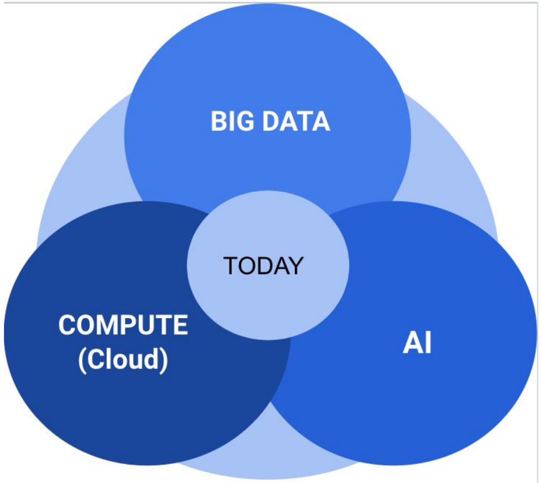
Data & AI

Two words in one breath

Why now?

Software eats the world & AI eats Software ….



We are witnessing massive explosions of **data** - IoT, Genomics & across every industry vertical

Rise of Cloud made JIT **compute** accessible to one & all - & gets more powerful - Moore's law

GenAI catapulted **AI** & democratized it to every organization - big & small - at different levels of maturity - trend towards Point of Singularity

Data drives business use cases in every industry

Threat Detection Prevention

Health and Life Sciences

Autonomous Vehicles

Connected Factory

Personalizations

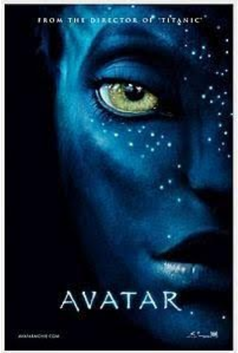
Gaming/Entertainment

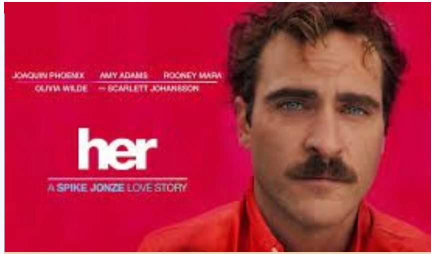
Smart Farming

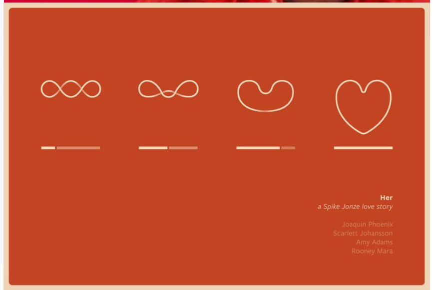
Banking

Forecasting

Homage to SciFi AI





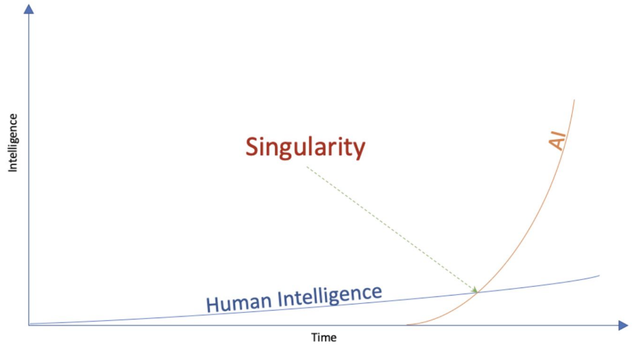


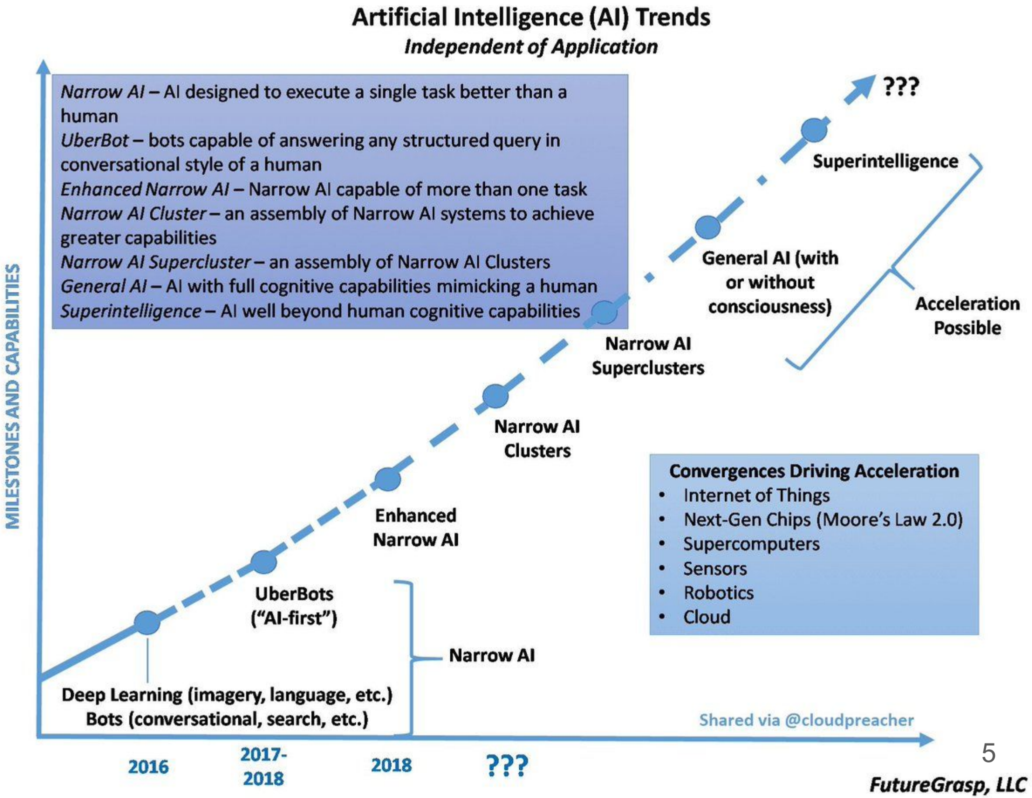
Evolution into Point of Singularity

The point where AI can improve itself faster than humans can

Narrow Vs Global Intelligence

AI can pass the Turing test without GAI





Poll

- How many have used chatGPT
- How may use chatGPT at least 3 times a day
- How many are working on Gen AI projects
- How many are using Gen AI products

Agenda

- Data-centric ML
- ABC of Generative AI
- Role of LLMs in modern data & AI landscape
- Fit for purpose of LLMs - how to pick the right one for your needs?
- Deep dive into the ‘RAG’ Architecture
- LLM Ops
- Legal & Ethical considerations

Introductions

Lead Solutions Architect – Databricks

“If Data has arrived, it better be served!”

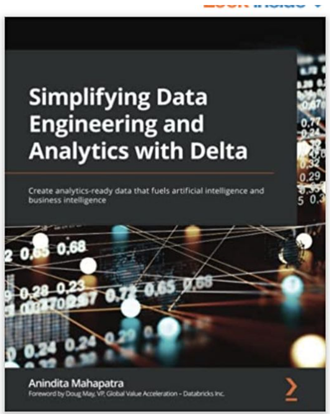
- Past Experience in Big Data
  - Teradata/Think Big Analytics
  - Nokia/Microsoft

- MS in Computer Science - Boston University
- Master of Liberal Arts & Management - Harvard Extension

- I teach a graduate course on Data Engineering (CSCI E-103) at the Harvard Extension School.

- I've authored the book “Simplifying Data Engineering and Analytics with Delta: Create analytics-ready data that fuels artificial intelligence and business intelligence”

Anindita Mahapatra



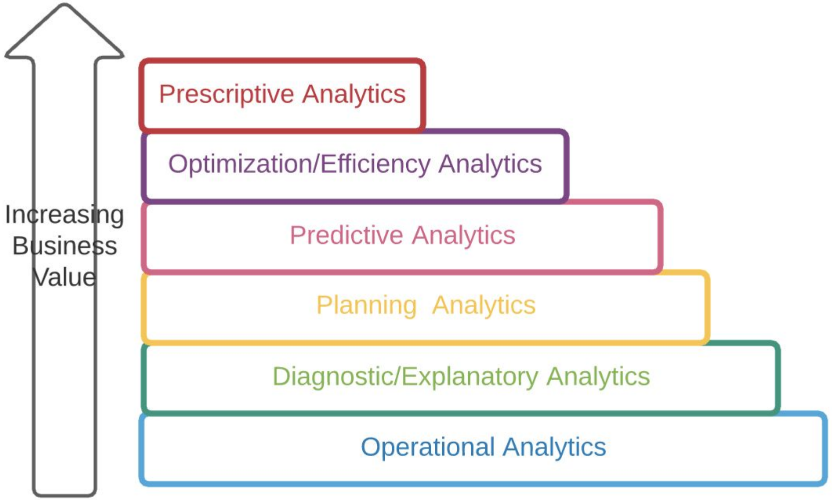
ISBN-13: 978-1801814867

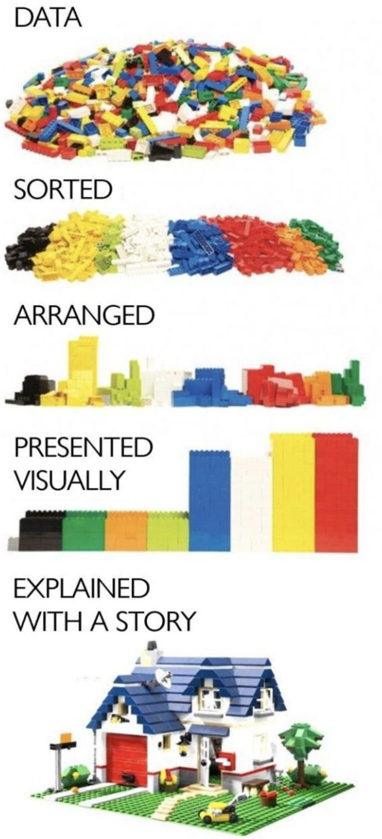
ISBN-10: 1801814864

Data-Centric ML

Data -> Analytics

From a LinkedIn post:
"The Difference between Raw Data and the Stories Data can tell."





The analogy used is that of cutting carbon to create a diamond.

Raw data is the carbon that gets increasingly refined.

The longer the processing layers, the more refined and curated is the value of

However it is more time consuming and expensive to produce the artifact

Hardest Part of ML isn't ML, it's everything else

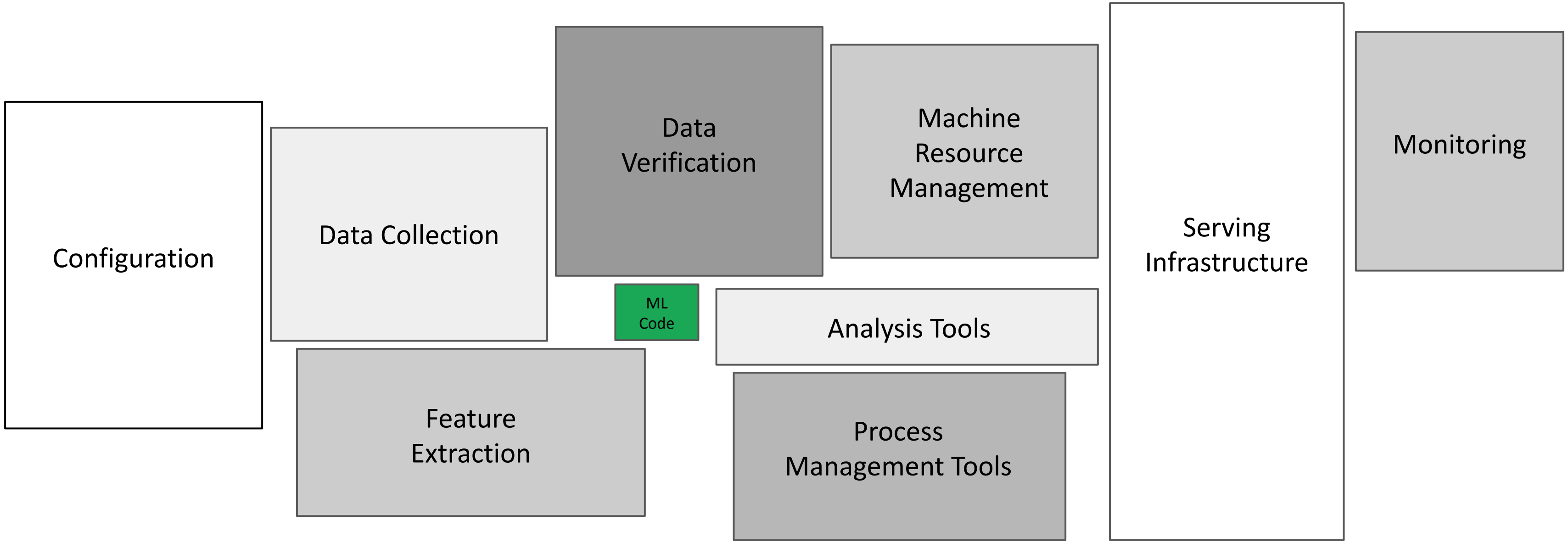
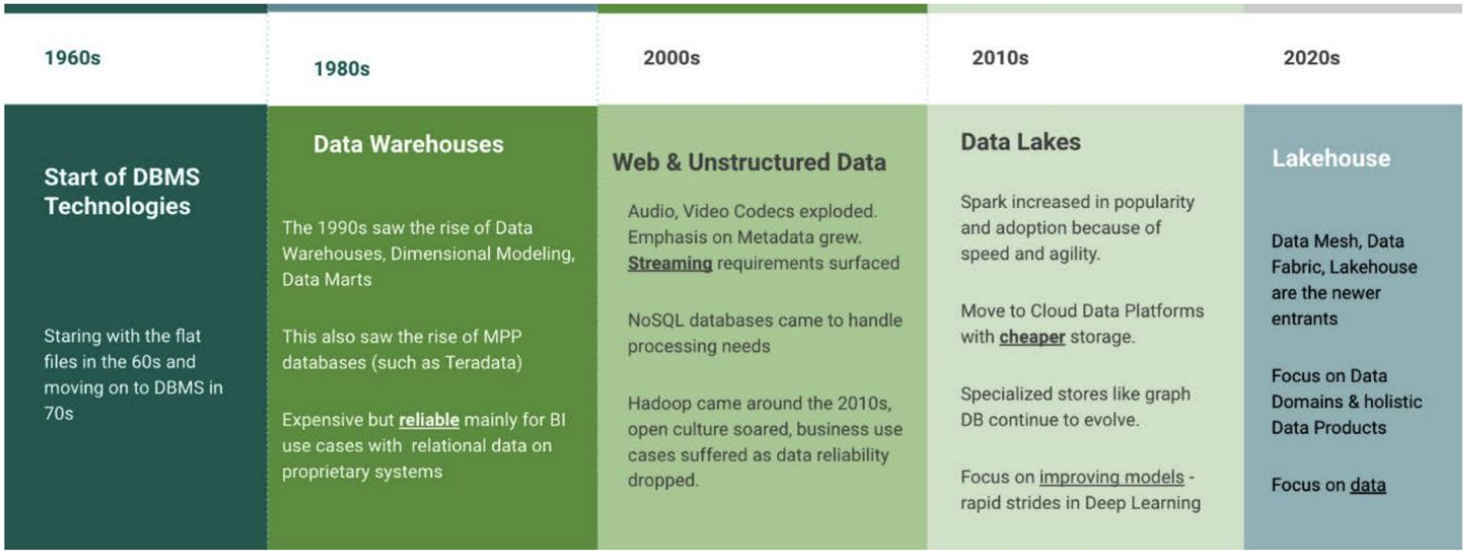
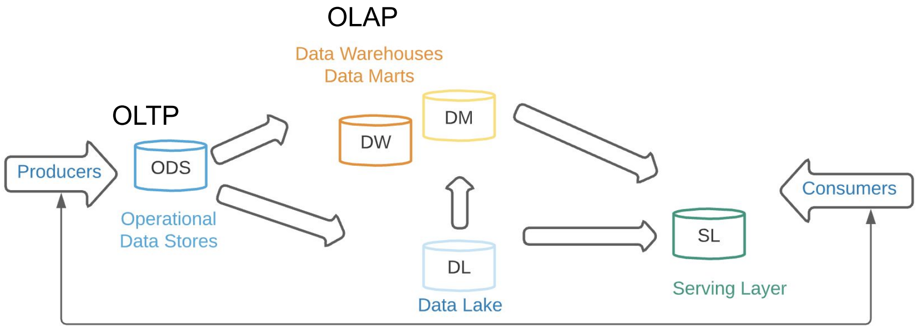


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small green box in the middle. The required surrounding infrastructure is vast and complex.

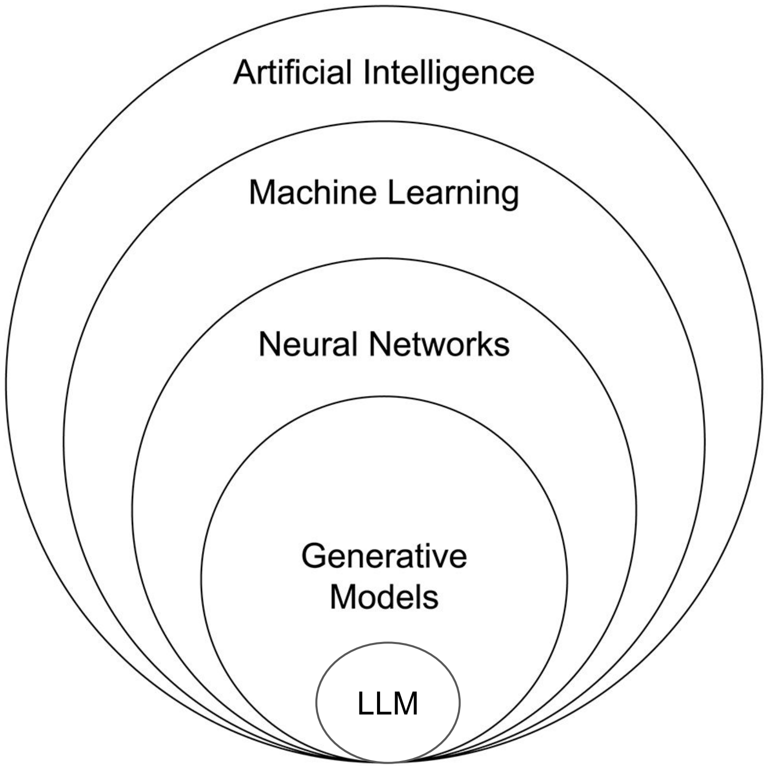
Evolution of Data Platforms





Different Consumers tap into the data at different stages

Evolution of AI



- Rule Based Systems
- Classical ML
- Deep Learning (unstructured data)
- Gen AI
- LLM(language), GAN(image)

While Traditional AI aims to perform specific tasks based on predefined rules and patterns, Generative AI goes beyond this limitation and strives to **create** entirely new data that resembles human-created content

ABC of Generative AI

Is it a threat or an opportunity for my business
Can it be used for gaining a competitive advantage
How can I use data securely with GenAI

Why are LLMs so powerful?

LLMs are very capable because they are trained on massive amounts of data – giving them a grasp of how language works <u>and</u> a significant amount of knowledge

- Training data such as “all text on the internet”

LLMs excel at language related tasks, such as:

- Answering questions or chatting
- Summarizing longer form content
- Writing computer code such as writing SQL or HTML or Java
- Generating content such as marketing copy
- Translation

There are 1000s of different LLMs, each with different skills & capabilities

- GPT family (e.g. ChatGPT), BERT, T5, BLOOM

ML/AI has been around for a while – why should I care now?

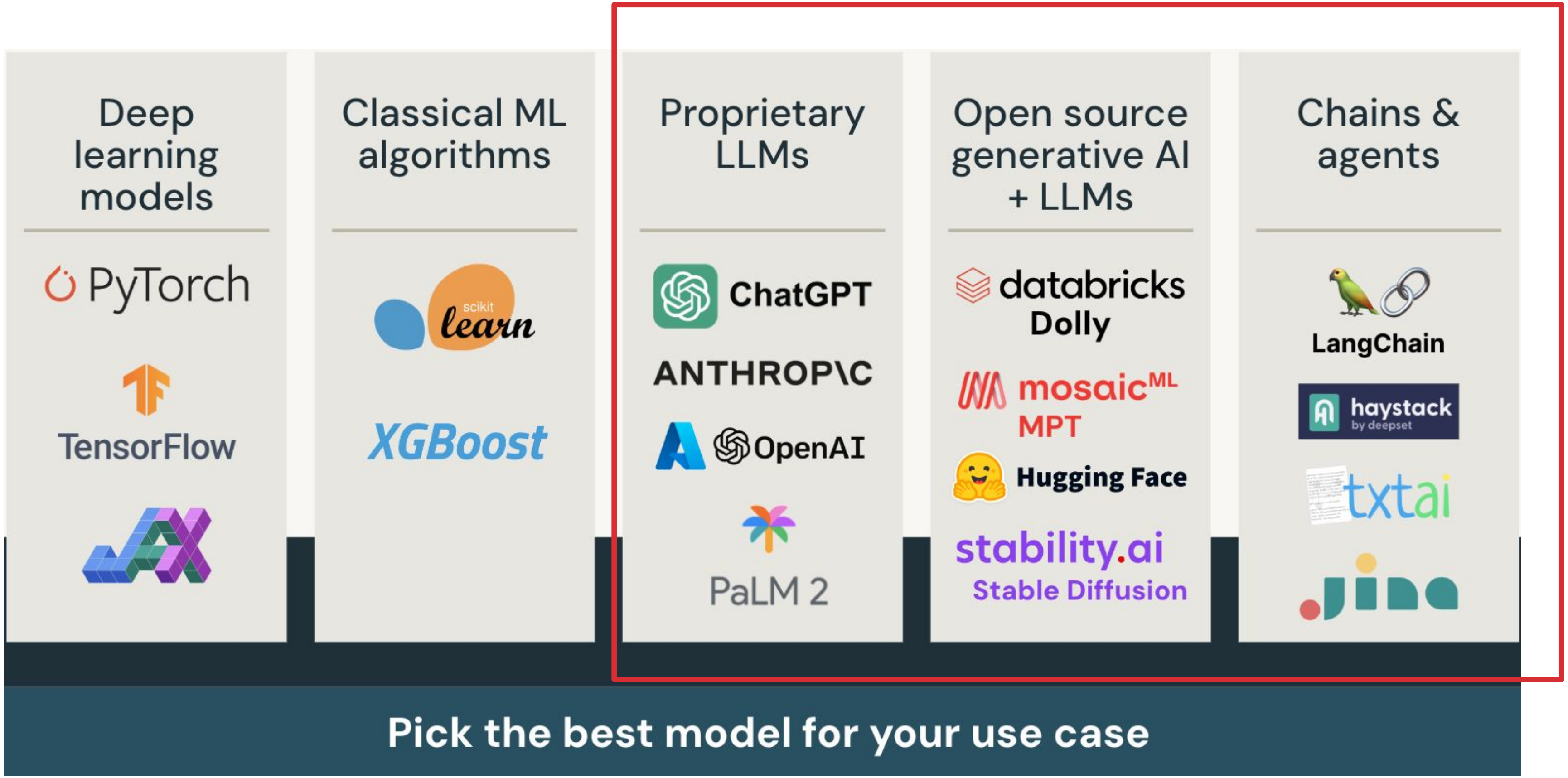
LLM accuracy and effectiveness has hit a tipping point

- Powerful enough to enables use cases not feasible even a year ago
- Yet economical enough to access and use - even by non-technical business users

LLMs and tooling are readily available

- Many LLMs are open source and customizable
- Requires powerful GPUs, but are available in the cloud

LLMs are an addition to the existing ML arsenal



Gen AI Terminology (I)

- **LLM** - Large Language Model (NLP and beyond)
- **GAN** - Generative Adversarial Network (images)
- **Diffusion** - simulate the dynamics of complex systems over time (Lip sync)
- **Foundational LLM** - a pre-trained lang model that is the starting point for more specific models
- **Hallucination** - a confident response by an AI that it has not been trained on (Temperature)
- **Grounding** - process of associating words with their real-world entities and concepts.
- **Prompt Engineering** - process of designing effective NL prompts for use with LLMs
- **Zero-shot Learning**: An input text + prompt that describes the expected output from the model
- **Few-shot Learning**: Zero-shot + few examples of in/out

Gen AI Terminology (I)

- **Chain of Thought**: improves the reasoning ability of LLMs by prompting them to generate a series of intermediate steps that lead to the final answer of a multi-step problem.
- **Modality** - Multiple types of data - text, image, audio, video
- **Transformers**: NN arch for NLP - Encoder, Decoder, Embedding(transform from high dim to lower)
- **Tuning**: Instruct/Fine
- **RLHF** - Reinforcement Learning with the Human Feedback

Examples
Ready for democratization

LLMs generate output for NLP tasks
Code Generation & Developer Productivity

Gen AI examples

Synthetic yet realistic content generation

Image generation

- Generate realistic/artistic high-quality images
- Virtual agent generation

Video Synthesis

- Animation
- Scene generation

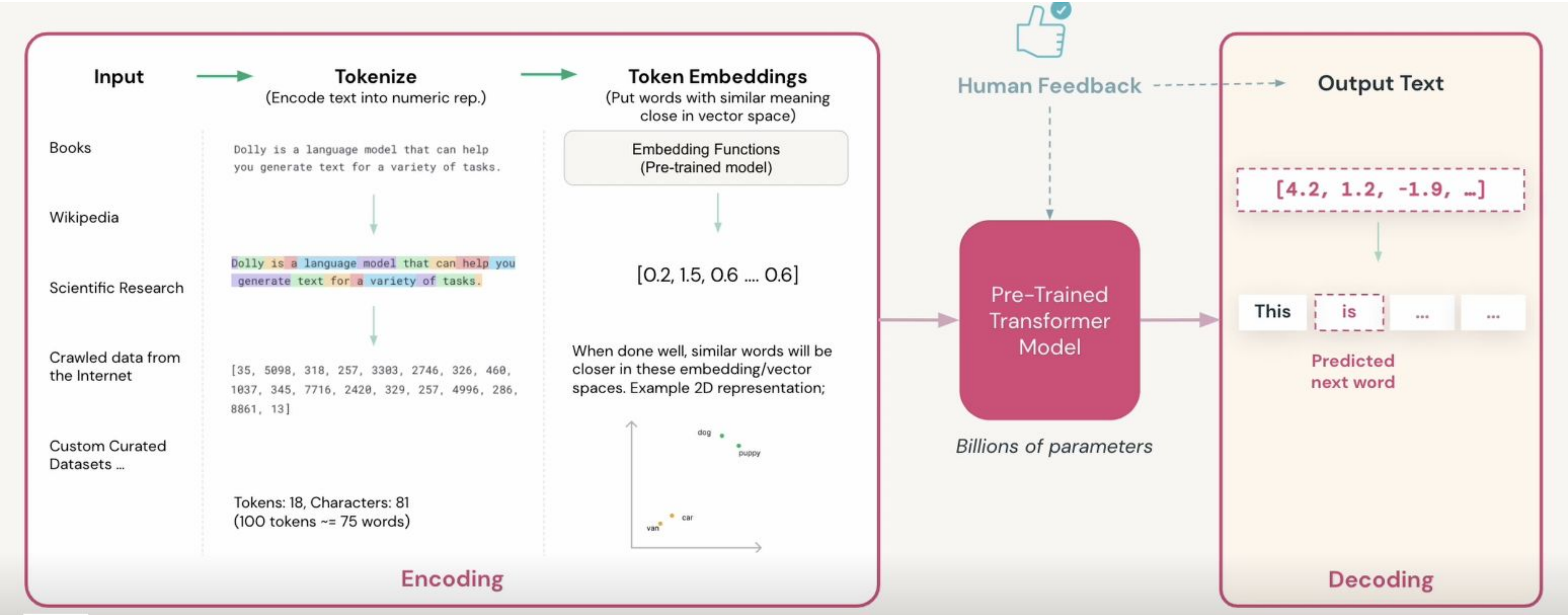
3D Generation

- Object, character generation
- Animations

Audio Generation

- Narration
- Music composition

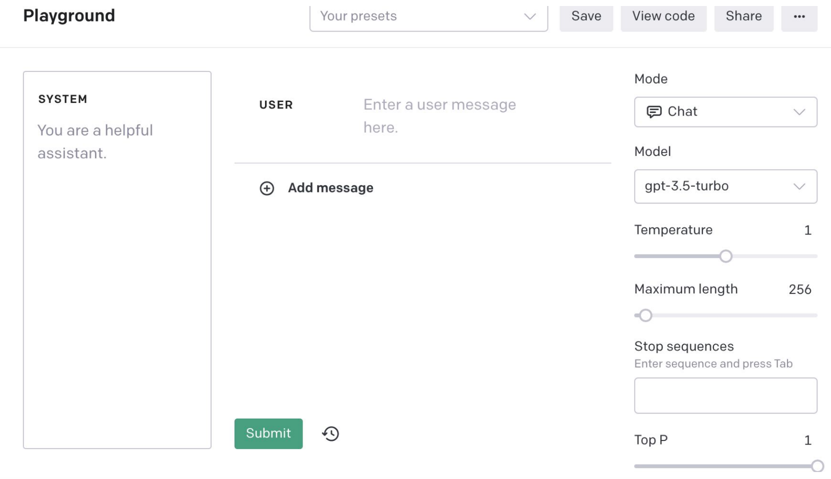
How are LLMs trained



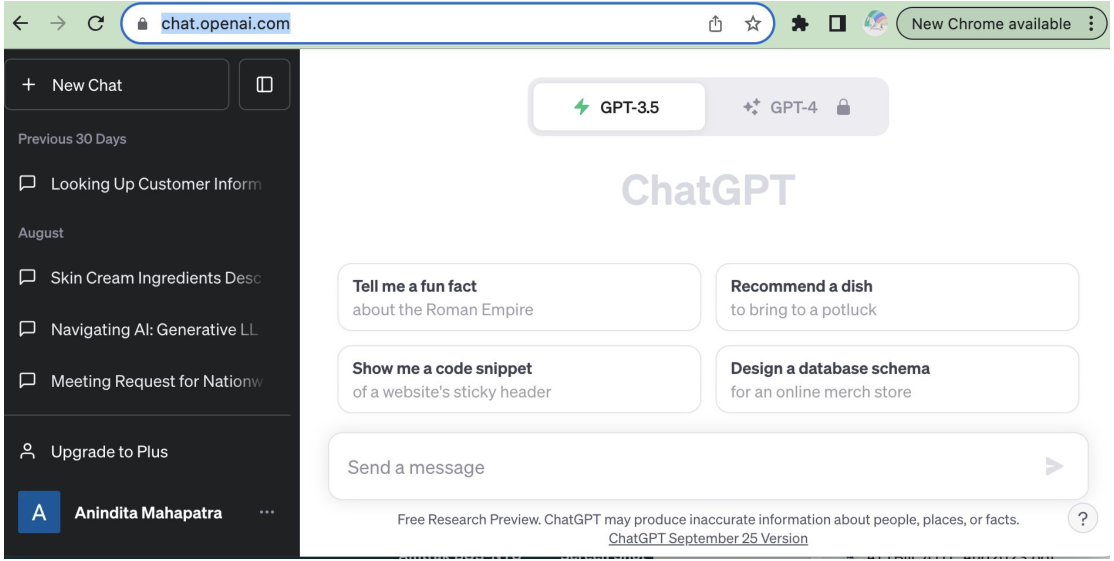
Open AI

[Playground]

[ChatGPT]



Whisper: Audio -> Text

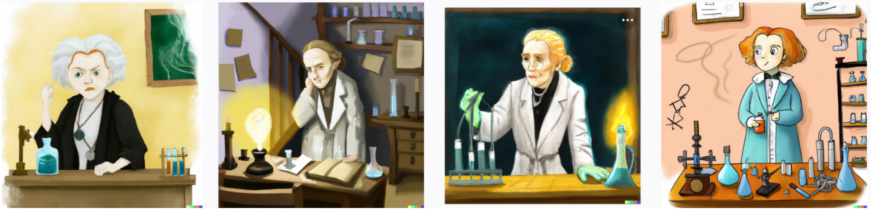


DALL-E (Text to Image)

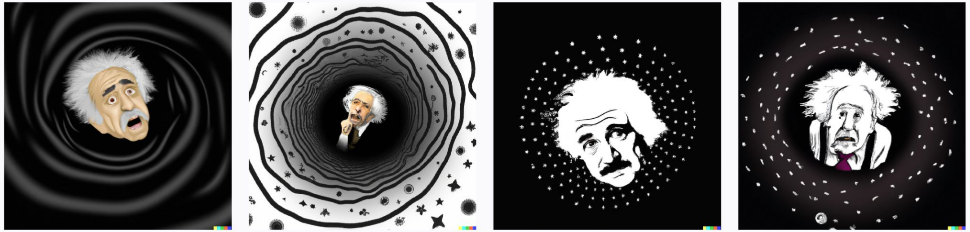
https://openai.com/dall-e-2

Phoenix DAMA chapter attending a talk on GenAI

Madame Curie in her lab



albert einstein in a black hole



A Digital Human to answer Qs from this session

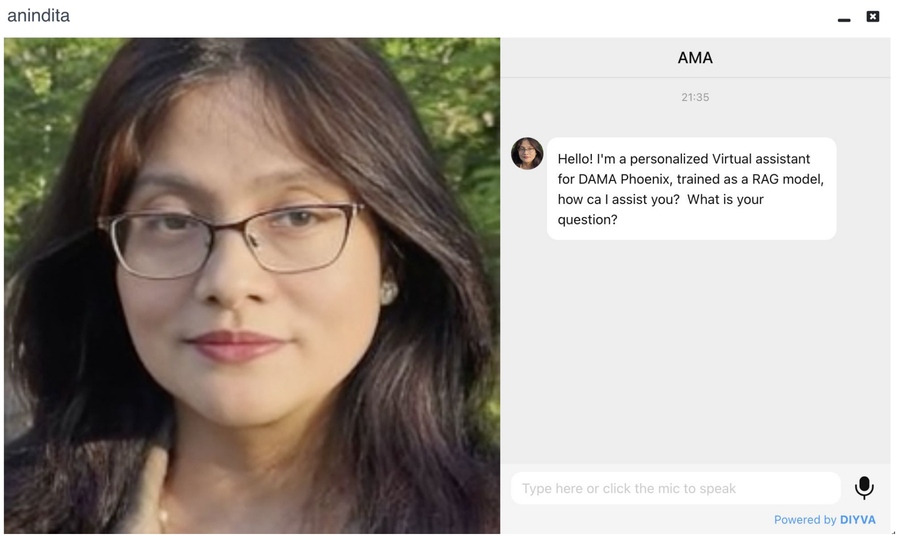
Anindita’s Digital Avatar

RAG model for Q/A

Text & Voice

Translation

Transcription



Metaverse - 3D Experience
Integrate physical & digital realities

Virtual Augmented Reality beyond video calls!

Mixed Reality, Mixed hangouts with holograms

Emotions before words, Multi-thread humans

You can be anywhere With anyone, breaks geographic barriers not just games, social, work!

Scans, Collect expressions (headsets)

Mark Zuckerberg: First Interview in the Metaverse | Lex Fridman Podcast #398

Role of LLMs in Modern Data/AI Landscape

Some Examples of How do LLMs enhance use cases

Data Q&A: democratize access to knowledge	Simplify structured insights about unstructured data	Improve efficiency of knowledge worker's basic tasks	Improve existing machine learning models
Enable call center staff to ask questions of all previous support tickets Let users ask which Delta table best meets their analysis needs	What are the 5 top issues based on the call center transcripts this week Which customer reviews mentioned an issue with defects? Has that spiked in the last 2 weeks	Ask a data question, get a draft SQL query Describe a landing page, get draft HTML code Automated personalized marketing messages	Include customer forum posts in our fraud models Tune our product recommendation model based on customer's written feedback

Typical LLM Use Cases

Chatbot for Q/A, Smart Search, Assistant	**Content Generation**
Q/A from complicated policy docs Aid the analyst by looking up info Debugging/Coding	Summarize policies, reports, technical documents Generate reports, decks Human-readable explanations of difficult-to-parse commands
Classification	**Automation**
Basic triaging Social Engineering Detection Sentiment Analysis	Translation & Response automation Pattern finding Automate mundane tasks via prompt Engineering

Fit for Purpose

Why would you investing in your own Gen AI models?

Business Advantage	Inference Economics	Increase Accuracy	Regulation and Privacy
Data is Your Moat	Train Smarter, Smaller Models	Domain-specific, Proprietary Data	Data & Model Ownership

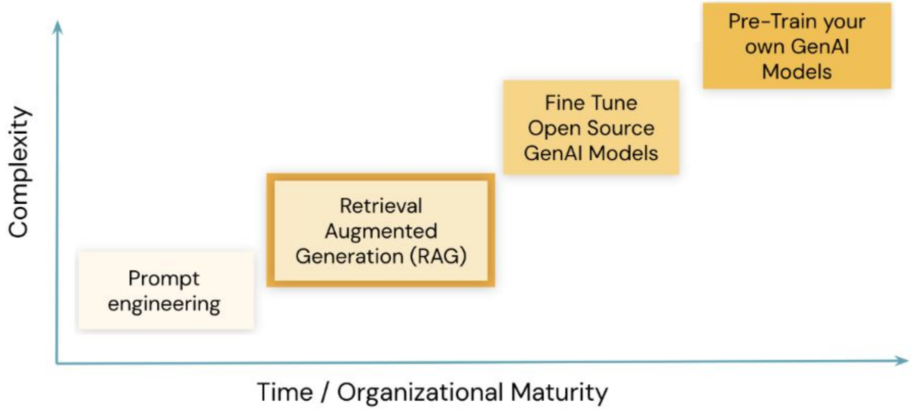
First cut
Model & deployment type

- Open source models (as is/tune, commercial/non)
    - Pros
        - Data/model stays within your control, fine-tuning, faster inferencing
        - Quality is rapidly improving
    - Cons
        - Larger modes/datasets, in-house expertise
    - Llama2 (FB), MPT(Mosaic ML), Dolly(Databricks)
- Proprietary model
    - LLM-as-a-service, May be fine-tuned - usually hosted elsewhere, you send the data
    - Pros
        - Faster development, better quality on routine tasks
    - Cons
        - Pay per request, Data privacy concerns, vendor lock-in
    - Eg. OpeAI, Anthropic

Pre-train Vs Fine-tune

- Pre-train produces foundational LLMs
  - Trained on very large corpus of data
- Fine tuning is used for domain adaptation of a base foundation model
  - To learn a new task

LLM Progression



LLM Type	Word Volume	Quality	Cost	Latency	Privacy
Prompt Eng	(No domain data)			No control	
RAG	100s of K				
Fine Tune	Millions/Billions				
Pre-Train	Billions/Trillions			Max control	Most secure

Challenges implementing LLMs

- Need to move quickly
  - Your competitors are also jumping into LLMs, and you need ensure you aren't left behind your peers—how to quickly tackle high value use cases?
- Need to customize, control, and secure your LLMs
  - Using proprietary SaaS LLMs requires you to send your data to 3Ps and may leave you without a competitive edge. How to customize LLMs that you own & control with your proprietary data?
- Need to connect LLMs with your existing data
  - Just like other forms of machine learning, LLMs require a tight coupling with your existing data strategy—how to best connect LLMs with all your existing data sources?

Known Limitations of LLMs

- Hallucination (can be controlled by temperature, prompt)
- Bias (limited or biased training data)
- Adversarial Tokens (inaccurate tokens fed to cause malfunction)
- Malicious content authoring and social engineering
- Train an LLM for Malicious Reward Hacking or train an LLM for Malicious Reward Hacking - LLMs have the possibility of finding loopholes in real world systems, but rather than fix them, might end up exploiting them.

Path to LLM Implementation

1 Use Case Prioritization

What this will look like
- 2 hour in-person or virtual session
- Interactive discussion to understand business needs and scope

What we'll need from you
- Participation from relevant technical and strategy/roadmap owners
- Prepared list of use cases to sort & rank
- Generic idea of business impact and feasibility/data availability

Select the optimal use case to prove business value and serve as a template for future projects

2 Use Case Deep Dive

What this will look like
- 2 hour in-person or virtual session
- Interactive discussion to understand the business problem and scope impact

What we'll need from you
- Participation from relevant technical and business teams/stakeholders
- Current state overview and process description

Understand how a technical solution will satisfy the core business needs at hand

3 Solution & Architecture Design

What this will look like
- 2-3 hour in-person or virtual session
- Interactive diagramming and documentation to create future state infrastructure

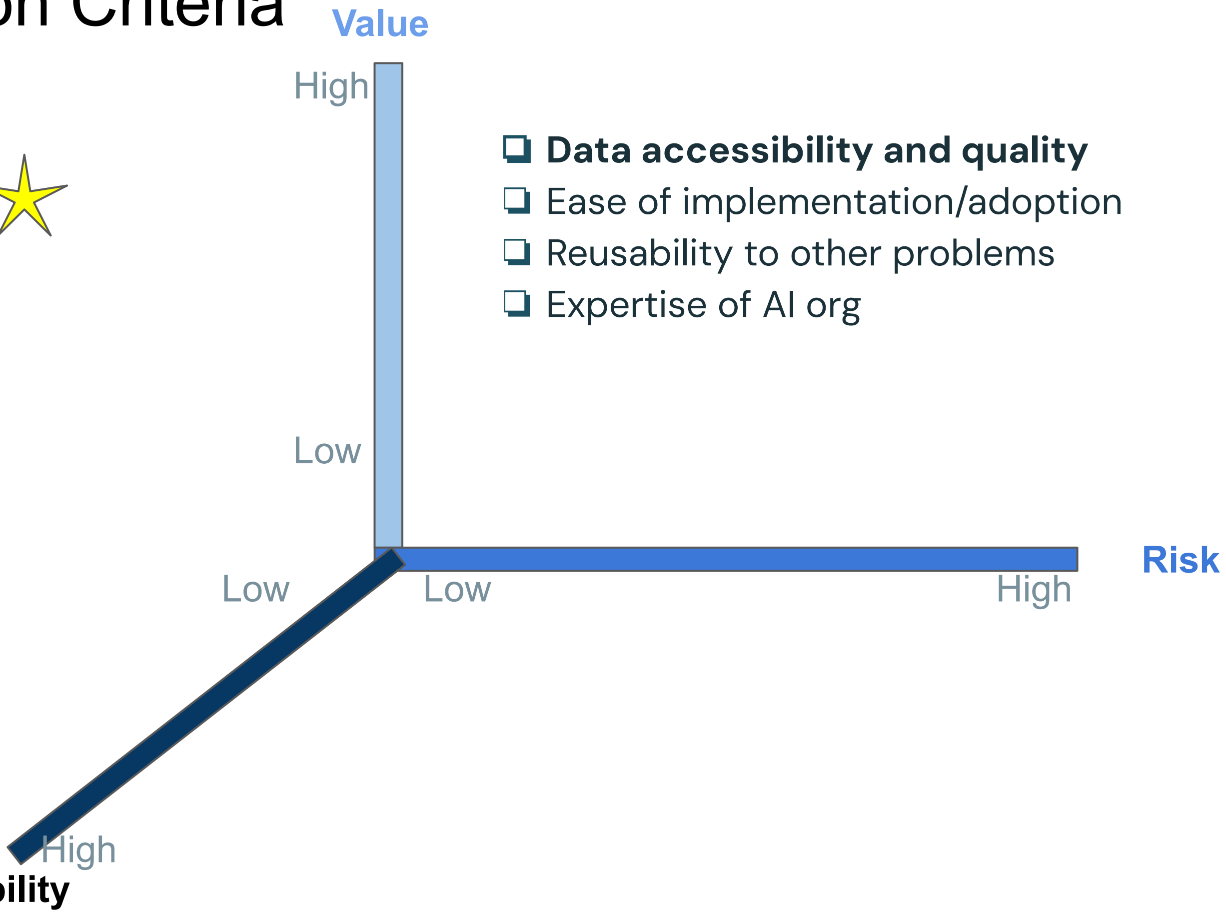
What we'll need from you
- Participation from relevant technical and business teams/stakeholders
- Current state process description and documentation of all impacted systems

Generate a blueprint for immediate implementation of a beta-version of the technical solution

Use Case Selection Criteria

Lighthouse Use Case Desired Outcomes

- ★ Learning
- ★ Testing
- ★ Templating
- ★ Piloting
- ★ Creating excitement
- ★ Educating employees



RAG Architecture Deep Dive

Gen AI Terminology (II)

- Memory
- Index
- Vector Search
- Chains
- Agents

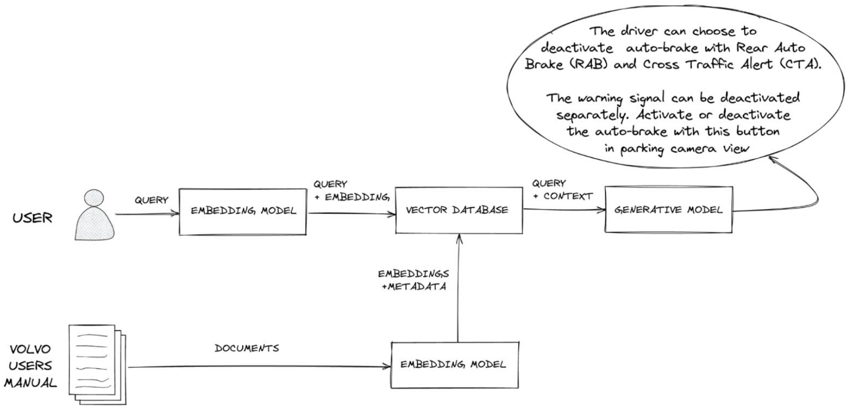
RAG - Retrieval Augmented Generation

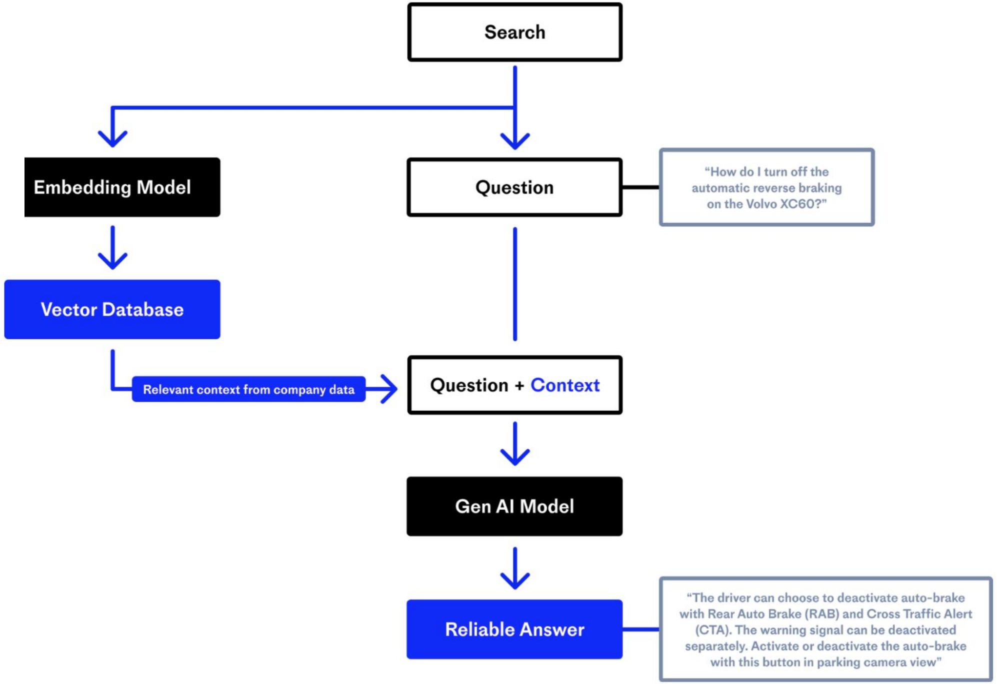
RAG is a potential solution for LLM limitations

- LLMs are "stuck" at a particular time (of training) - It is not feasible to update their gigantic training datasets - but RAG can bring them into the present.
- LLMs are trained for generalized tasks, meaning they do not know your company's private data.
- It's not easy to understand which sources an LLM was considering when they arrived at their conclusions.
- Few organizations have the financial and human resources to produce and deploy foundation models.

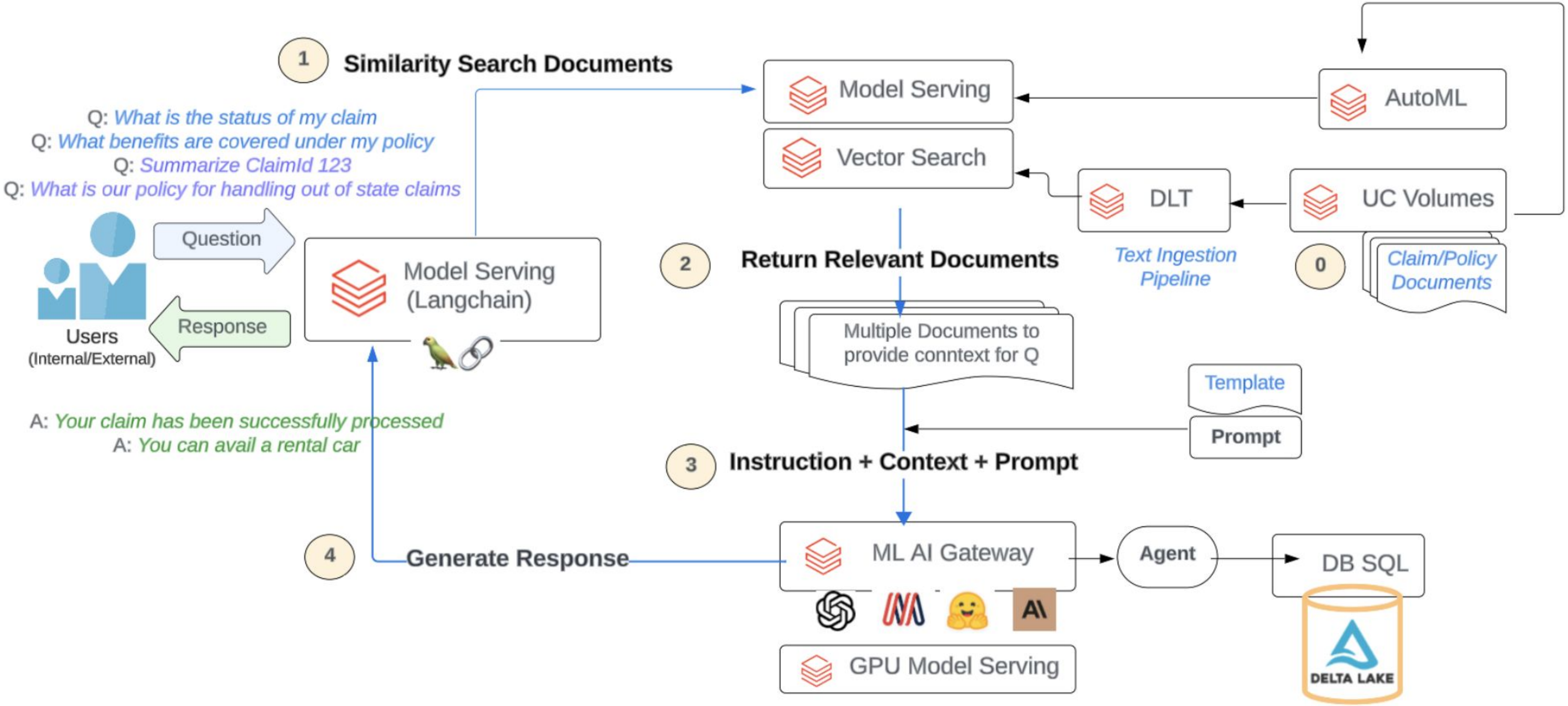
RAG is one of the most cost-effective, easy to implement, and lowest-risk path to higher performance for GenAI applications.

Reference Architecture for RAG





Transform documents into a Knowledge Engine for Q&A

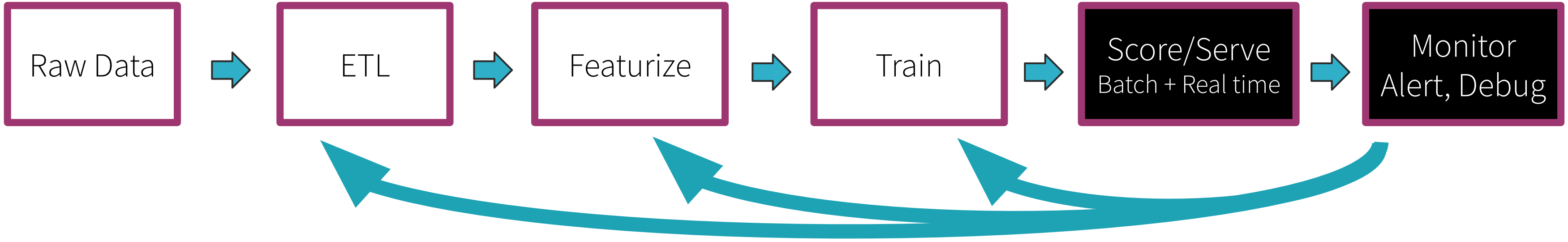


LLM Ops

Improving LLM

- Accuracy of responses
- Latency of response
- Prevent Hallucination
- Continue training for newer more relevant information
- Combine structured and unstructured data

ML Lifecycle and Challenges



Zoo of Ecosystem Frameworks

Tuning

Deploy

Model Mgmt

Collaboration

Scale

Governance

- Feature Repository
- Experiment Tracking
- AutoML, Hyper-p. search
- Remote Cloud Execution
- Project Mgmt (scale teams)
- Model Exchange
- A/B Testing
- CI/CD/Jenkins push to prod
- Orchestration (Airflow, Jobs)
- Lifecycle mgmt.
- Data Drift
- Model Drift

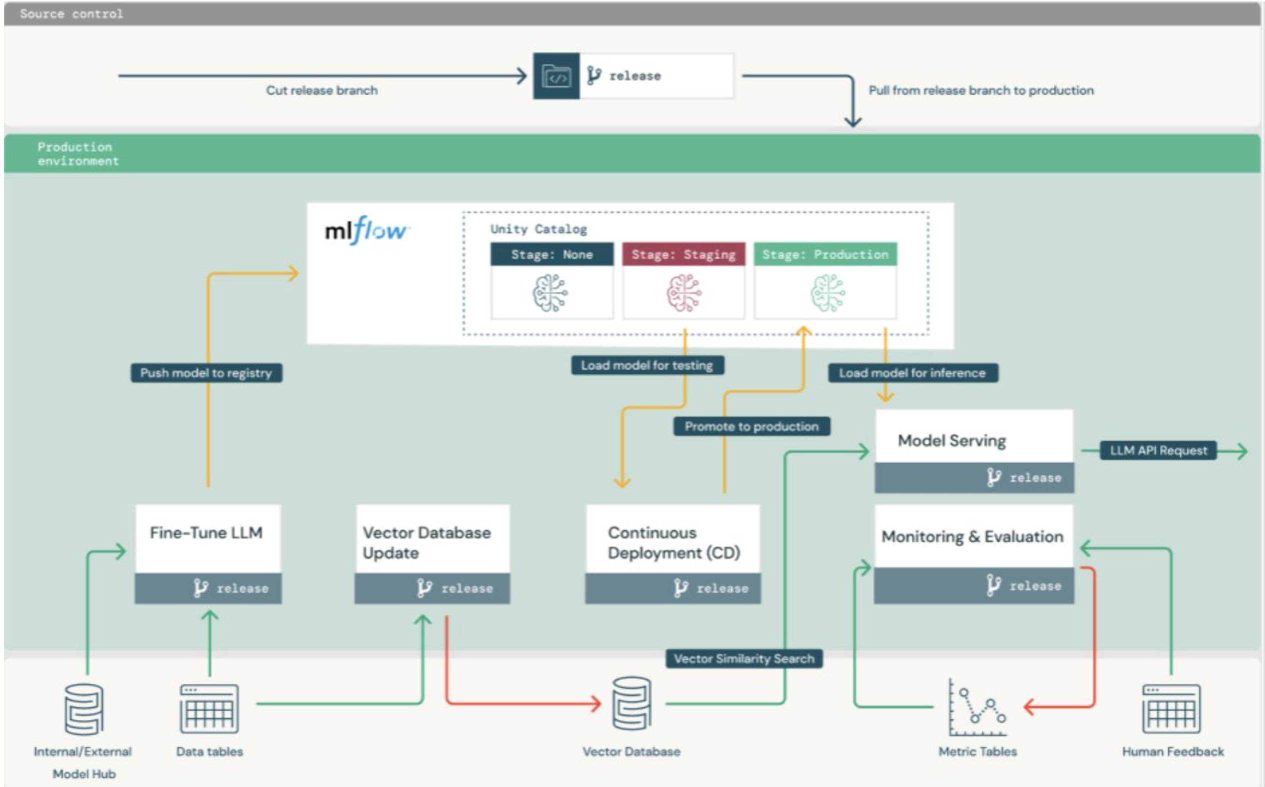
DataOps + MLOps + GenAI = LLMOps

LLM Operations for end-to-end production

- Databricks unifies LLMOps with traditional MLOps & DevOps
- Teams need to learn mental model of how LLMs coexist with traditional ML in operations

Differences to MLOps

- Internal/External Model Hub
- Fine-Tuned LLM
- Vector Database
- Model Serving
- Human Feedback in Monitoring & Evaluation



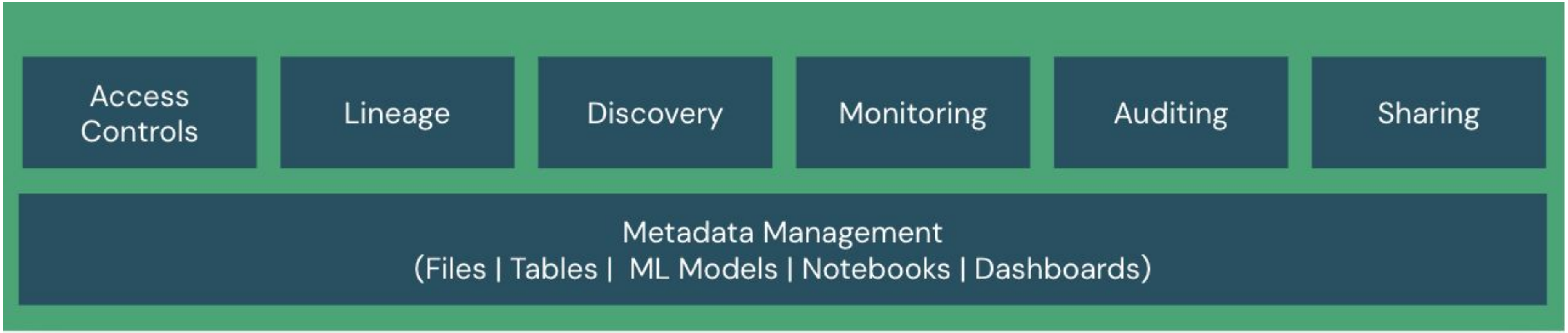
Legal & Ethical Considerations

The Imperative of Insuring LLMS

Governance

Governance

- All Data
  - structured/semi/Un-structured
- All Data Assets
  - Model
  - Dashboard
  - Services & Products



The usual suspects & more …

- Potential for biased predictions
  - Gender, age, ethnicity
- Risk of misuse (more Black box)
  - Explainability is imp
  - Generation of harmful content (toxicity)
  - Hallucination
- Breaches of Privacy
  - Regulatory/Compliance
  - IP
  - GDPR

Discrimination, exclusion, and toxicity
Information hazards
Misinformation harms
Malicious uses
Human-computer interaction harms
Automation, access and environmental harms

Laws can be enforced but not Ethics

As a society, we have a collective moral responsibility

Focus on ESG score of a company- (Env/Social/Governance) By incentivizing organizations to prioritize fairness, transparency, privacy, and accountability, policies contribute to building ethical LLMs that benefit society as a whole.

Potential Safety Nets & Band-aids via iterative process

- Collect Interaction details
- Model Monitoring (Output/Results)
  - Automatic ML Scoring
  - RLHF
- Guardrail Models
  - To vet training data & possibly responses from ML
- Careful Prompt Design
  - Explicit instructions to be factual
  - Set temperature to be 0
- EU AI Act - Regulated by Disclosure of Data, Compute, Model, Deployment

Key Takeaways

- Every organization will be a Data & AI company in the future
- ‘Your Data’ & ‘Your Model’ will set you apart from your competition
- There are various levels of complexity of LLMs & You can adopt the one that best suits your needs and maturity
- A repo of narrow purpose fit LLMs are more useful to an organization as compared to a hunking large one for now
- Models are improving and it is getting cheaper to create them, so it is possible to have own foundational model in the near future
- LLMs in educational context can help promote plagiarism- but benefits far outweigh
- LLMs do pose the risk of automatic some routine tasks but the human is not going to be eliminated completely, at least not yet …