

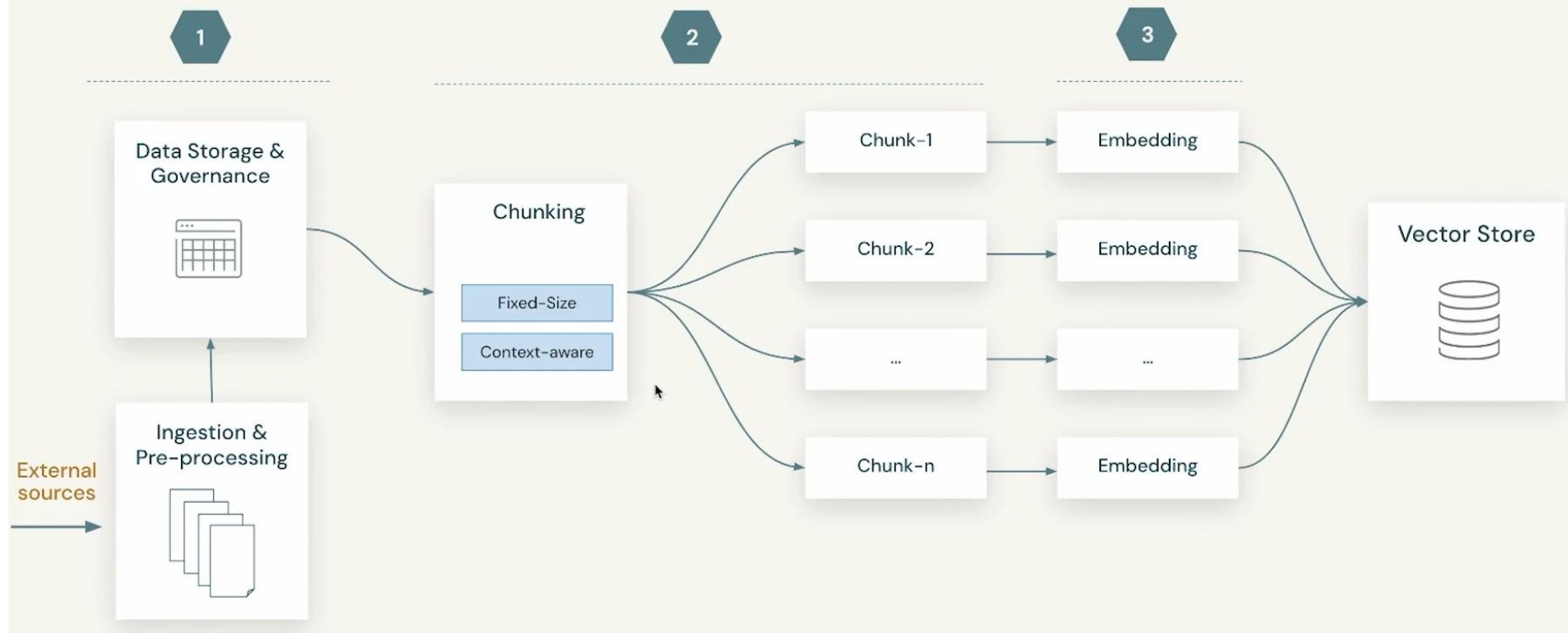
Why is Data Prep Important for RAG?

Potential **issues** when data is prepared improperly

- **Poor quality model output:** If data is inaccurate, incomplete, or biased, the RAG system is more likely to produce misleading or incorrect responses.
- **“Lost in the middle”:** In long context, LLMs tend to overlook the documents placed in the middle. ([Related Research Paper](#) and [needle in haystack test](#) repo).
- **Inefficient retrieval:** Poorly prepared data would decrease the accuracy and precision of retrieving relevant information from knowledge base.
- **Exposing data:** Poor data governance could lead to exposing data during the retrieval process.
- **Wrong embedding model:** Wrong embedding model would decrease the quality of embeddings and retrieval accuracy.

Data Prep Process Overview

A simple data prep process



How to Chunk Data?

How should we organise it?

Neural network

Article Talk

From Wikipedia, the free encyclopedia

Read Edit View history Tools

in

For other uses, see *Neural network* (disambiguation).

A **neural network** can refer to either a neural circuit of biological neurons (sometimes also called a *biological neural network*), or a network of artificial neurons or nodes in the case of an *artificial neural network*.^[1] Artificial neural networks are used for solving artificial intelligence (AI) problems; they model connections of biological neurons as weights between nodes. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred to as a *linear combination*. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be -1 and 1.

These artificial networks may be used for predictive modeling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks, which can derive conclusions from a complex and seemingly unrelated set of information.^[2]

Overview [edit]

A biological neural network is composed of a group of chemically connected or functionally associated neurons. A single neuron may be connected to many other neurons and the total number of neurons and connections in a network may be extensive. Connections, called *synapses*, are usually formed from axons to dendrites, though dendrodendritic *synapses*^[3] and other connections are possible. Apart from electrical signalling, there are other forms of signalling that arise from neurotransmitter diffusion.

Artificial intelligence, cognitive modelling, and neural networks are information processing paradigms inspired by how biological neural systems process data. Artificial intelligence and cognitive modelling try to simulate some properties of biological neural networks. In the artificial intelligence field, artificial neural networks have been applied successfully to speech recognition, image analysis and adaptive control, in order to control software agents (in computer and video games) or autonomous robots.

Historically, digital computers evolved from the von Neumann model, and operate via the execution of explicit instructions via access to memory by a number of processors. On the other hand, the origins of neural networks are based on efforts to model information processing in biological systems. Unlike the von Neumann model, neural network computing does not separate memory and processing.

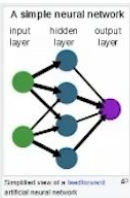
Neural network theory has served to identify better how the neurons in the brain function and provide the basis for efforts to create artificial intelligence.

History [edit]

The preliminary theoretical base for contemporary neural networks was independently proposed by Alexander Bain^[4] (1873) and William James^[5] (1890). In their work, both thoughts and body activity resulted from interactions among neurons within the brain.

For Bain,^[6] every activity led to the firing of a certain set of neurons. When activities were repeated, the connections between these neurons strengthened. According to his theory, this repetition was what led to the formation of memory. The general scientific community at the time was skeptical of Bain's^[7] theory because it required what appeared to be an inordinate number of neural connections within the brain. It is now apparent that the brain is exceedingly complex and that the same brain "wiring" can handle multiple problems and inputs.

James^[8] theory was similar to Bain's,^[6] however, he suggested that memories and actions resulted from electrical currents flowing among the neurons in the brain. His model, by focusing on the flow of electrical currents, did not require individual neural connections for each memory or action.



Title

Section

Diagram

Context-aware Chunking:

- Chunk by sentence/paragraph/section
- Leverage special punctuation (i.e. '.', '\n')
- Include/Inject metadata/tags/title(s)

&/OR

Fixed-size Chunking:

- Divide by a specific number of tokens
- Simple and computationally cheap method

Chunking Strategy is Use-Case Specific

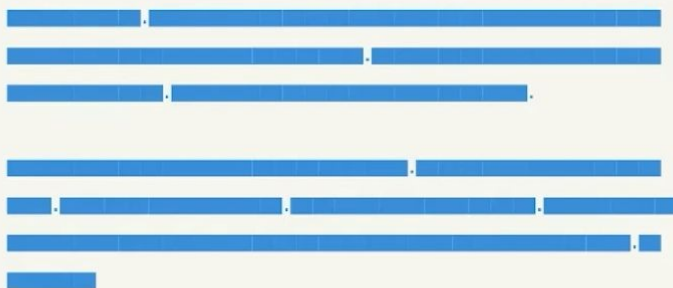
Another iterative step! Experiment with different chunk sizes and approaches

- How long are our documents?
 - 1 sentence?
 - N sentences?
- If 1 chunk = **1 sentence**, embeddings focus on specific meaning
- If 1 chunk = **multiple paragraphs**, embeddings capture broader theme
 - How about splitting by headers?

Chunking by sentence:



Chunking by Paragraph:



Chunking Strategy is Use-Case Specific

Another iterative step! Experiment with different chunk sizes and approaches

- Chunk **overlap** defines the amount of overlap between consecutive chunks, ensuring that no contextual information is lost between them.

- **Windowed summarization** is a 'context-enriching' chunking method where each chunk includes a 'windowed summary' of previous few chunks.

- Prior knowledge of user's query patterns can be helpful (*i.e. query length?*)
 - While long queries may have better aligned embeddings to returned chunks, shorter queries could be more precise

Chunk overlap:



Windowed summarization:



Advanced Chunking Strategies

Summarization with metadata

Neural network

Article · Talk
From Wikipedia, the free encyclopedia

For other uses, see [Neural network \(disambiguation\)](#).

A **neural network** can refer to either a neural circuit of biological neurons (sometimes also called a biological neural network), or a network of artificial neurons or nodes in the case of an artificial neural network.^[1] Artificial neural networks are used for solving artificial intelligence (AI) problems; they model connections of biological neurons as weights between nodes. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred to as a *linear combination*. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be -1 and 1 .

These artificial networks may be used for predictive modeling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks, which can derive conclusions from a complex and seemingly unrelated set of information.^[2]

Overview

A biological neural network is composed of a group of chemically connected or functionally associated neurons. A single neuron may be connected to many other neurons and the total number of neurons and connections in a network may be extensive. Connections, called *synapses*, are usually formed from axons to dendrites, though *dendrodendritic synapses*^[3] and other connections are possible. Apart from electrical signaling, there are other forms of signaling that arise from *neurotransmitter diffusion*.

Artificial intelligence, cognitive modeling, and neural networks are information processing paradigms inspired by how biological neural systems process data. Artificial intelligence and cognitive modeling try to simulate some properties of biological neural networks. In the artificial intelligence field, artificial neural networks have been applied successfully to speech recognition, image analysis and adaptive control, in order to construct software agents (in computer and video games) or autonomous robots.

Historically, digital computers evolved from the von Neumann model, and operate via the execution of explicit instructions via access to memory by a number of processors. On the other hand, the origins of neural networks are based on efforts to model information processing in biological systems. Unlike the von Neumann model, neural network computing does not separate memory and processing.

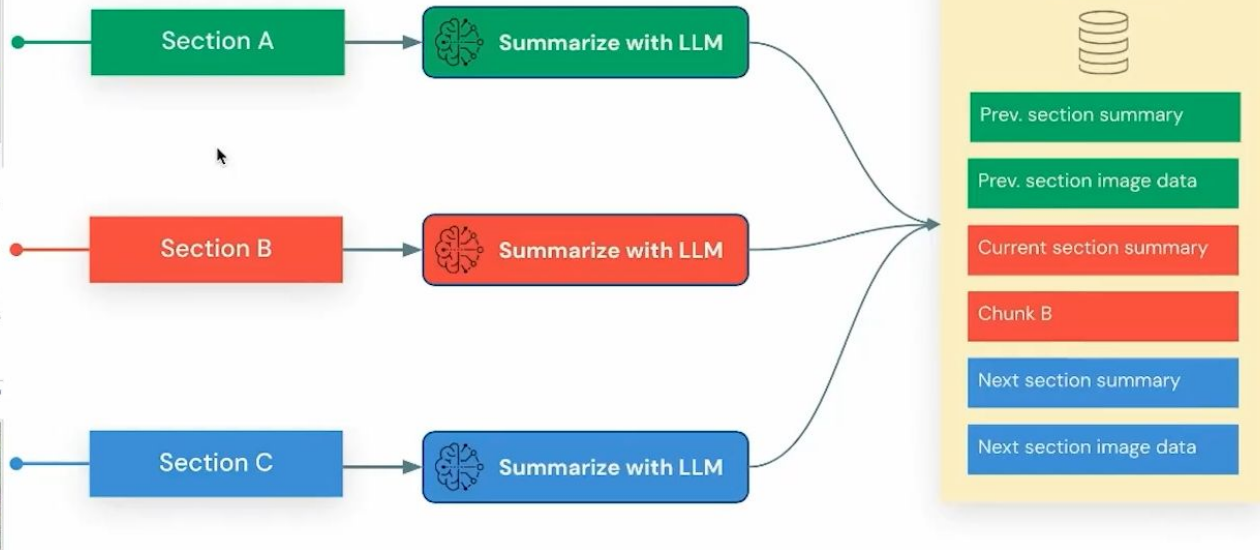
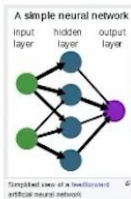
Neural network theory has served to identify better how the neurons in the brain function and provide the basis for efforts to create artificial intelligence.

History

The preliminary theoretical base for contemporary neural networks was independently proposed by Alexander Bain^[4] (1873) and William James^[5] (1890). In their work, both thoughts and body activity resulted from interactions among neurons within the brain.

For Bain,^[6] every activity led to the firing of a certain set of neurons. When activities were repeated, the connections between these neurons strengthened. According to his theory, this repetition was what led to the formation of memory. The general scientific community at the time was skeptical of Bain's^[7] theory because it required what appeared to be an inordinate number of neural connections within the brain. It is now apparent that the brain is exceedingly complex and that the same brain "wiring" can handle multiple problems and inputs.

James^[8] theory was similar to Bain's,^[9] however, he suggested that memories and actions resulted from electrical currents flowing among the neurons in the brain. His model, by focusing on the flow of electrical currents, did not require individual neural connections for each memory or action.



Data Extraction and Chunking Challenges

Working with complex documents

**HOLIDAY
PACKAGES**

Our exclusive range of holiday packages has been specially designed with you in mind and features a choice of accommodation and local experiences. These packages are a perfect introduction to Bali and offer great value for money.

Relax on a beautiful beach, explore lush green rice terraces, or visit the local markets. Whether you're looking for adventure and fun for the whole family or a quiet, romantic getaway, Bali is the perfect holiday destination.

Be inspired by one of our fantastic holiday packages and discover all that Bali has to offer.



Image



**UBUD SPA & WELLNESS RETREAT
4 NIGHTS**

Spending a day at The Yoga Barn, located in the heart of Uluwatu, and experience the holistic healing of this full sensory yoga studio.

INCLUDES

- *Experiments*

- Full breakfast daily
 - Indonesian and international on-site evening lectures (homestay)
 - Afternoon tea at Royal Kailash Hotel and daily
 - Yoga class and scheduled cultural activities daily
 - Yoga and Guided Trek leader Experience
 - Multiple services to 11 local centres
 - Welcome drink, full breakfast and gift on arrival
 - Return private car transfers from Airport for International Return
- Notes to apply: To be more details on this program

... ..

from \$1055*

*Price includes 1 free night and 10% early bird discount (book 45 days prior to travel, based on 4 night package, valid 1 Apr - 30 Jun, 1 Sep - 18 Dec 20, 17 Jan - 31 Mar 21. And your travel agent has access to our online travel and booking tools.



Text



BEST OF BALI BEACHES
7 NIGHTS

Enjoy the sun and sand at Legian and Canggu beaches for the ultimate Bali beach getaway.

With tennis, a golf course, a swimming pool, a tennis club, a large beach, and a wide range of facilities and a great beachfront location, spend an afternoon on a relaxing sunset dinner cruise with live entertainment and a buffet dinner.

A two-hour drive from Legian, you'll arrive at Candikaya, the perfect base to explore northern Bali's villages and countryside. Clear water and crisp red muds in this destination make it a must-visit.

INCLOSURE

- 2 nights 4 site accommodation in a Deluxe room at Legian Beach Hotel
 - 2 nights 4 site accommodation in a Deluxe Garden room at Candl Beach Resort & Spa • Full breakfast daily
 - 2 Hour Sunset Dinner Cruise from Legian
 - Daily guided morning walks from Candl Beach Resort & Spa
 - Return private car transfers from Ngurah Rai International Airport
 - Private car transfers between Legian Beach Hotel and Candl Beach Resort & Spa
- Refer to pages 21 and 24 for more details on these properties

10254

From \$935*

^aBased on 7 night packages, valid 1 Apr - 14 Jun, 16 Oct - 20 Dec 20, 6 Jan - 21 Mar 21. Ask your travel agent for prices for other dates and room types.



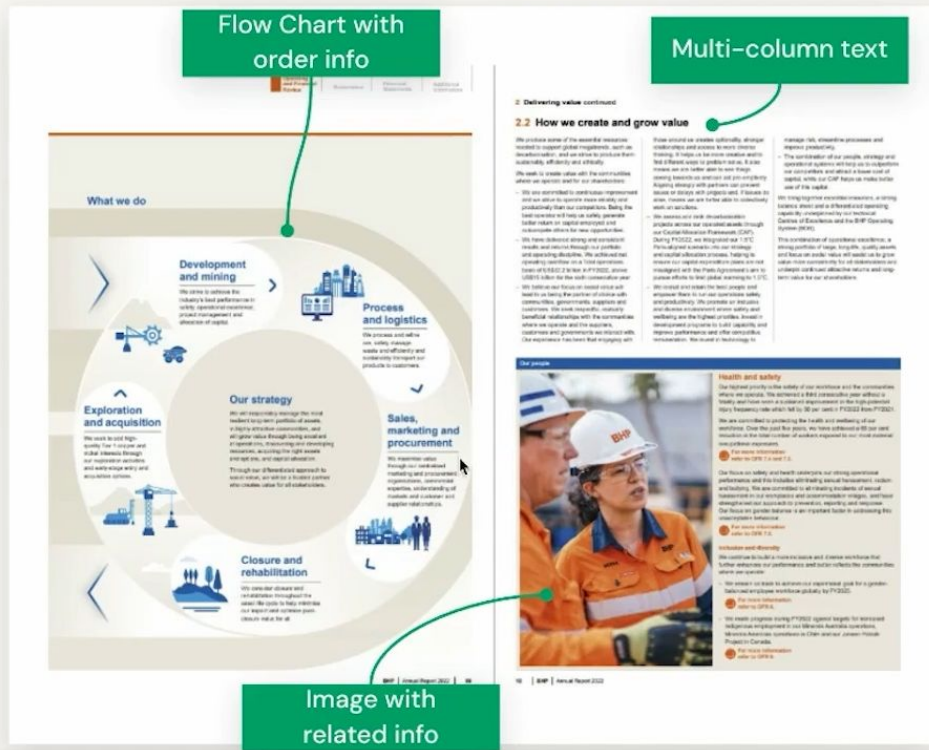
Price and disclaimer

Other challenges:

- Text mixed with image
- Irregular placement of text
- Color encoded focus (*Important for context*)

Data Extraction and Chunking Challenges

Working with complex documents



Other challenges:

- Chart with hierarchical information. Keeping the order of the information is critical.
- Multi-column text and the order of columns if crucial.
- Keeping images with related information is crucial.

General Approaches

Approaches to address unstructured/complex raw text documents

Traditional Approach

Libraries:

- PyMuPDF
- PyPDF

Features:

- Breaks down text to into raw constructs
- Very low level requires hard coding rules

Use a layout model

Libraries:

- Hugging Face
 - LayoutLMv3
- doctr
- Donut
- Unstructured

Features:

- Apply Deep learning models built to do text extraction and context extraction

Multi-Modal Models

Models:

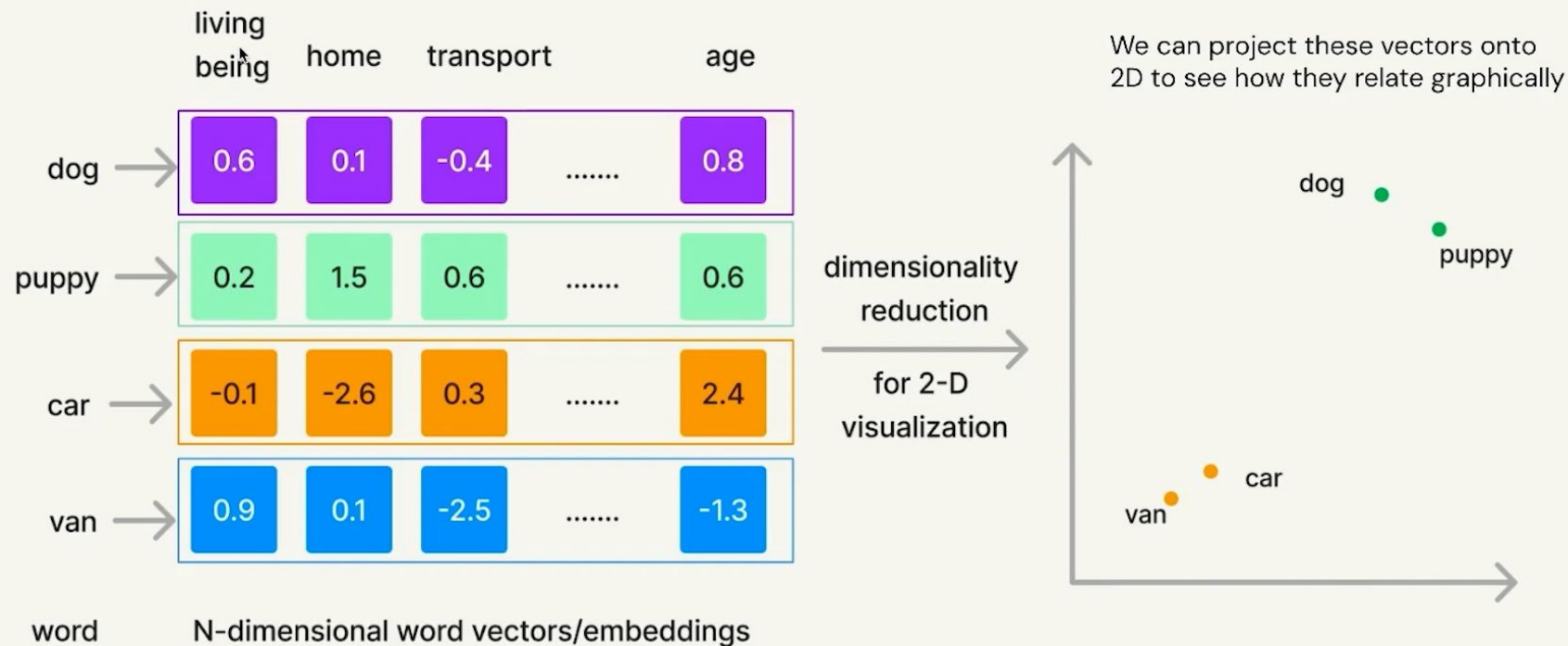
- OpenAI's GPT-4o (and beyond)
- Alphabet's Gemini1.5 (and beyond)
- Other OSS models (i.e. Dolphin's Series, OpenFlamingo, Llava, OLMo)

Features:

- Multimodal LLMs intrinsically understand images but are still more experimental at this stage

Refresher: Representing Words with Vectors

Embedding: A numerical representation of content



Embedding Models

Choosing the **right model for your application**

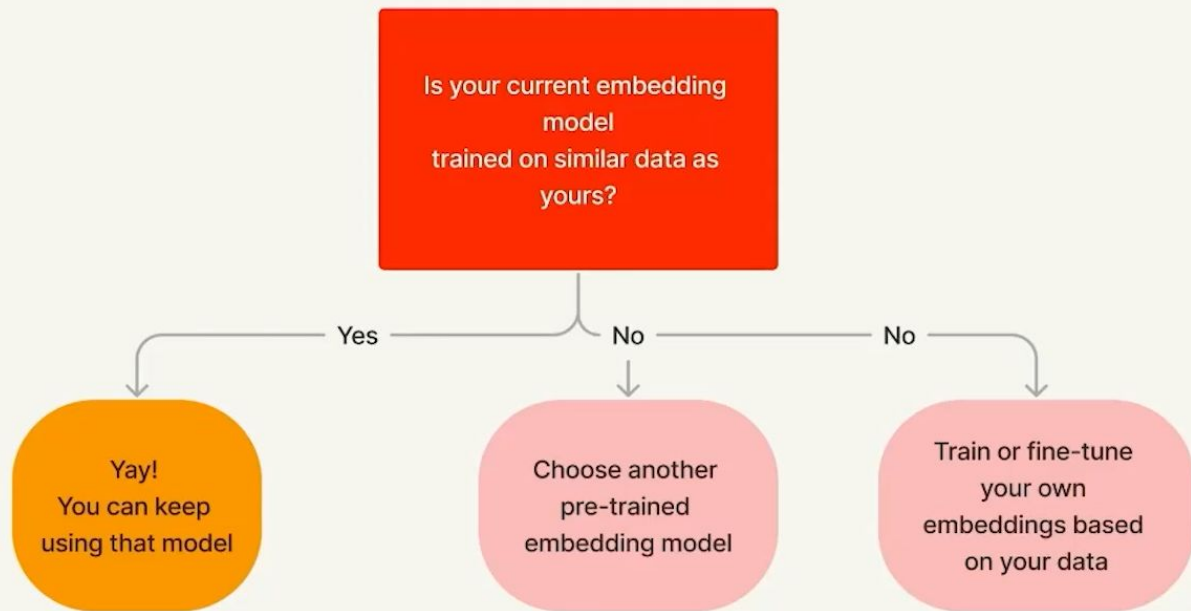
- Data/Text properties:
 - Vocabulary size in your text/documents (some models handles more diverse words)
 - Domain/Topic (i.e. finance, medical, news etc.)
 - Text length: typical length of chunks/docs to be embedded
- Model capabilities:
 - Multi-Language support
 - Embedding dimensions/size: more storage cost for higher dimensions

Practical considerations:

- Be aware of context window limitations. Many embedding models will **ignore text beyond their context window limits**.
 - Privacy and cost/licensing when using proprietary API-based models.
- ⇒ benchmark multiple models & choose the one that strikes the best balance.

Tip 1: Choose Your Embedding Model Wisely

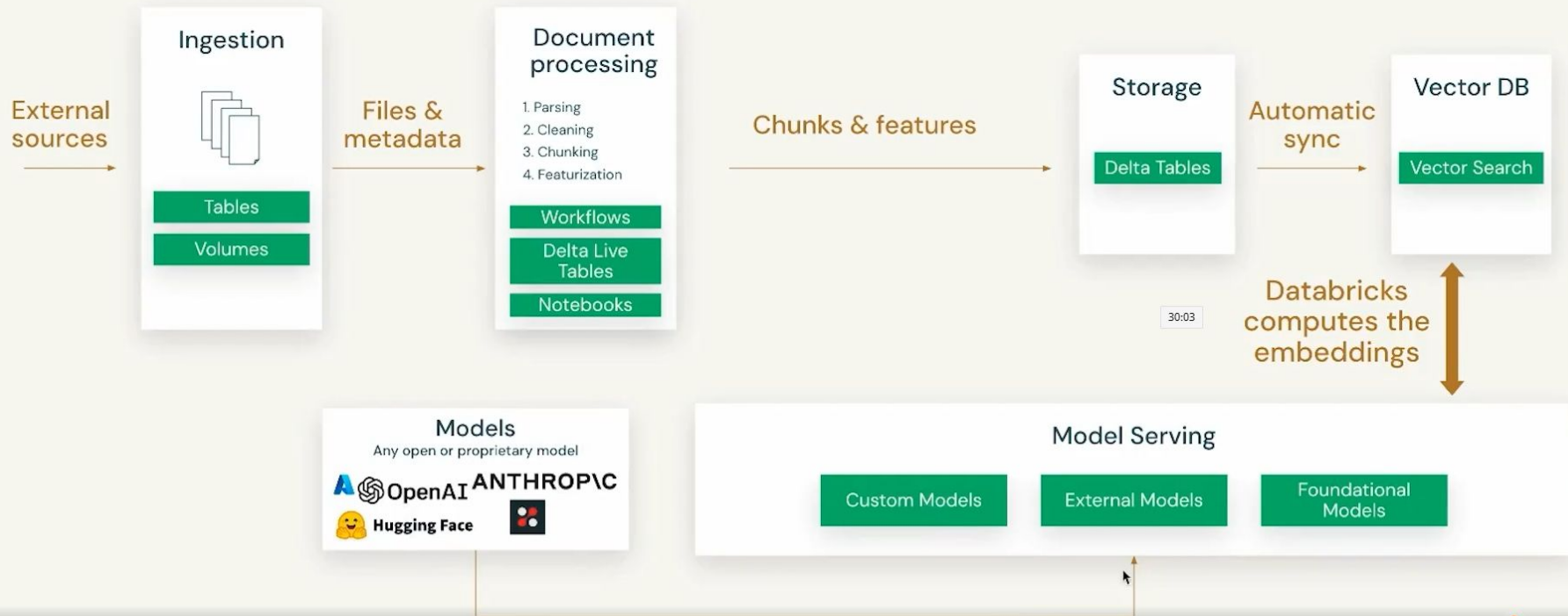
The embedding model should represent **BOTH** queries and documents



This practice has been around for years in NLP.
Example: Fine-tune BERT embeddings

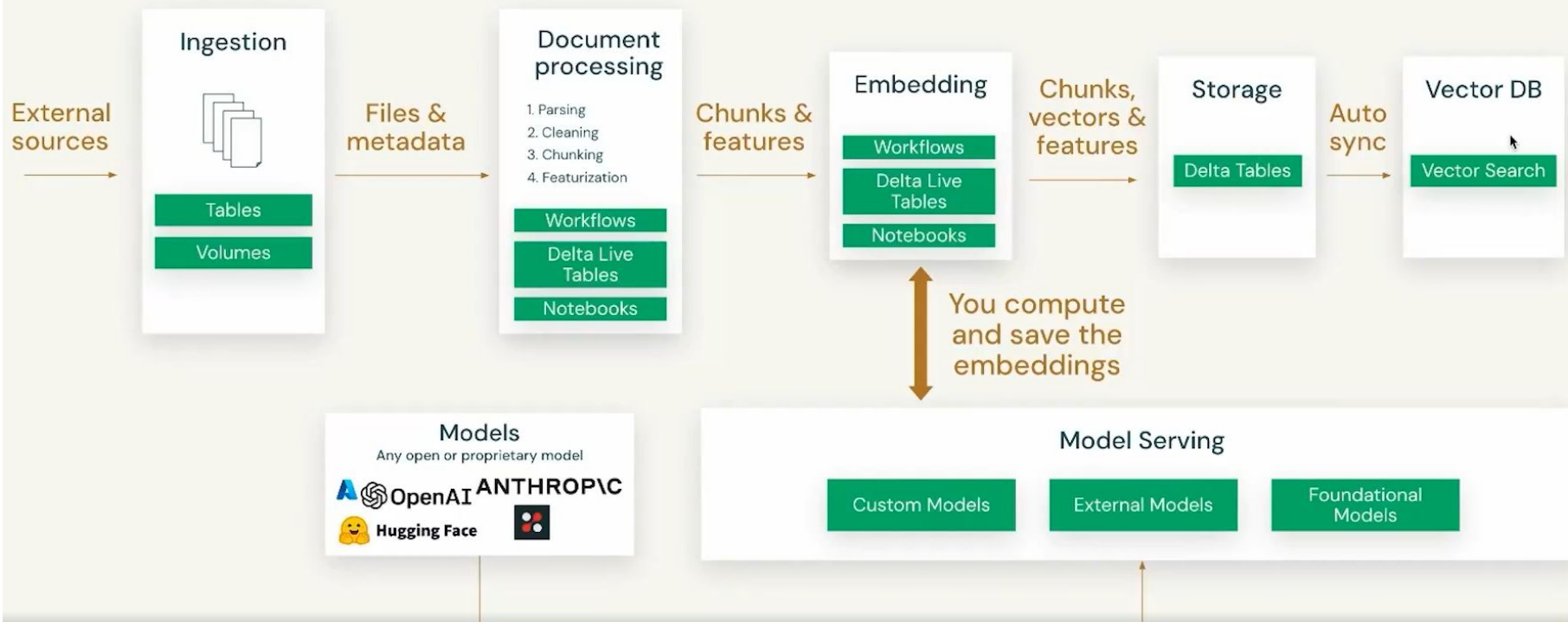
Unstructured Data Prep

Vector Search with Databricks-managed embeddings



Unstructured Data Prep

Vector Search with user-managed embeddings



Structured Data Prep

Feature Serving and Online Tables

