Pinecone

Better Semantic Search with Hybrid Vector Search

Roie Schwaber-Cohen, Developer Advocate

Agenda

- What are embeddings?
- The Evolution of Semantic Embeddings
- Sparse Embeddings
- Dense Embeddings
- Dense Semantic Vector Search
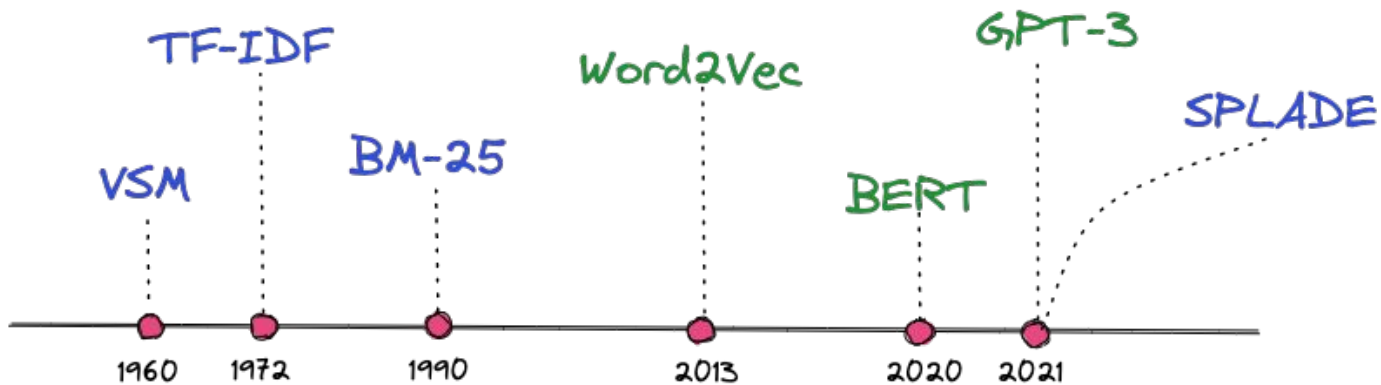- Hybrid Vector Search
- Quick Demo

What are embeddings?

Embeddings are **numerical representations** that capture the **essential features** and **relationships** of discrete objects, like **words** or **documents**, in a continuous **vector space**.
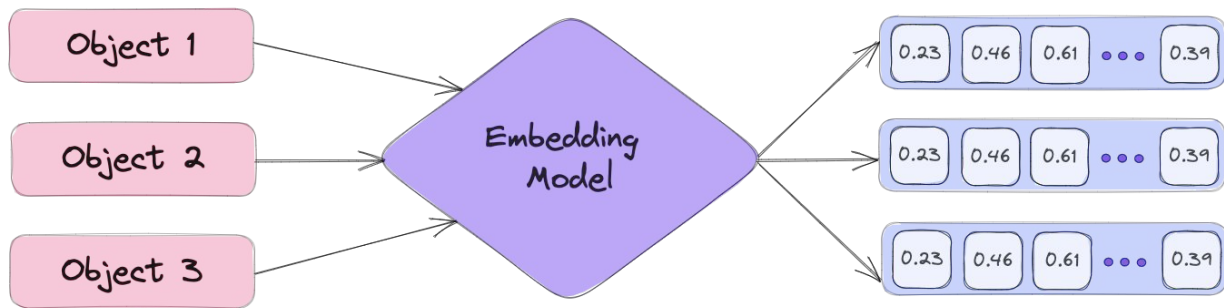
The of evolution **Semantic Embeddings**

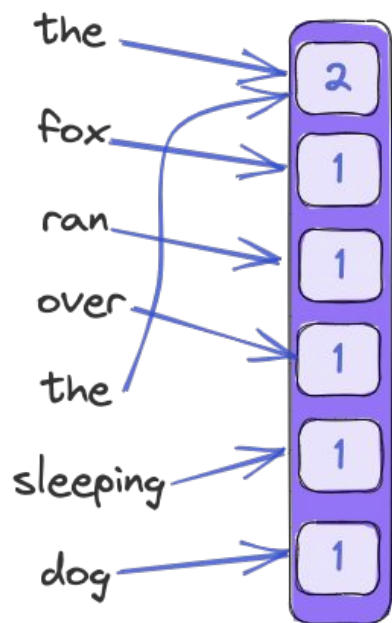# The evolution of **Semantic Embeddings**

The era of keyword-based search: **Sparse Embeddings**

Sparse Embeddings

[0 0 0 0 0 .3 0 0 0 0 0 0 .1 0 ...]

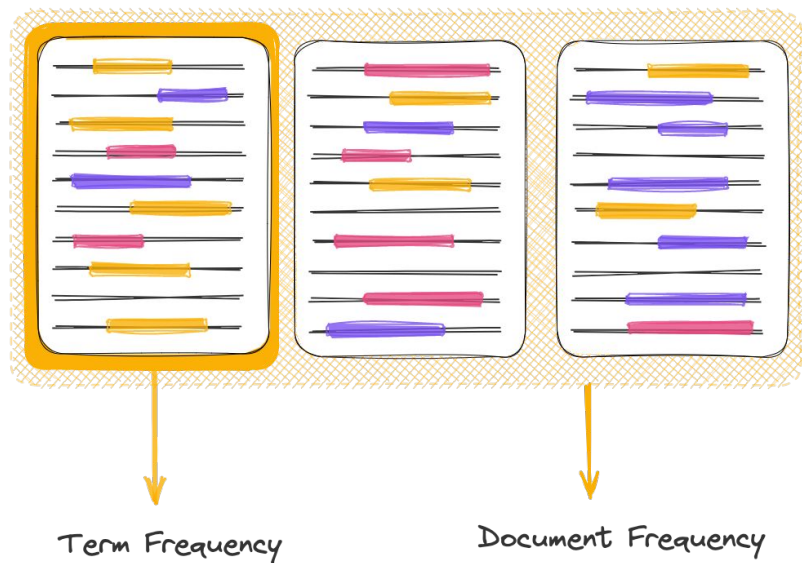# Sparse Embeddings

the fox ran over the sleeping dog
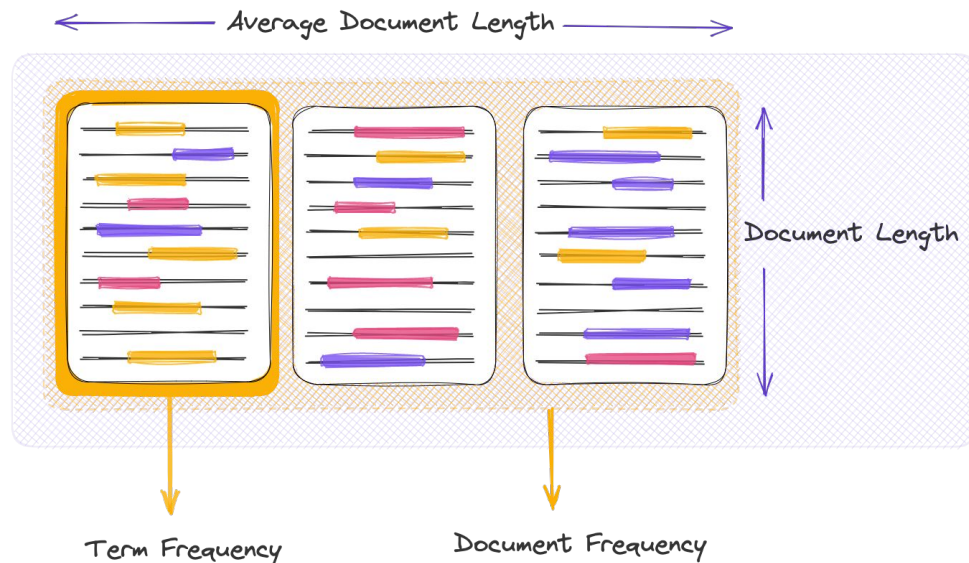
the frog jumped over the log

The era of keyword-based search: Sparse Embeddings

Sparse Embeddings: Pros and Cons

Pros:
- **Efficient** and **performant**
- **Low cost** - no fine tuning required
- Geared towards **exact search**

Cons:
- **Highly** dimensional
- Baseline performance **can't be improved**
- **Meaning isn't** encapsulated
- **Vocabulary mismatch** problem

Pinecone

The rise of Semantic Search: **Dense Embeddings**

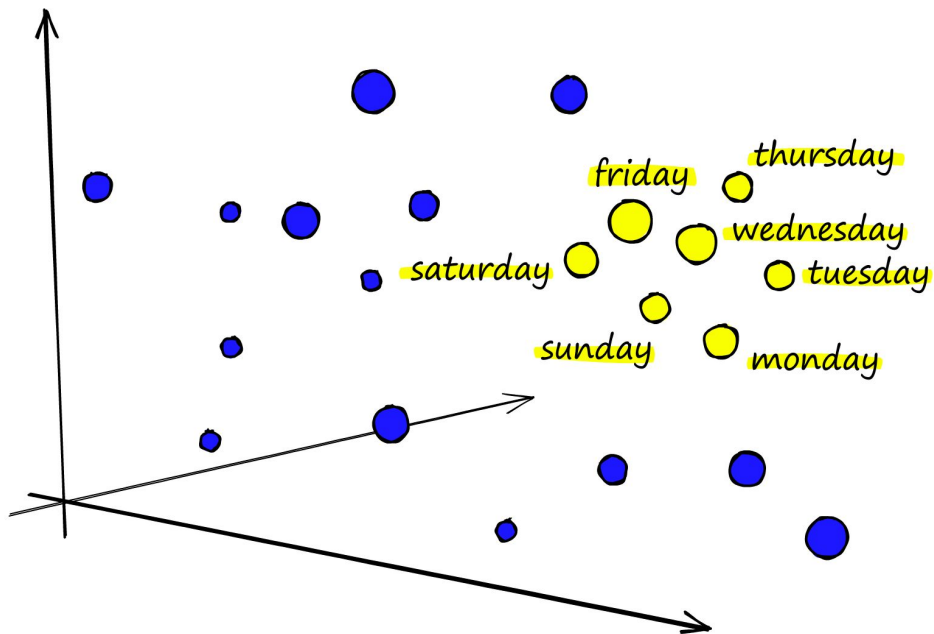# The rise of Semantic Search: **Dense Embeddings**

Image by James Briggs
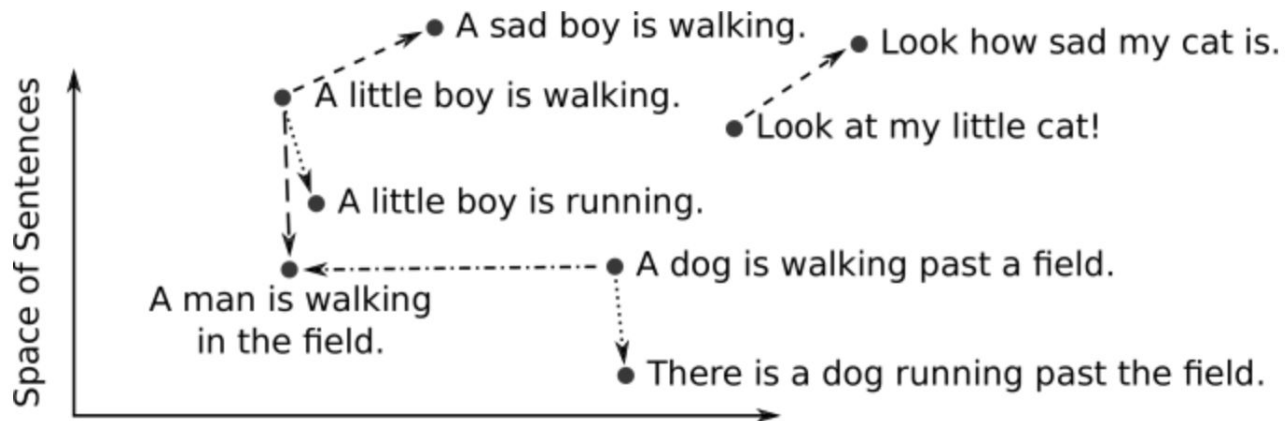
The rise of Semantic Search: **Dense Embeddings**

Image by Deep AI

The rise of Semantic Search: **Dense Embeddings**

Pros:
- Captures **relationships** between words and their meanings
- Can **outperform** sparse representations
- More **compact** representations than sparse embeddings
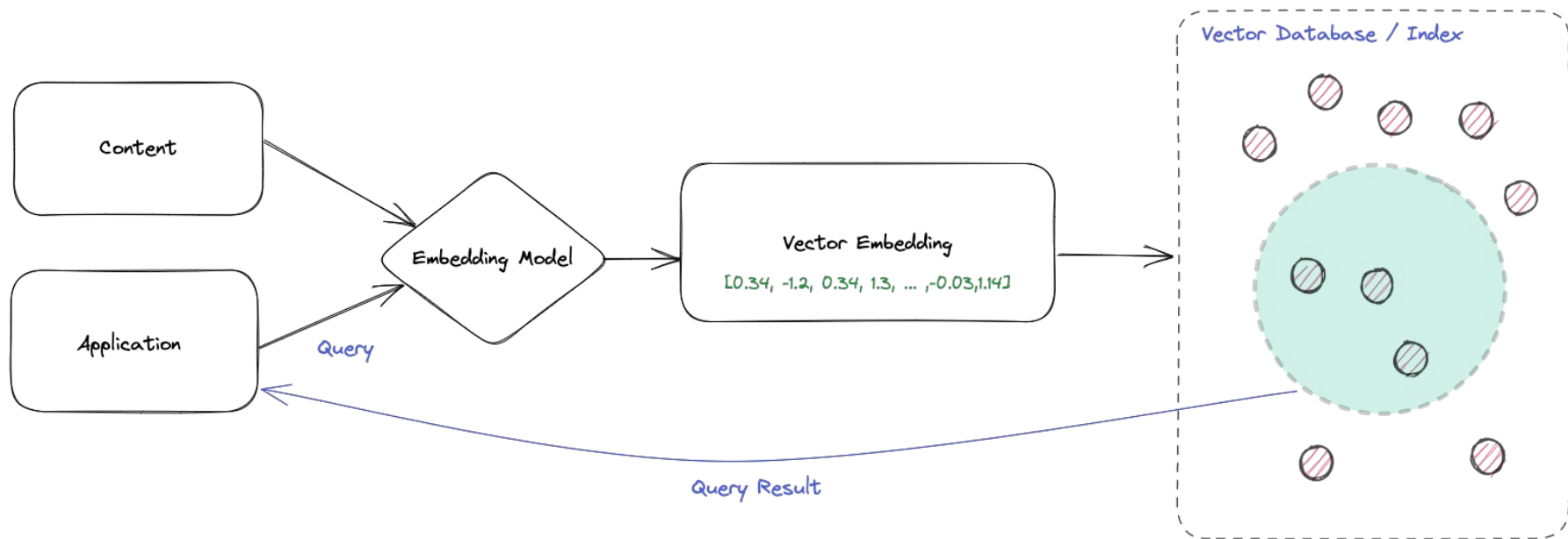- Can be **fine-tuned**

Cons:
- Requires **training**
- **Don't generalize** very well
- Require **more compute** for inference
- Can't do exact matching
- **Challenging** to interpret

Semantic Search using **Dense Embeddings**

Application Architecture

Content → Embedding Model

Application → (Query) → Embedding Model → Vector Embedding [0.34, -1.2, 0.34, 1.3, ... ,-0.03,1.14] → Vector Database / Index

Query Result

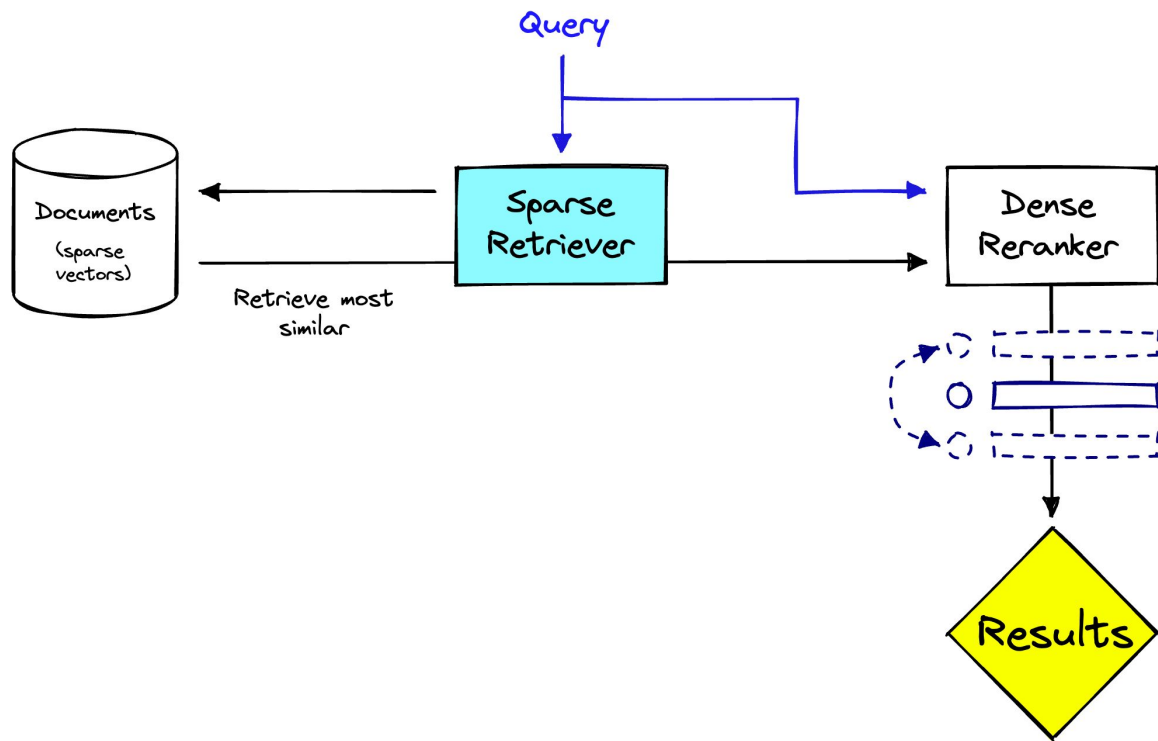# These powers combined: **Hybrid Vector Search**

Two-Stage Retrieval

Image by James Briggs

What if we don't want a subset selection?

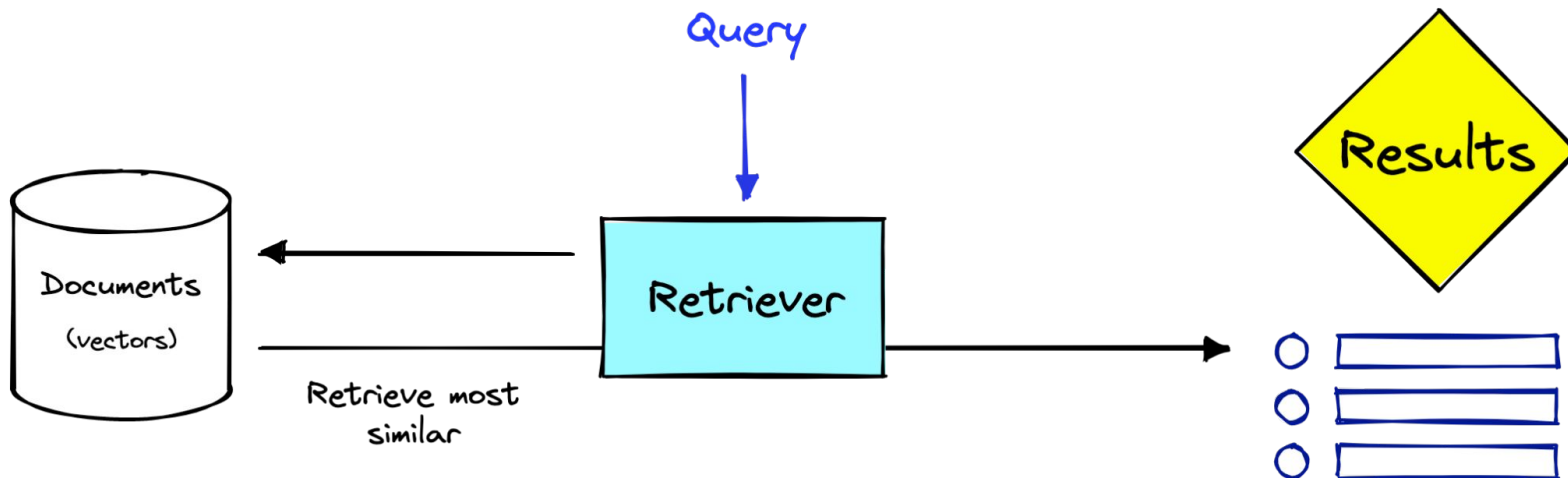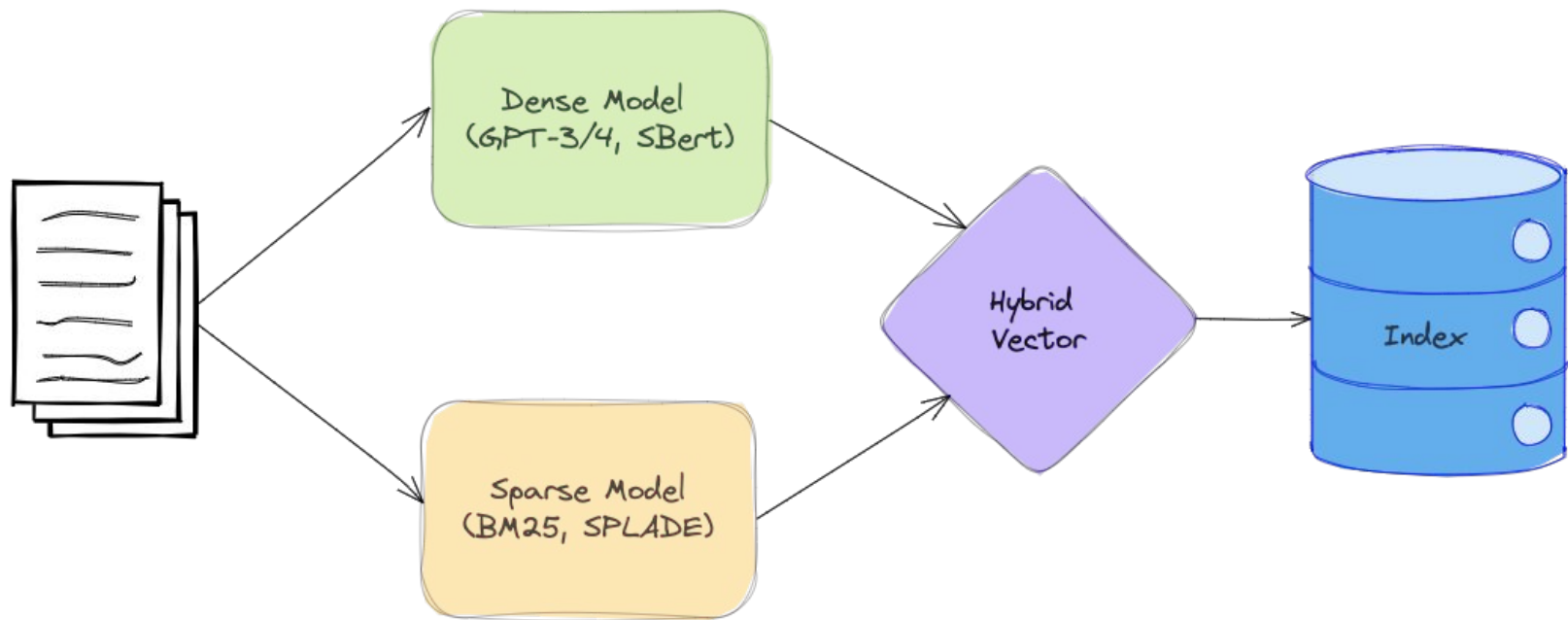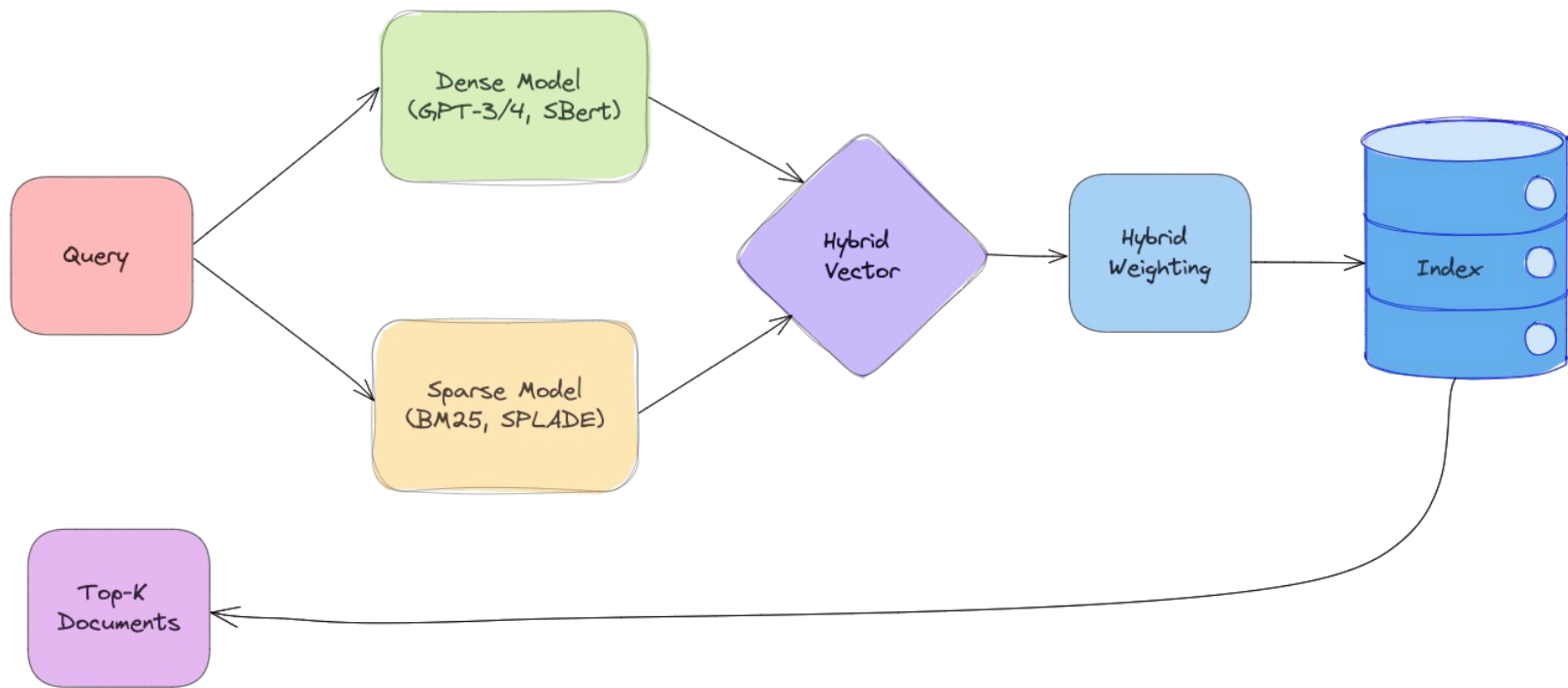Image by James Briggs

Hybrid Vector Search: Storage

Hybrid Vector Search: Query

Quick Demo