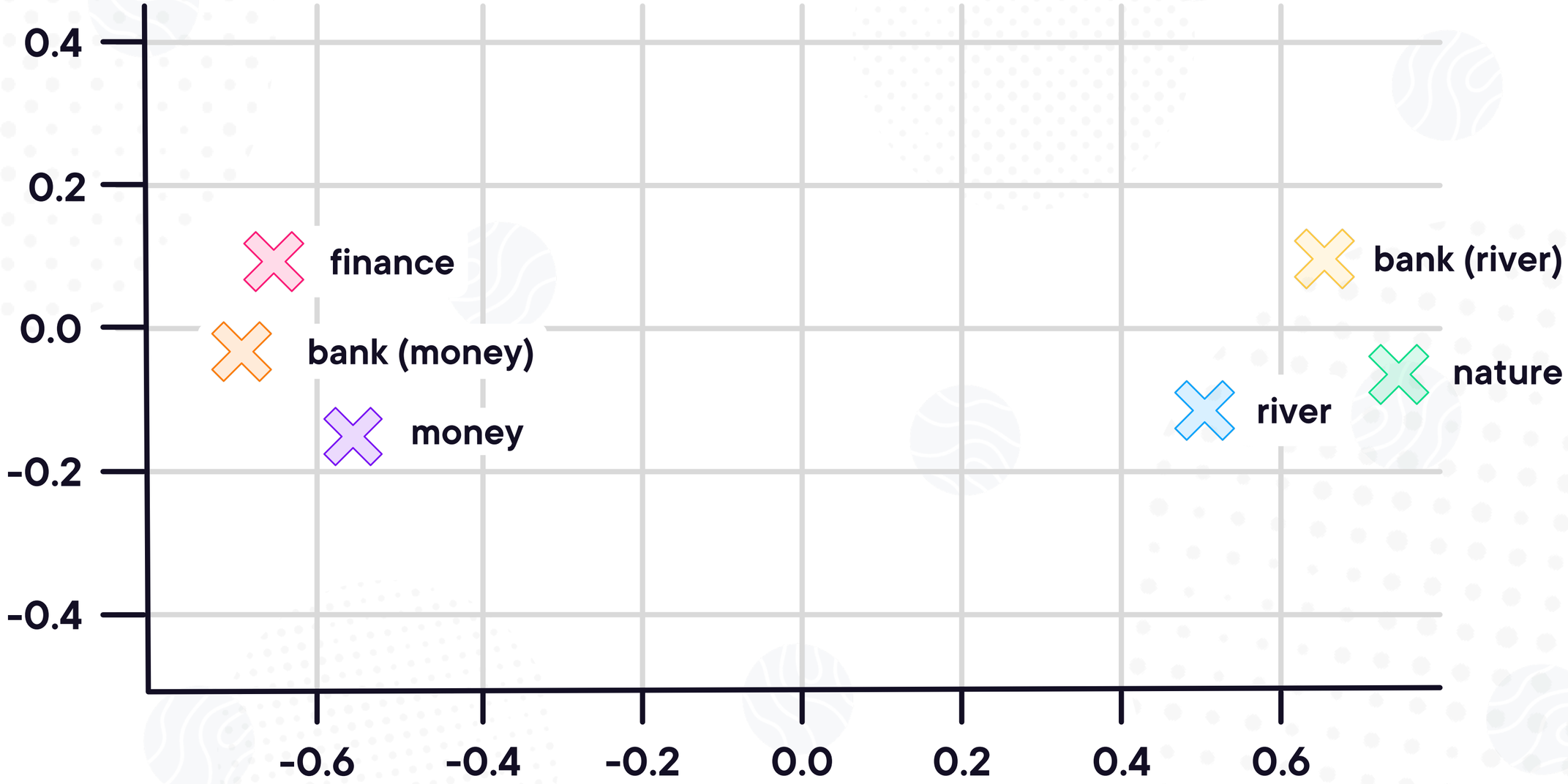
Retrieval-Augmented Generation (RAG)

Unstructured text

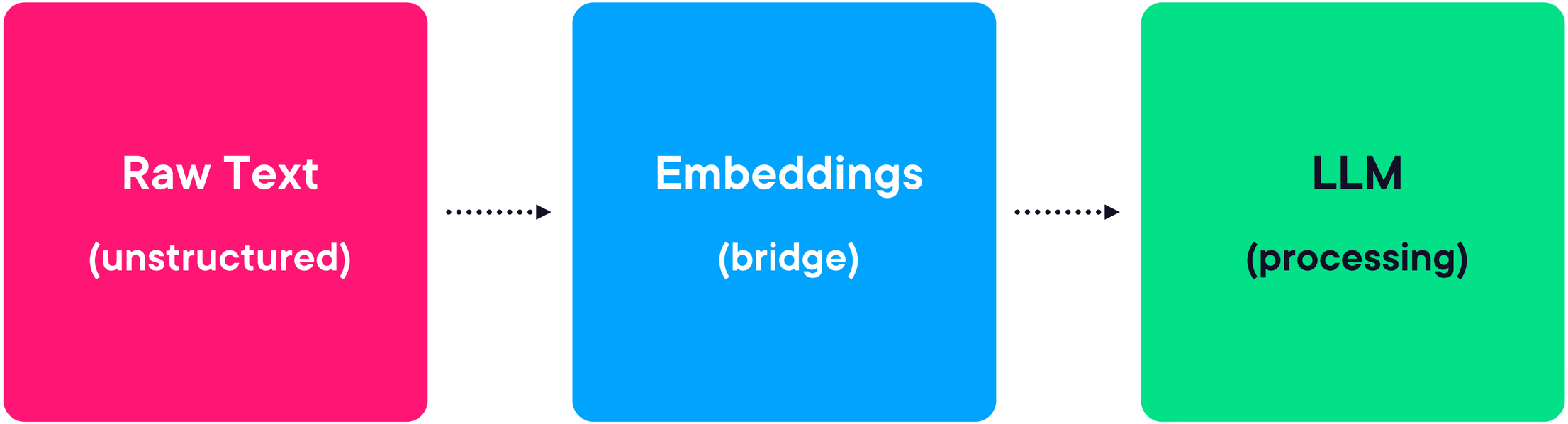
Large language model

Embeddings in Vector Space

Word Meaning as a Vector



Using Embeddings



The Translator Bridge

Raw Text

(e.g., ‘car’, ‘automobile’)

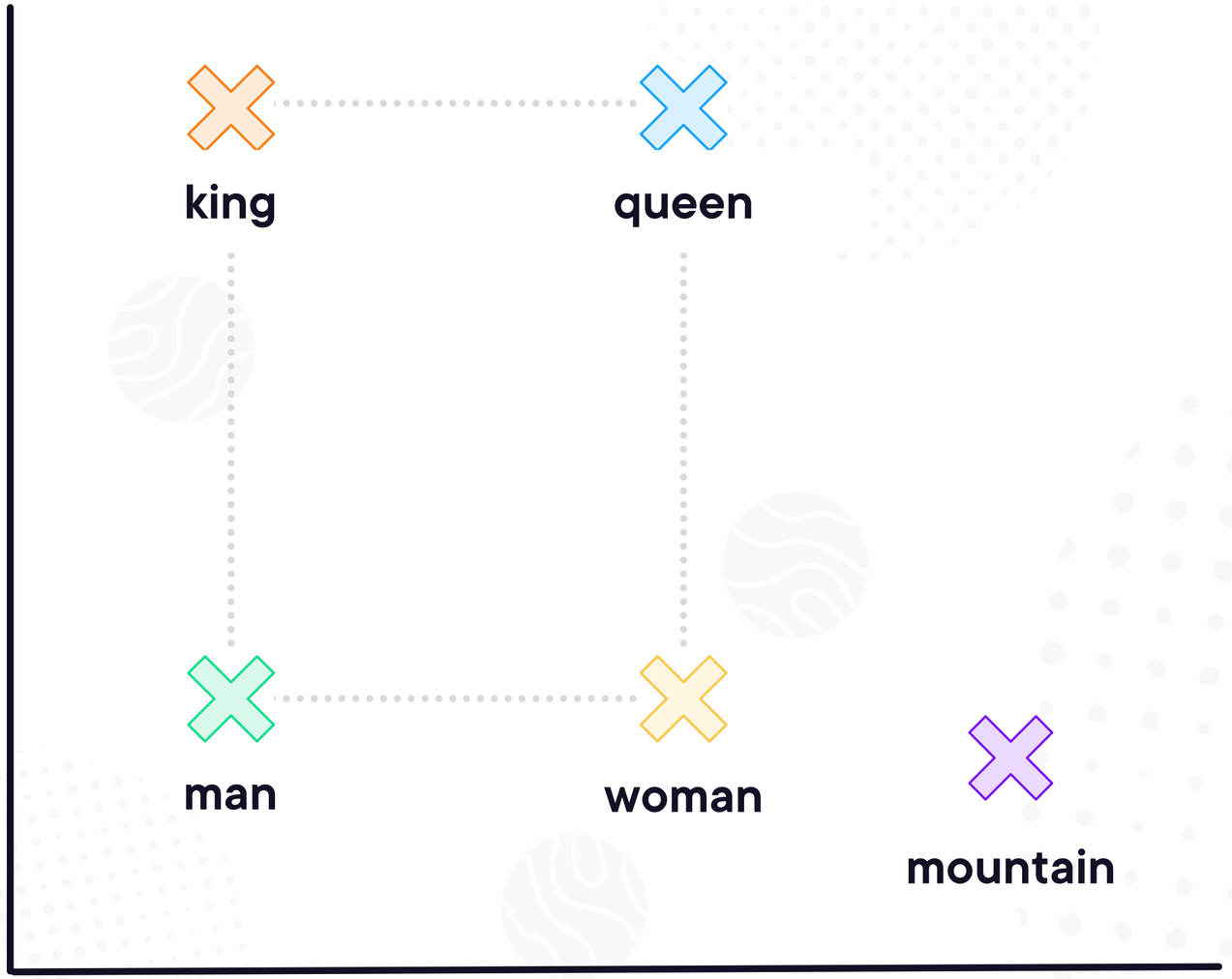
Embeddings

(Translator → vectors)

LLM

(Understands numbers)

The Map and Coordinates



Comparing Traditional and Modern Embedding Models

TF-IDF (Term Frequency–Inverse Document Frequency)

Term Frequency (TF)

- TF = 5/100 = 0.05

Inverse Document Frequency (IDF)

- IDF is high
- IDF is very low

TF-IDF score = TF × IDF

TF-IDF Score

TF-IDF high score

TF-IDF low score

Word2Vec

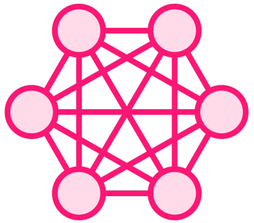
Analyzing large amounts of text

Semantic meaning of words

Word2Vec Analogies

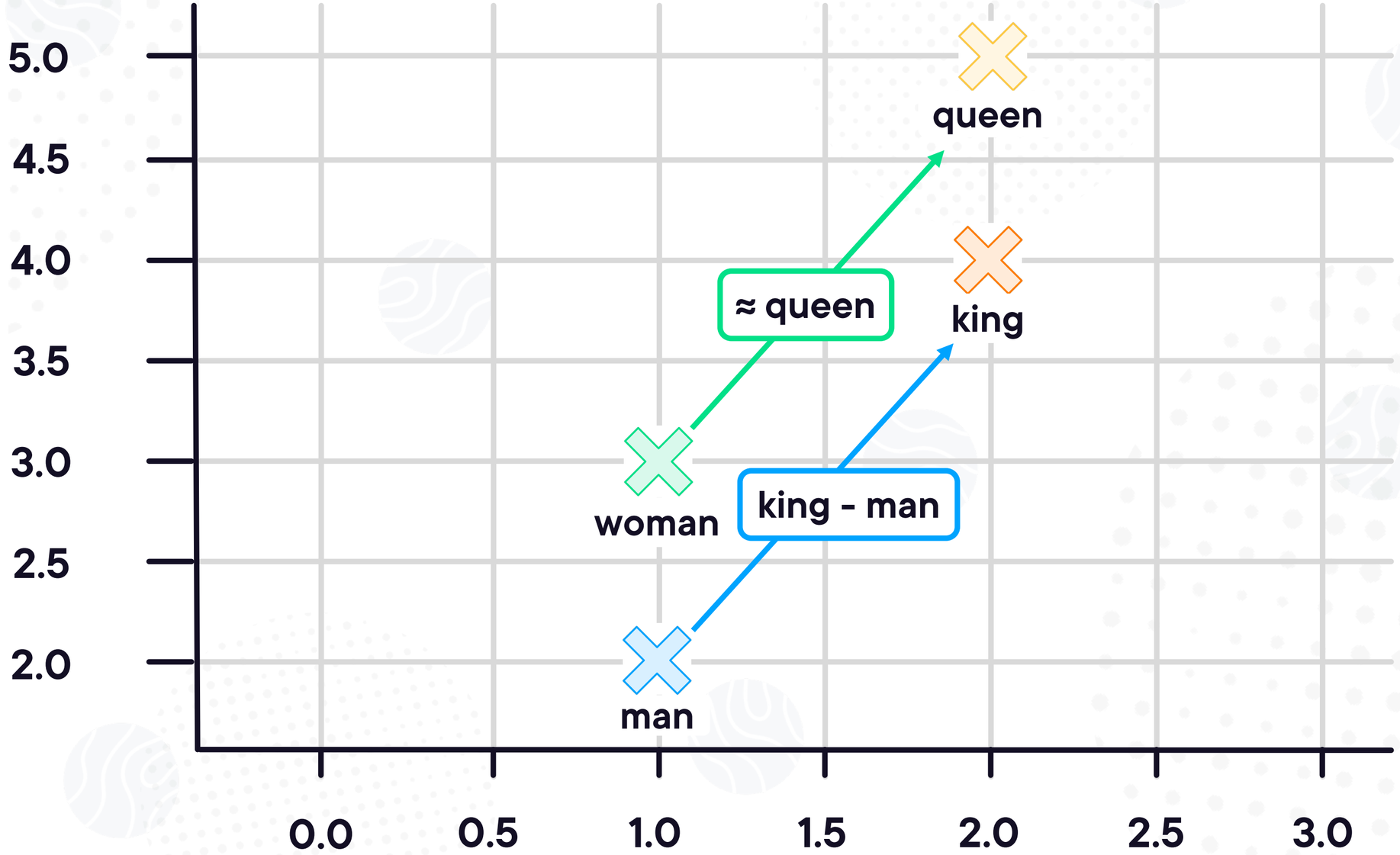
vector("king") – vector("man") + vector("woman") ≈ vector("queen")

Professor/student



Romania/Bucharest

The Map and Coordinates



TF-IDF

Text as sparse vectors of word frequencies weighted by rarity

Simple, fast, good for keyword-based search

Ignores word order and context, cannot capture semantic meaning

Word2Vec

Dense vector representations of words, contexts in large corpora

Captures semantic relationships

Low-dimensional dense vectors

Similarity between words

Static embedding

Limited to handle longer text

Embedding Models

Transformer architecture

Attention mechanism

Text is converted to numerical representations

Importance of a component in a sequence relative to the other components

Contextualized vectors

Embedding Models

Semantic search

Question answering

Retrieval-Augmented Generation

Long documents, relationships across sentences

Computationally expensive, large pre-trained models

Transformer Architecture

Input embeddings

Positional encodings

Self-attention mechanism

Feed-forward neural network

Output layer