Microsoft

Microsoft Java Developer Conference 2024

Code. Cloud. Community.

Enter the Brave New World of GenAI with Vector Search

Mary Grygleski
Senior Developer Advocate
𝕏 @mgrygles

March 2024

Agenda

- Who's Mary?
- A Brief Background of AI
- Understanding the "Players" in the GenAI Era
- The Generative Pre-Trained Transformers (GPTs)
- Natural Language Processing (NLP)
- Large Language Models (LLMs)
    - Prompt Engineering
- Vector DB - Vector Search
    - Vector Embeddings
    - Approximate Nearest Neighbor (ANN)
- *A Quick Demo (if time permits)*
- Benefits
- Challenges
    - Hallucinations
    - Ethical issues
    - Real-time / RAG
- Resources

Who is Mary?

Passionate
Advocate

Java Champion

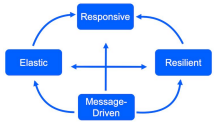
mgrygles

mary-grygleski

mgrygles

mgrygles

PULSAR



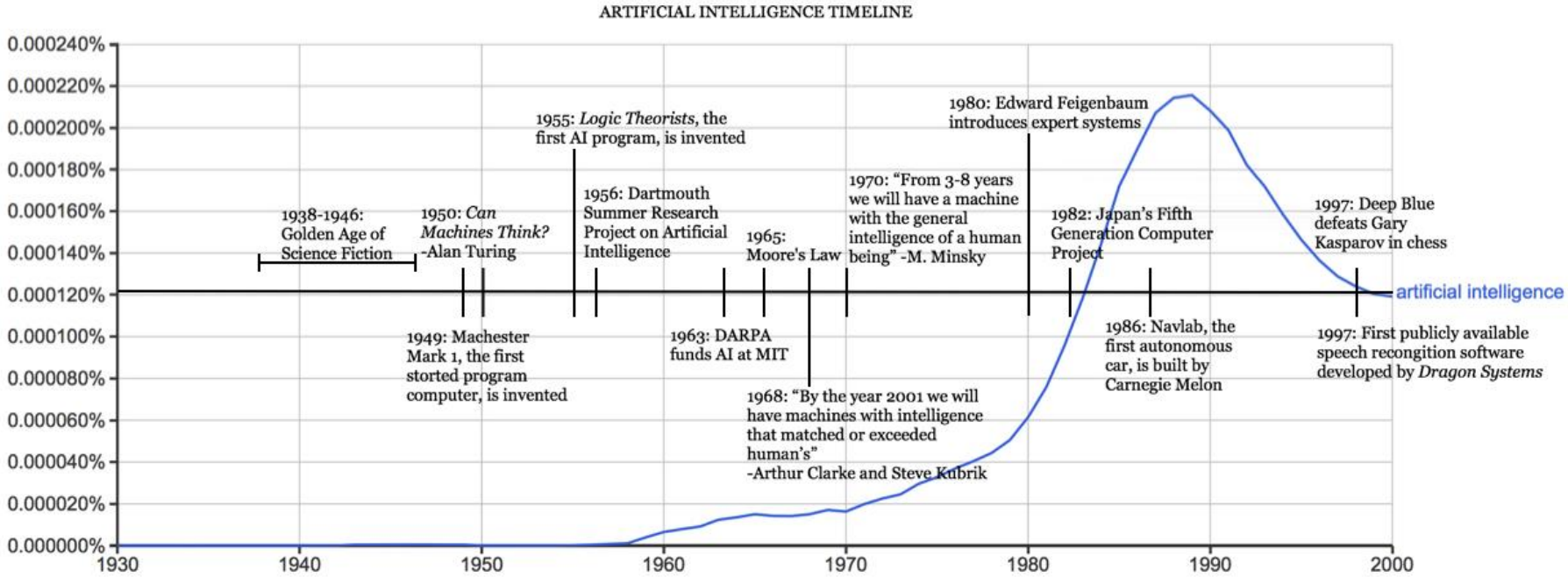
Open
Liberty

- Streaming
- Distributed Systems
- Reactive Systems
- IoT/MQTT
- Real-Time AI/ML

Senior Developer Advocate

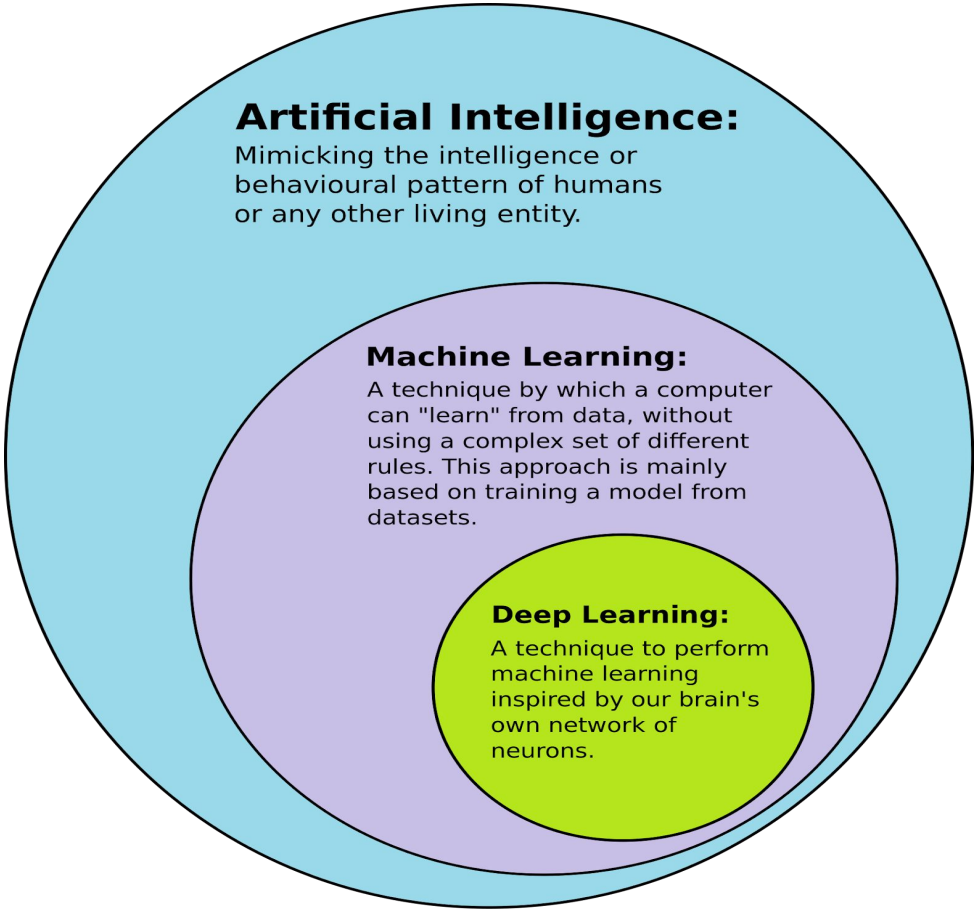
A Brief Background of AI

Near “old times” – 20th century



https://i0.wp.com/sitn.hms.harvard.edu/wp-content/uploads/2017/08/Anyoha-SITN-Figure-2-AI-timeline-2.jpg

It’s all about
AUTOMATION


Artificial Intelligence:
Mimicking the intelligence or behavioural pattern of humans or any other living entity.
Machine Learning:
A technique by which a computer can "learn" from data, without using a complex set of different rules. This approach is mainly based on training a model from datasets.
Deep Learning:
A technique to perform machine learning inspired by our brain's own network of neurons.

A fascinating look at the GenAI “era”

What is Generative AI (GenAI) ?

- A “disruptive” field in AI
- Has the potential to change the way we create and consume content
- Generate new contents based on prompts
- Uses a combination of machine learning and deep learning to produce contents
- Tends to be on the “creative” side: generating code, writing an article, designing new fashion, composing a new song... especially when compared with Predictive AI which tends to be more strictly about business, marketing, and weather forecasting.

Since the new millennium (2000)...

2003: Yoshua Bengio and his team develop the first **feed-forward neural network** language model

2011: Apple brings AI and NLP assistants to the masses by releasing its **first iPhone with Siri**.

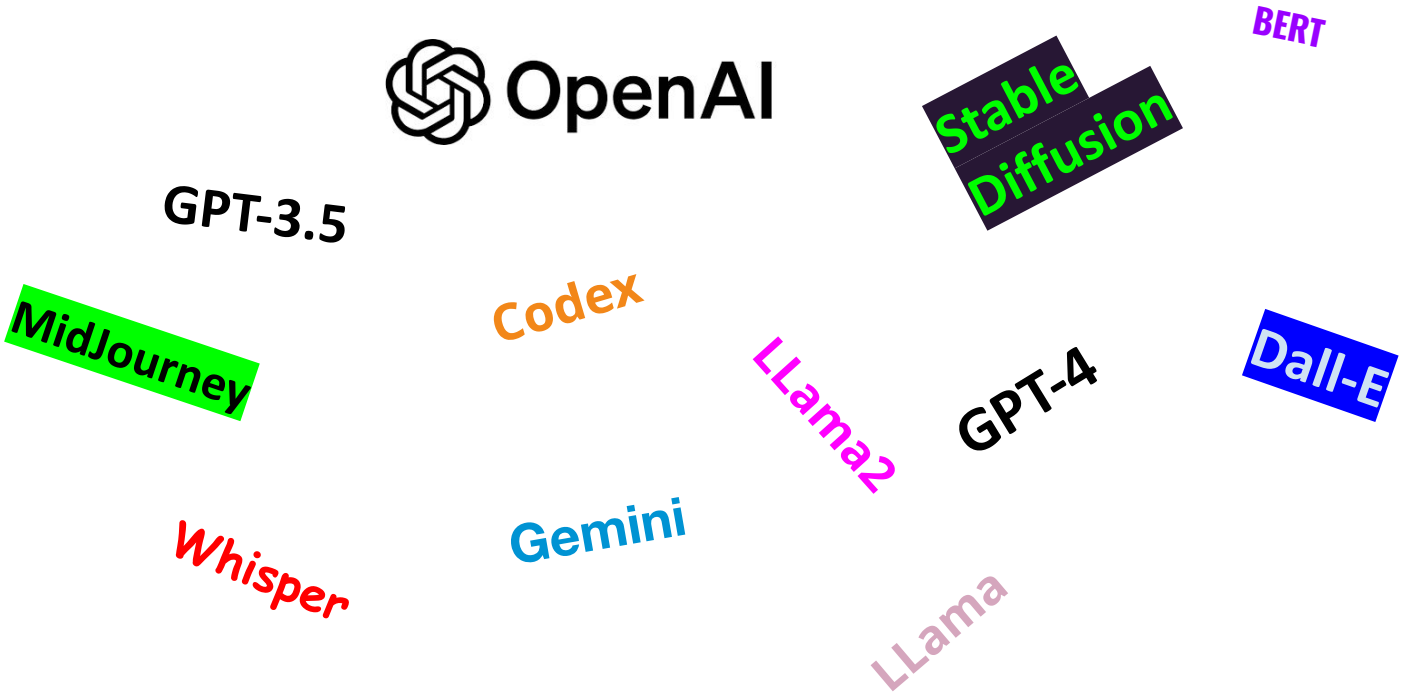
2013: A group of Google researchers led by Tomas Mikolov create **Word2vec**, a technique for natural language processing that uses a neural network to learn word associations from a large set of text

2017: A team of Google researchers led by Ashish Vaswani propose a new simple network architecture, **the Transformer**

…

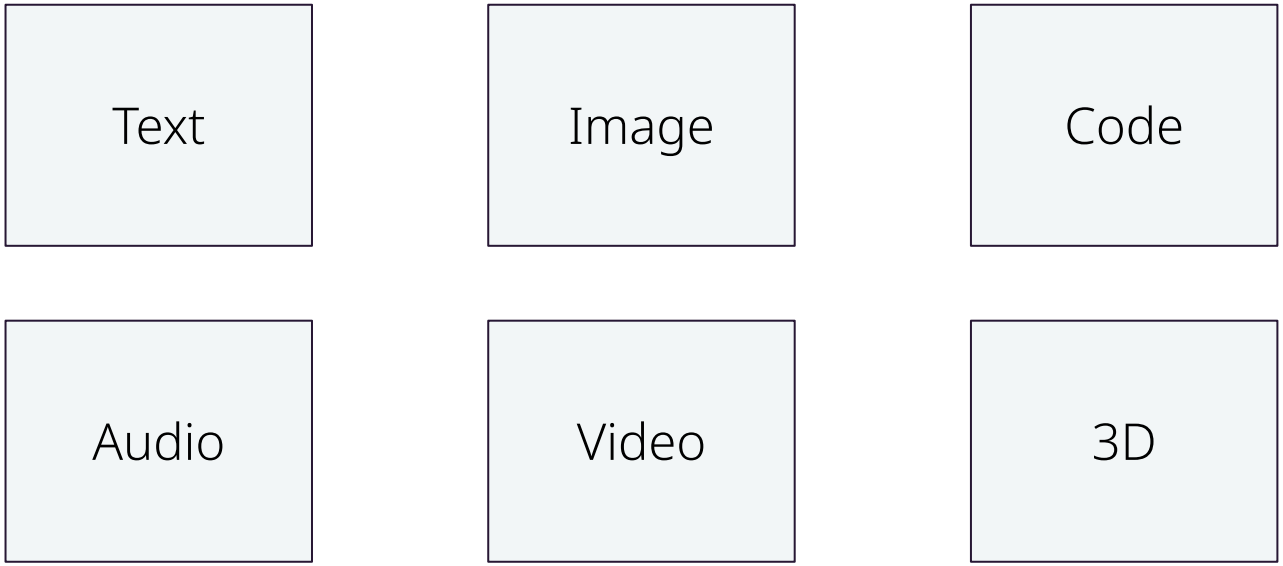
Understanding the New “Players” in the GenAI Era

Generative Models

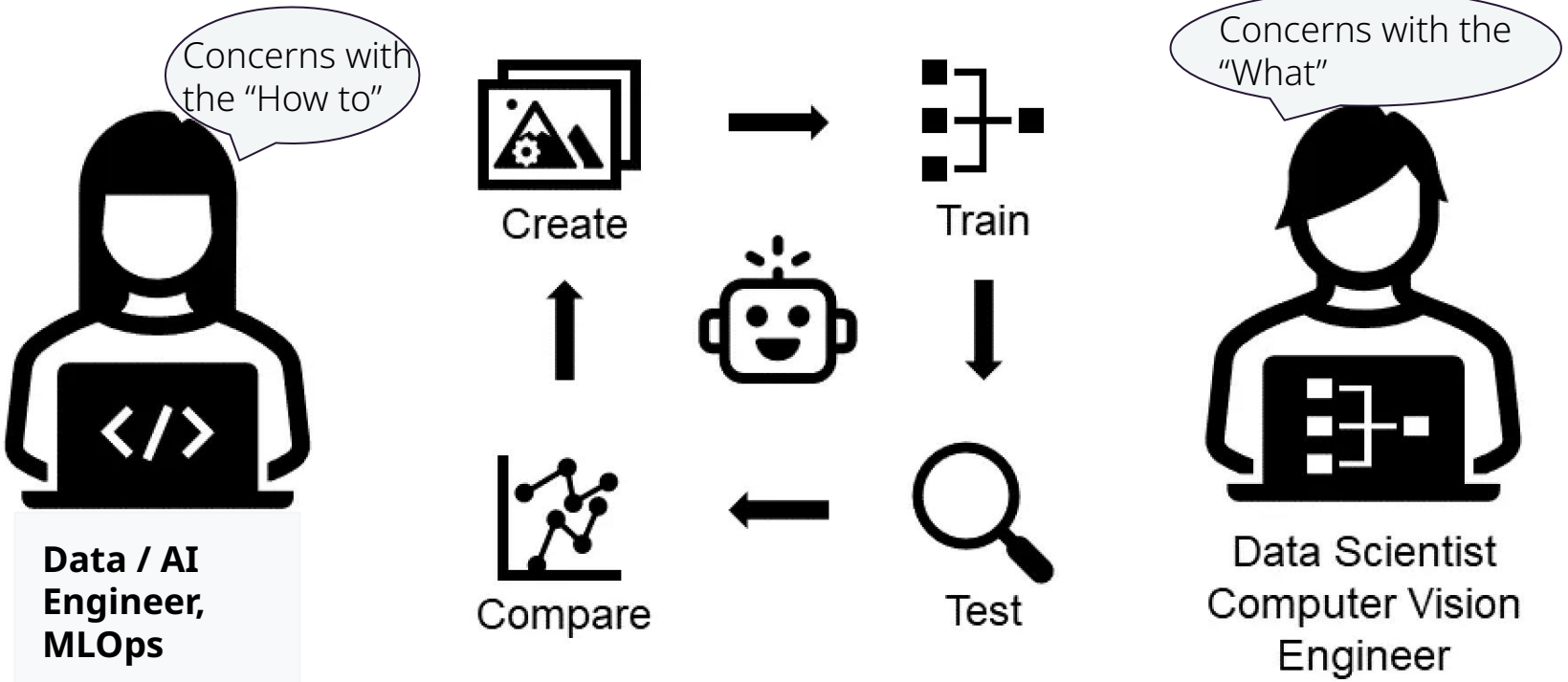


Generative Apps

Expanding Modality / Multi-Modal



What about the People players ?



The Generative Pre-trained Transformer (GPT)

What is Generative Pre-Trained Transformer (GPT) ?

- Takes simple prompts (in natural human language) as input
- Pattern matching (also commonly being called as “search”)
- Answers questions for the **prompts**
- Produce contents such as: a new essay, a blog post, a new computer program

GPT started showing up around 2018

2018: Alec Radford's paper on **generative pre-training (GPT)** of a language model is republished on OpenAI's website, showing how a generative language model can acquire knowledge and process dependencies unsupervised based on pre-training on a large and diverse set of data

2019: OpenAI releases the complete version of its **GPT-2 language model**, which was trained on a dataset of more than nine mission documents — including text from URLs shared in Reddit posts with at least three upvotes.

GPT started showing up around 2018

2018: Alec Radford's paper on **generative pre-training (GPT)** of a language model is republished on OpenAI's website, showing how a generative language model can acquire knowledge and process dependencies unsupervised based on pre-training on a large and diverse set of data

2019: OpenAI releases the complete version of its **GPT-2 language model**, which was trained on a dataset of more than nine mission documents — including text from URLs shared in Reddit posts with at least three upvotes.

2020s: ChatGPT - a fast-growing AI Chatbot

- 2022: Startup company Stability AI develops Stable Diffusion, a deep learning text-to-image model that generates images based on text descriptions. This leads to the rise of other diffusion-based image services, such as DALL-E and Midjourney.

- 2022: ChatGPT releases GPT-3.5, an AI tool that reached one million users within five days. The tool can access data from the web from up to 2021.

2023: ChatGPT continues...

- 2023: The generative AI arms race begins. Microsoft integrates ChatGPT technology into Bing, a feature now available to all users. Google releases its own generative AI chatbot, Bard. And OpenAI releases yet another version of their bot, GPT-4, along with a paid “premium” option.

- 2023: OpenAI releases a beta version of its browser extension for ChatGPT (now available to all ChatGPT Plus subscribers), which has potentially unbounded access to current data on the web — something no other generative AI tool currently offers. It also announces the availability of third-party plugins.

Natural Language Processing (NLP)

What is Natural Language Processing (NLP)?

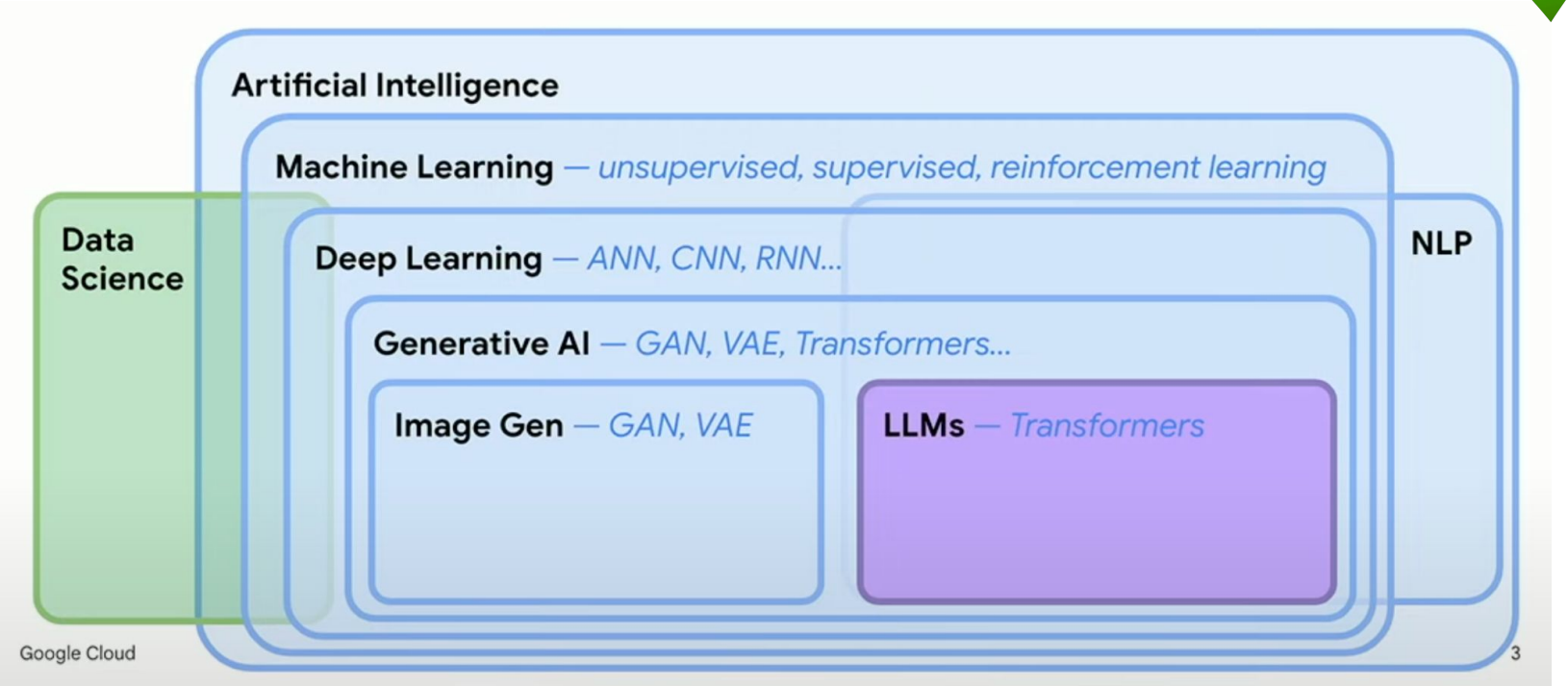
- An interdisciplinary sub-field of linguistics of computer science
- Primary concern is to process natural language datasets (as such text corpora or speech corpora)
- Uses rule-based or probabilistic machine learning approaches
- Enables computer to learn from contents, including the contextual nuances of the language itself
- Ideally to draw insights from the documents

Large Language Models (LLMs)

What is Large Language Models (LLMs)

- A type of Machine Learning Model
- Foundation type of model
- Typically the pre-training consumes a humongous amount of resources: $$$$$$$ GPUs, multiple weeks of processing
- Performs NLP tasks
- Generates and classifies texts
- Answers questions (**prompts**) just like a human: analyze sentiments, chatbot conversations, etc.

What is an LLM ?



How can I work with LLMs?

- 🦜🔗 LangChain (https://www.langchain.com/)
- LlamaIndex  LlamaIndex (https://www.llamaindex.ai/)
- Semantic Kernel (https://learn.microsoft.com/en-us/semantic-kernel/
- PaLM  (https://developers.generativeai.google/)
- Hugging Face 🤗 (https://huggingface.co/)

Java-based API Frameworks

- Semantic Kernel - Java SDK

  (https://learn.microsoft.com/en-us/semantic-kernel/)

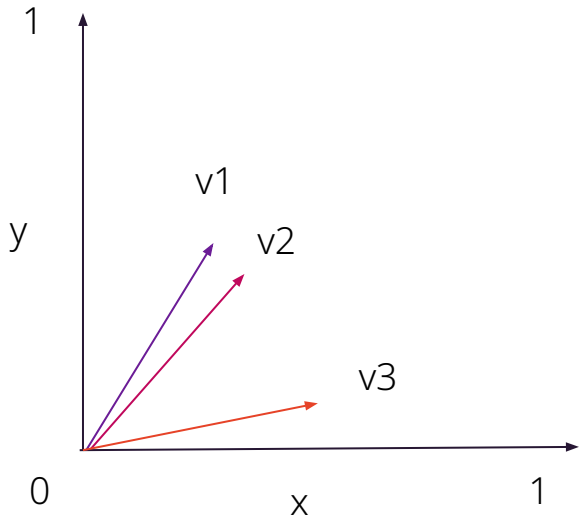
- JLama (https://github.com/tjake/Jlama)
- JVector (https://github.com/jbellis/jvector)
- Langchain4J (https://github.com/langchain4j)
- Llama2.java (direct port from Llama.c)

Vector DB and Vector Search

What is Vector Database (DB)

- A purpose-built database that serves up vector data type for complex machine learning purposes
- Relies on vector embeddings which are numerical representations of the data that are stored in vector DB
- An automatic “feature engineering”
- Approximate Nearest Neighbor (ANN)

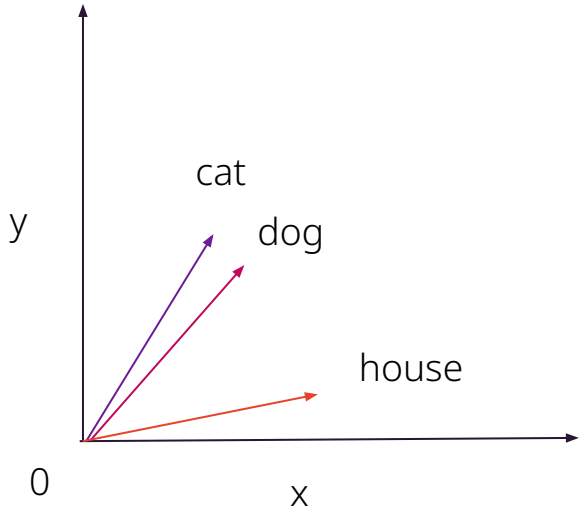
Mechanism: What is a vector/embedding ?



2 dimensions normalised vectors

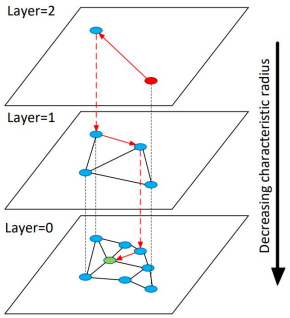
- An embedding model transforms a text into a vector called an embedding.
- The embedding can be N dimensions. For instance OpenAI's embeddings are 1536 dimensions.
- Similarity: v1 is more similar to v2 than v3. This is a simple mathematical formula.

(cont'd): What is a vector/embedding ?



- The vector captures the essence of a word or a block of text within its context.
- The dimensions are the result of the LLM training.

Vector search



Vector stores / vector databases

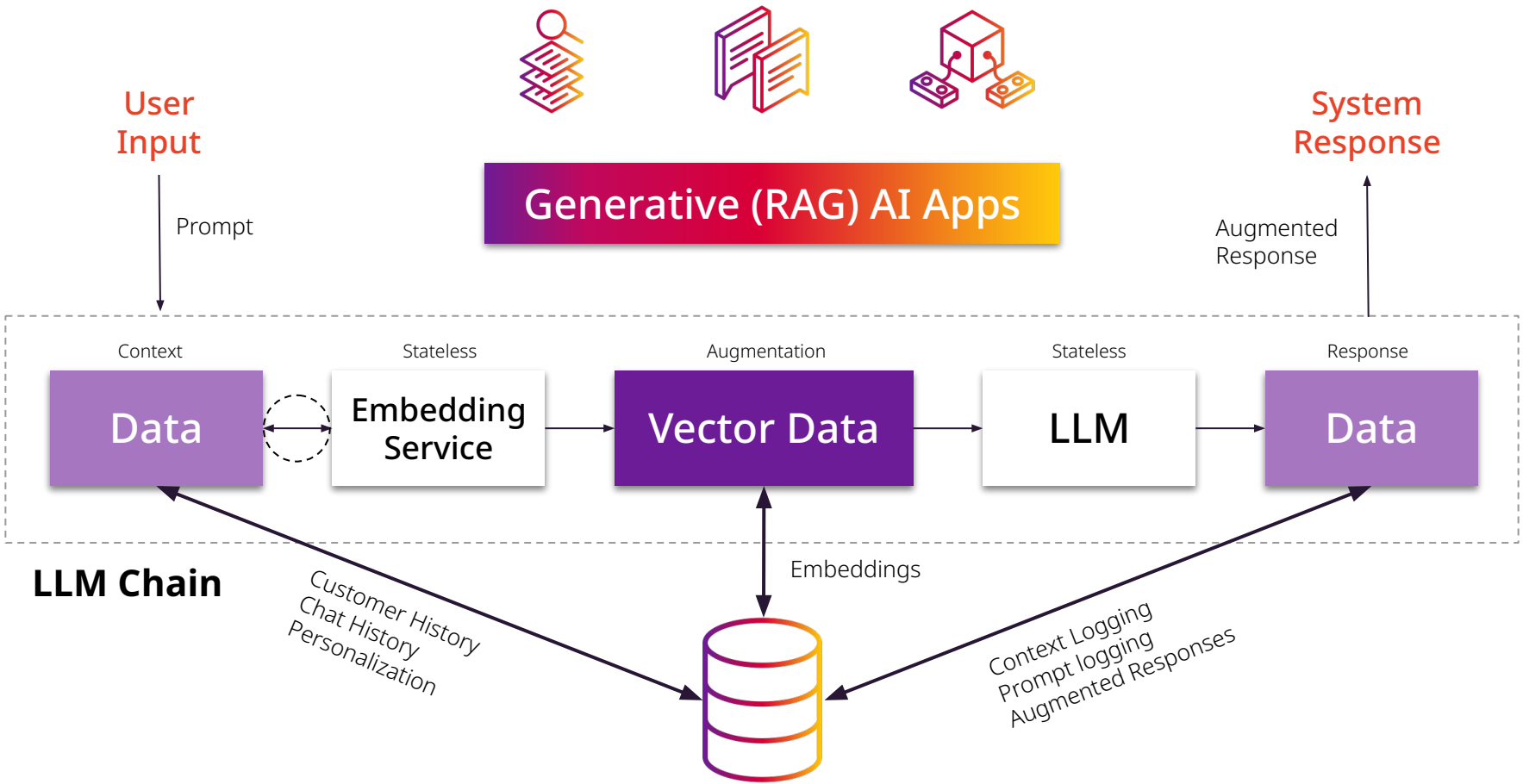
- Embeddings storage (with or without metadata depending on the DB)
- Built-in algorithms for fast retrieval of so-called “nearest-neighbors” embeddings (eg. HNSW, JVector-Cassandra/Astra, ...)
- Vectors are a new type of data supported in established databases (DataStax AstraDB, Pinecone, Weaviate, PgVector, Milvus ...)

What are vector embeddings being used for?

- Search (where results are ranked by relevance to a query string)
- Clustering (where text strings are grouped by similarity)
- Recommendations (where items with related text strings are recommended)
- Anomaly detection (where outliers with little relatedness are identified)
- Diversity measurement (where similarity distributions are analyzed)
- Classification (where text strings are classified by their most similar label)

The Problem with “Traditional” DB in AI

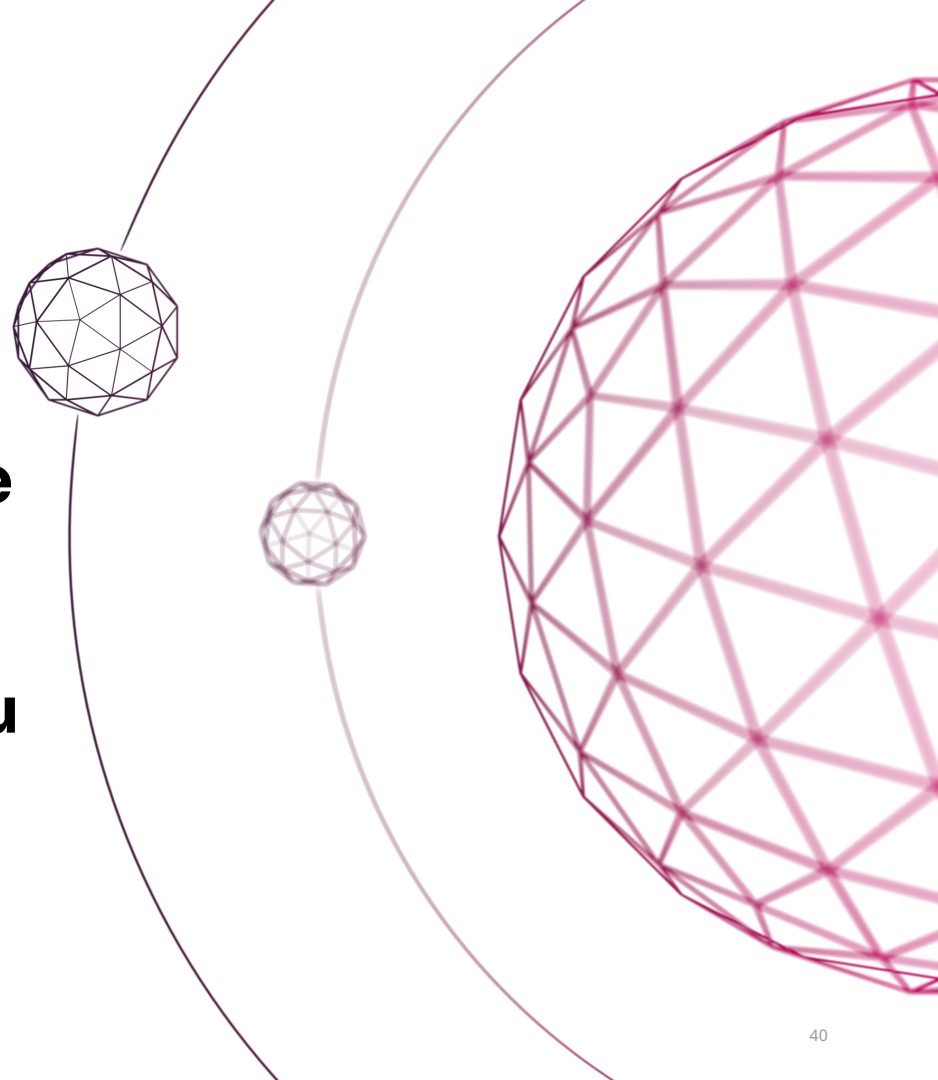
- Unable to handle the complex data that’s required in AI to handle the dimensions, patterns and relationships
- Should function like human memories but not so
- Essentially we need to provide the context for GenAI processing
- Cannot be used to store and querying of high dimensional vector data


User Input
Prompt
Generative (RAG) AI Apps
System Response
Augmented Response
Context
Stateless
Augmentation
Stateless
Response
Data
Embedding Service
Vector Data
LLM
Data
LLM Chain
Customer History
Chat History
Personalization
Embeddings
Context Logging
Prompt logging
Augmented Responses

Benefits

Ask and you shall receive !! But make sure to ask wisely

Work being done for you by the “bot” – much faster!!



Challenges

Be aware of the following issues (partial list)

- Hallucinations
- Ethical concerns / Potential misuse
  - Currently no one is there to oversee its usages 
- How about real-time up-to-date data??
  - RAG pattern for LLMs

Resources

This slide deck can be accessed here:

https://bit.ly/48EP14i

Follow Mary's Stream

[**Different topics:** GenAI/ChatGPT, Java, Python, JS/TS, Open Source, Distributed Messaging, Event-Streaming, Cloud, DevOps, etc]

Wed|Thurs|Fri-afternoon-US/CST

https://twitch.tv/mgrygles

https://youtube.com/@marygrygleski9271

Thank You

Mary Grygleski

https://www.linkedin.com/in/mary-grygleski/

@mgrygles

https://discord.gg/RMU4Juw