DATASTAX

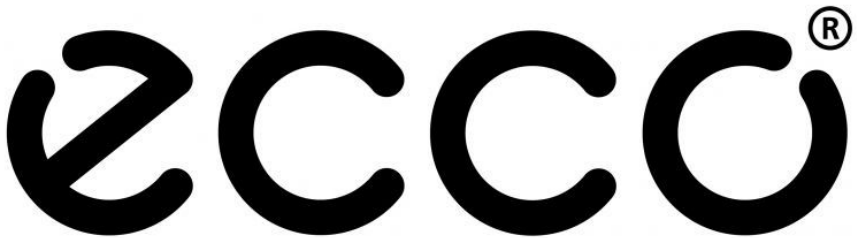
Mastering RAG : Build & Deploy Your AI Agent

Hands-on Workshop

Copenhagen, January 30th, 2024

- 4.00 pm - Welcome
- 4.05 pm - Introduction to Generative AI
- 4:30 pm - Hands-on: Build Your Own RAG
- 6.00 pm - Networking, food and drinks
- 7.00 pm - We go home proudly with a working

OUR SPONSOR

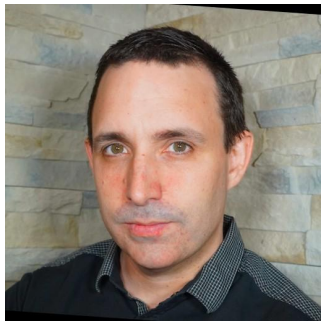


Welcome!

Kai Kühnel
Sales

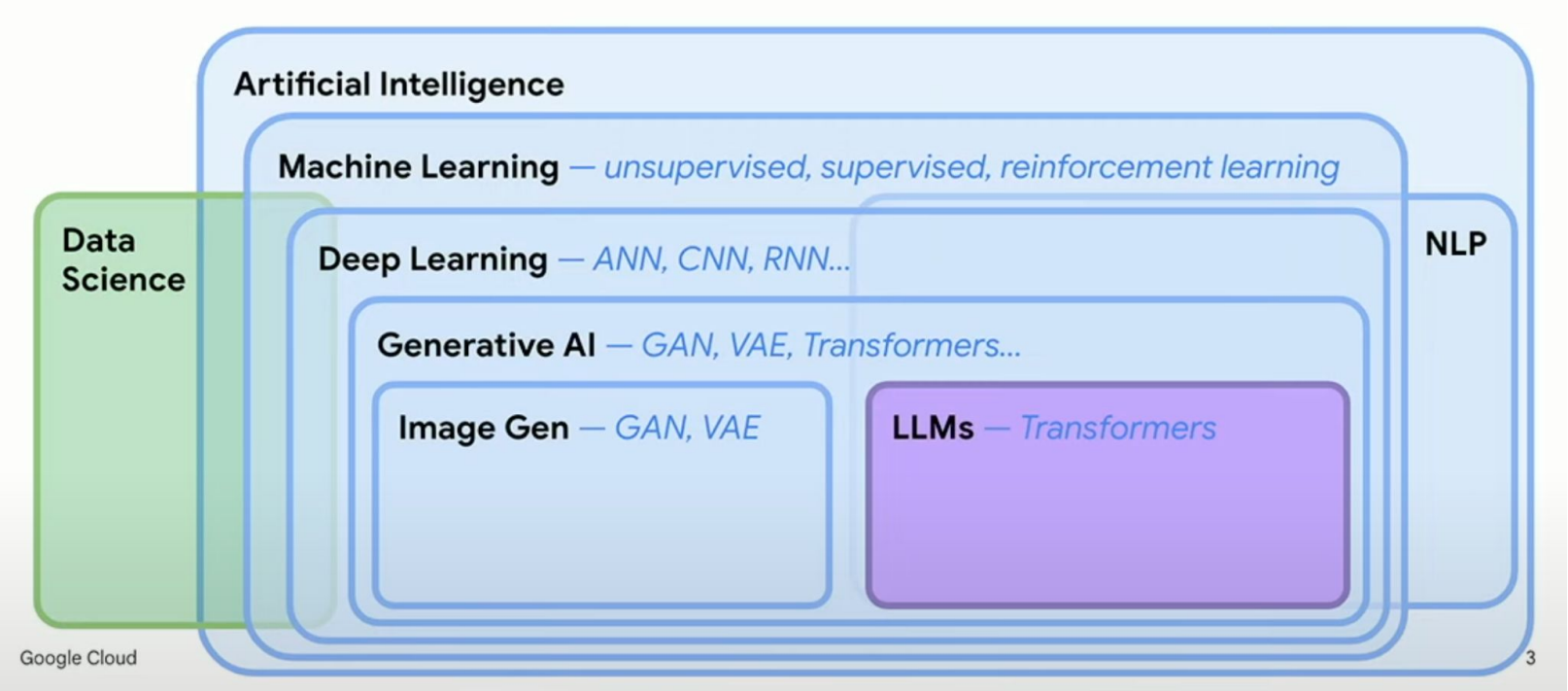
Sami Kaksonen
Solution Engineer

Cédrick Lunven
Software Engineer

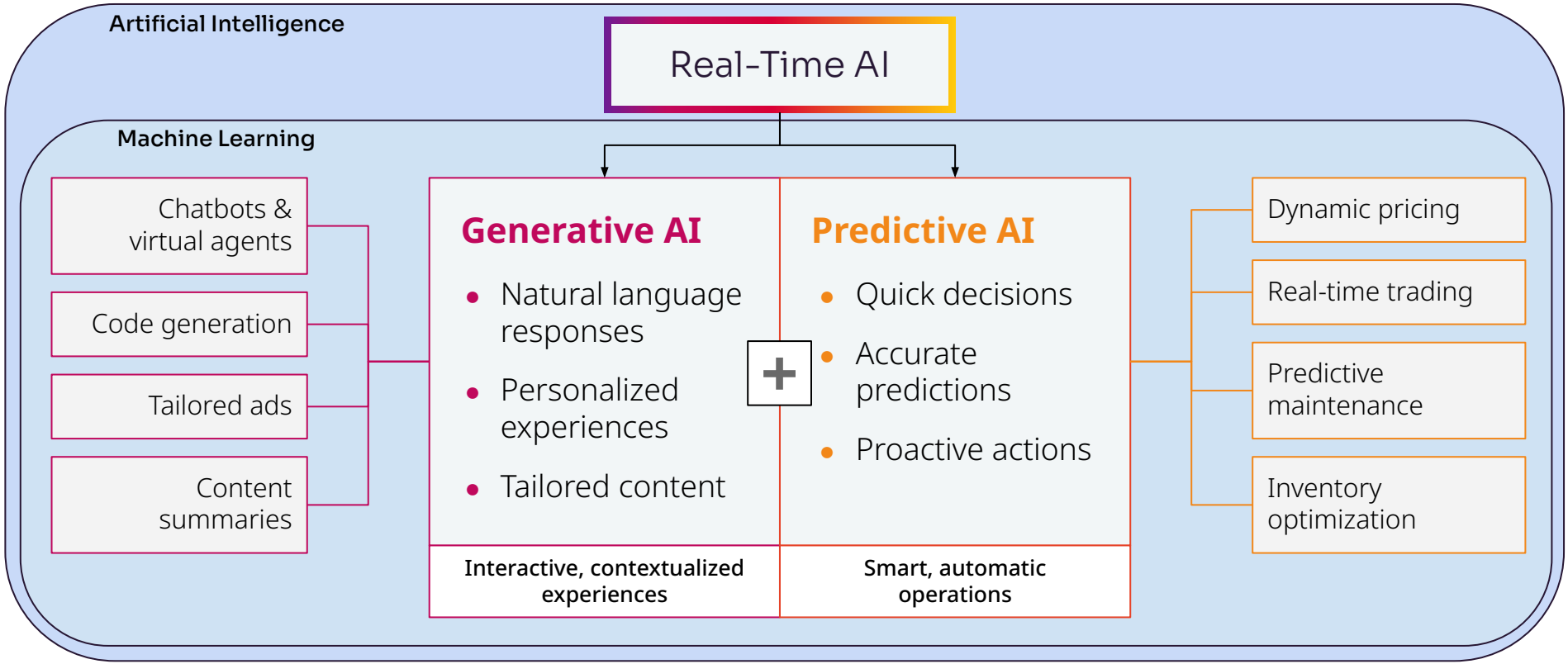


Introduction to Generative AI

The Big picture



Predictive vs Generative AI



Large Language Models (LLM)

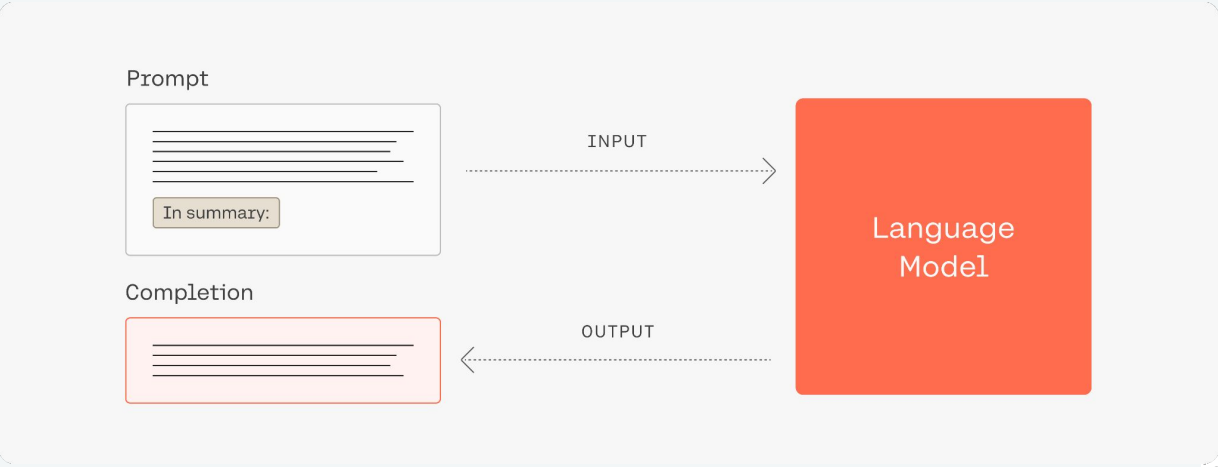


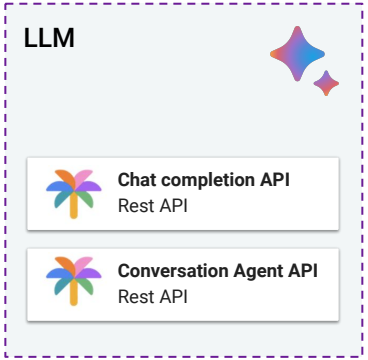
image by cohere

Large language models are trained on massive amounts of text data.

Designed to process and understand natural language, such as human speech and text.

This allows the model to learn patterns and relationships between words, phrases, and sentences, enabling it to generate coherent and meaningful language output.

Limitations of “LLM only” mode



- LLM .....can be outdated
- LLM .....Does not know *your* data
- LLM .....is not tuned = hard steerability
- LLM .....Hallucinating if not properly prompted
- LLM .....works with limited Input windows (tokens)

There is no AI without Data

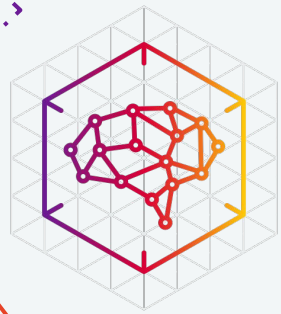
Adding Enterprise Context to GenAI

Context added at run-time
Data not retained

Proprietary & Personalized Information

Retrieval-Augmented Generation

Context Window



Domain Specific Enterprise Data

Fine-Tuning

Public Data

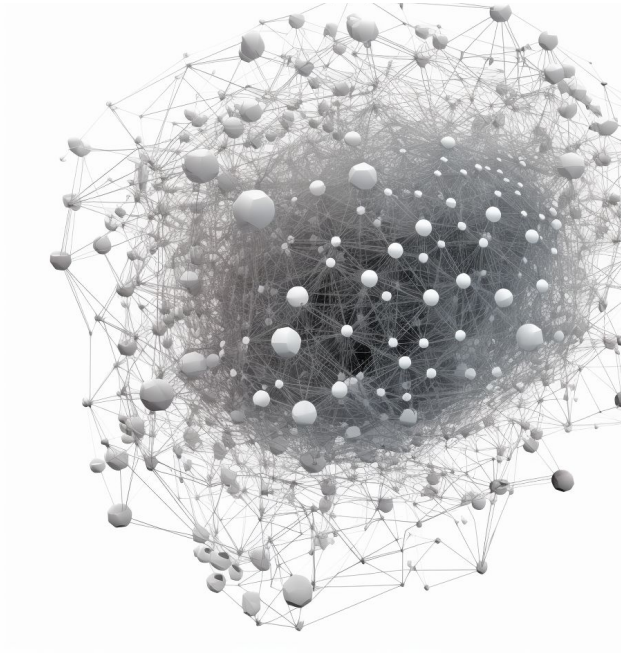
Pre-Training

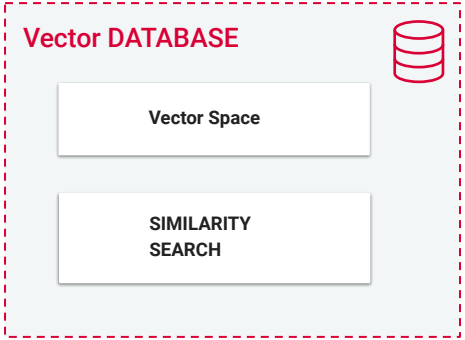
Model Parameters

Context added in training
Data is retained

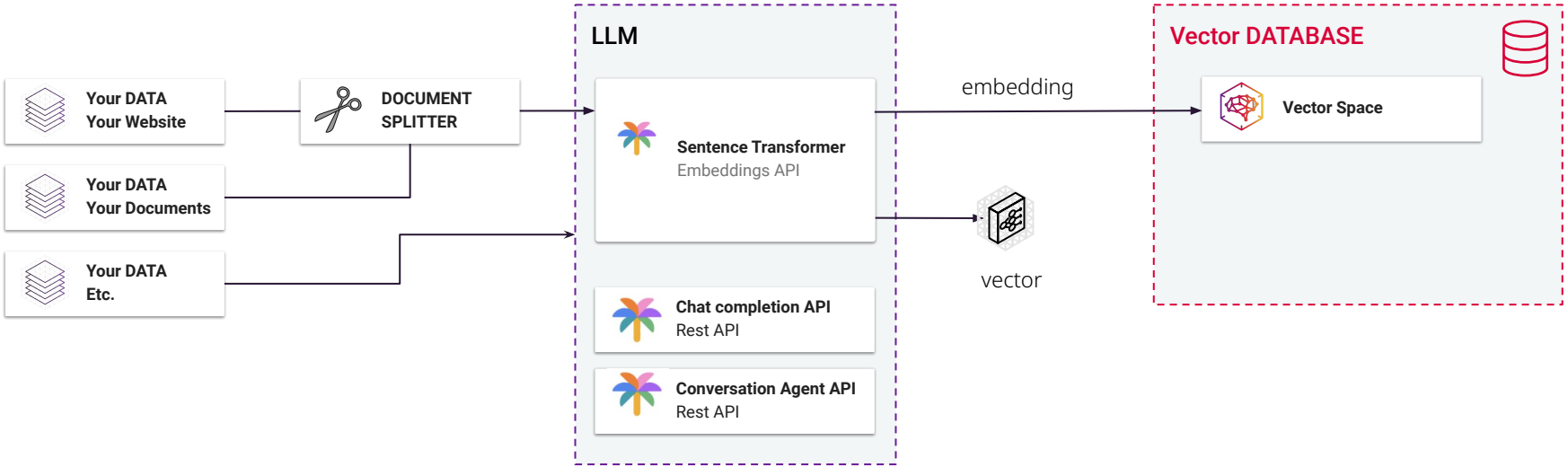
Good Vector Database ?

- Handling **A LOT** of vector
- Effective Search Algorithms
- Performant, Resilient
- Dynamic (vector sizes)
- Meta Data Filtering
- Keyword search
- Semantic Caching
- Chat History
- Key Value cache



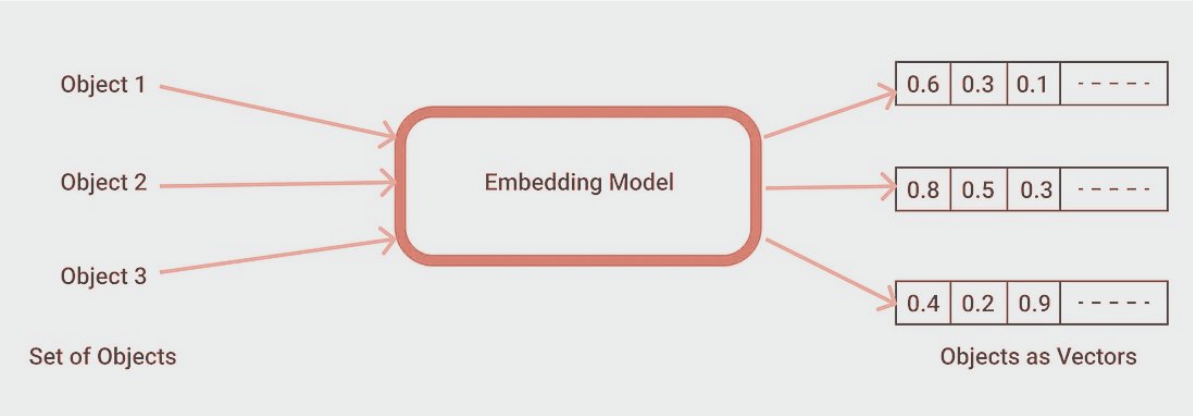


Adding Context: **Vectorization**



What is Embedding?

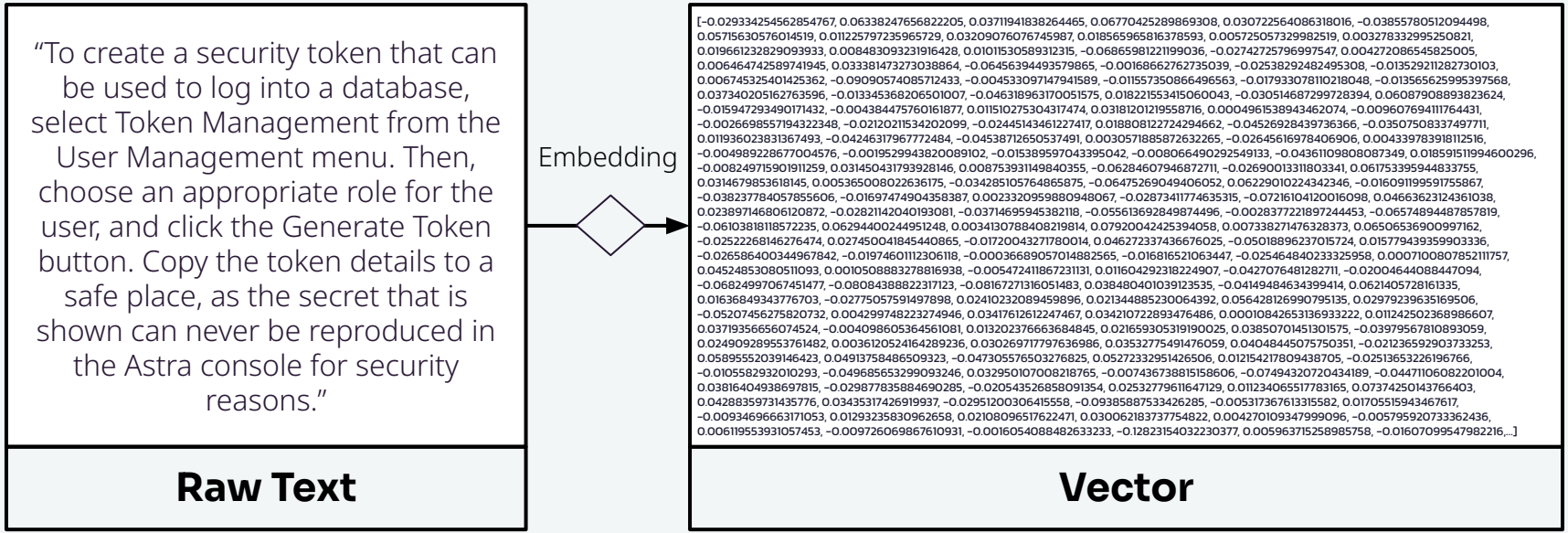
Take source objects (text, images, sound, movies) and create Vector format represenation of the context.
This allows for Similarity Search on the Database finding Semantically comparable objects.



Goal is to capture semantics and context.
Processed by a ML model

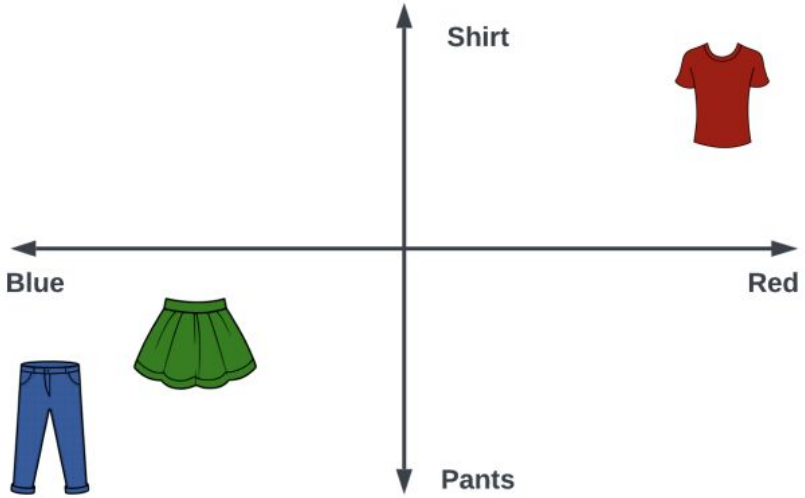
What is Vector?

Vector is a multi-dimensional numerical representation of text/image/video

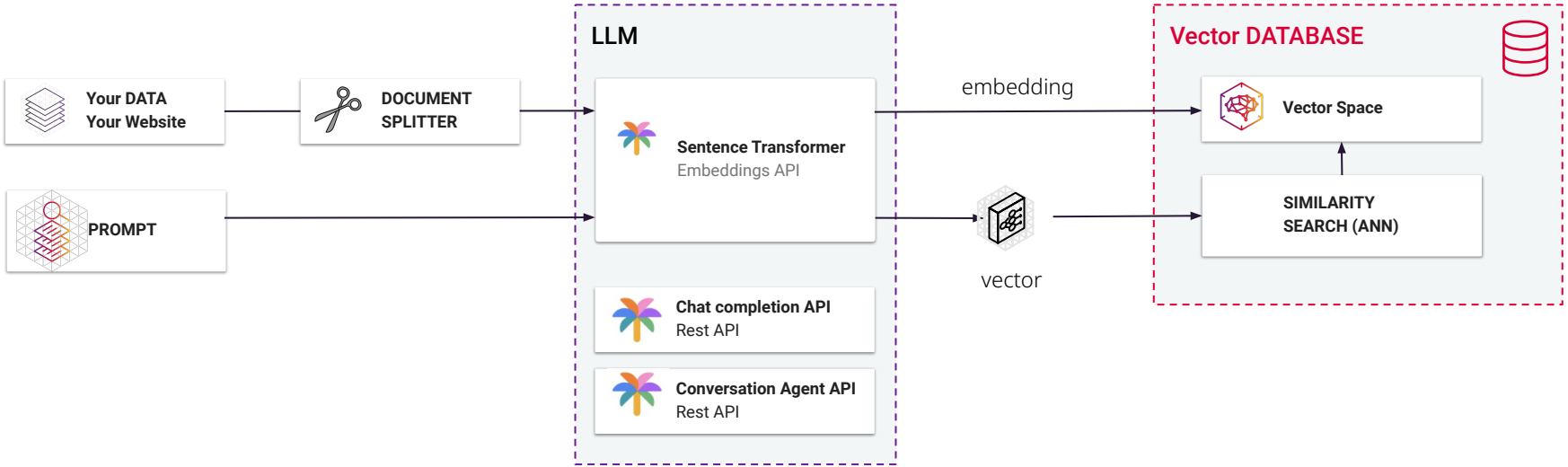


What is Vector Search?

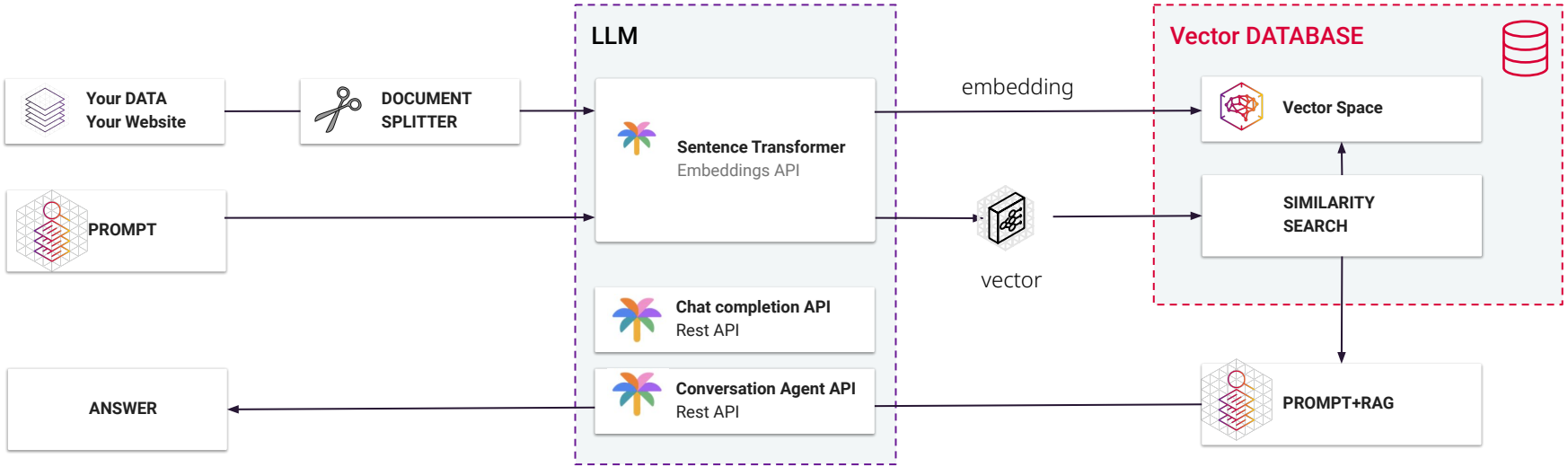
- Vector search finds objects that have similar meaning
- Vector search understands MEANING
- Vectors created from EXISTING data through EMBEDDINGS



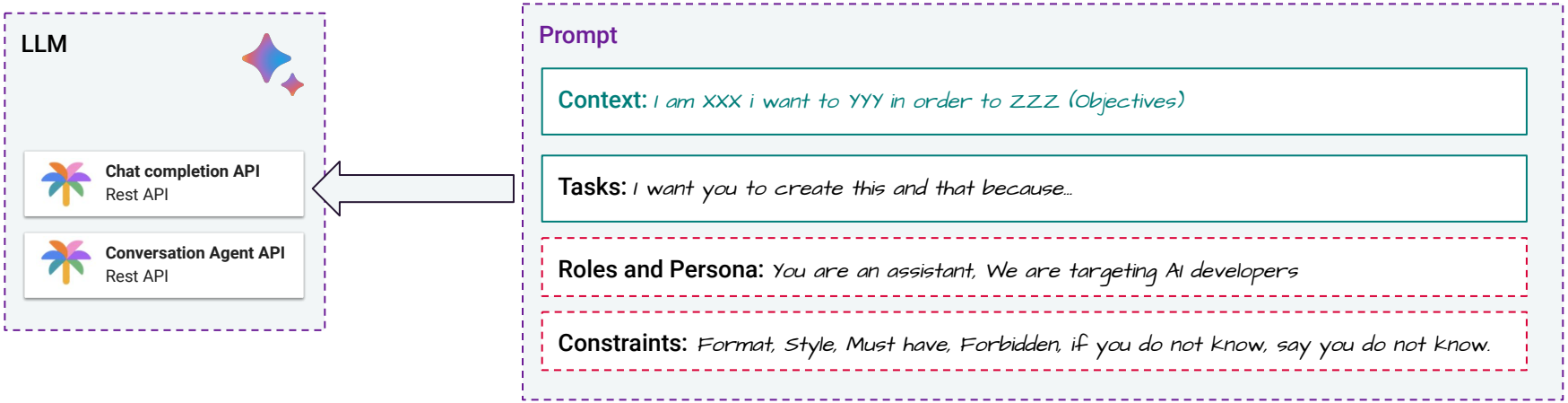
Retrieving Context: **Semantic Search**



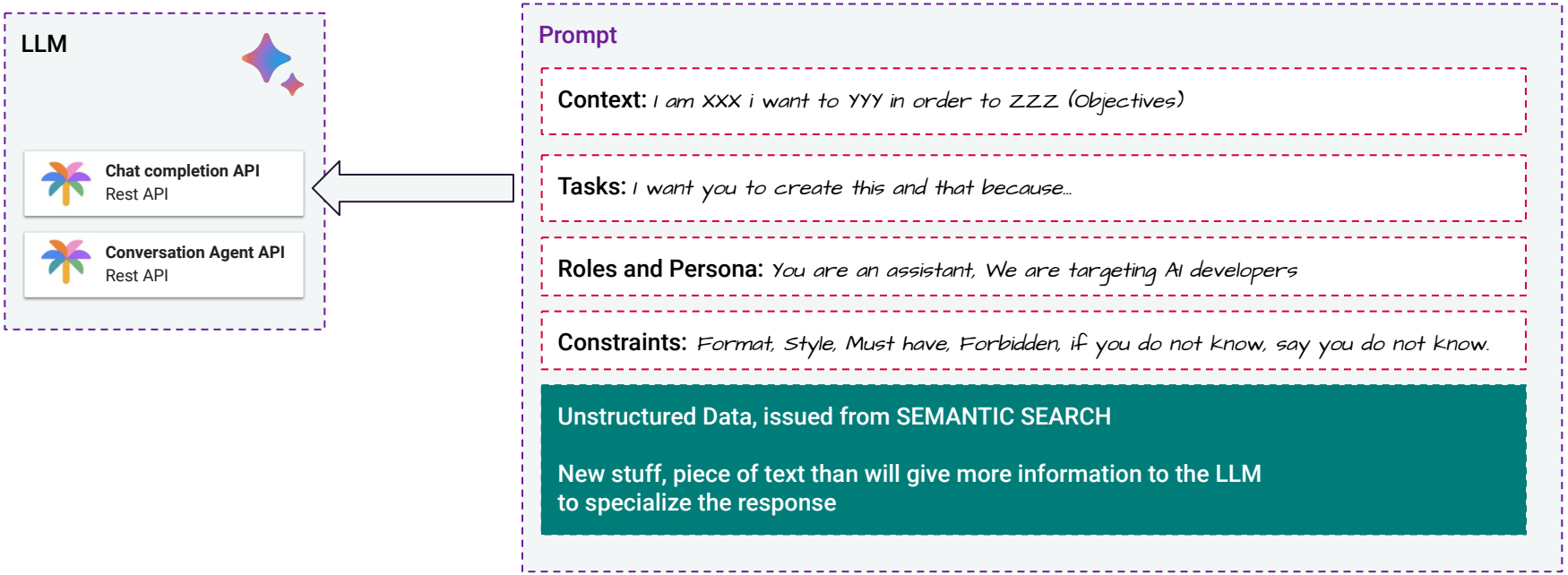
RAG: **Prompt Engineering**



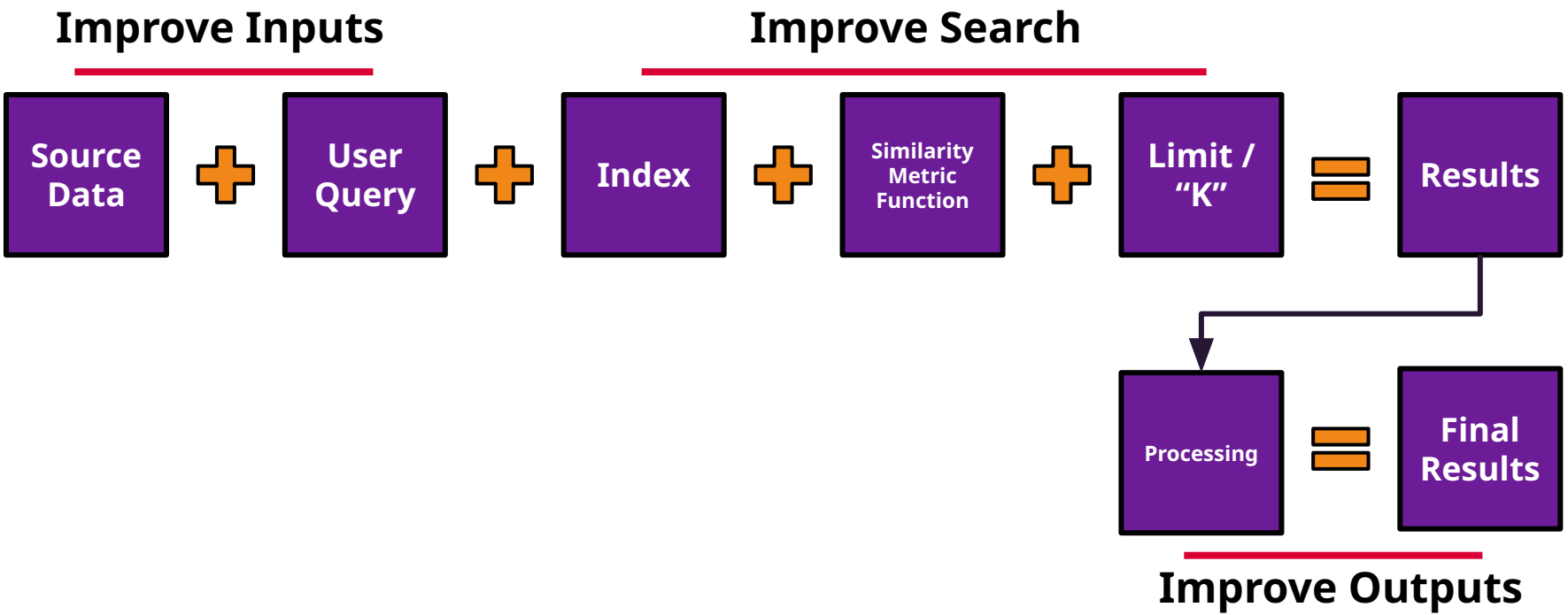
Prompt Engineering



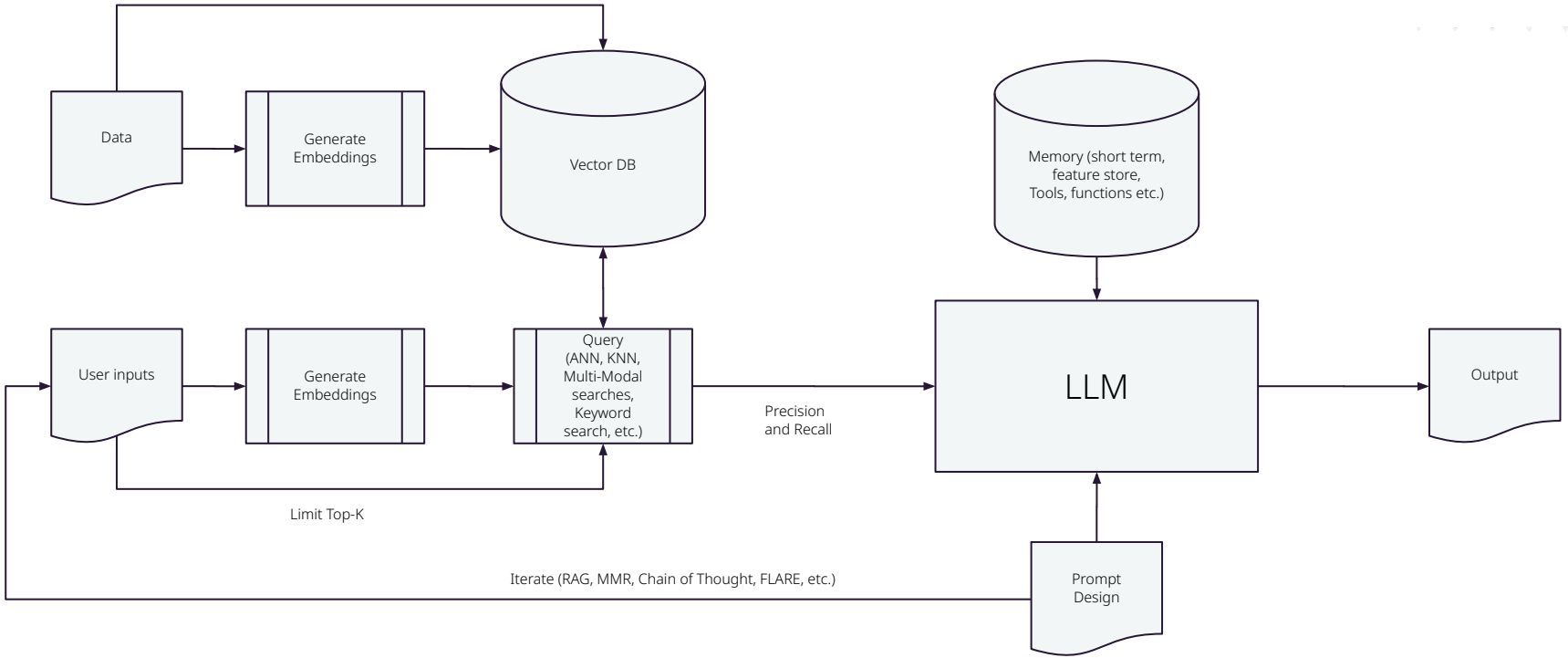
Prompt Engineering



How to improve relevance

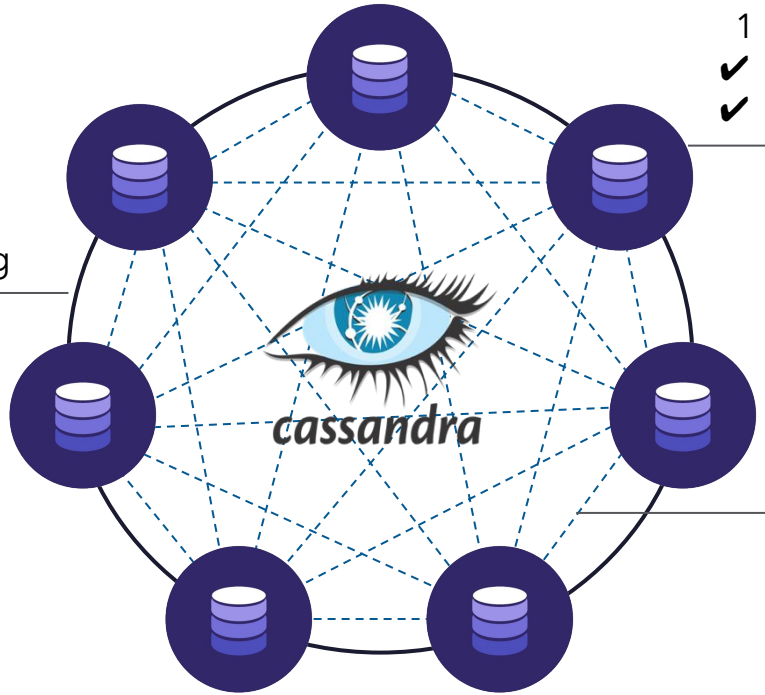


Putting it all together



Apache Cassandra™ as a vector database

› **Nosql** Distributed database



1 Installation = 1 NODE
- ✔ Capacity = ~ 2-4TB
- ✔ Throughput = LOTS Tx/sec/core

DataCenter (DC) | Ring

Communication:
- ✔ Gossiping
- ✔ No Master (peer-to-peer)

Apache Cassandra®

Undisputed Leader of Scale and Reliability

DATASTAX

Why Astra DB as a vector database?

DataStax Astra DB

- Apache Cassandra® in the cloud!
- https://astra.datastax.com/
- Free tier available!

Fast and Easy

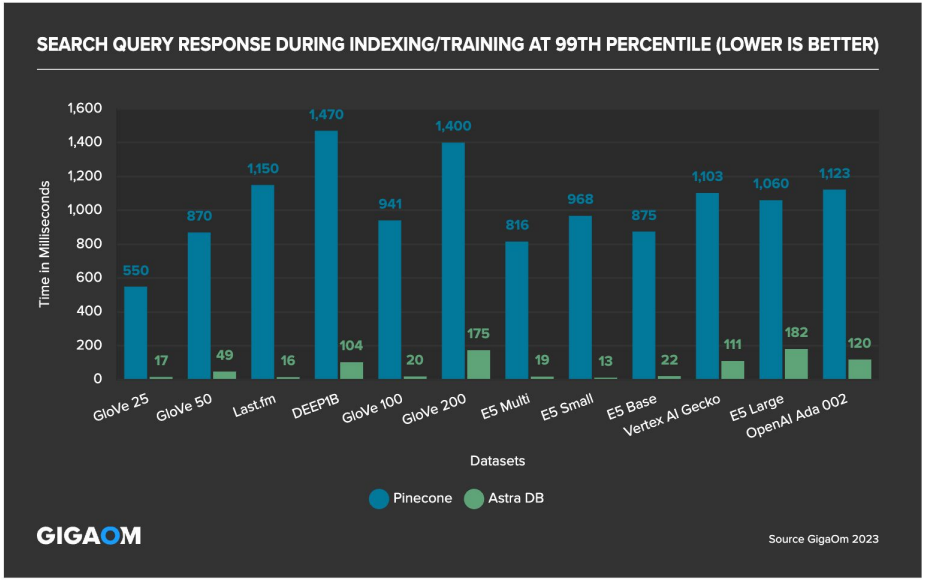
Things you don't have to worry about

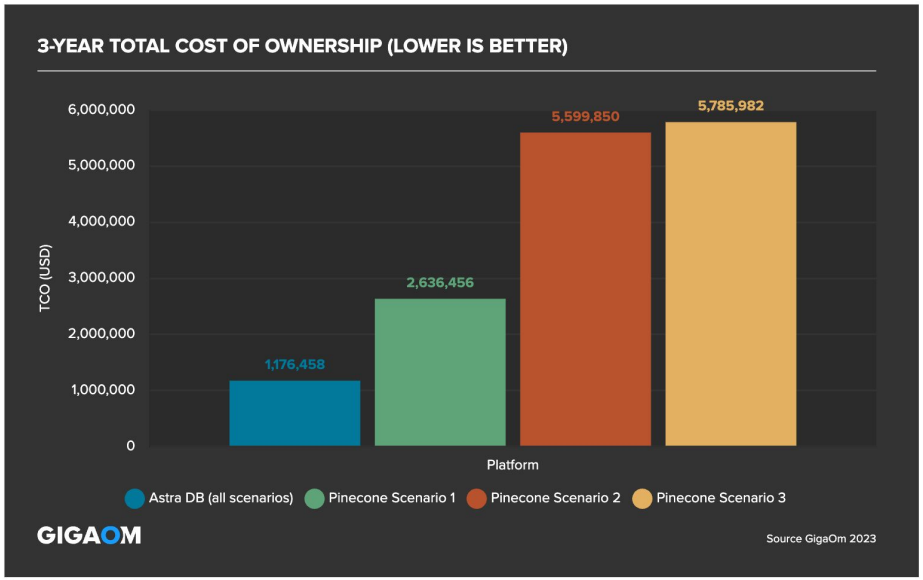
Scale
Cost Efficiency
Security
Reliability
Low latency
Performance

10x Faster than any Vector DB

Scales to Billions of Rows

Industry Leading Performance & TCO

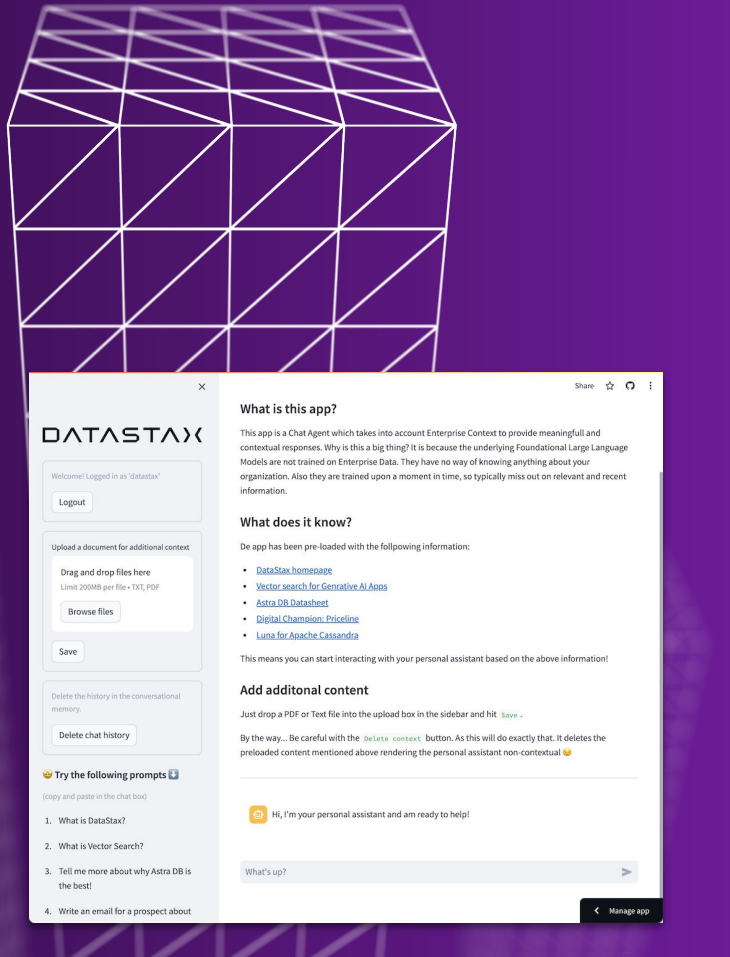




...For AI Apps & Services

Full report: https://gigaom.com/report/benchmarked-vector-search-databases/

Now let's build your own GenAI application



Prerequisites status check

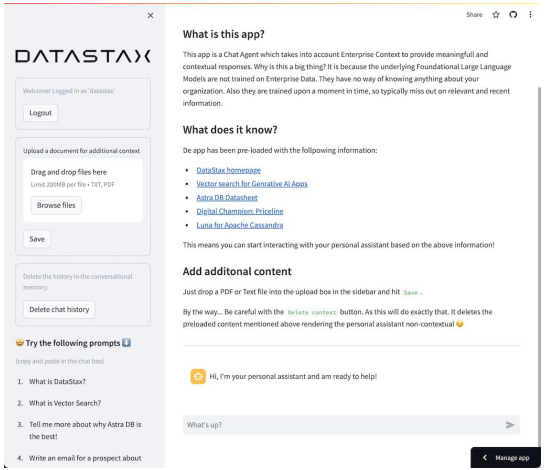
- Go to https://github.com/clun/build-your-own-rag-chatbot

- A Github account
- Google Colab

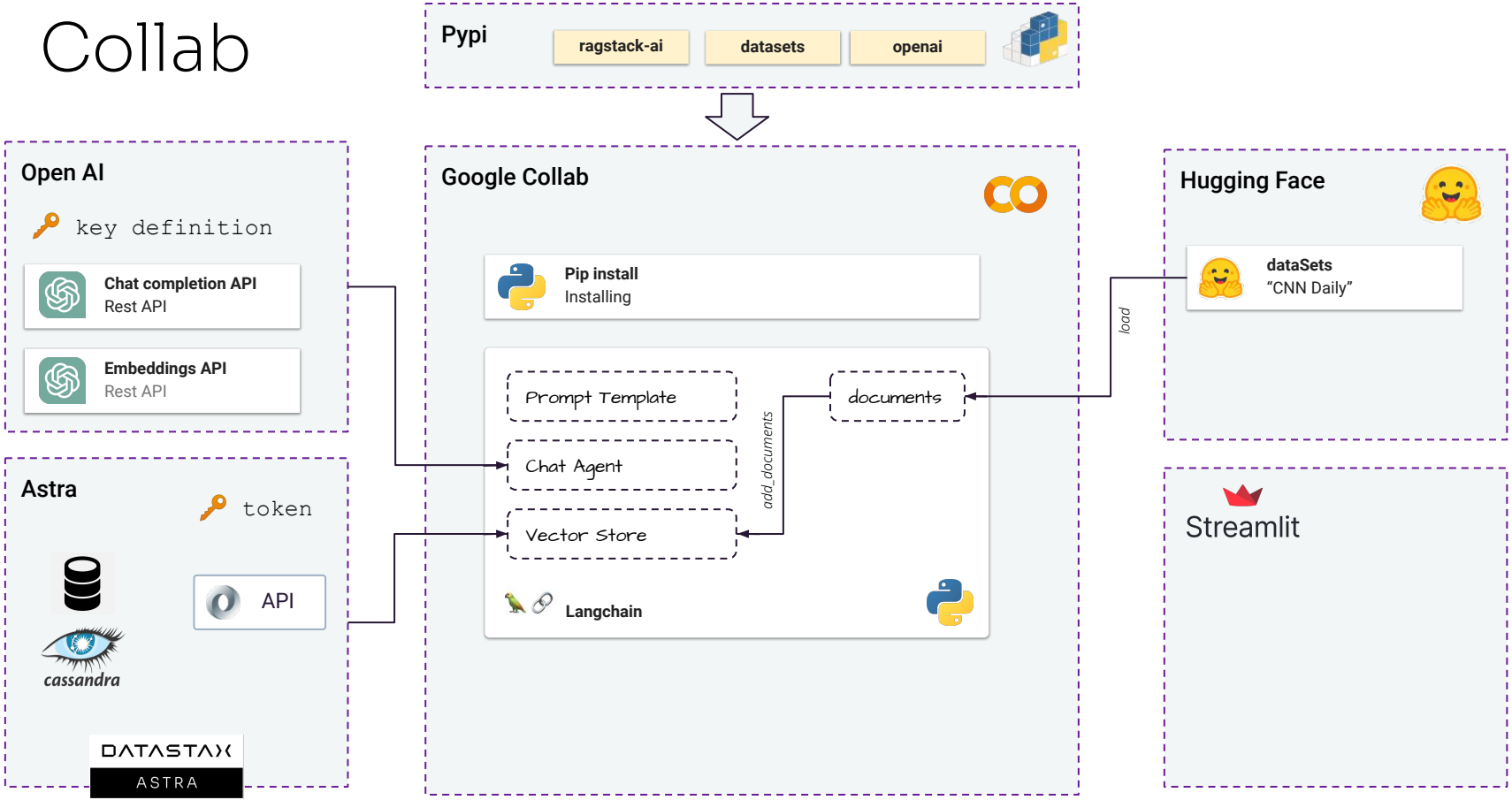
- Accounts created during the session:
  - DataStax Astra DB
  - OpenAI account
  - Streamlit

What you'll learn

- How to leverage DataStax RAGStack for production-ready use of the following components:
  - The Astra DB Vector Store for Semantic Similarity search
  - LangChain for linking OpenAI and Astra DB
- How to use OpenAI's Large Language Models for Q&A style chatbots
- How to use Streamlit to easily deploy your awesome app to the internet for everyone to see!



Collab



LLM Framework

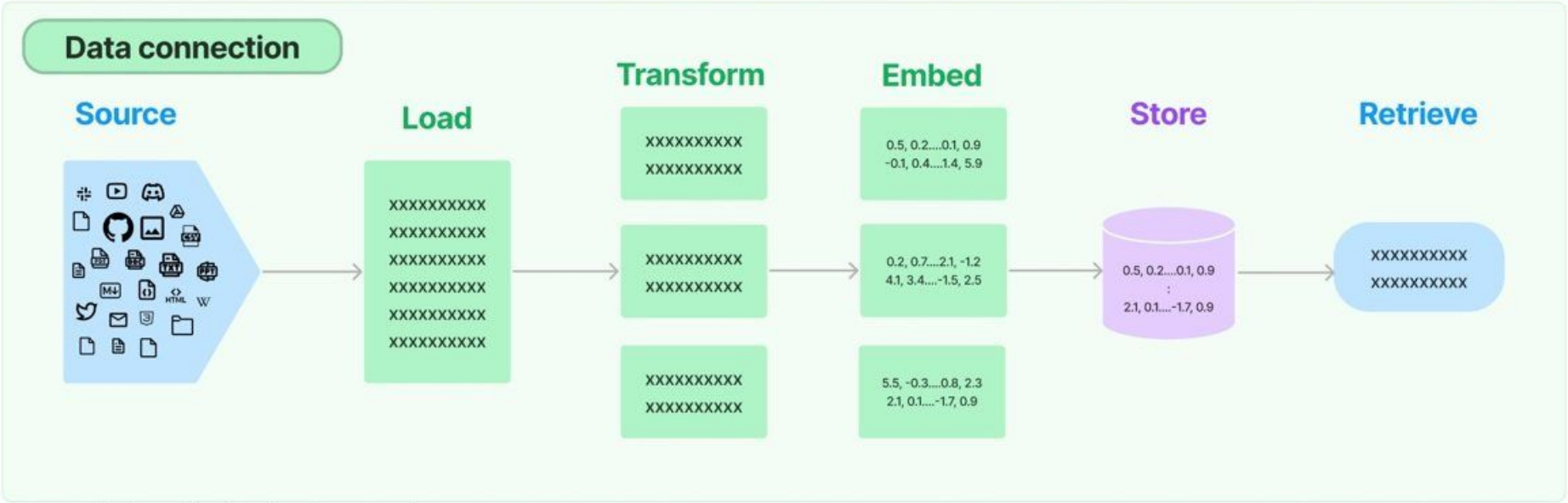
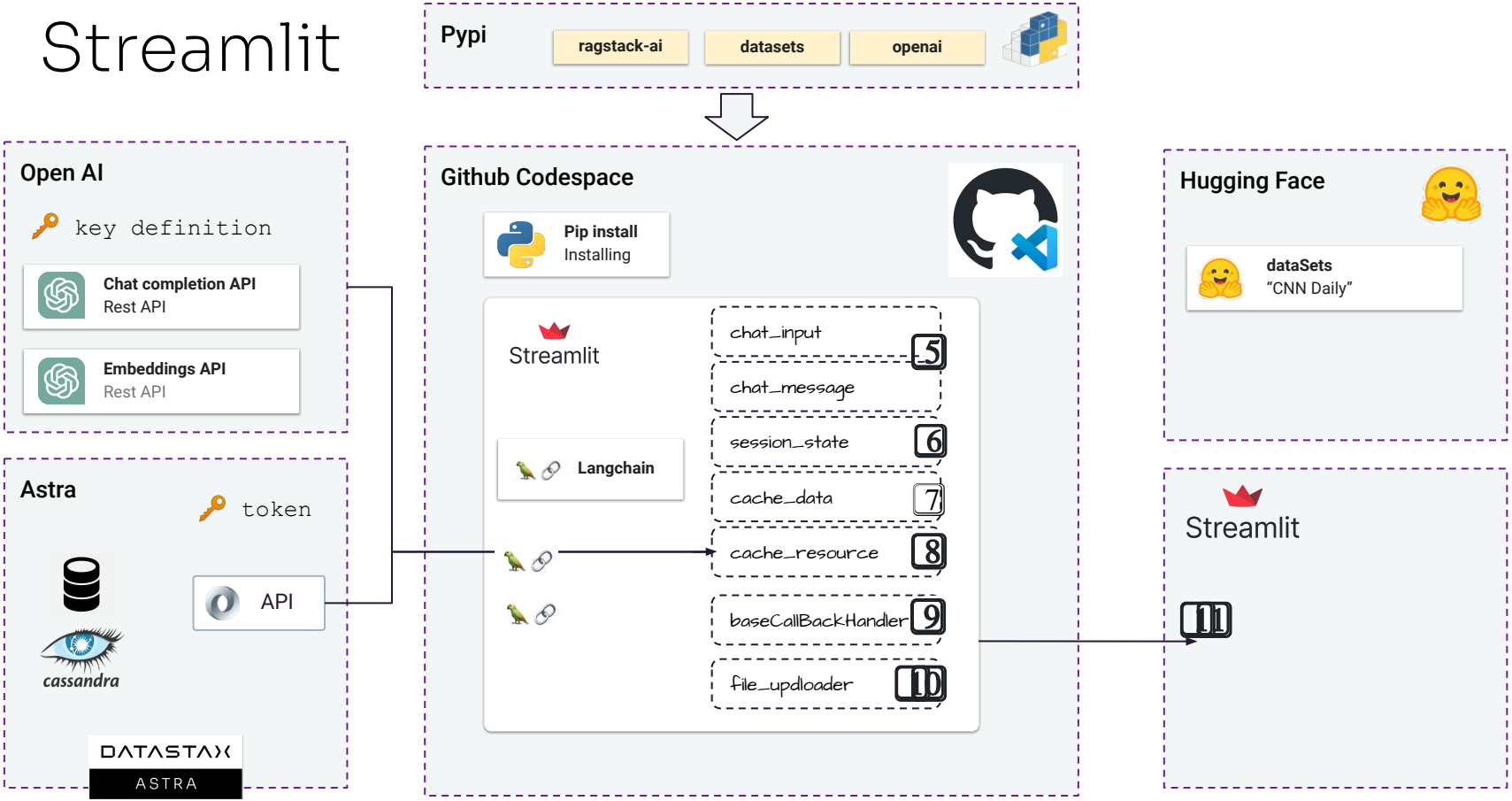


Image Credit: https://python.langchain.com/

Streamlit



DATASTAX

Thank You


DATASTAX