

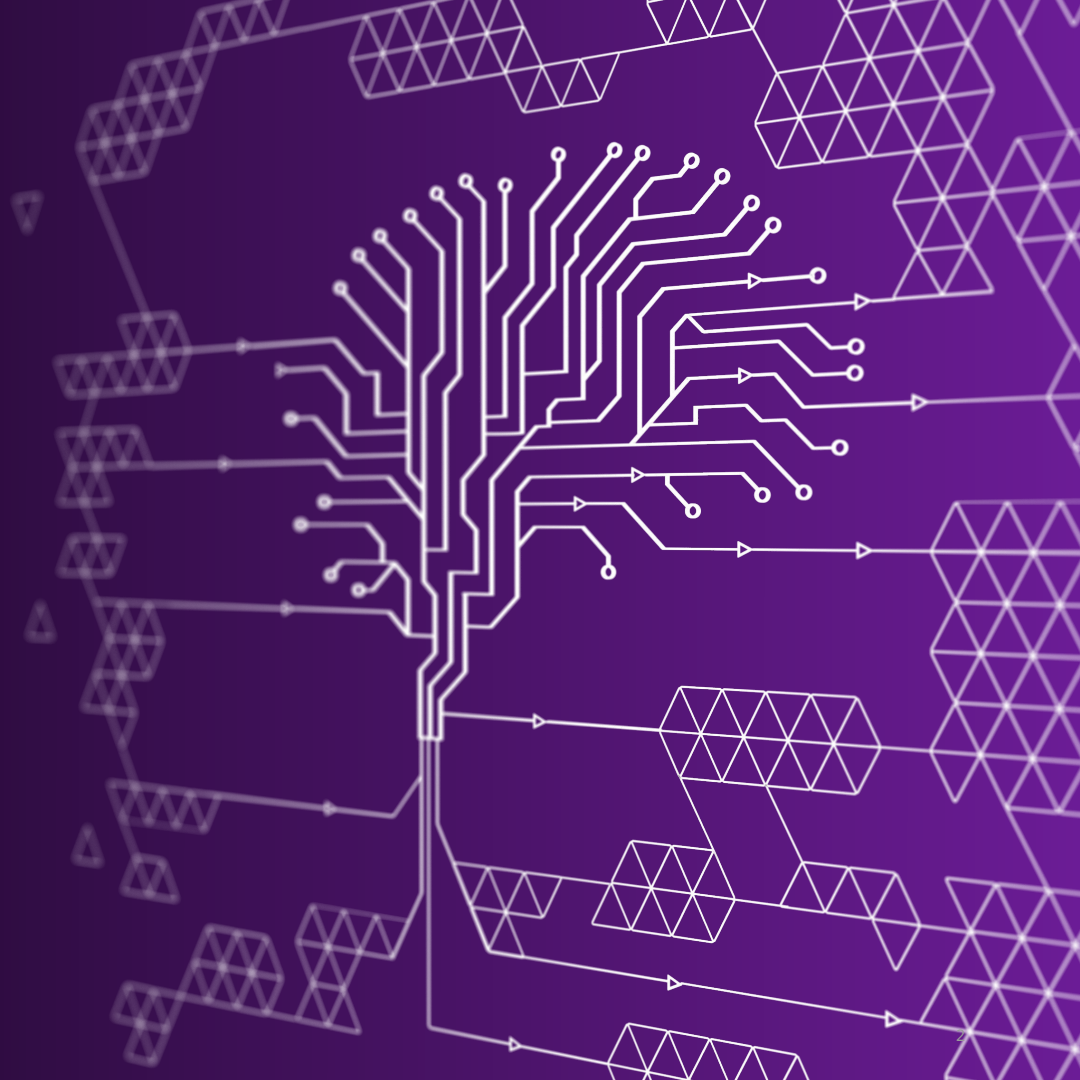
DATASTACK

Building a Petabyte-Scale Vector Store: Powering Future AGI

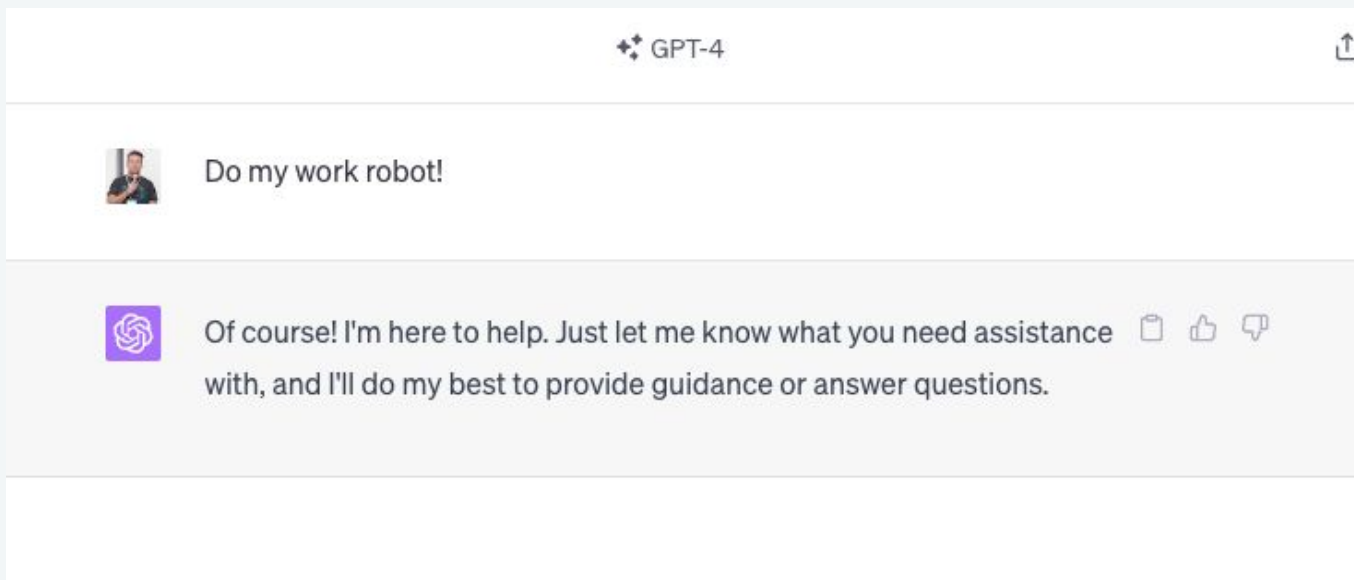
Patrick McFadin, Apache Cassandra Committer, Developer Advocate



Where We Are



› Gen AI is a... ChatBot?



› Agents!

Andrej Karpathy - OpenAI

"I think it's very obvious to a lot of people that AGI will take the form factor of some kind of an AI agent.

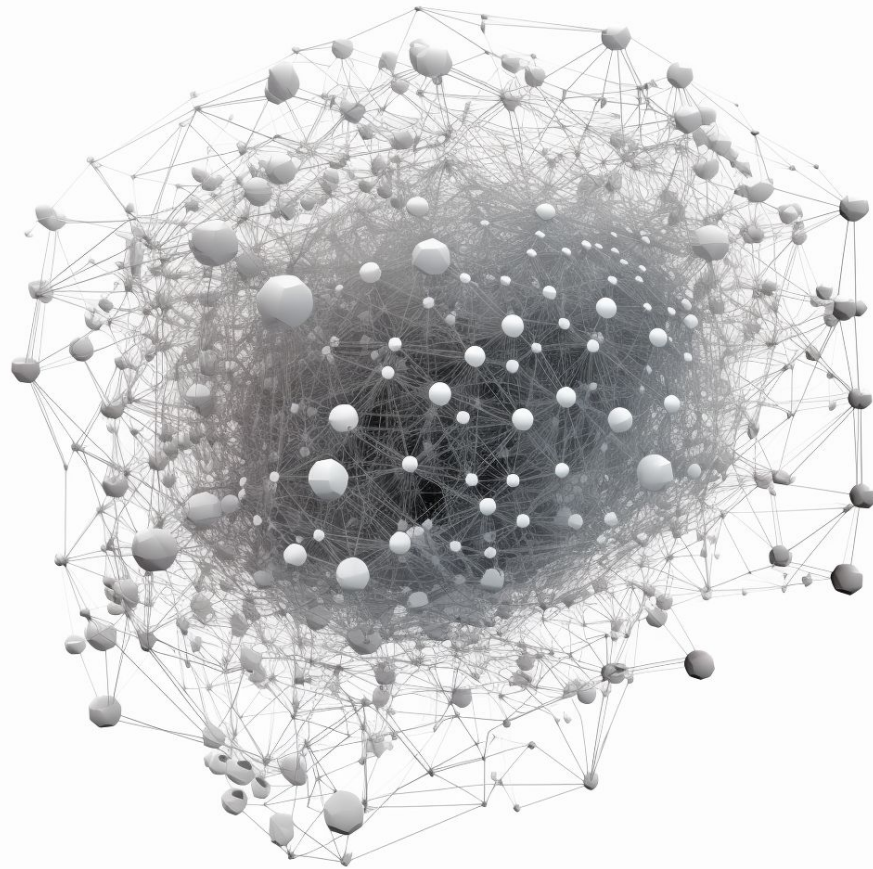
It's not just going to be a single agent thing.

There's going to be many agents."

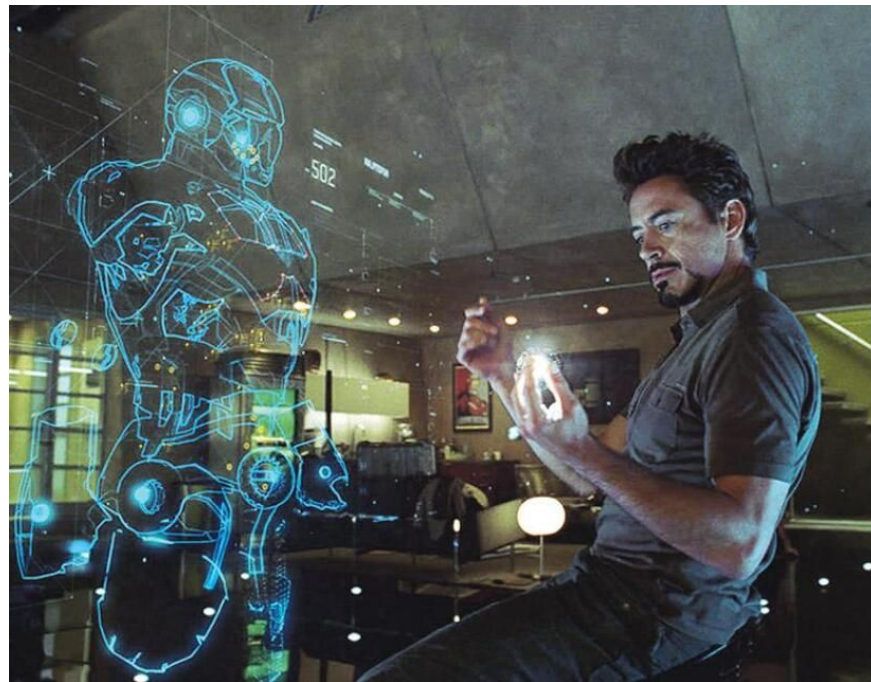
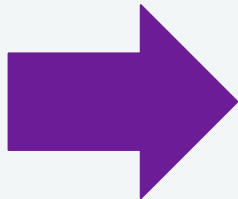


» Large Scale Agent Networks

- Loosely connected
- Dynamic
- Require shared context



➤ The Real Use Case



➤ Vector Databases



› Real talk about LLMs

Great at communication and reasoning

Terrible about knowing stuff



› Context

“Find out if my kids are going to be home for dinner and if not, order doordash from that Indian place. Otherwise, just order a couple of our normal pizzas”

› Context

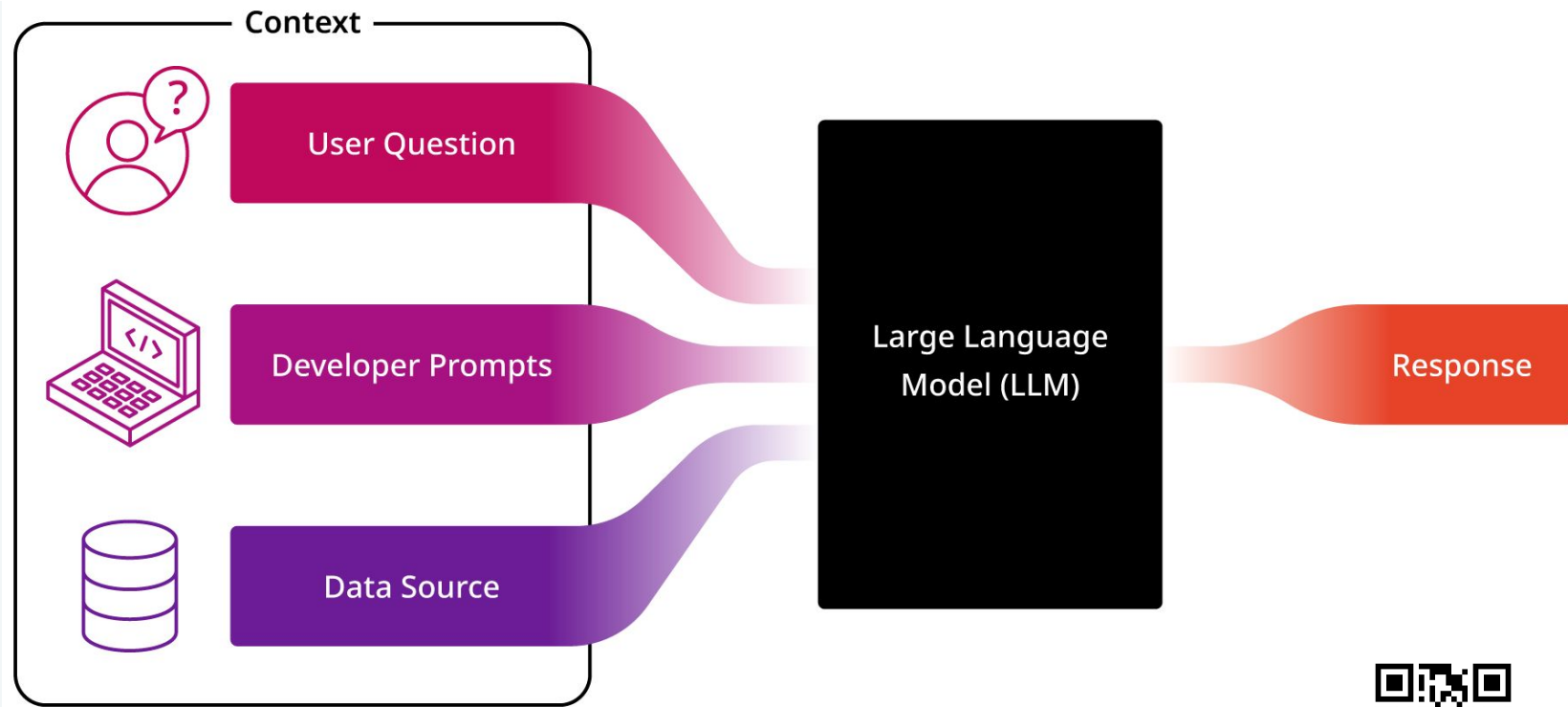
“Find out if **my kids** are going to be home for dinner and if not, order doordash from **that Indian place**. Otherwise, just order a couple of our **normal pizzas**”

ChatGPT (GPT4) can remember around 10000 words

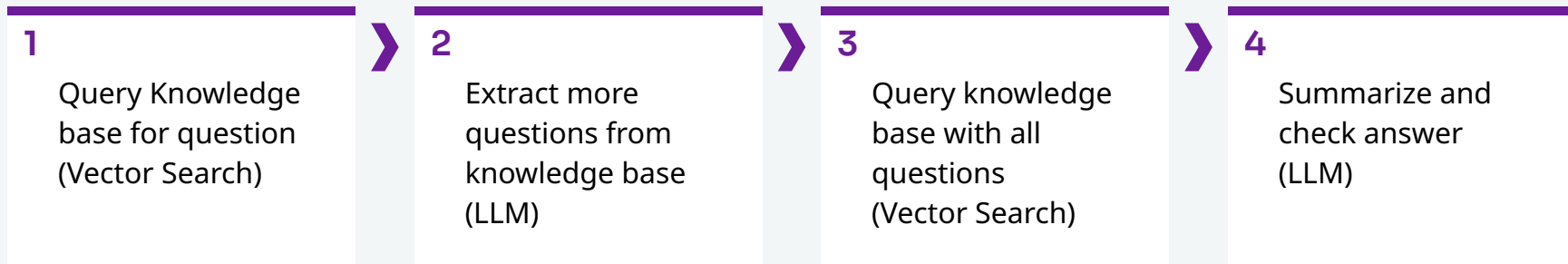
› Multi-Turn Autonomous Agents



› Retrieval-Augmented Generation



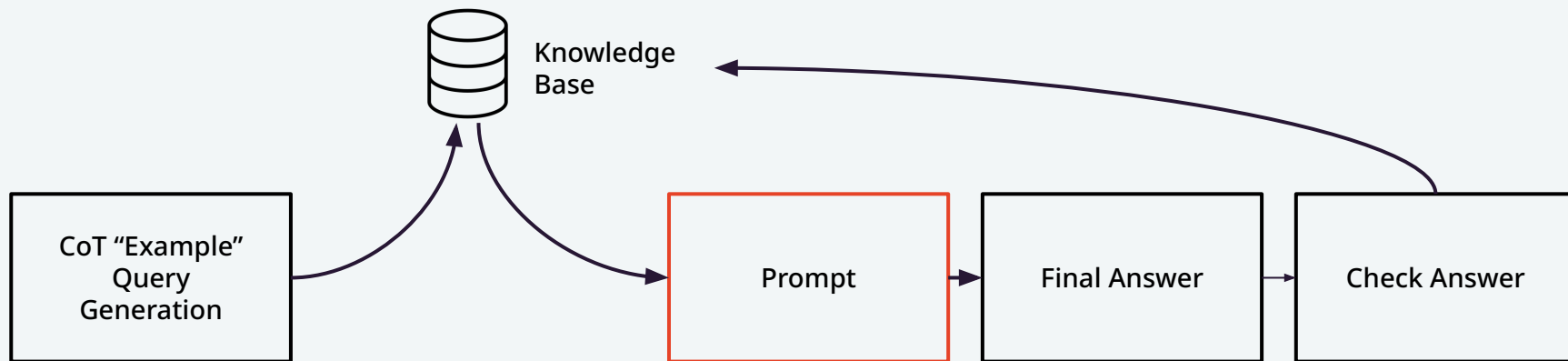
Forward Active Retrieval Generation



Repeat 1–4 until answer looks good



› Reasoning Workflow





› Multimodal Memory

- Models are becoming multi-modal
- Humans are multi-modal
- Modalities:
 - Text
 - Images
 - Sound

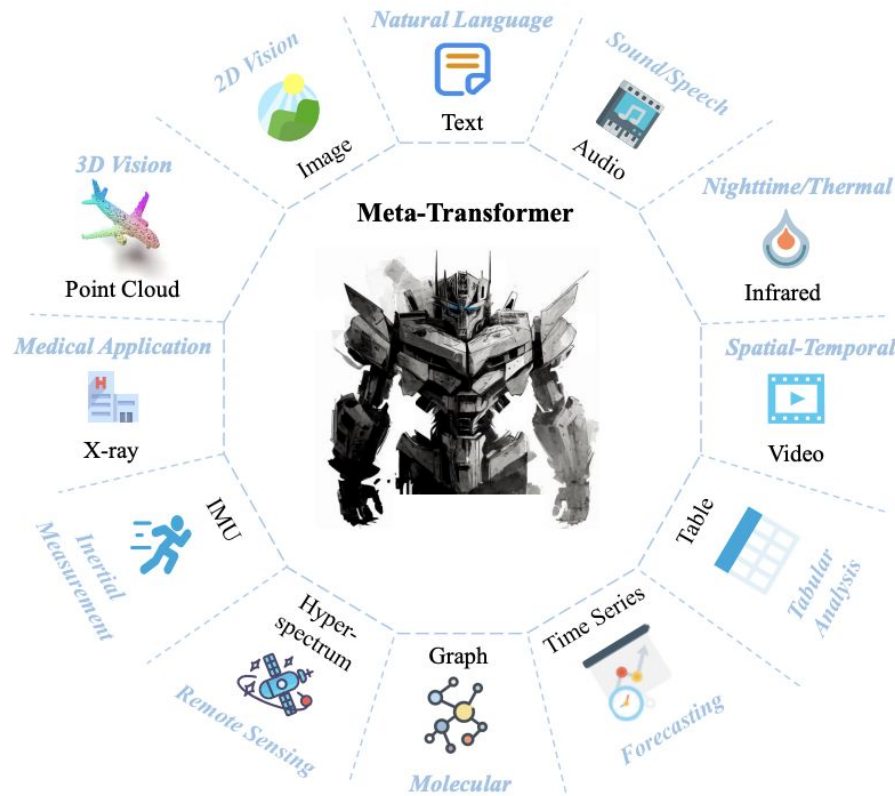
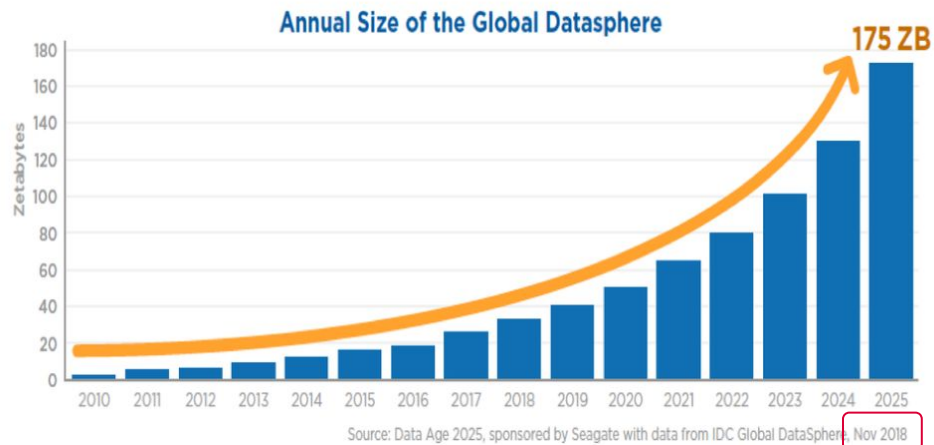


Figure 1: **Unified Multimodal Learning**. Meta-Transformer utilizes the same backbone to encode natural language, image, point cloud, audio, video, infrared, hyperspectral, X-ray, time-series, tabular, Inertial Measurement Unit (IMU), and graph data. It reveals the potential of transformer architectures for unified multi-modal intelligence.



More Data!



November 2018

› Specialty Vector Databases

- A feature trying to be a database
- No experience with scale
- Will be distracted

vector search is an undifferentiated feature



➤ Hello Cassandra



Apache Cassandra®

Undisputed Leader of Scale and Reliability

Apache Cassandra at Apple
Scale and Scope

- Over three hundred thousand instances
- Hundreds of petabytes of data
- Over two petabytes per cluster
- Millions of queries per second
- Thousands of clusters
- Thousands of applications

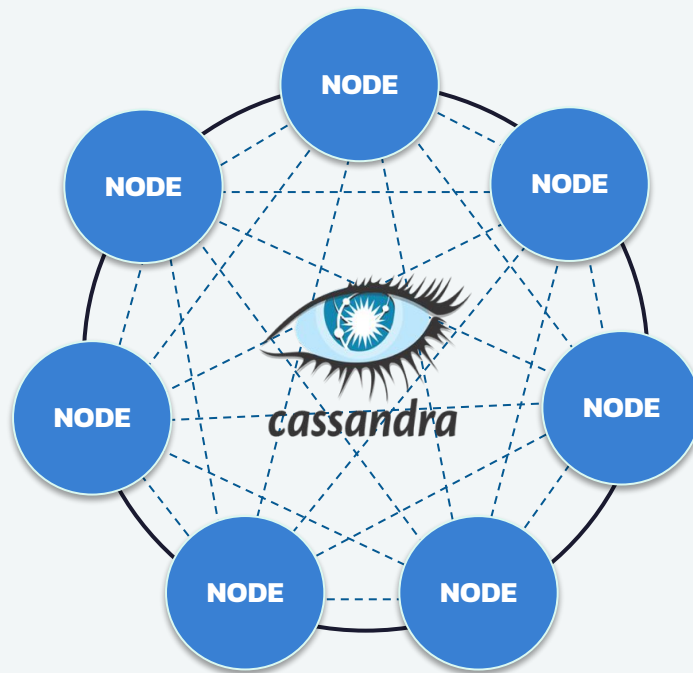
Instances Storage Density

QPS Clusters Applications

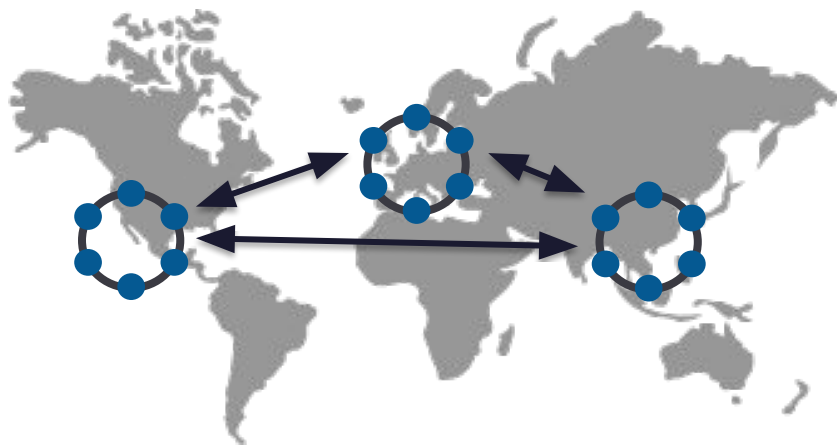
APACHECON

» The Invincible Brain

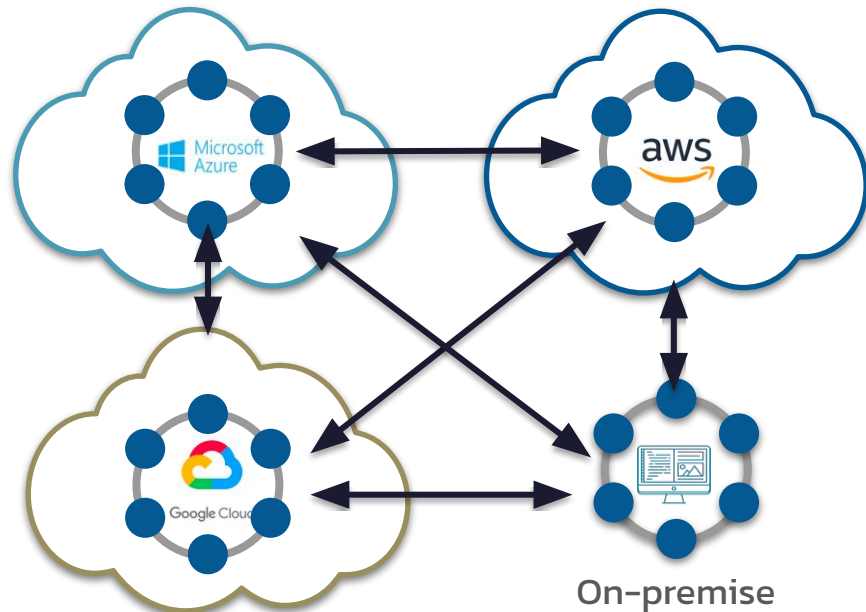
- Shared nothing
- Linear scale
- Fully replicated

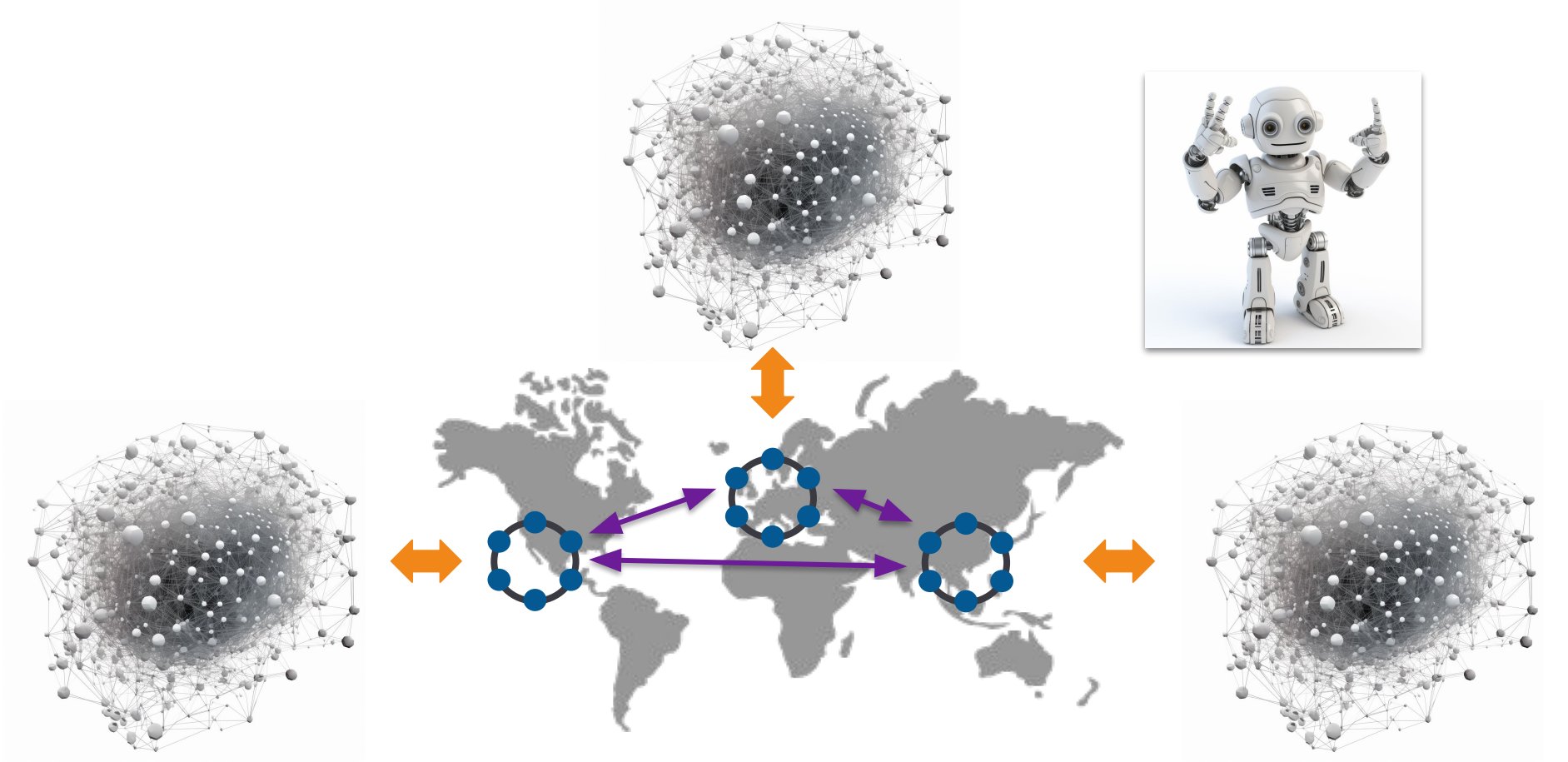


Global



Multi-Cloud



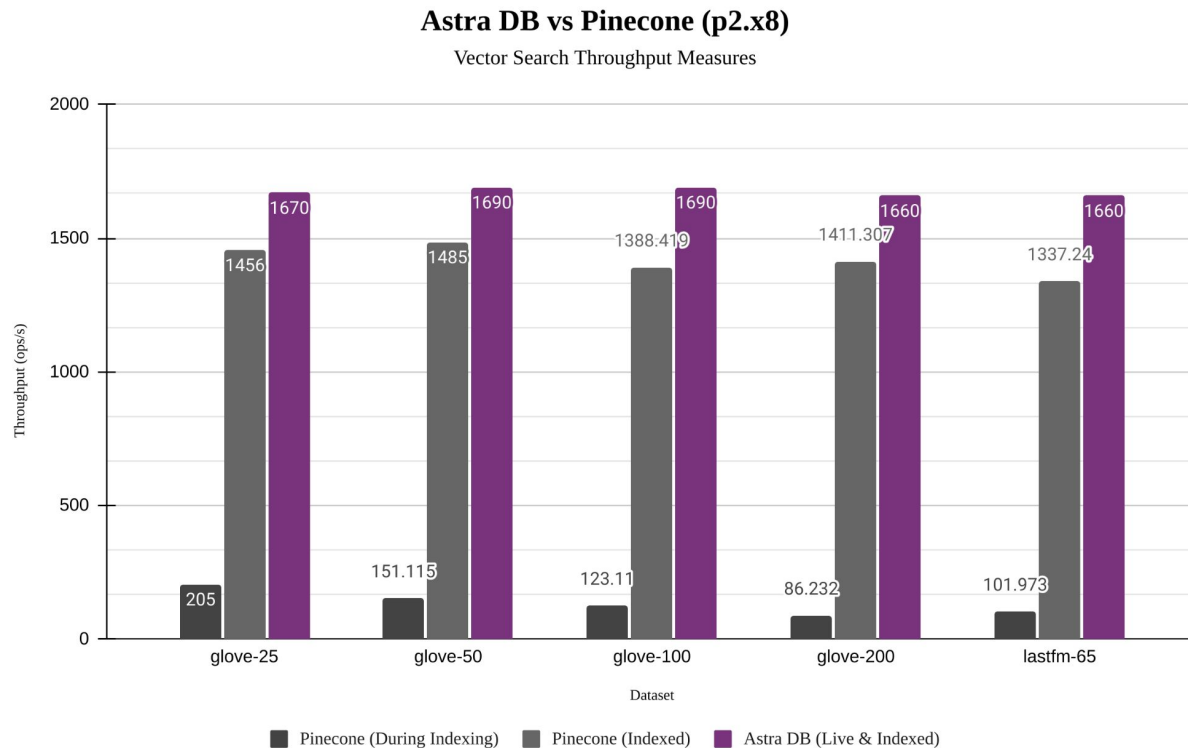


› 5 Hard Problems We're Solving

- **Scale-Out Capabilities:** No upper limits
- **Garbage Collection:** Pruning obsolete index information
- **Effective Use of Disk:** Enabling high throughput
- **Composability:** Predicates, term-based searches. Aka Hybrid Search
- **Concurrency:** Non-blocking, multi-threaded index construction

<https://thenewstack.io/5-hard-problems-in-vector-search-and-how-cassandra-solves-them/>

» Concurrency is Hard



» New Data Model

```
CREATE TABLE IF NOT EXISTS vsearch.products (  
  id int PRIMARY KEY,  
  name TEXT,  
  description TEXT,  
  item_vector VECTOR<FLOAT, 5> //5-dimensional embedding  
);
```


› Creating a Vector Search Index

```
CREATE CUSTOM INDEX IF NOT EXISTS ann_index  
ON vsearch.products(item_vector)  
USING 'StorageAttachedIndex';
```

› Searching for Neighbors

```
SELECT * FROM vsearch.products  
ORDER BY item_vector ANN OF [0.15, 0.1, 0.1, 0.35, 0.55]  
LIMIT 1;
```

id	description	item_vector	name
5	A deep learning display that controls your mood	[0.1, 0.05, 0.08, 0.3, 0.6]	Vision Vector Frame

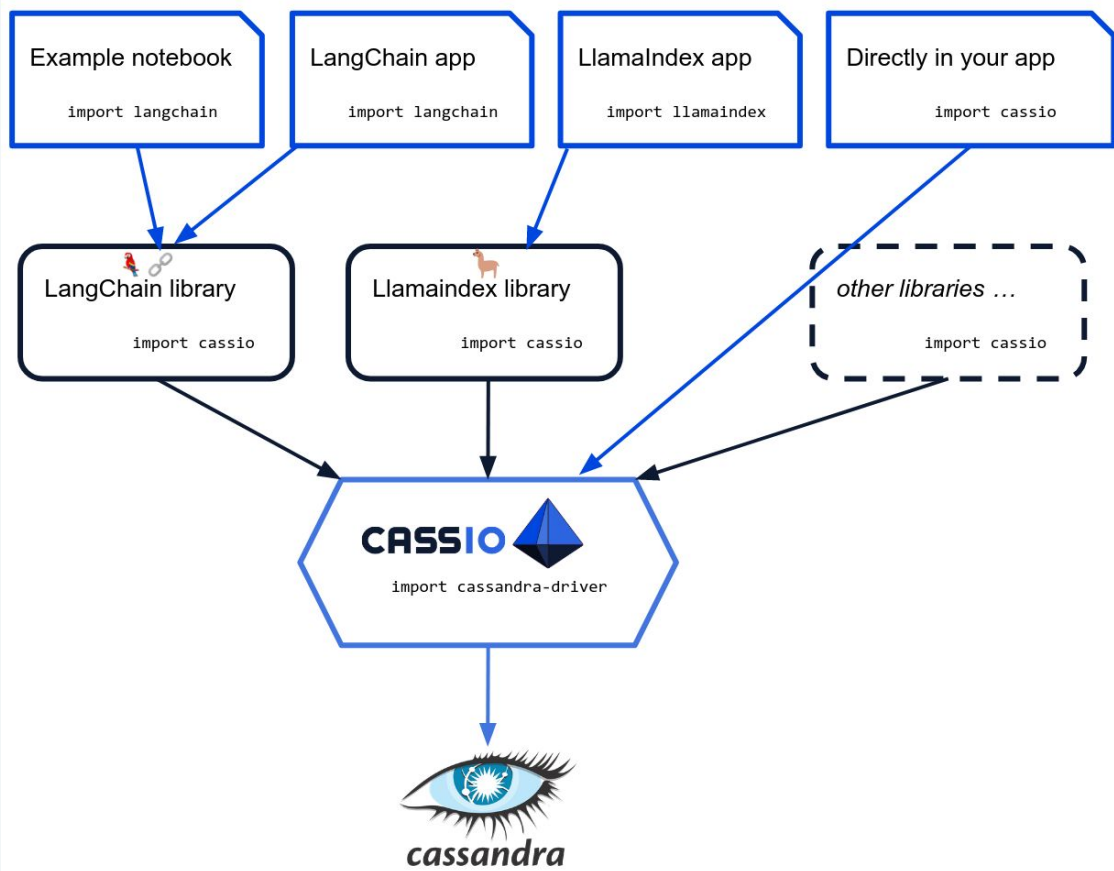


Cassio

Python library for GenAI
and Cassandra



<https://cassio.org/>





Everyone

- **Stressed out**
- **Feeling behind**
- **Boss be all “We need an AI!”**

› Fast and Easy



Things you don't have to worry about

Scale

Security

Wasting Money

Reliability

» Easy to Get Started



Welcome to Astra, Patrick 🙌

Accelerate your workflows, access recently visited resources, and explore Astra's integrations and documentation!

Vector Search is now production ready 🚀

Supercharge your AI Applications and Agents with Vector Search. Create a database or explore the quickstart or [example demos](#).

Create Database

Read the Blog [↗](#)



Guides & Examples

Q&A Search with LangChain ⚡

Perform a text similarity search on HuggingFace datasets using Astra DB, LangChain and CassIO.

View Quickstart [↗](#)

Retrieval Augmented Generation (RAG) for AI Chatbots 🗣️

Execute a vector similarity search to add supplemental context to an LLM call.

View Example [↗](#)





Vector Search using the JSON API 🧠

Perform a text similarity search against a movie dataset using Astra DB and the Mongoose powered JSON API for JavaScript.

View Example [↗](#)

» Easy to Grow

Examples

Example	Description	Topic
Retrieval Augmented Generation for AI Chatbots 	RAG performs a vector similarity search to add supplemental context to a LLM.	Chatbot Large Language Model Python RAG
Image Search with the Contrastive Language Image Pre-training 	Generate image embeddings using the CLIP model, store them in Astra DB, and utilize Astra Vector Search to find and display a particular image from a set of images.	CLIP Image Embeddings Python
Query vector data with CQL	To query data using Vector Search, use a SELECT query as shown in the documentation.	Cassandra Query Language (CQL) Code examples
Philosophy quote finder & generator with Vector Search & RAG using CQL  or Using CassIO 	Use OpenAI's vector embeddings and DataStax Astra DB as the vector store for data persistence.	Cassandra Query Language (CQL) RAG Vector Search Large Language Model CassIO
Quick start with LangStream to build chatbot	Use LangStream, streaming AI agents, and OpenAI to build streaming GenAI apps faster.	LangStream Vector Search

docs.datastax.com

› Do this today!

<http://astra.datastax.com>



Featured



Try Astra with Vector Search

Available in Public Preview, Astra now includes the ability to create a database with Vector Search capabilities. Try it out on your next generative AI project.



Create Database

or

Interactive Guide

Use your business email address and get from \$1000- \$3000 in free credits and consulting with your subscription.

DATASTAX