AI/ML Overview (2025)

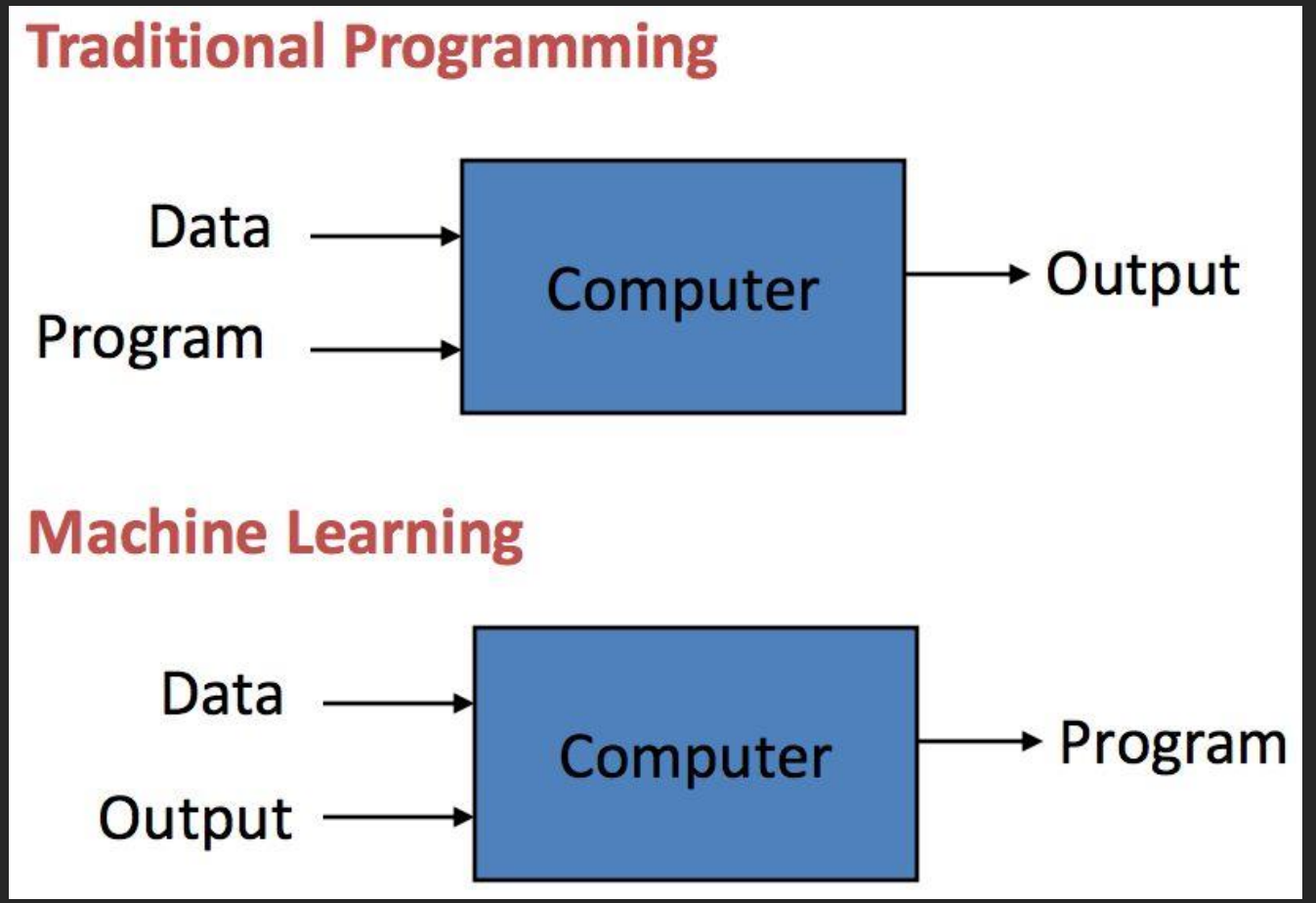
By: Pranpaveen Laykaviriyakul

Agenda

- Introduction to AI/ML
- Deep Dive into Machine Learning
- Tools / Platforms
- Current Trends in AI/ML
- ML related Careers

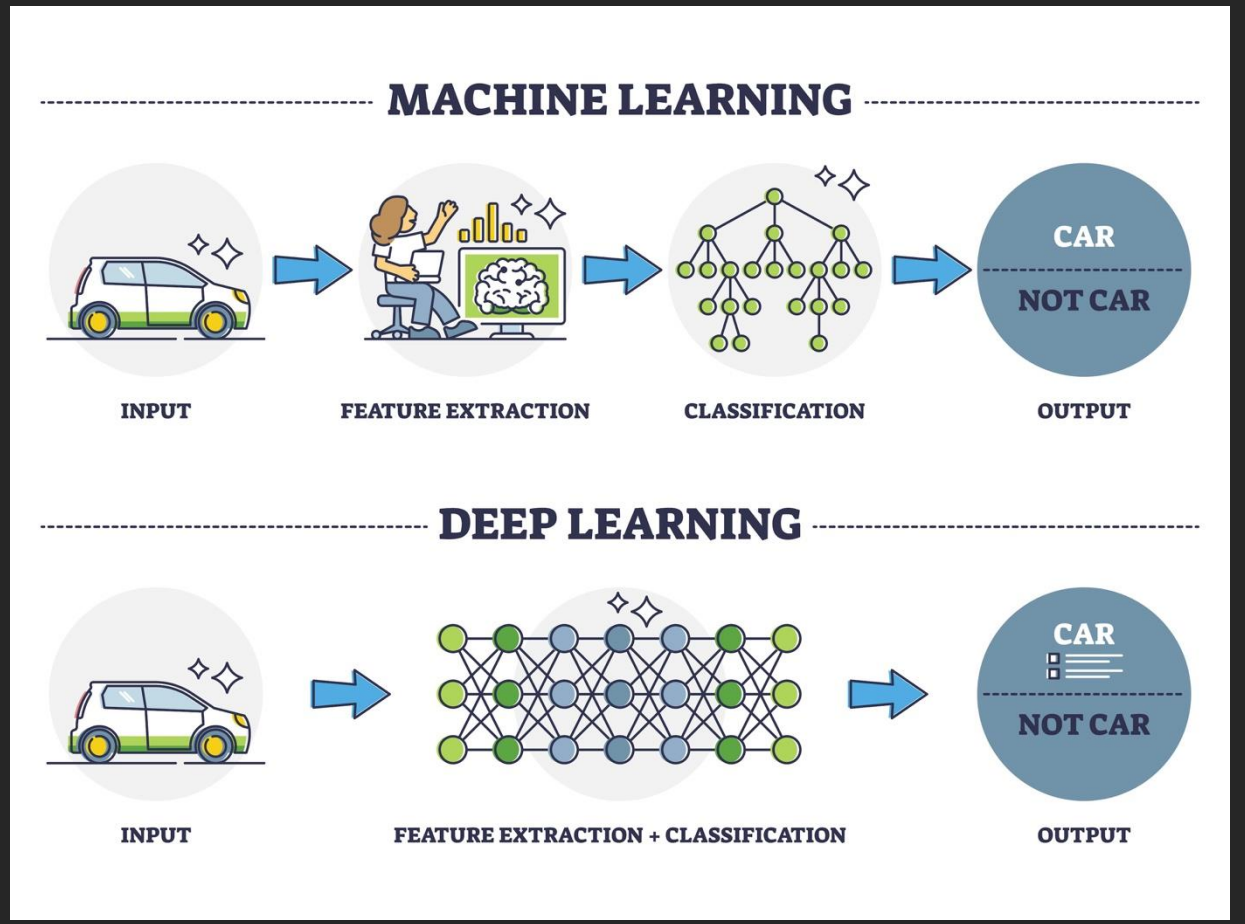
Introduction to AI/ML

Machine Learning



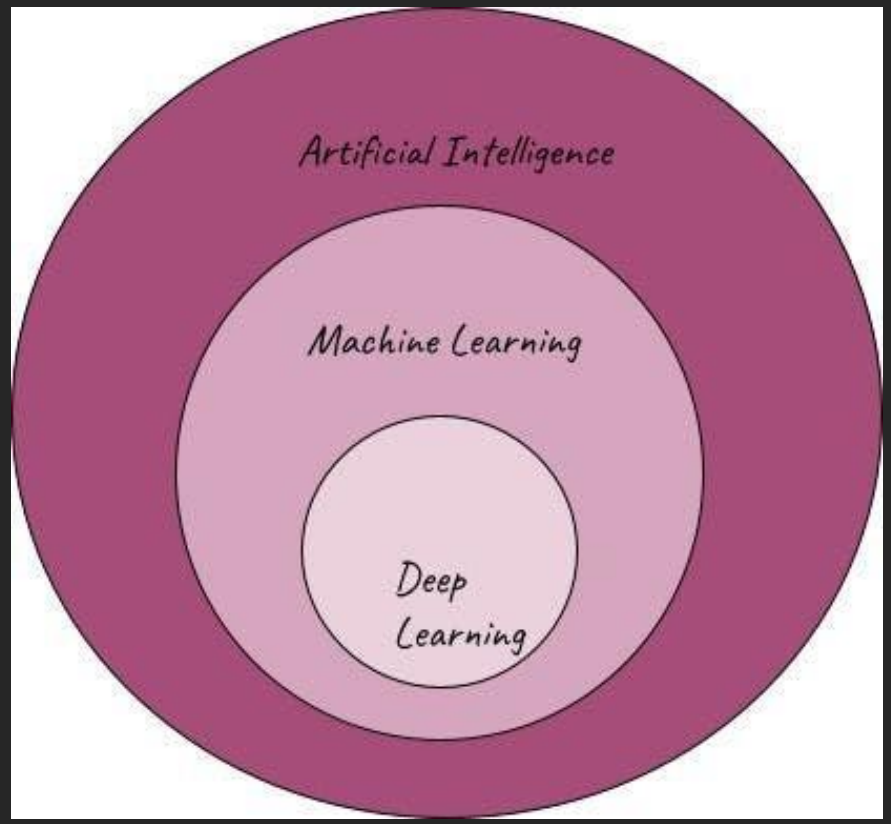
https://vitalflux.com/difference-between-machine-learning-traditional-software/

ML vs DL



https://www.turing.com/kb/ultimate-battle-between-deep-learning-and-machine-learning

AI, ML, and DL

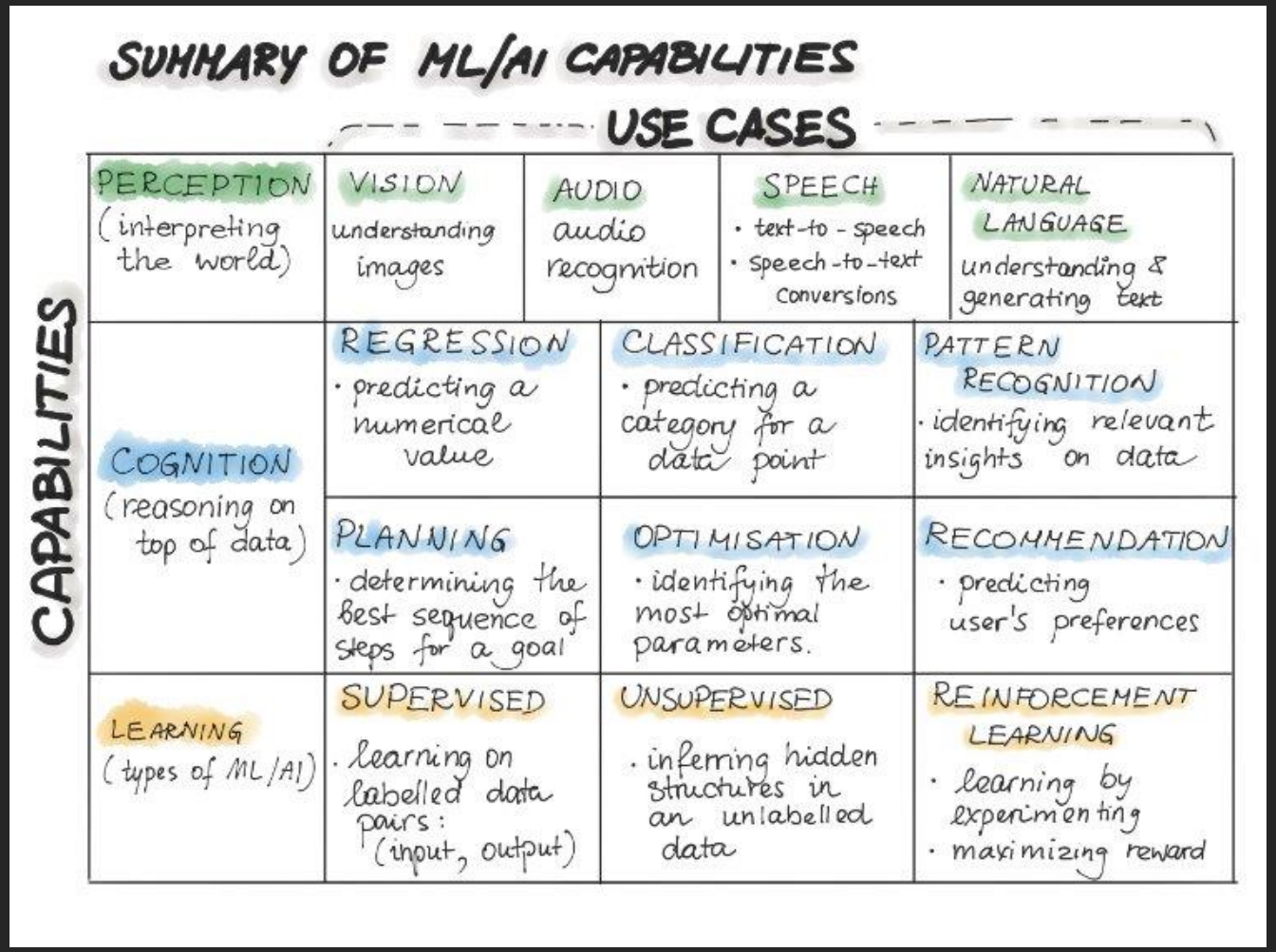


AI: Technique which enables computer to mimic the intelligence or behavior of human.

Machine learning (ML): Subset of AI techniques which enable computer to
improve from experiences without being explicitly programmed.

Deep learning (DL) Subset of ML which uses deep artificial neural networks as a model and automatically build a hierarchy of data representation .

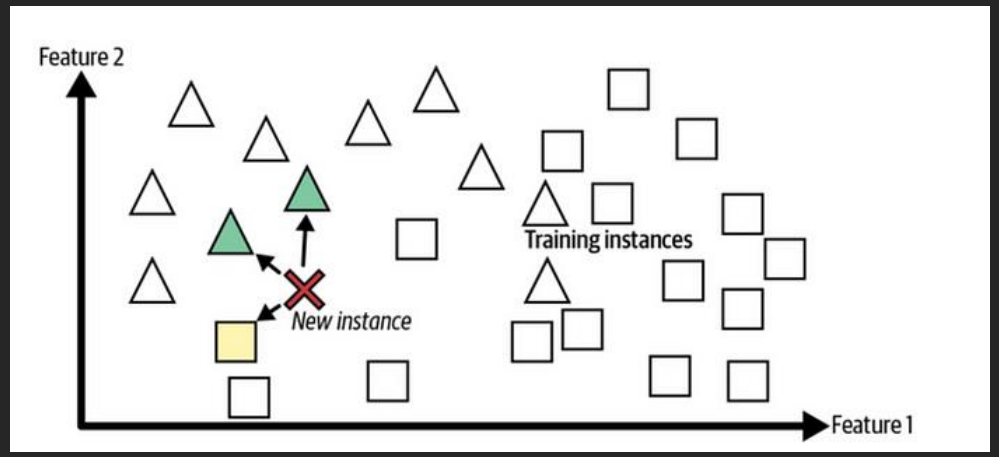
ML/AI capabilities



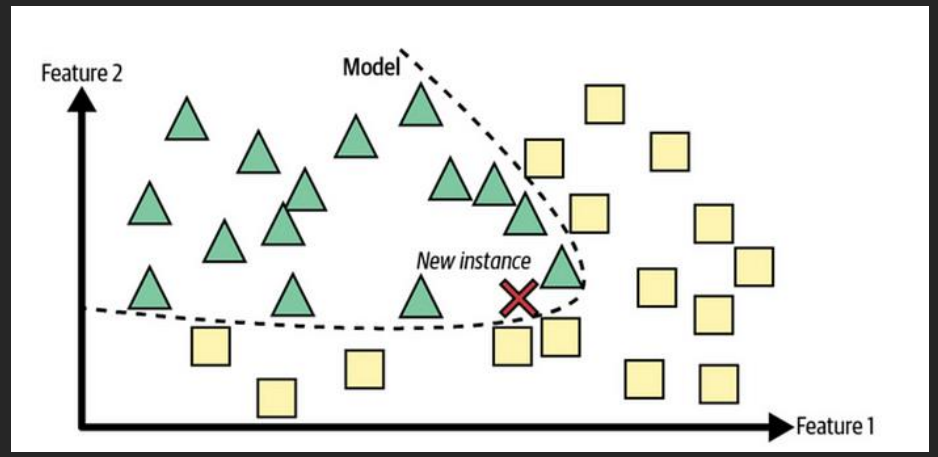
Types of Machine Learning

Instance-based vs Model-based learning

Instance-based

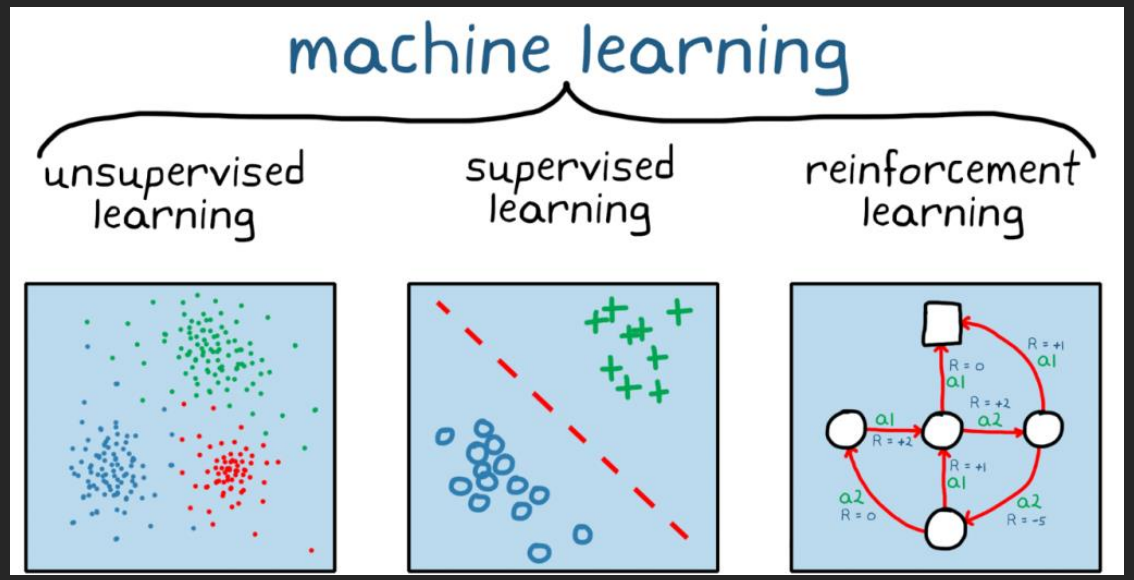


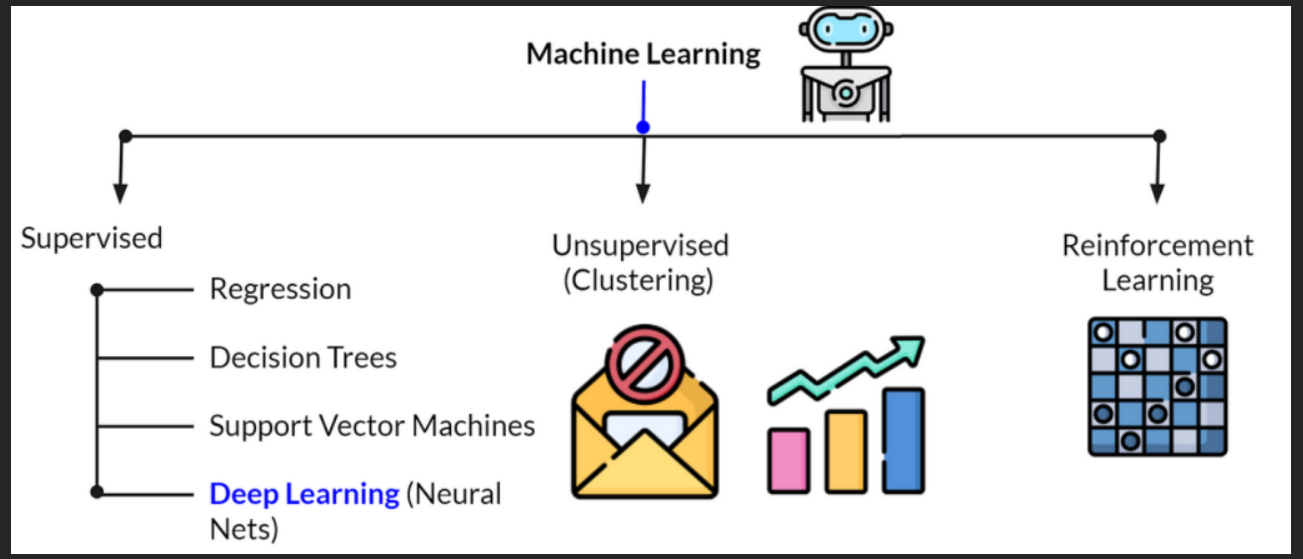
Model-based



Types of Machine Learning

Learn style

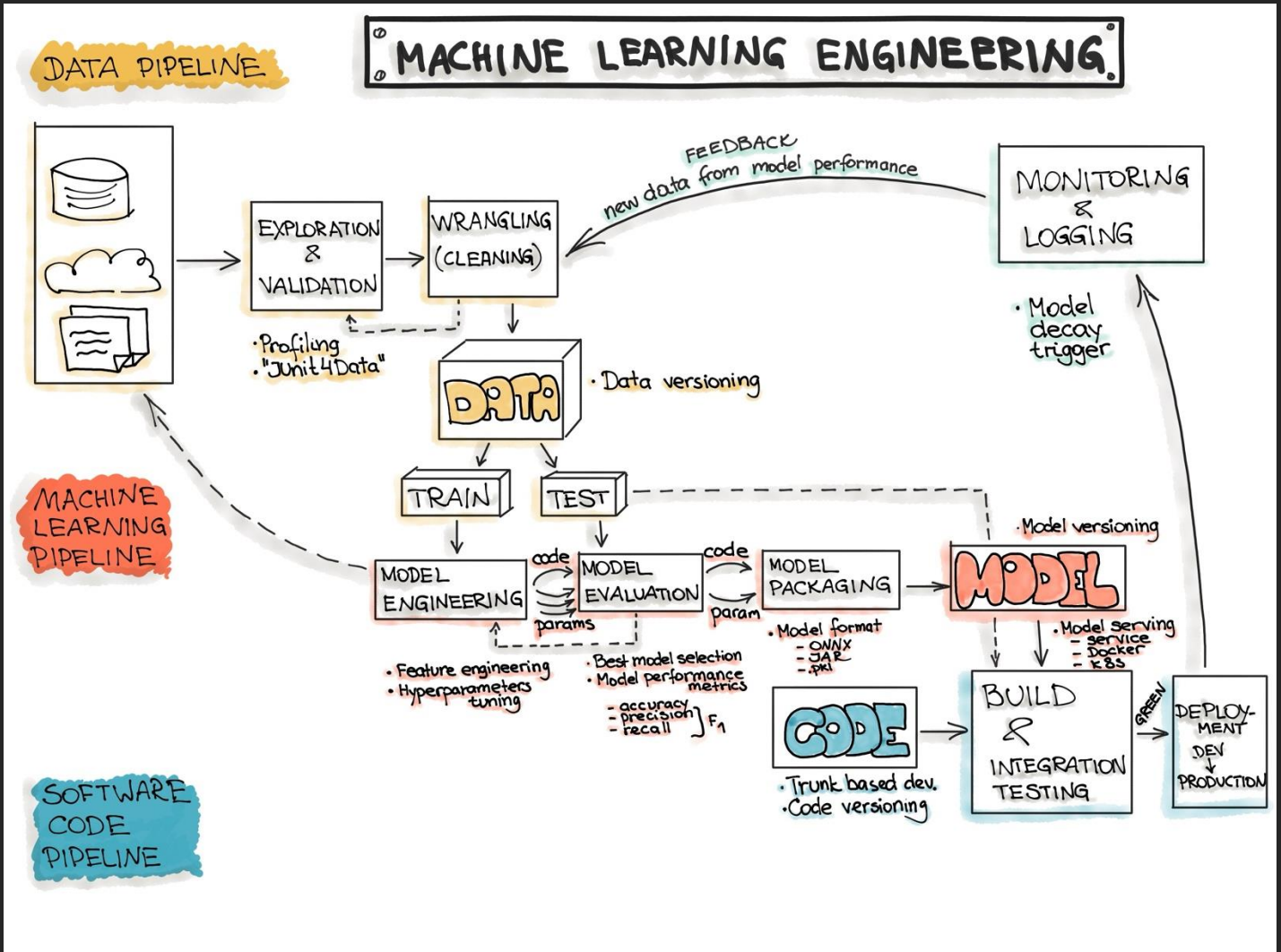




https://www.devopsschool.com/blog/machine-learning-compare-supervised-learning-vs-unsupervised-learning-vs-reinforcement-learning/

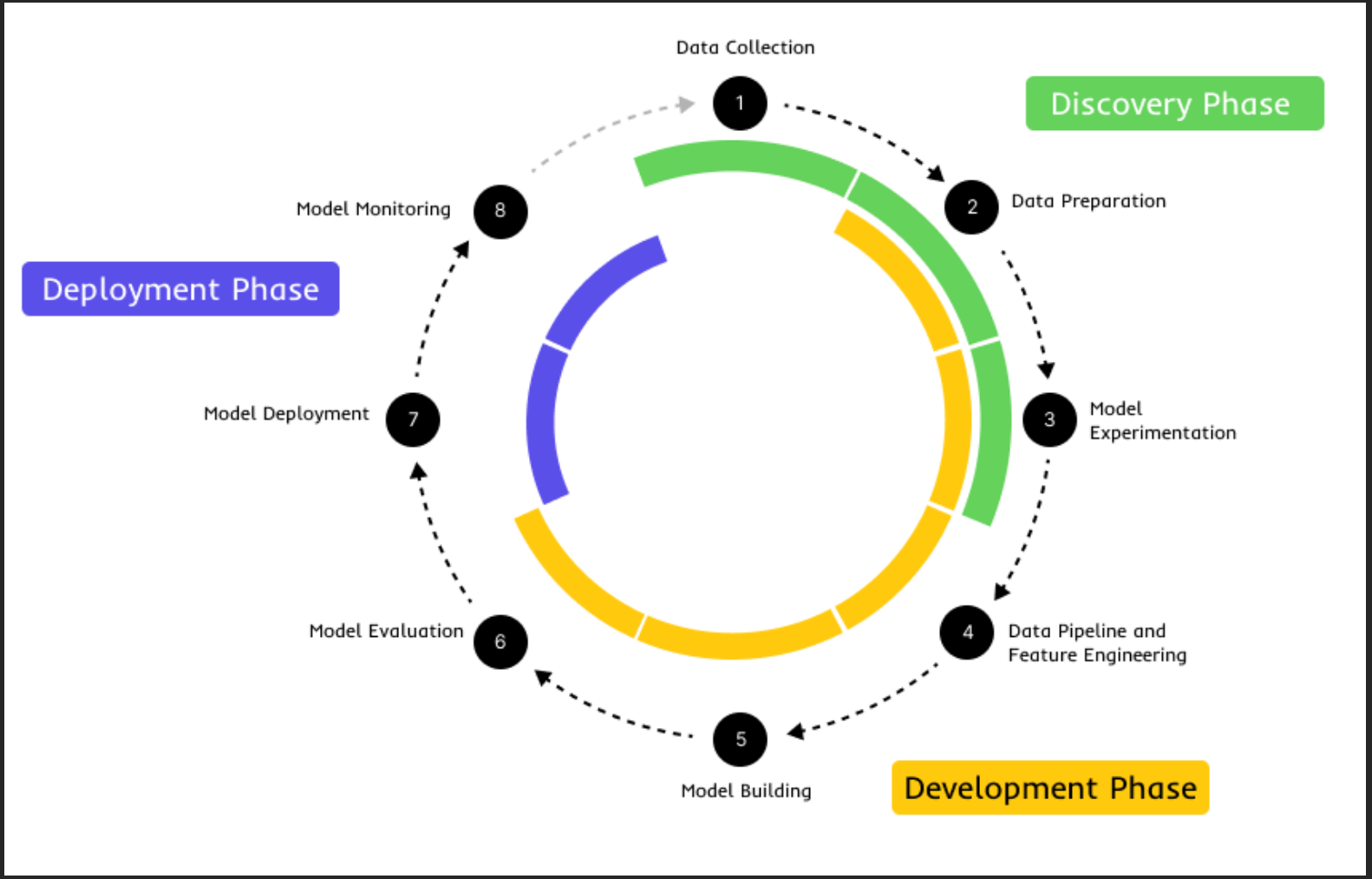
Deep Dive into Machine Learning

Machine Learning Workflow

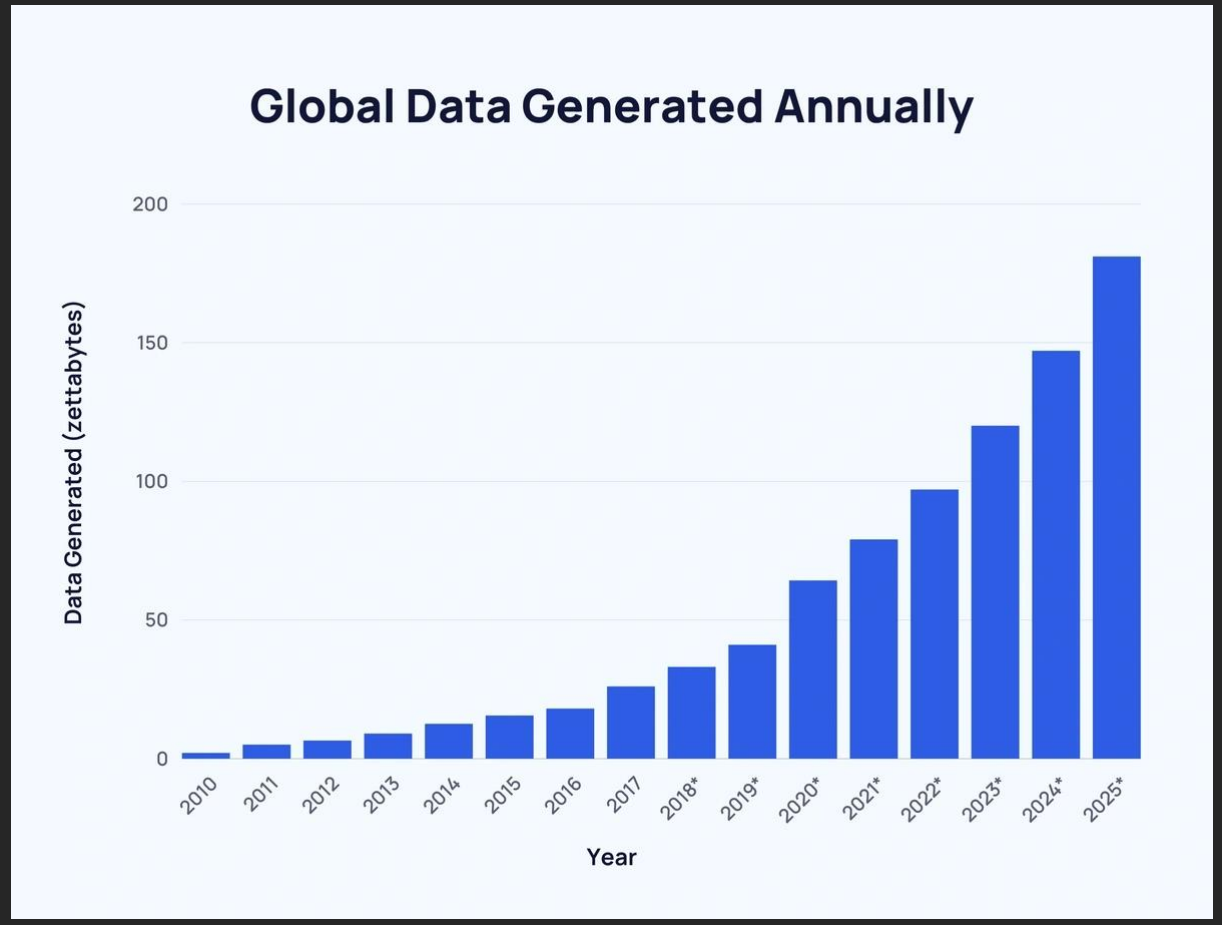


- Data Engineering
  - data acquisition & data preparation
- ML Model Engineering
  - ML model training & serving, and
- Code Engineering
  - integrating ML model into the final product.

Machine Learning Lifecycle

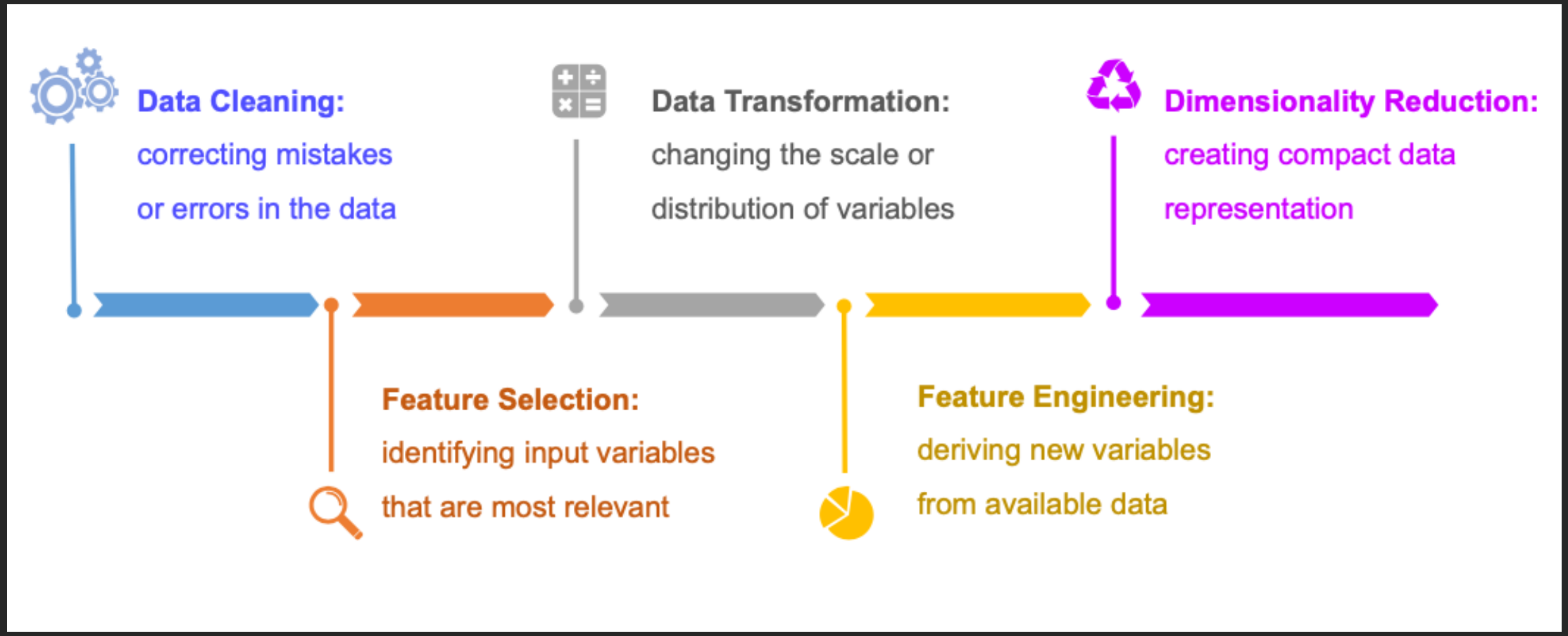


Data Is the New Gold

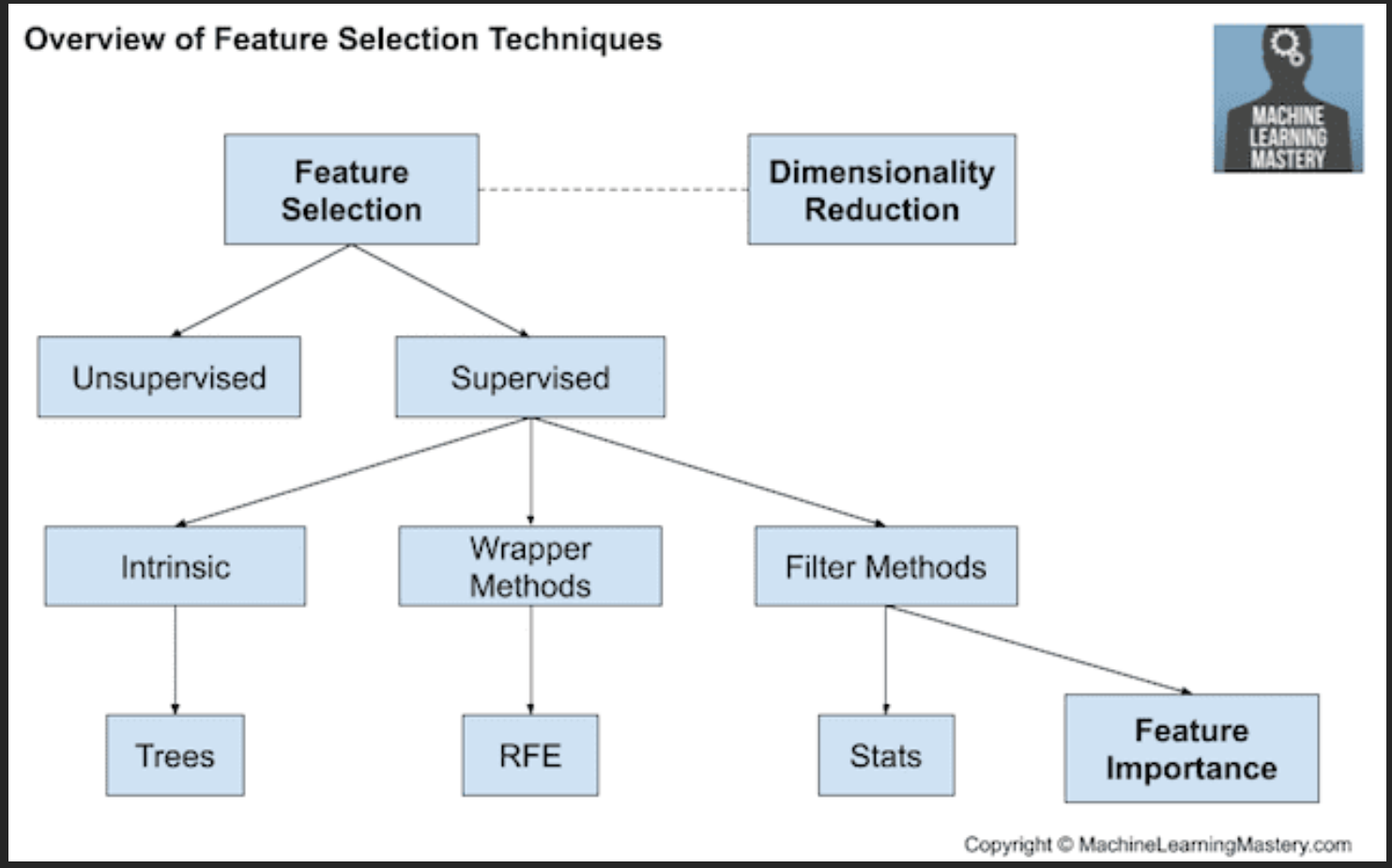


https://explodingtopics.com/blog/data-generated-per-day

Data preparation

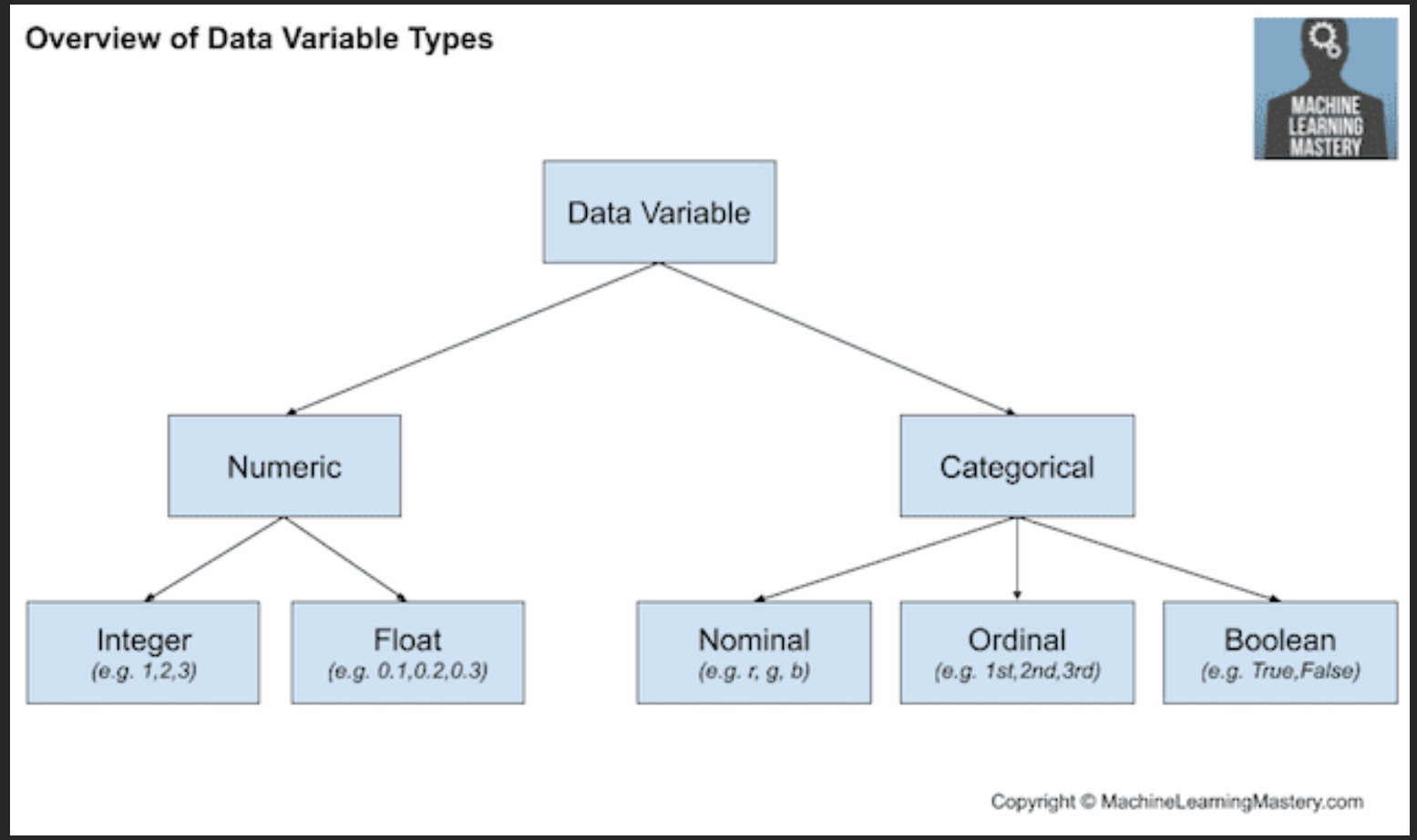


Feature Selection



https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Data Transformation



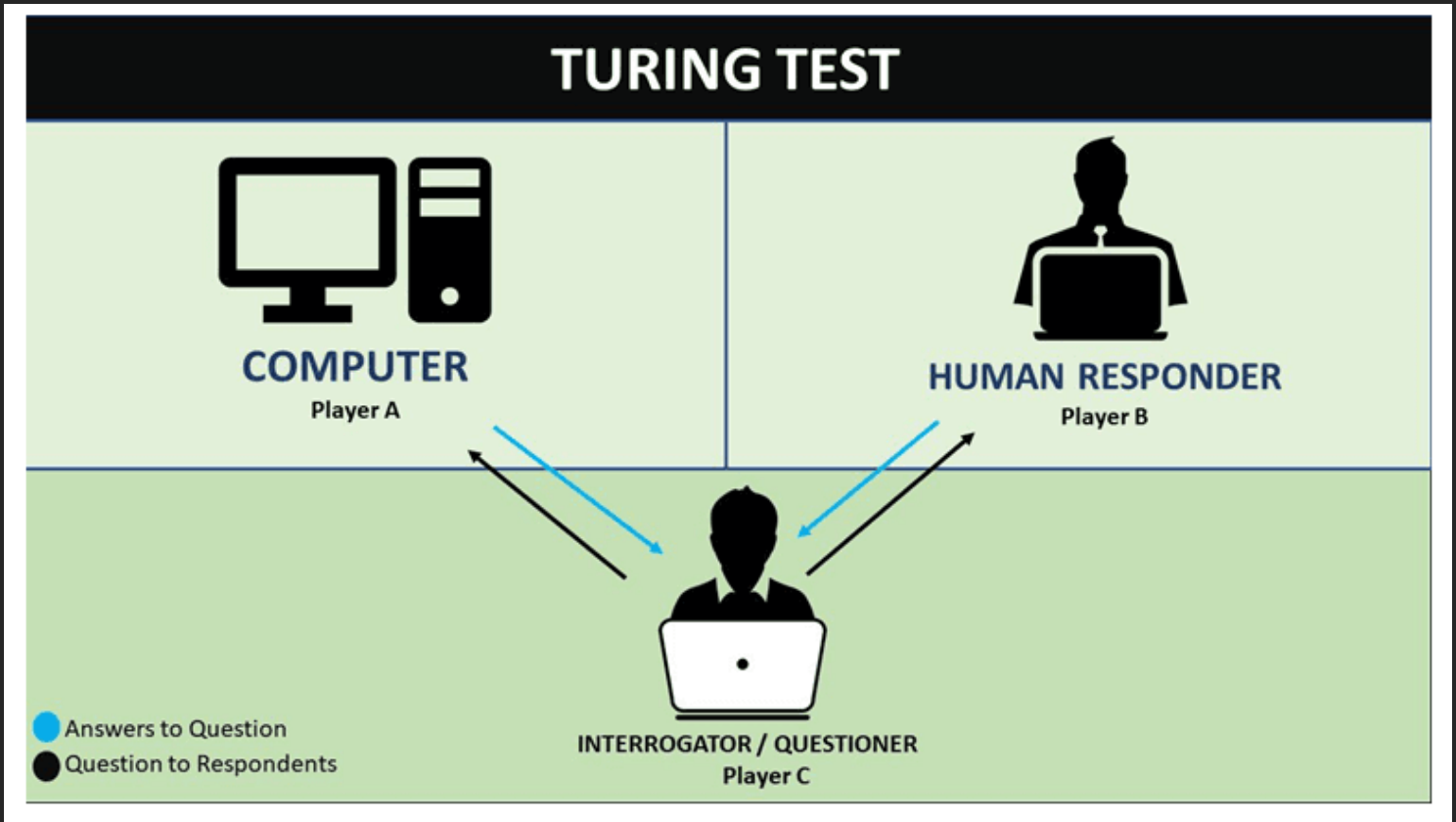
https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Features Engineering

We should clarify how should the input data be represented.

- How do we extract features from raw sources?
- Consider to include domain experts to specify what data aspects are most important for the particular ML task.

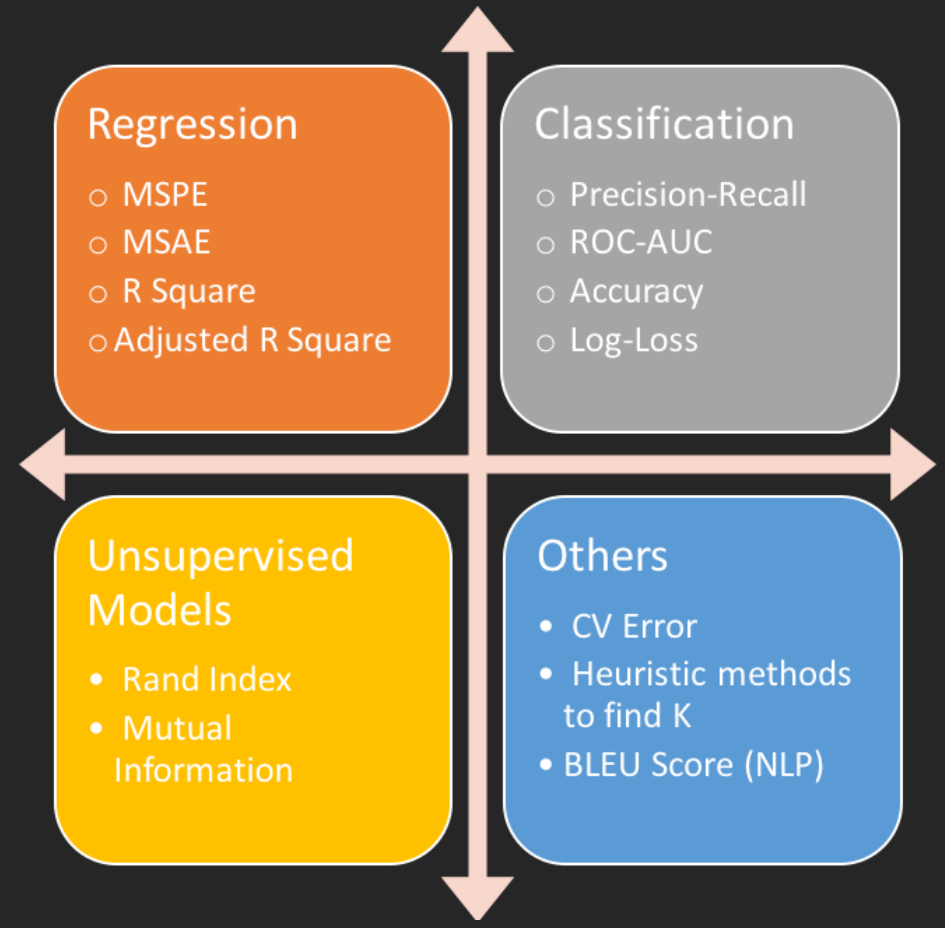
Evaluation



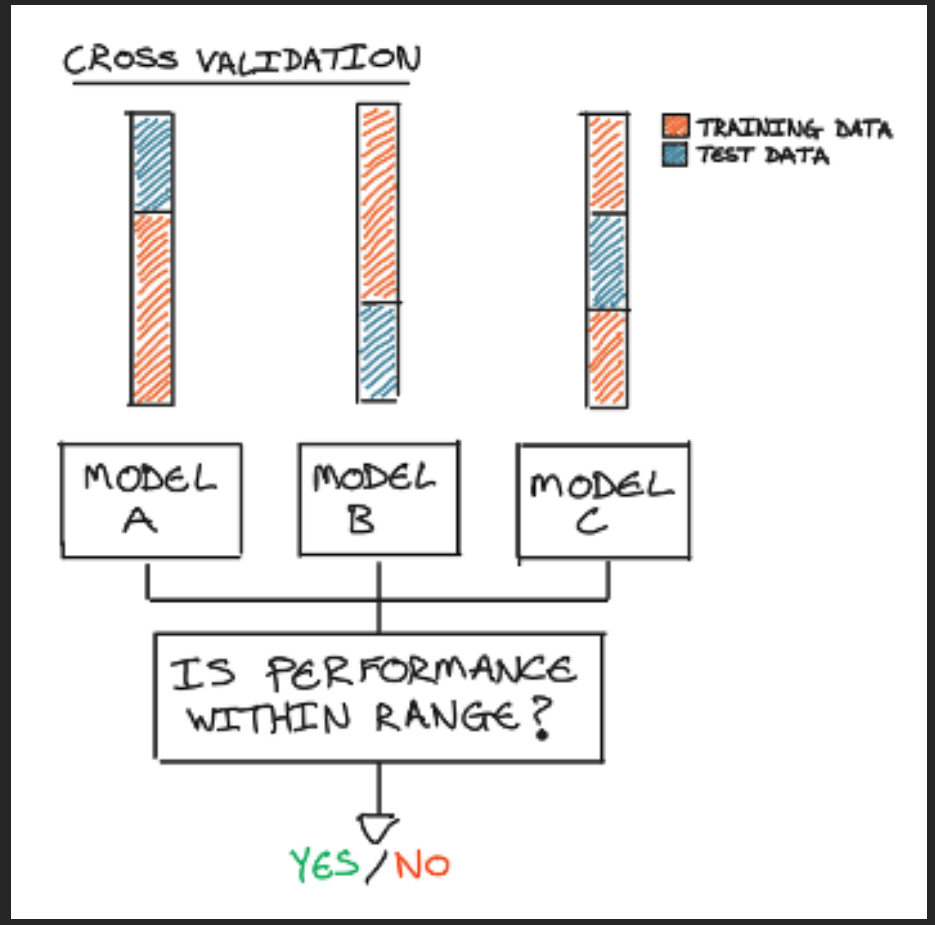
- The Turing Test, proposed by Alan Turing in 1950
- Assesses a machine's ability to exhibit human-like intelligence
- If a human evaluator cannot distinguish between the machine and a human based on their responses, the machine passes the test

https://botpenguin.com/glossary/turing-test

Evaluation - Metrics



Cross-validation



https://towardsdatascience.com/the-5-stages-of-machine-learning-validation-162193f8e5db/

Experiments Tracking - Test

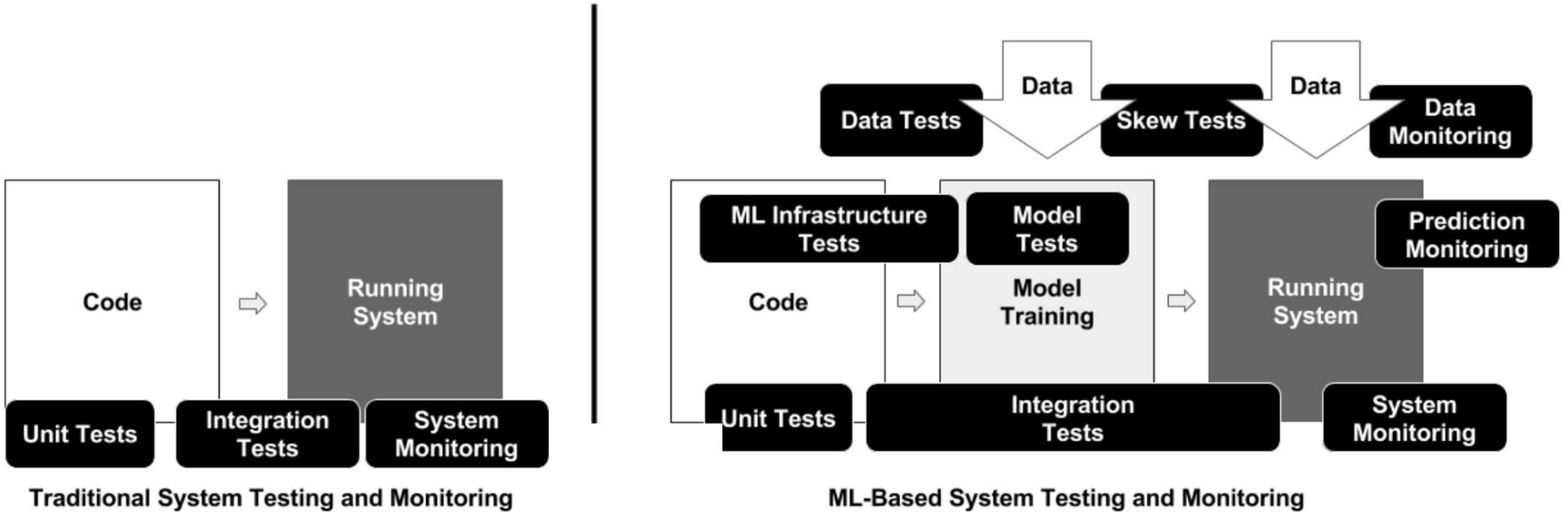
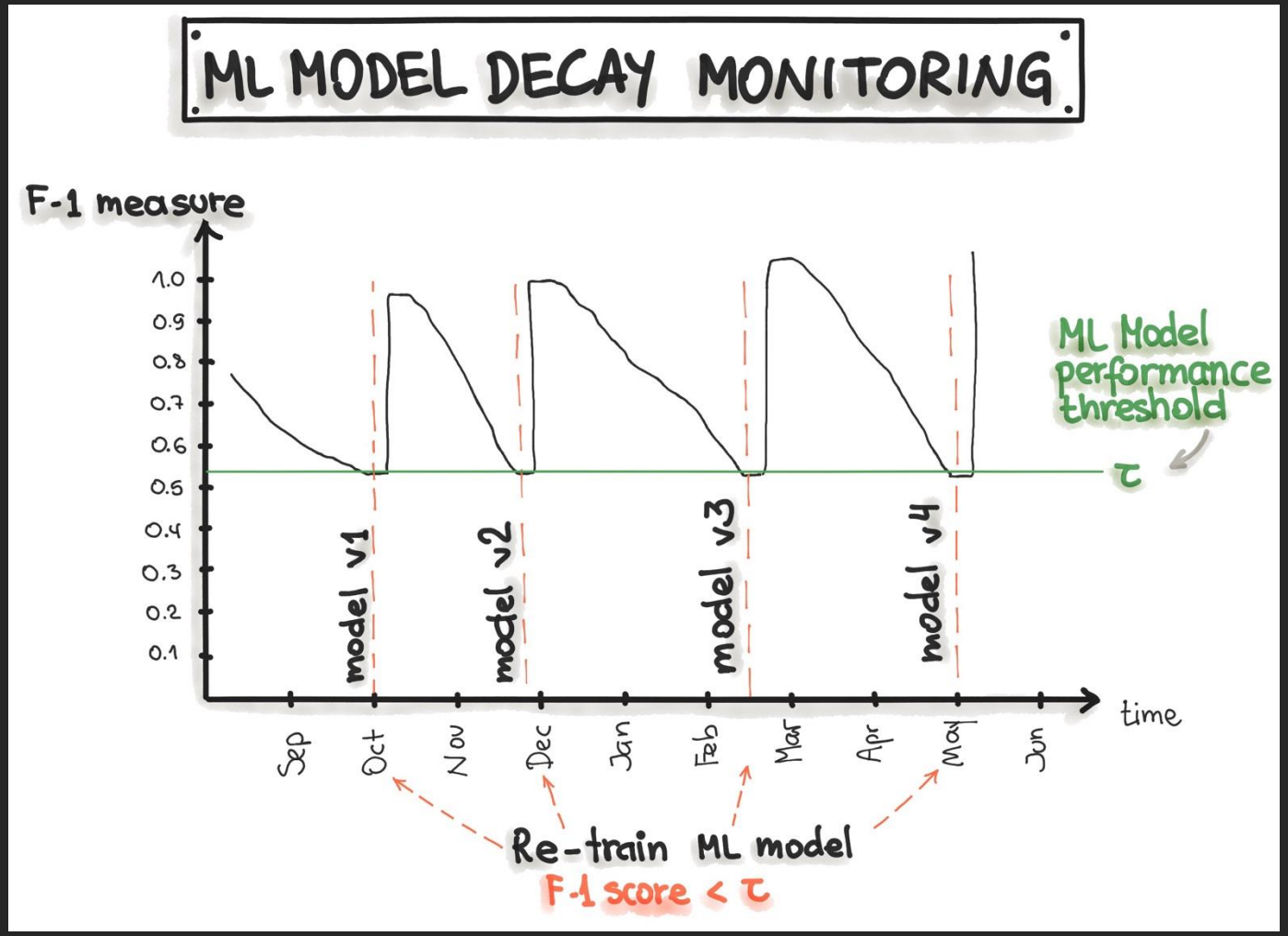
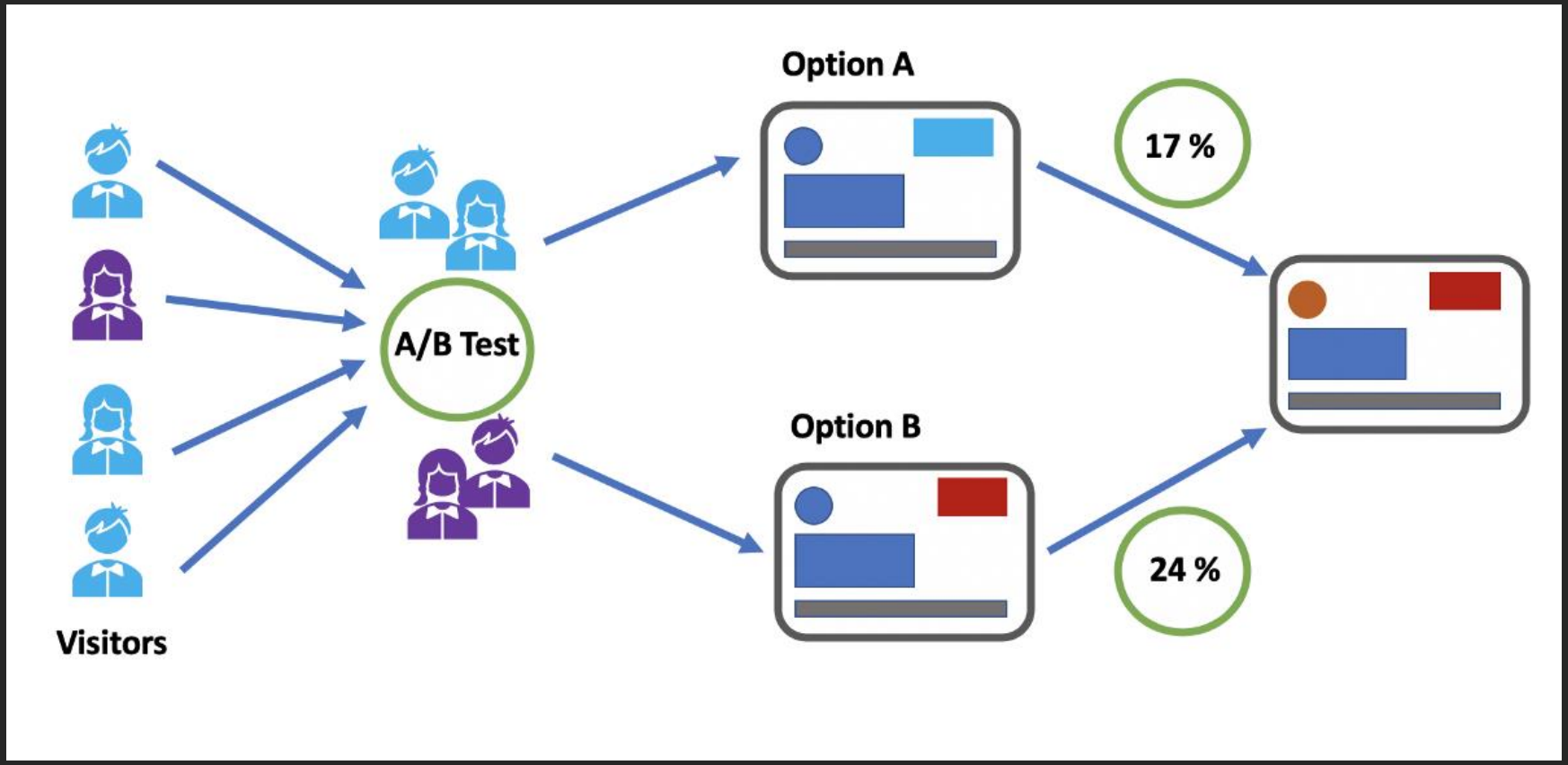


Figure 1. **ML Systems Require Extensive Testing and Monitoring.** The key consideration is that unlike a manually coded system (left), ML-based system behavior is not easily specified in advance. This behavior depends on dynamic qualities of the data, and on various model configuration choices.

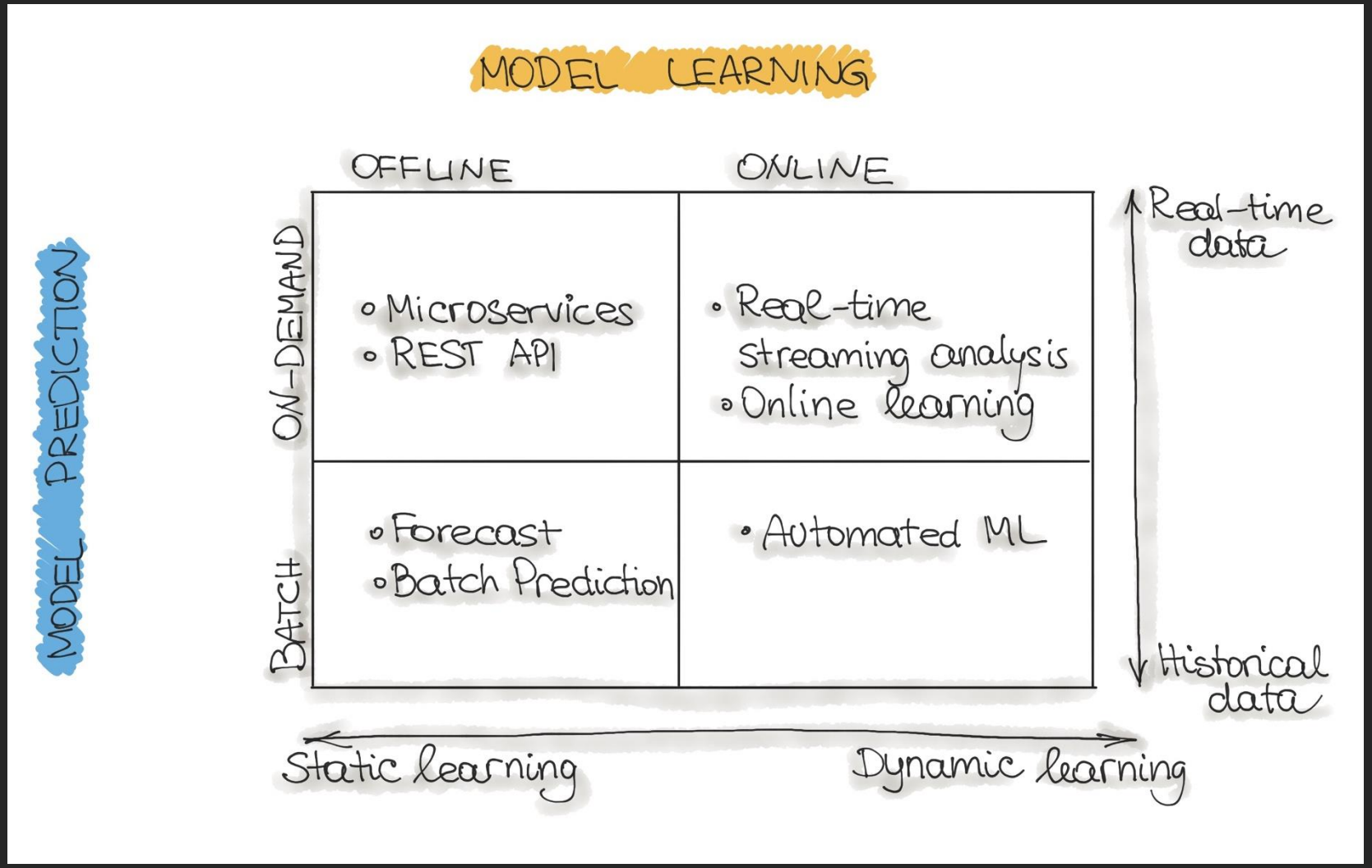
Experiments Tracking - Monitoring



Experiments Tracking - A/B Testing



Model Serving Patterns



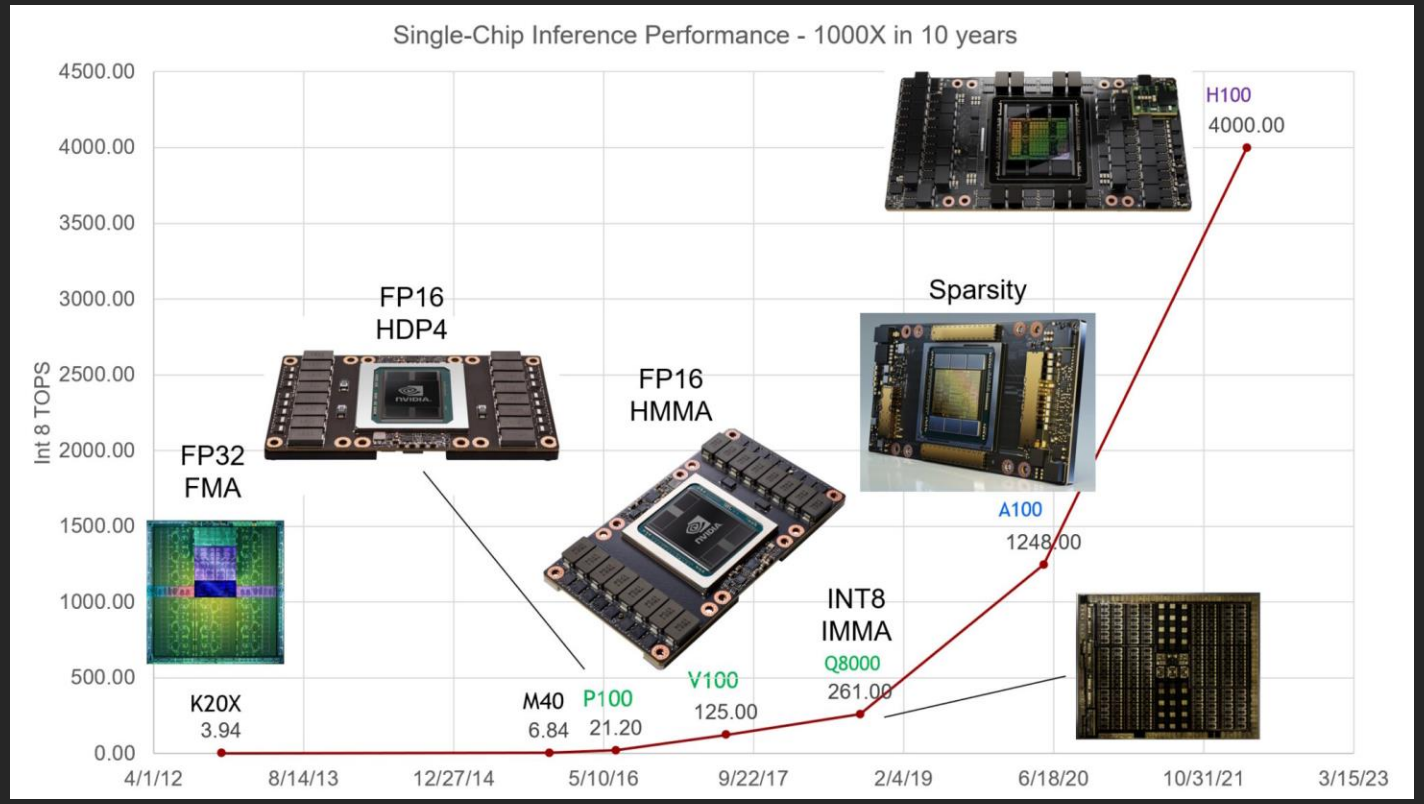
Tools / Platforms

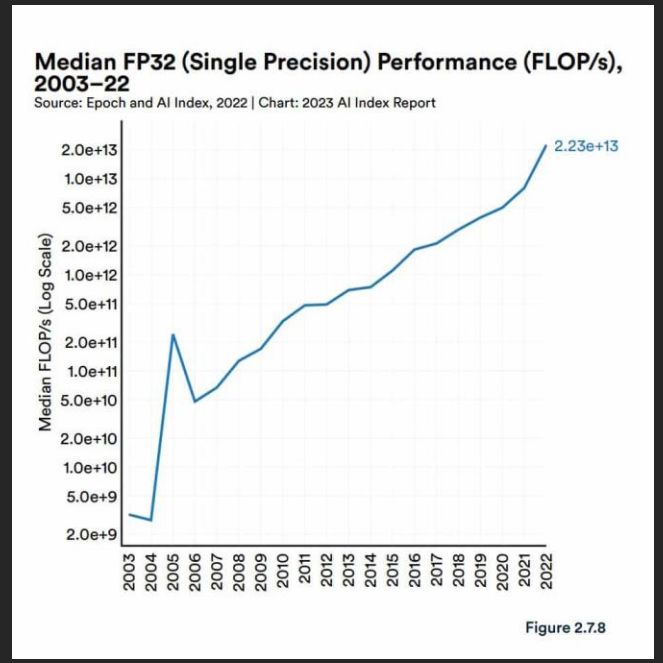
GPU - What is a GPU?

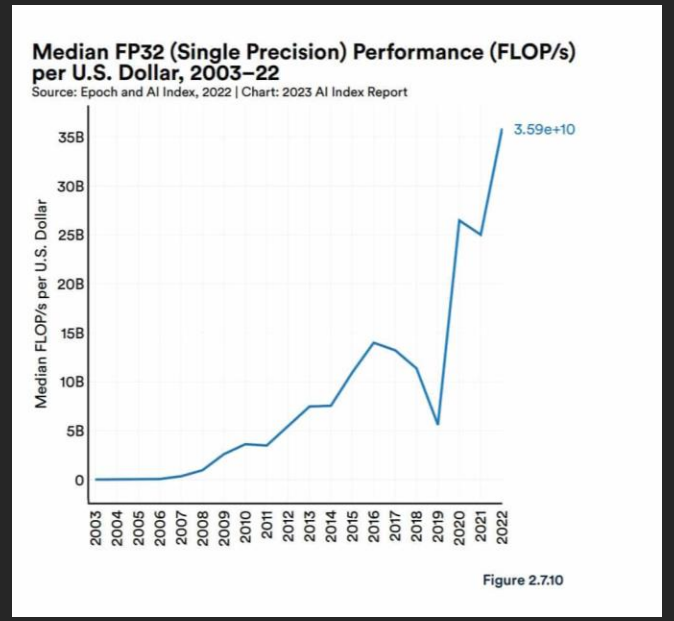
- A Graphics Processing Unit (GPU)
- Specialized processor originally designed to accelerate graphics rendering
- Unlike CPUs (Central Processing Units), which handle general-purpose tasks sequentially
- GPUs are built to process many operations in parallel, making them ideal for certain compute-intensive tasks.

GPU

Models Grow, Systems Expand
The complexity of AI models is expanding a whopping 10x a year.

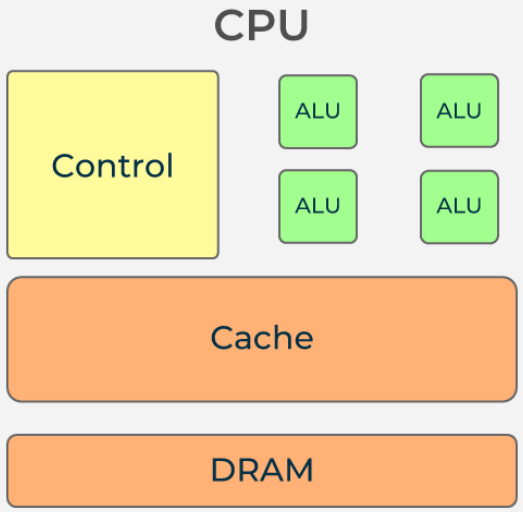






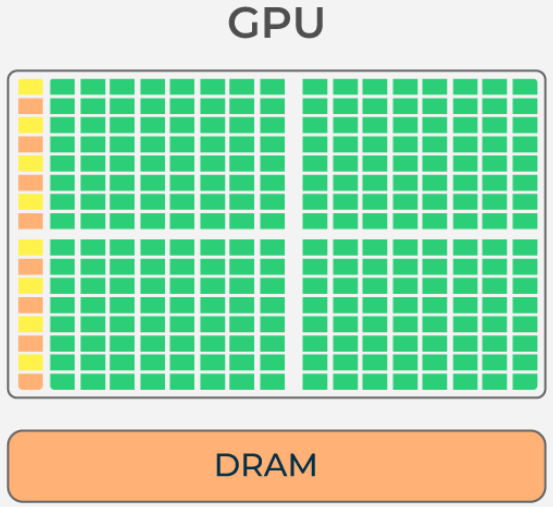
https://blogs.nvidia.com/blog/why-gpus-are-great-for-ai/?utm_source=chatgpt.com

GPU vs CPU



- **CPU (Central Processing Unit)**

- Executes general-purpose tasks (arithmetic, logic).
- Handles complex instructions and task switching.
- CPU has a lower core count (4-8 cores) than others, focusing on delivering higher performance for tasks that rely on a single core.



- **GPU (Graphics Processing Unit)**

- Used for parallel processing tasks (graphics, AI).
- Has many cores (100s or 1000s of cores) for handling multiple operations simultaneously.
- GPUs use a specialized type of DRAM known as VRAM (Video RAM), specifically designed to handle graphical data and textures.

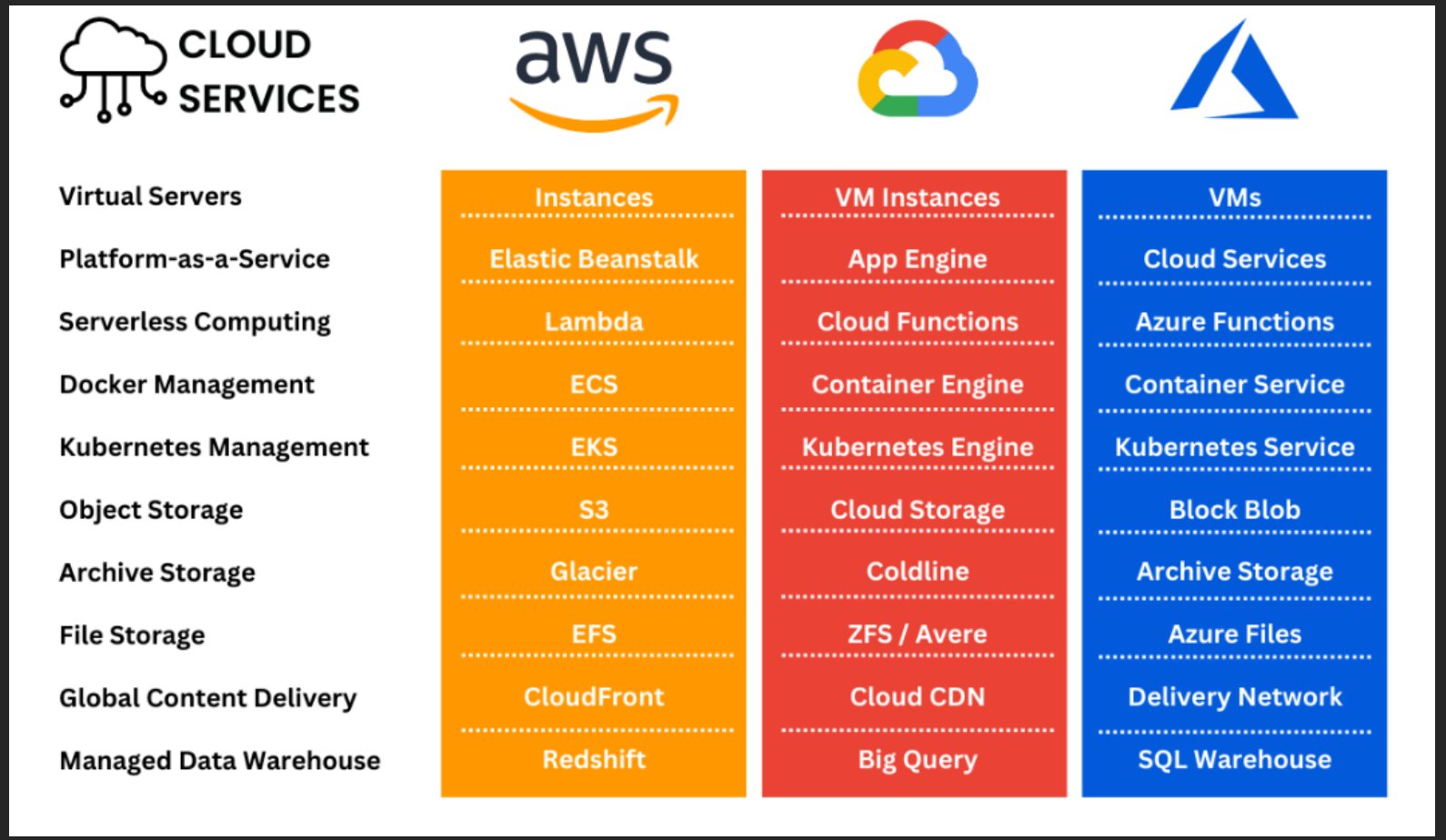
GPU vs CPU

CPU	GPU
Used for General-purpose computation.	Used for Specialized computation for graphics and parallel tasks.
Handles single-threaded, complex tasks.	Handles highly parallel tasks (e.g., graphics rendering).
Optimized for sequential processing.	Optimized for parallel processing.
Smaller cache memory (L1, L2, L3).	Larger memory (VRAM) optimized for high-speed data transfer.
More energy-efficient for general tasks.	Consumes more power due to parallel processing needs.
CPU emphasis on low latency.	While GPU emphasis on high throughput.
Runs operating system, applications, and tasks.	Handles graphics rendering, AI, machine learning.
Generally less expensive.	More expensive due to specialized hardware.

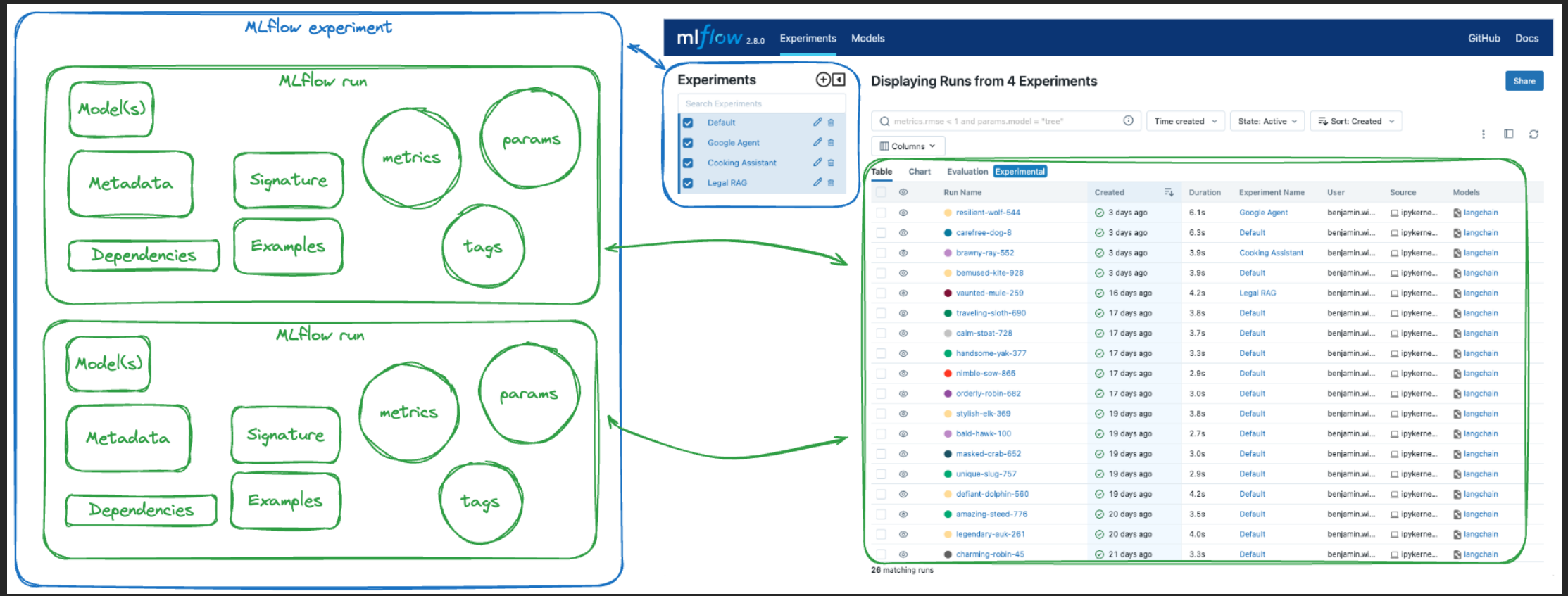
Cloud

- Massive Compute Power
- Scalability and Elasticity: Cloud infrastructure can elastically scale resources up or down based on AI workload demands.
- Cost Efficiency: Cloud's pay-as-you-go model reduces capital expenses and makes advanced AI capabilities accessible to both large enterprises and smaller organizations. (???)
- Collaboration and Accessibility
- Automation and Integration
- Hybrid and Multi-Cloud Flexibility

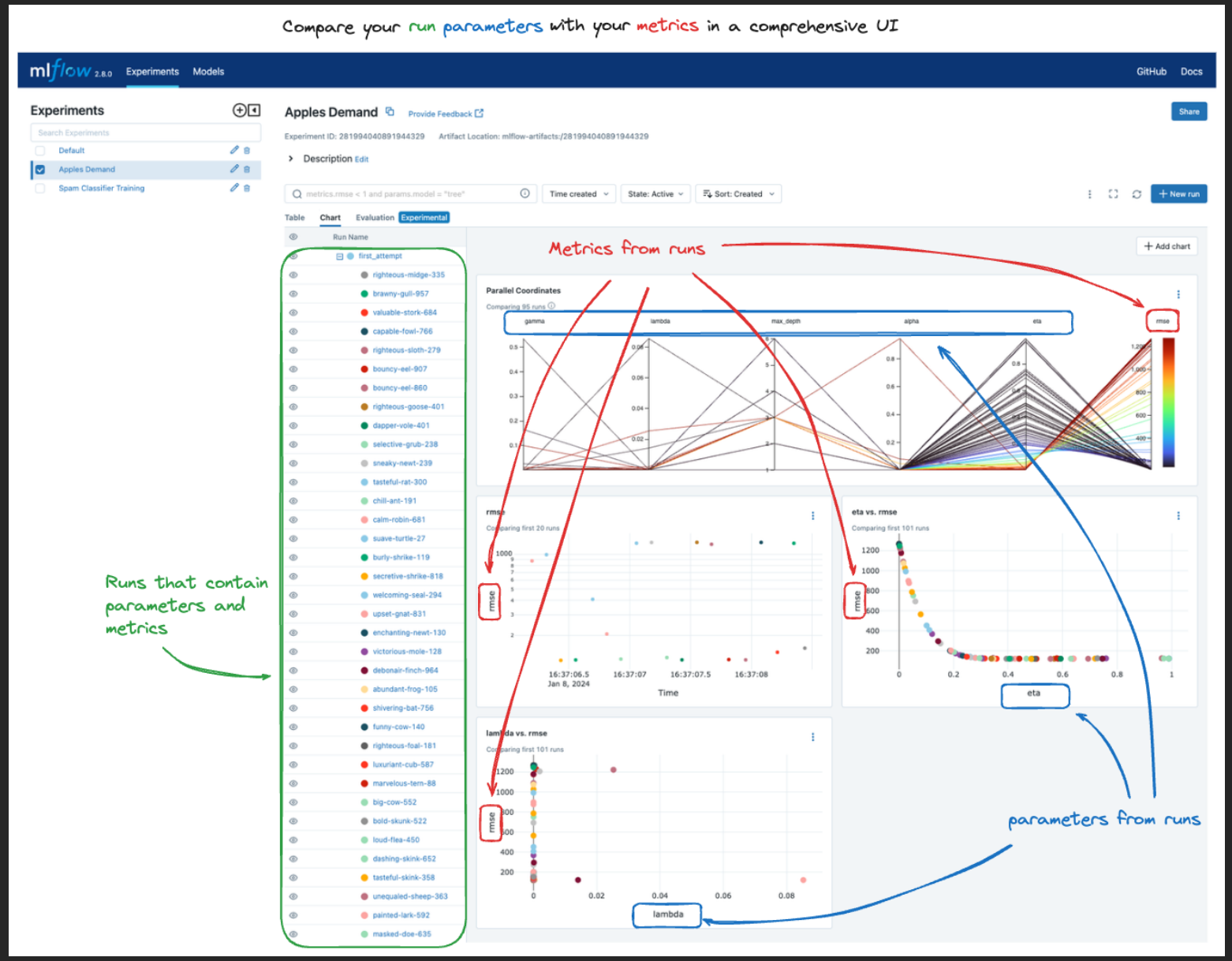
Cloud


CLOUD SERVICES
aws
Virtual Servers
Instances
VM Instances
VMs
Platform-as-a-Service
Elastic Beanstalk
App Engine
Cloud Services
Serverless Computing
Lambda
Cloud Functions
Azure Functions
Docker Management
ECS
Container Engine
Container Service
Kubernetes Management
EKS
Kubernetes Engine
Kubernetes Service
Object Storage
S3
Cloud Storage
Block Blob
Archive Storage
Glacier
Coldline
Archive Storage
File Storage
EFS
ZFS / Avere
Azure Files
Global Content Delivery
CloudFront
Cloud CDN
Delivery Network
Managed Data Warehouse
Redshift
Big Query
SQL Warehouse

MLflow

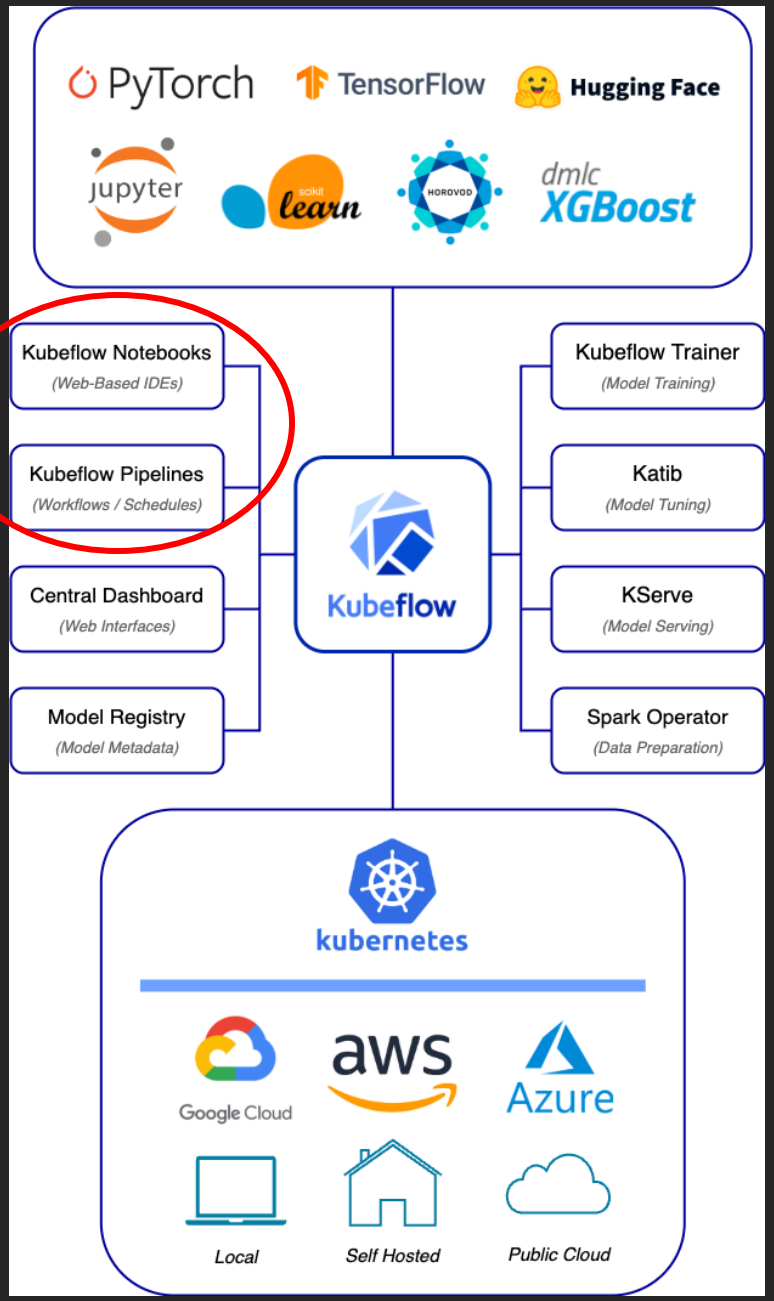

MLflow experiment
MLflow run
Model(s)
Metadata
Signature
metrics
params
Dependencies
Examples
tags
MLflow run
Model(s)
Metadata
Signature
metrics
params
Dependencies
Examples
tags
mlflow 2.8.0
Experiments
Models
GitHub
Docs
Experiments
Search Experiments
Default
Google Agent
Cooking Assistant
Legal RAG
Displaying Runs from 4 Experiments
Share
metrics.rmse < 1 and params.model = "tree"
Time created
State: Active
Sort: Created
Columns
Table
Chart
Evaluation
Experimental
Run Name
Created
Duration
Experiment Name
User
Source
Models
resilient-wolf-544 3 days ago 6.1s Google Agent benjamin.wi... ipykerne... langchain
carefree-dog-8 3 days ago 6.3s Default benjamin.wi... ipykerne... langchain
brawny-ray-552 3 days ago 3.9s Cooking Assistant benjamin.wi... ipykerne... langchain
bemused-kite-928 3 days ago 3.9s Default benjamin.wi... ipykerne... langchain
vaunted-mule-259 16 days ago 4.2s Legal RAG benjamin.wi... ipykerne... langchain
traveling-sloth-690 17 days ago 3.8s Default benjamin.wi... ipykerne... langchain
calm-stoat-728 17 days ago 3.7s Default benjamin.wi... ipykerne... langchain
handsome-yak-377 17 days ago 3.3s Default benjamin.wi... ipykerne... langchain
nimble-sow-865 17 days ago 2.9s Default benjamin.wi... ipykerne... langchain
orderly-robin-682 17 days ago 3.0s Default benjamin.wi... ipykerne... langchain
stylish-elk-369 19 days ago 3.8s Default benjamin.wi... ipykerne... langchain
bald-hawk-100 19 days ago 2.7s Default benjamin.wi... ipykerne... langchain
masked-crab-652 19 days ago 3.0s Default benjamin.wi... ipykerne... langchain
unique-slug-757 19 days ago 2.9s Default benjamin.wi... ipykerne... langchain
defiant-dolphin-560 19 days ago 4.2s Default benjamin.wi... ipykerne... langchain
amazing-steed-776 20 days ago 3.5s Default benjamin.wi... ipykerne... langchain
legendary-auk-261 20 days ago 4.0s Default benjamin.wi... ipykerne... langchain
charming-robin-45 21 days ago 3.3s Default benjamin.wi... ipykerne... langchain
26 matching runs

MLflow

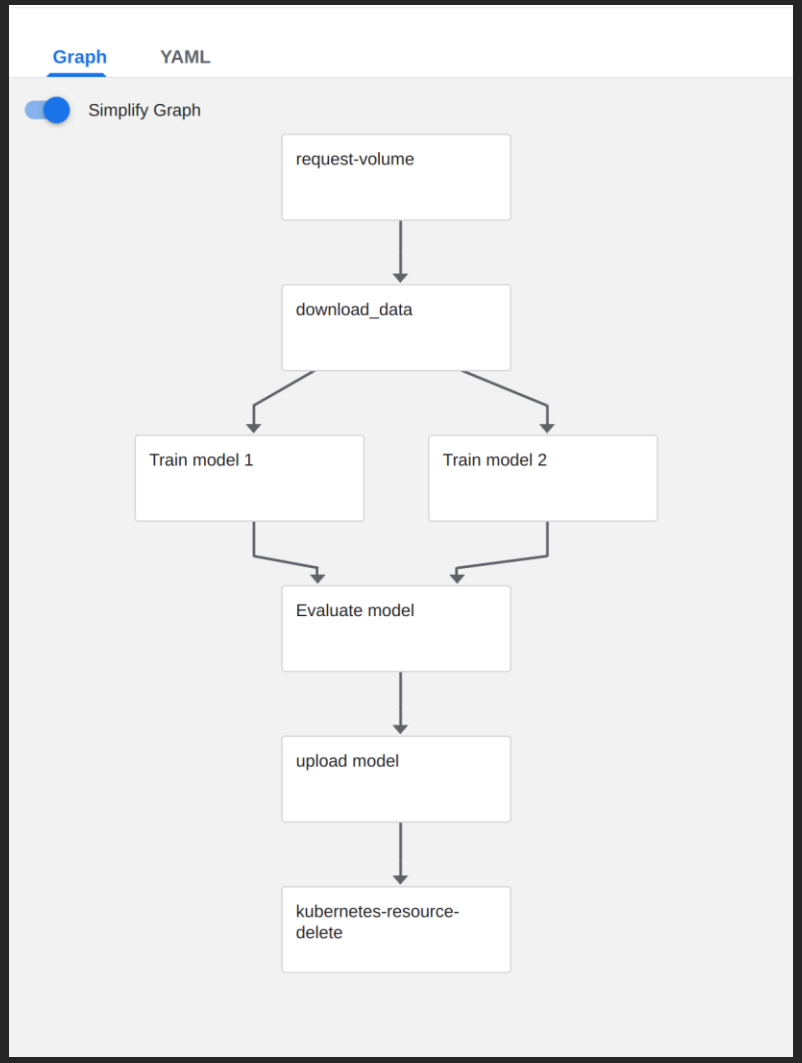

Compare your run parameters with your metrics in a comprehensive UI
mlflow 2.8.0
Experiments
Models
GitHub
Docs
Experiments
Default
Apples Demand
Spam Classifier Training
Apples Demand
Provide Feedback
Share
Description Edit
Time created
State: Active
Sort: Created
New run
Table
Chart
Evaluation
Experimental
Run Name
first_attempt
Metrics from runs
Add chart
Parallel Coordinates
gamma
lambda
max_depth
alpha
eta
rmse
Runs that contain parameters and metrics
Time
eta vs. rmse
lambda vs. rmse
parameters from runs

Kubeflow

- Kubeflow is a community and ecosystem of open-source projects to address each stage in the machine learning (ML) lifecycle with support for best-in-class open source tools and frameworks. Kubeflow makes AI/ML on Kubernetes simple, portable, and scalable.

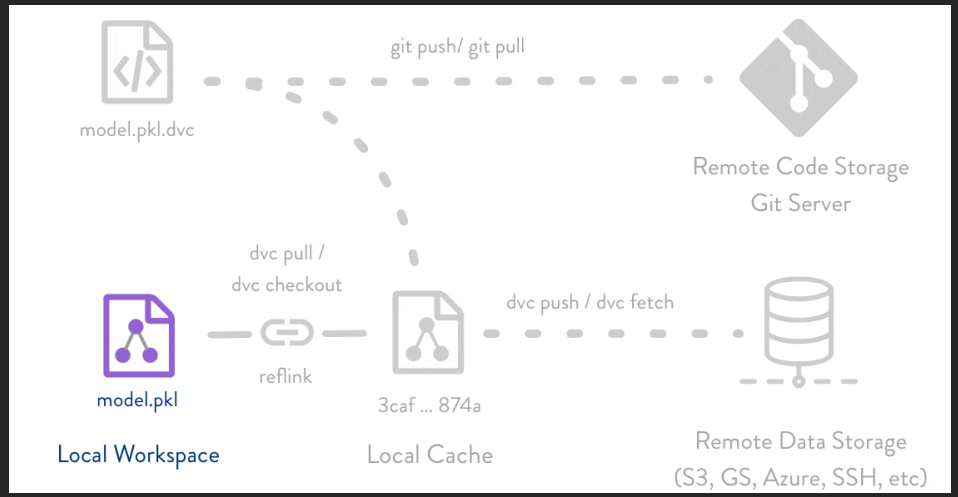


Kubeflow - Pipeline



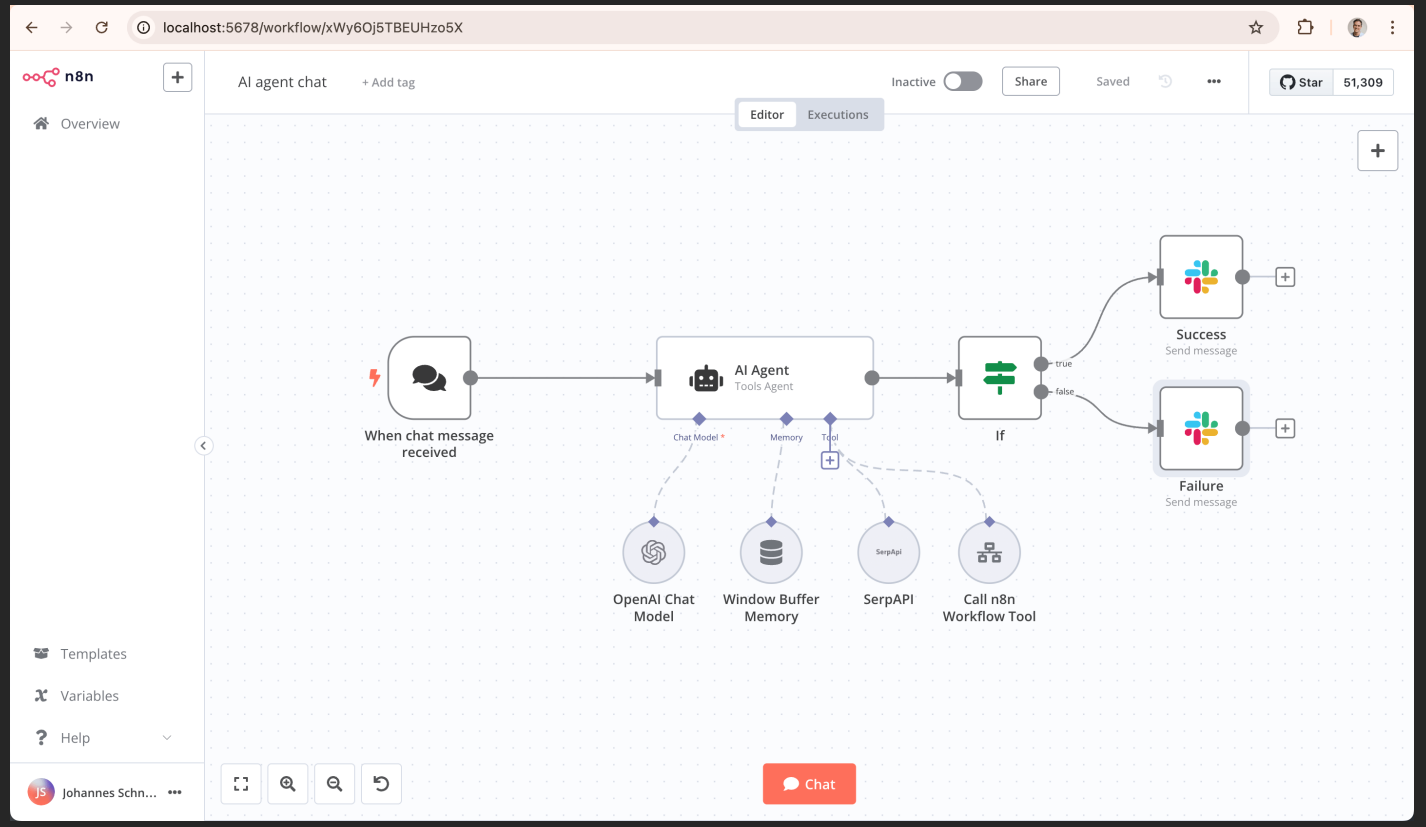
DVC (Data Version Control)

- Version your data and models. Store them in your cloud storage but keep their version info in your Git repo.
- Iterate fast with lightweight pipelines. When you make changes, only run the steps impacted by those changes.
- Track experiments in your local Git repo (no servers needed).
- Compare any data, code, parameters, model, or performance plots.
- Share experiments and automatically reproduce anyone's experiment.

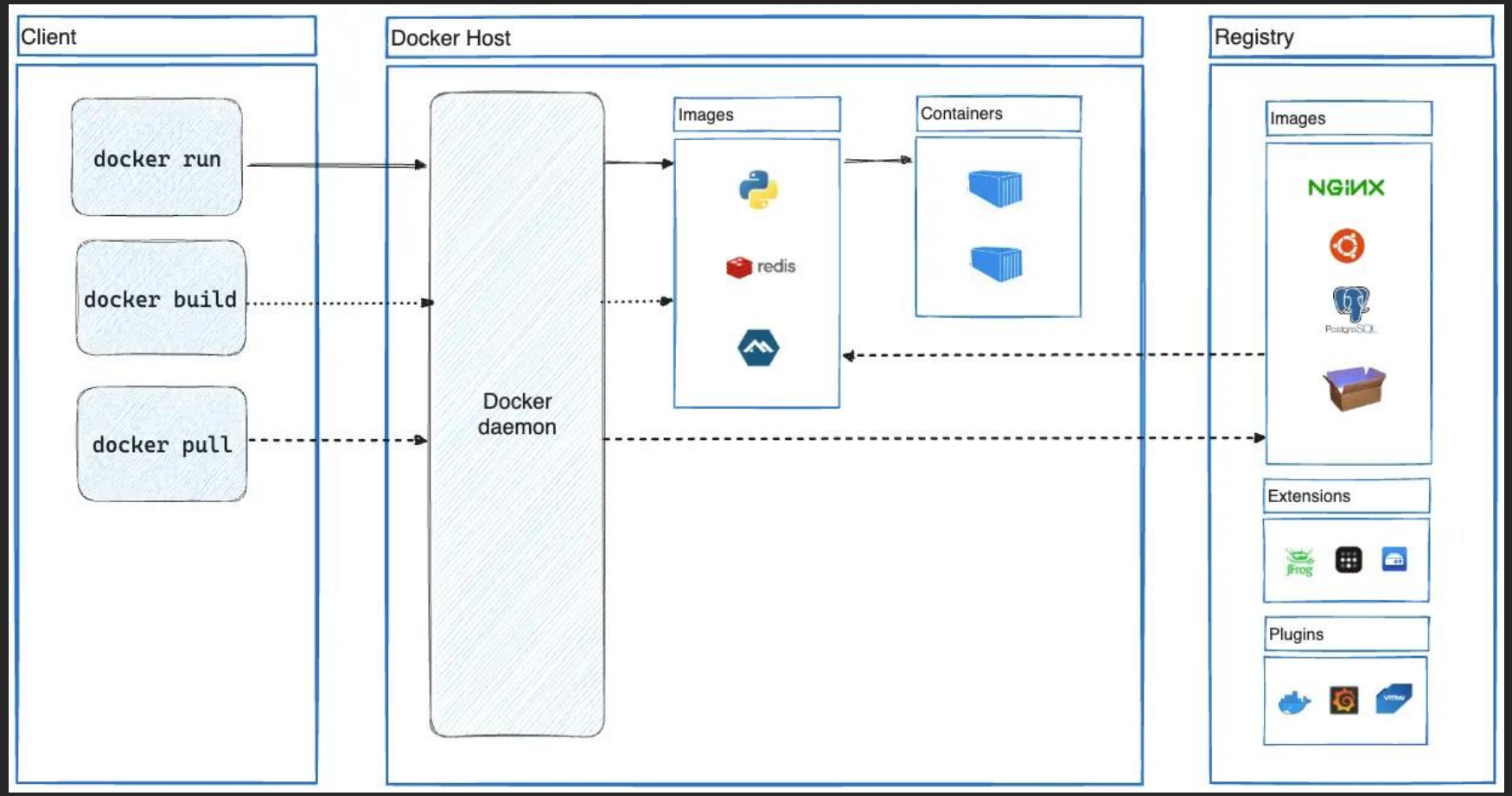


n8n

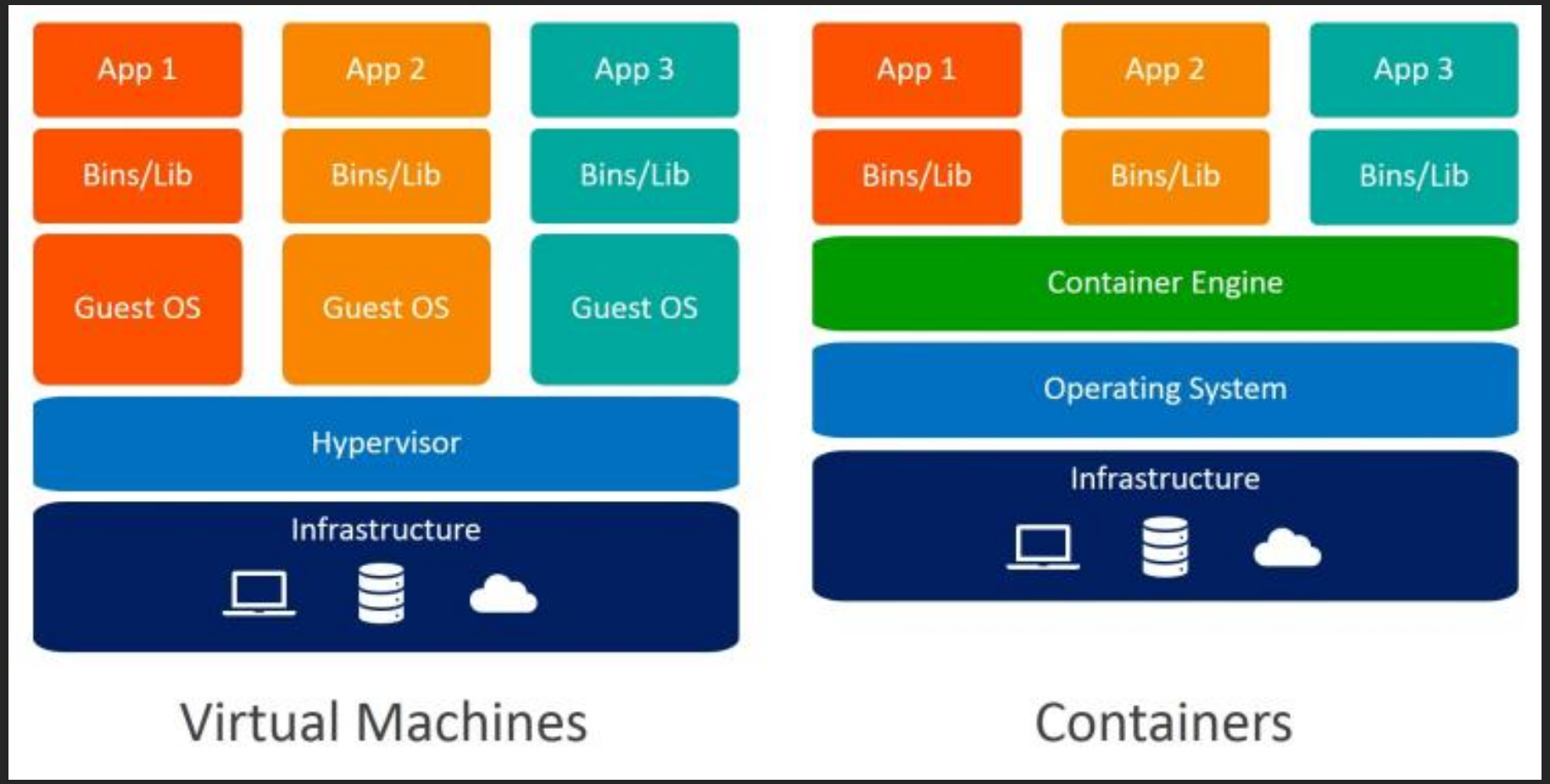
- AI agent workflow automation



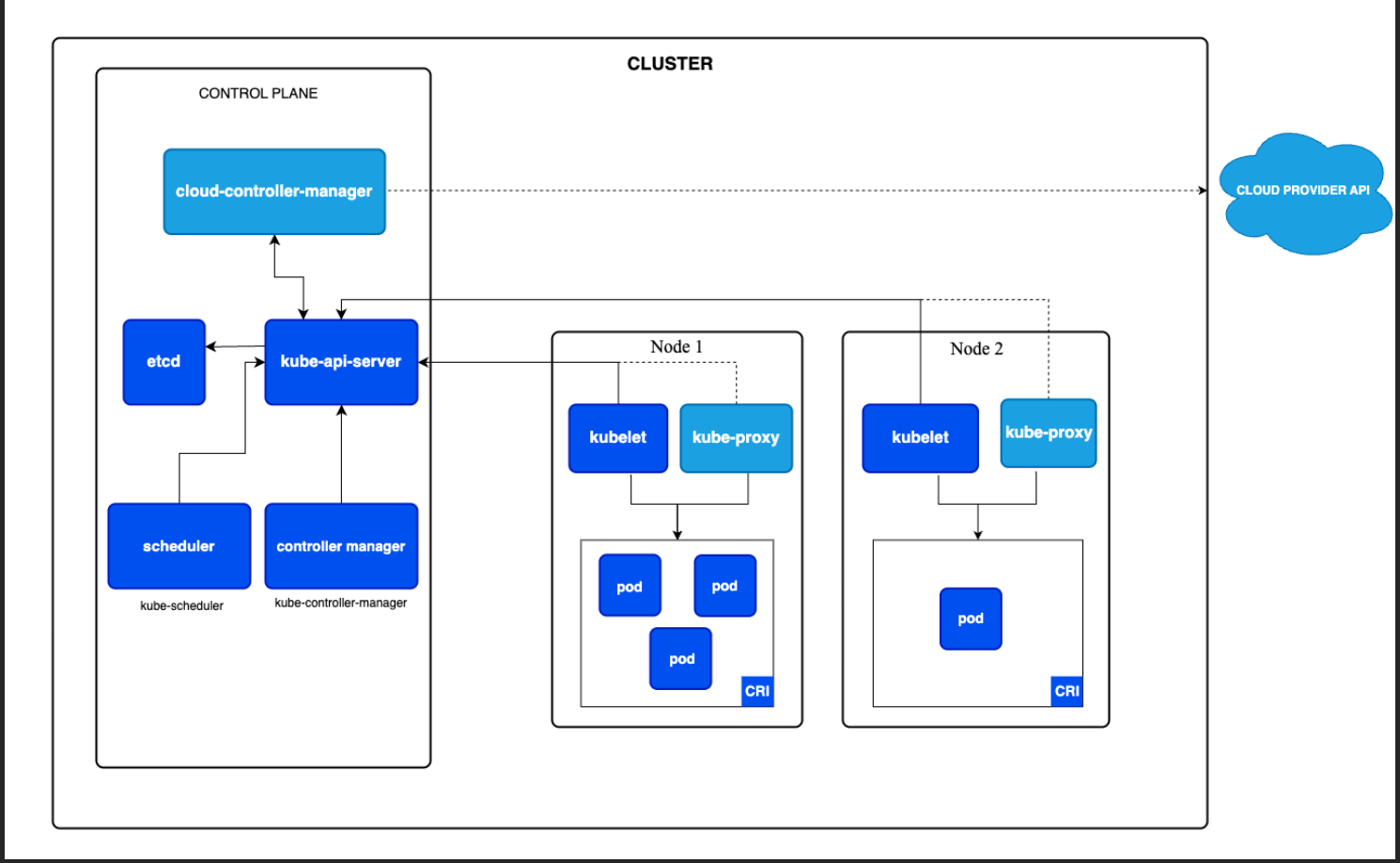
Docker (Container)



Docker (Container)

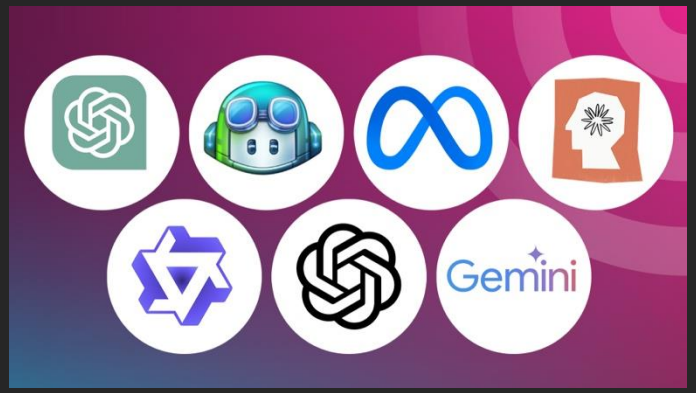


Kubernetes


CLUSTER
CONTROL PLANE
cloud-controller-manager
CLOUD PROVIDER API
etcd
kube-api-server
scheduler
controller manager
kube-scheduler
kube-controller-manager
Node 1
kubelet
kube-proxy
pod
pod
pod
CRI
Node 2
kubelet
kube-proxy
pod
CRI

Current Trends in AI/ML

Generative AI



Text: LLM, ChatGPT

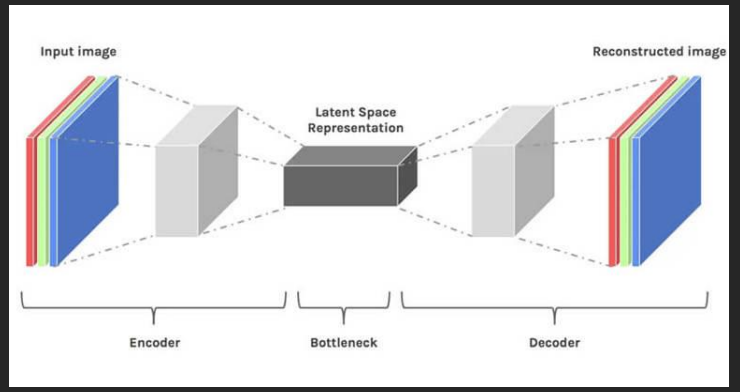
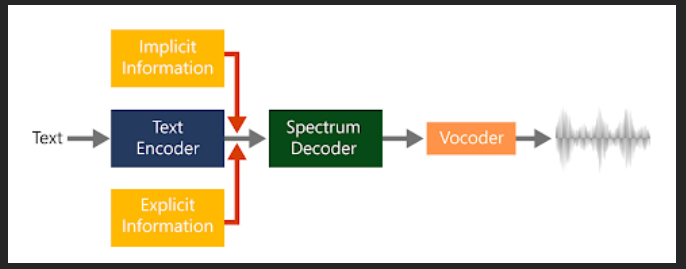


Image: Stable diffusion, DALL-E, Deepfake

Voice: Text-to-Spech, Retrieval-based Voice Conversion



LLM (Large Language Model)

- A large language model is an AI system trained on vast amounts of text data to understand and generate human-like text for various language tasks, though it can sometimes produce inaccurate or biased outputs.

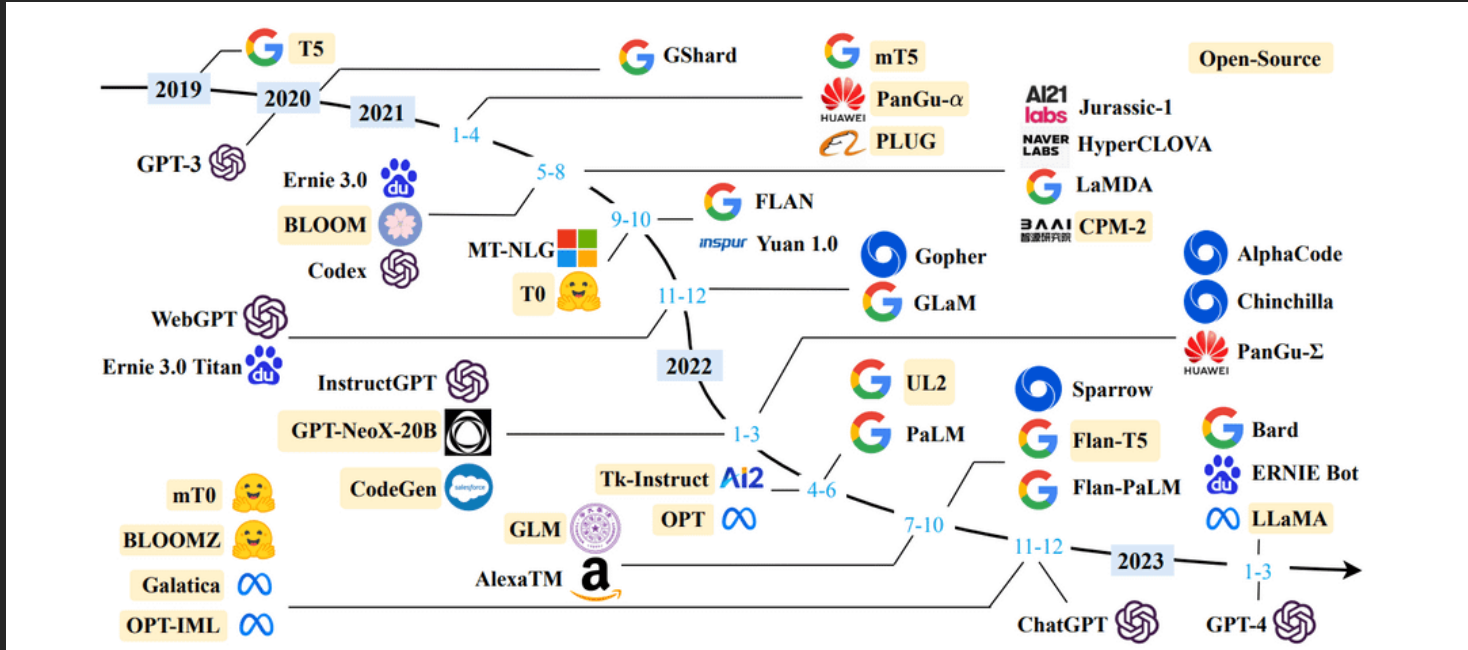
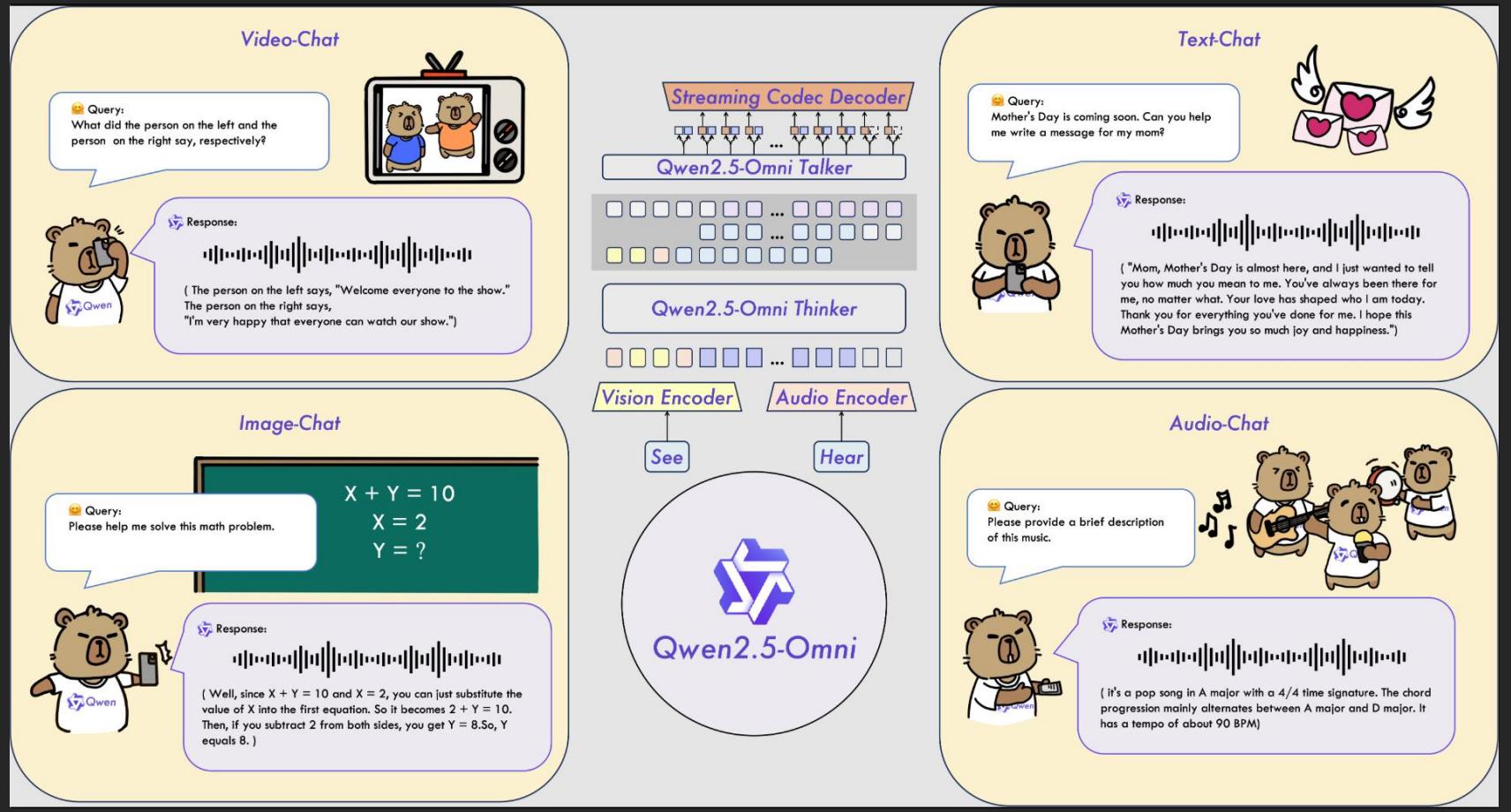


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

https://www.reddit.com/r/dataisugly/comments/12a8heb/a_timeline_of_large_language_models/?rdt=40625

LLM - OMNI (Multimodal)



https://github.com/QwenLM/Qwen2.5-Omni

LLM - Resoning

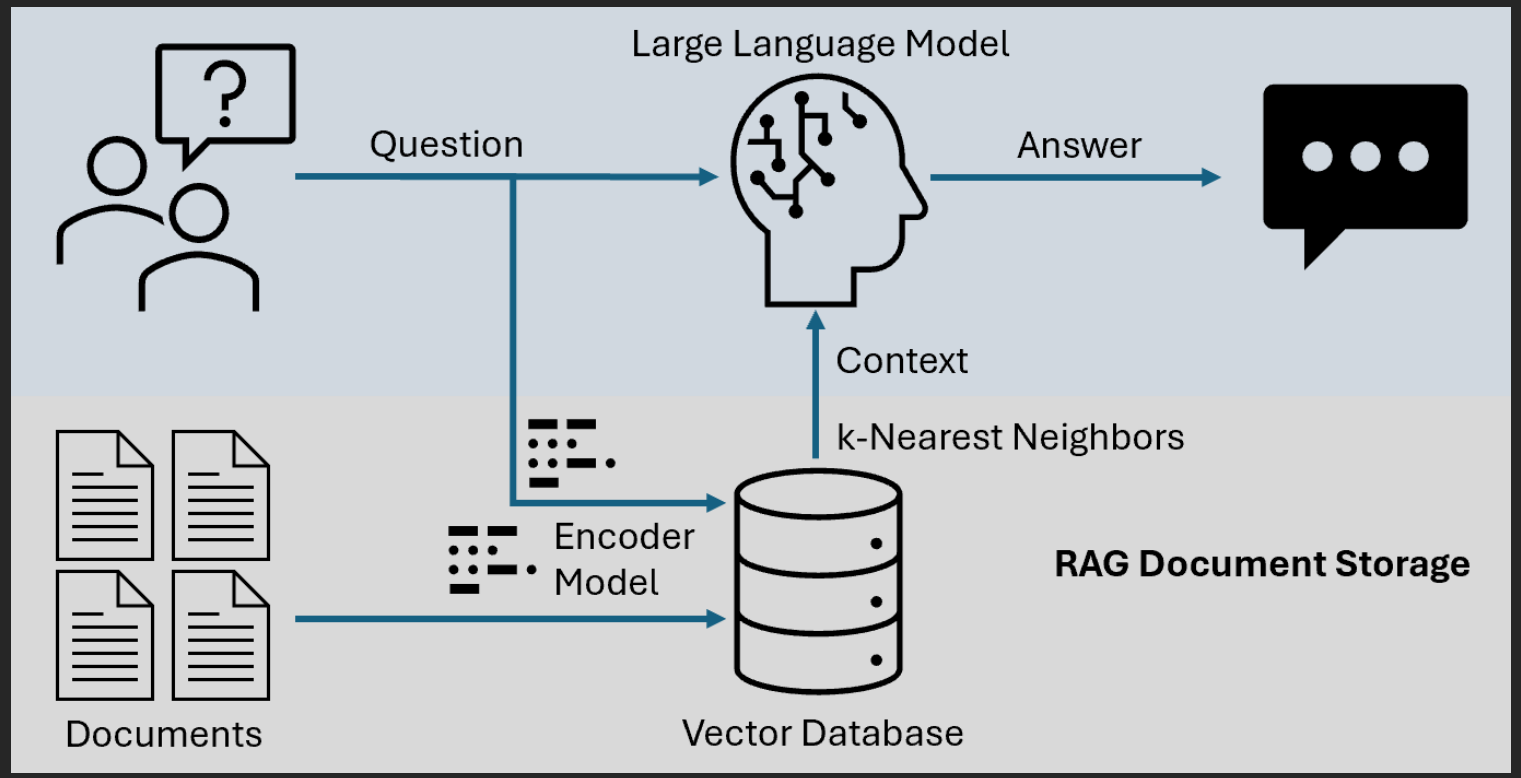
https://api-docs.deepseek.com/news/news250120

Prompt Engineering

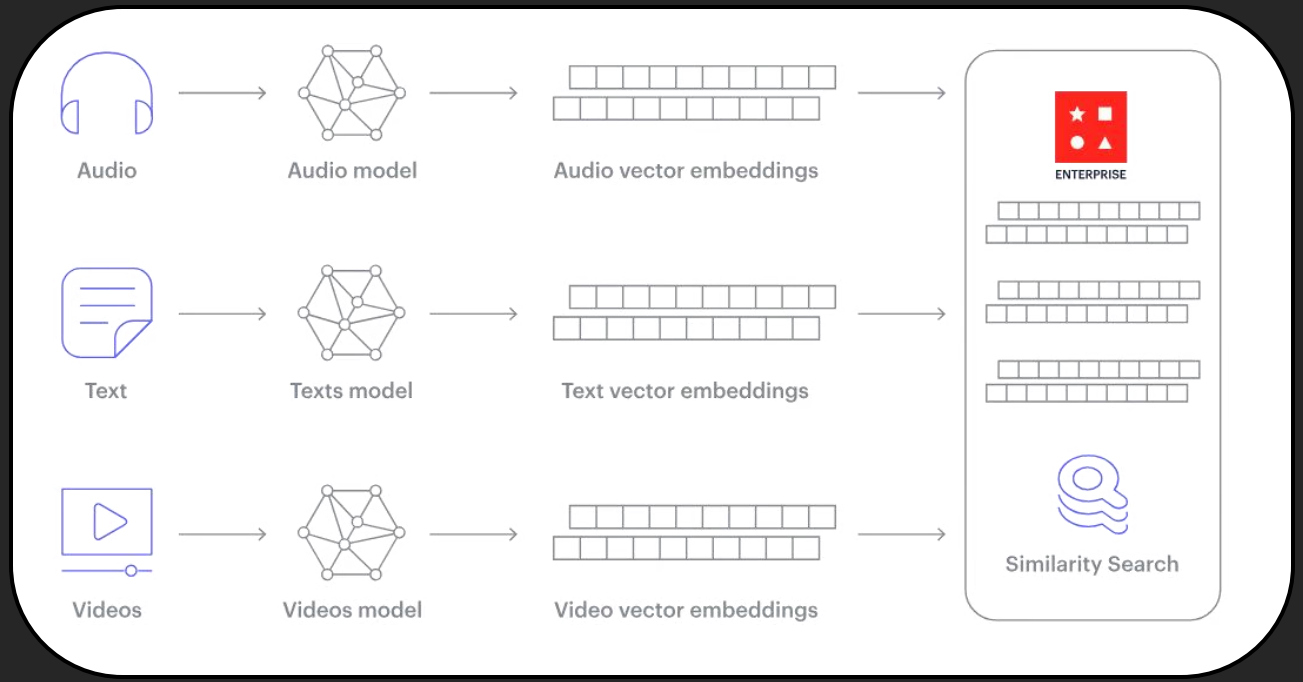
Crafting precise and effective prompts to communicate with AI models

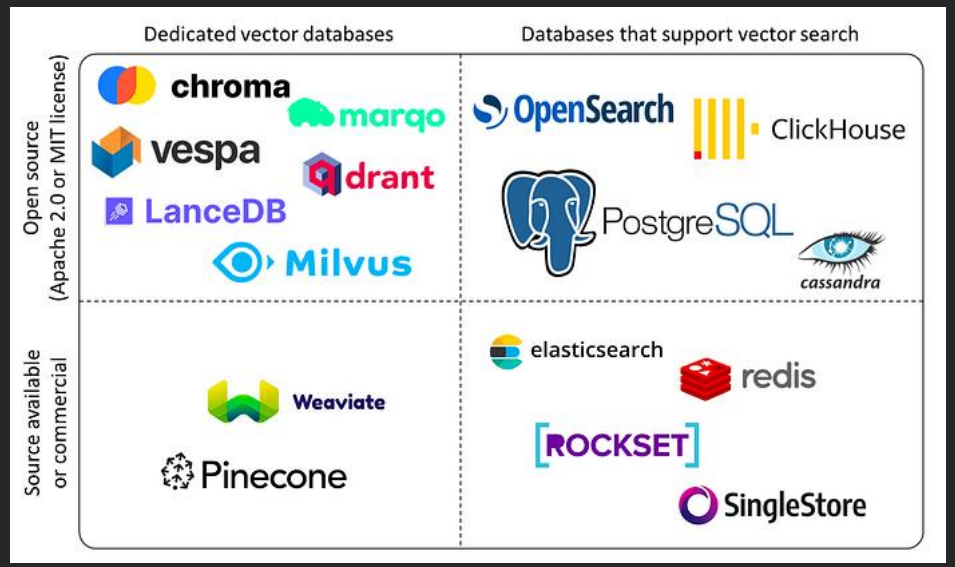
1. **Clarity and Specificity**: Ensuring prompts are clear and specific to reduce ambiguity.
2. **Contextual Information**: Providing background context to guide accurate responses.
3. **Iterative Refinement**: Testing and refining prompts to improve results.
4. **Model Understanding**: Knowing how AI models process text to inform better prompts.
5. **Ethical Considerations**: Being aware of biases and ethical implications in prompt design.
6. **Applications**: Used in areas like customer support, content creation, and coding.

RAG - Retrieval-Augmented Generation

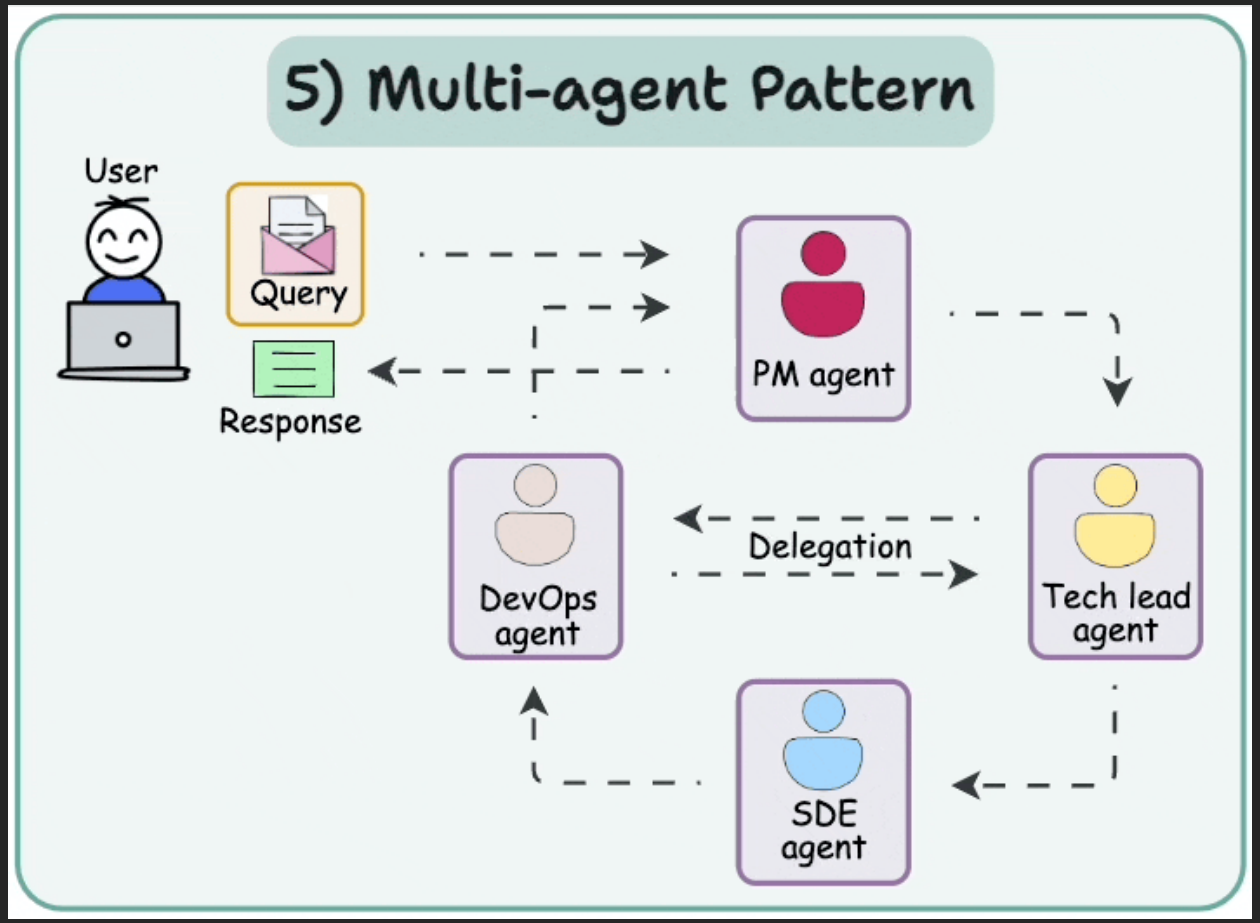


Vector Database





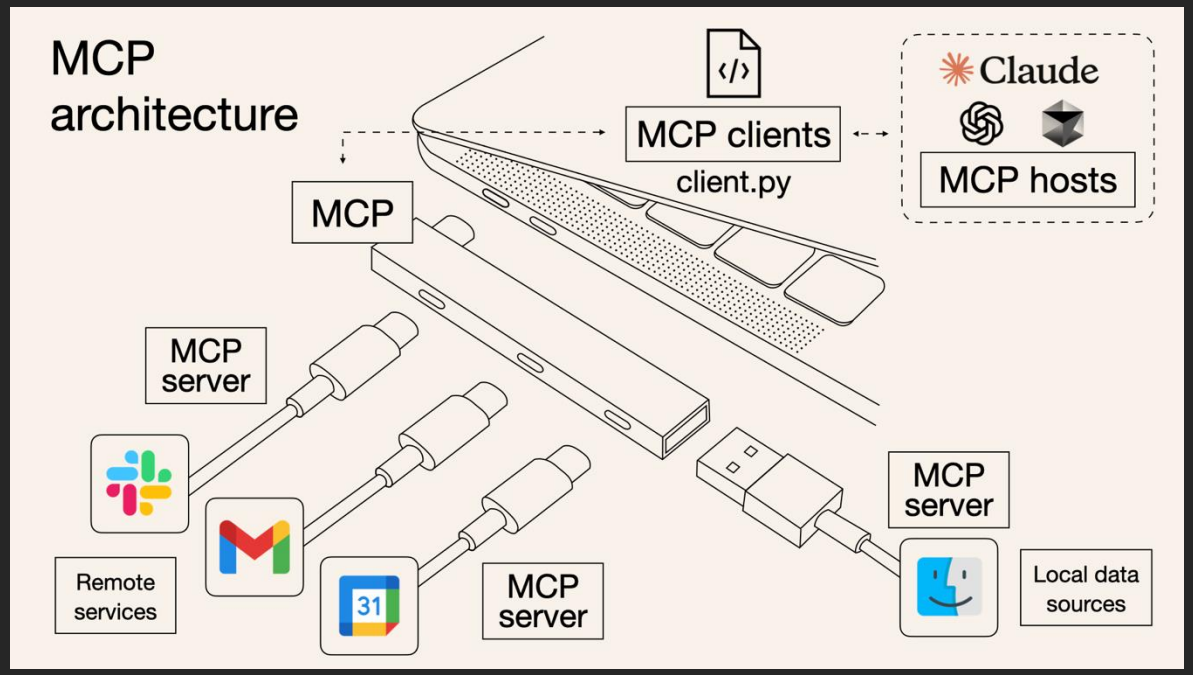
Agentic AI (Multi Agents)

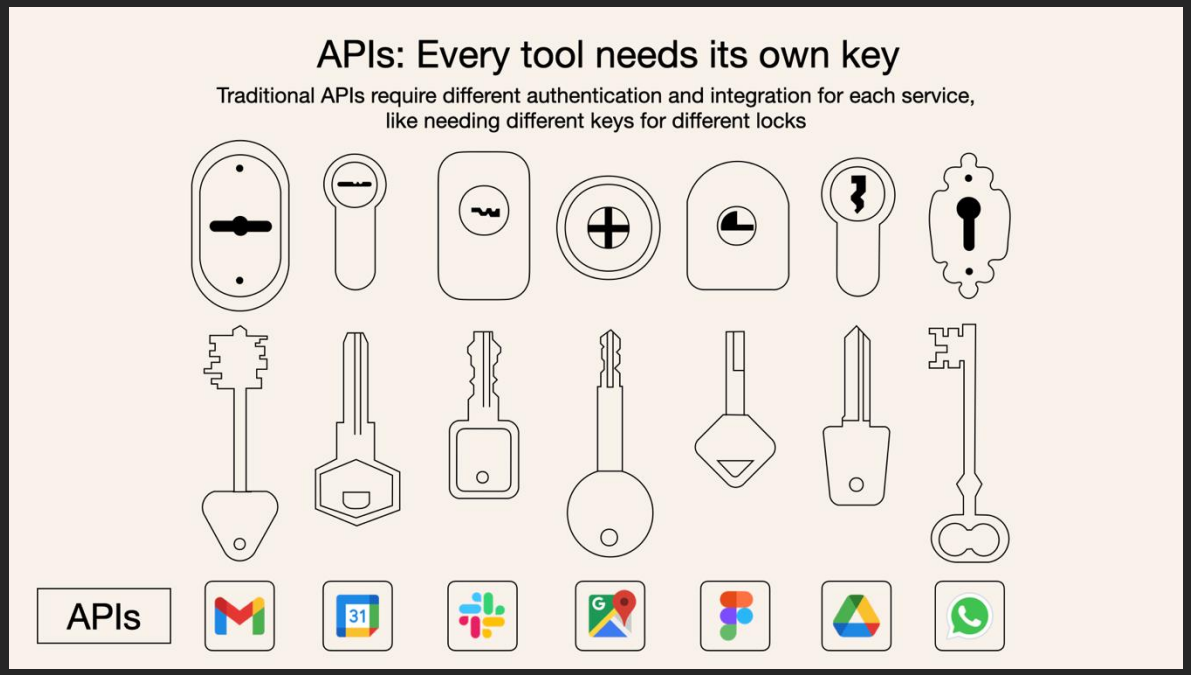


https://blog.dailydoseofds.com/p/5-agentic-ai-design-patterns

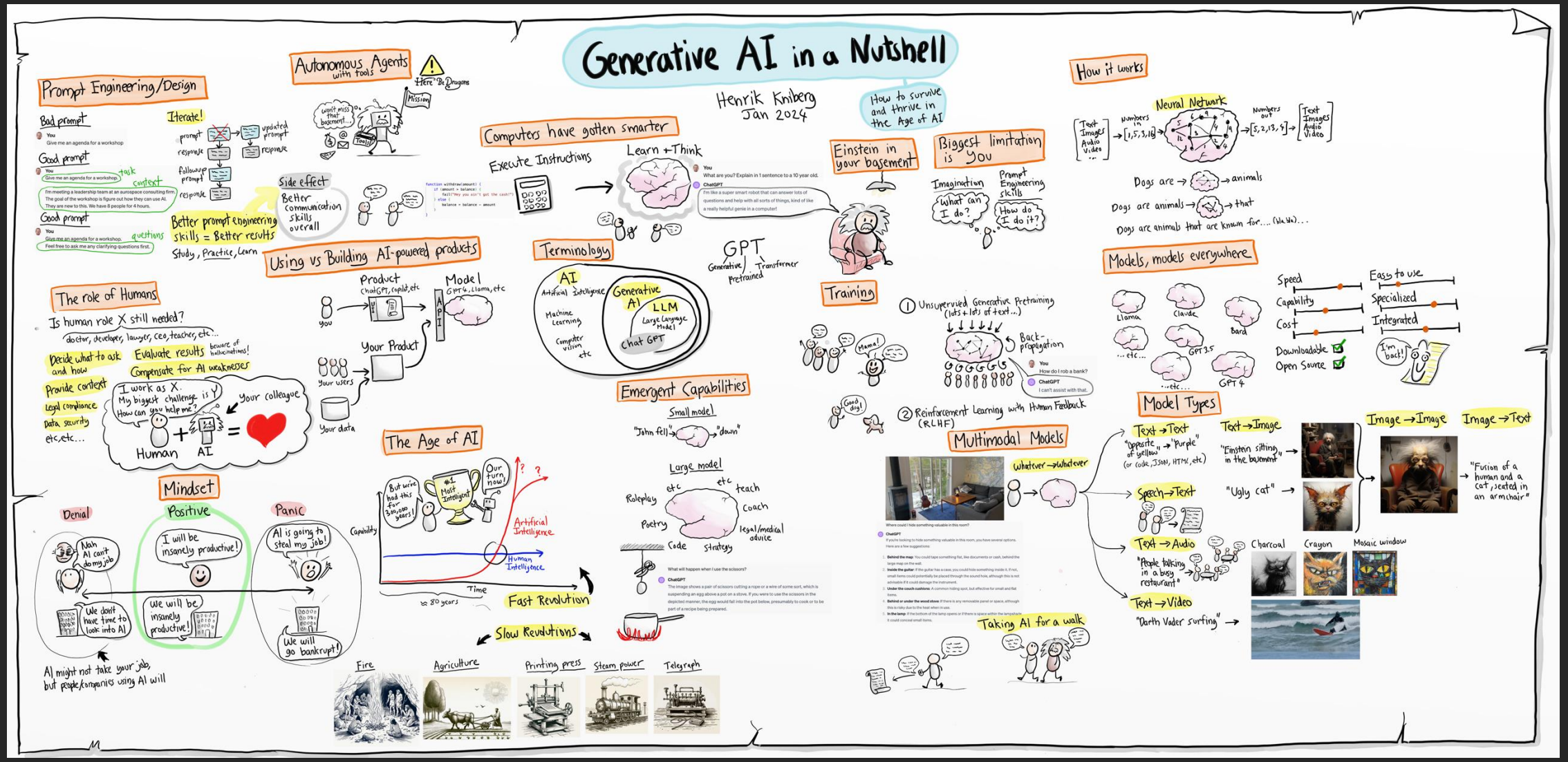
Model Context Protocol (MCP)

- MCP is an open protocol that standardizes how applications provide context to LLMs. Think of MCP like a USB-C port for AI applications.





Wrap-up


Generative AI in a Nutshell
Henrik Kniberg
Jan 2024
How to survive and thrive in the Age of AI
Prompt Engineering/Design
Bad prompt
Good prompt
Iterate!
Good prompt
Better prompt engineering skills = Better results
Study, Practice, Learn
Side effect
Better communication skills overall
Autonomous Agents with tools
Here Be Dragons
Computers have gotten smarter
Execute Instructions
Learn + Think
Einstein in your basement
Biggest limitation is You
Imagination
What can I do?
Prompt Engineering skills
How do I do it?
How it works
Neural Network
Dogs are → animals
Dogs are animals → that
Dogs are animals that are known for....
Using vs Building AI-powered products
Product
Model
Your Product
Your users
Your data
Terminology
AI
Artificial Intelligence
Machine Learning
Computer vision
etc
Generative AI
LLM
Large Language Model
Chat GPT
GPT
Generative
Pretrained
Transformer
Training
Unsupervised Generative Pretraining (lots + lots of text...)
Back-propagation
Reinforcement Learning with Human Feedback (RLHF)
Models, models everywhere
Llama
Claude
Bard
GPT 3.5
GPT 4
Speed
Capability
Cost
Downloadable
Open Source
Easy to use
Specialized
Integrated
I'm back!
The role of Humans
Is human role X still needed?
doctor, developer, lawyer, ceo, teacher, etc...
Decide what to ask and how
Provide context
Legal compliance
Data security
etc, etc...
Evaluate results
Beware of hallucinations!
Compensate for AI weaknesses
I work as X. My biggest challenge is Y. How can you help me?
Your colleague
Human
AI
Emergent Capabilities
Small model
"John fell" → "down"
Large model
Roleplay
Poetry
Code
etc
teach
Coach
legal/medical advice
strategy
The Age of AI
But we've had this for 300,000 years!
#1 Most Intelligent
Our turn now!
Artificial Intelligence
Human Intelligence
Capability
Time
≈ 80 years
Fast Revolution
Slow Revolutions
Fire
Agriculture
Printing press
Steam power
Telegraph
Mindset
Denial
Nah AI can't do my job
We don't have time to look into AI
Positive
I will be insanely productive!
We will be insanely productive!
Panic
AI is going to steal my job!
We will go bankrupt!
AI might not take your job, but people/companies using AI will
Multimodal Models
whatever → whatever
Taking AI for a walk
Model Types
Text → Text
"Opposite of yellow" → "Purple"
(or code, JSON, HTML, etc)
Speech → Text
Text → Audio
"People talking in a busy restaurant"
Text → Video
"Darth Vader surfing"
Text → Image
"Einstein sitting in the basement"
"Ugly cat"
Charcoal
Crayon
Mosaic window
Image → Image
Image → Text
"Fusion of a human and a cat, seated in an armchair"

Careers

PS. my personal opinion

Data/ML/AI Related Careers

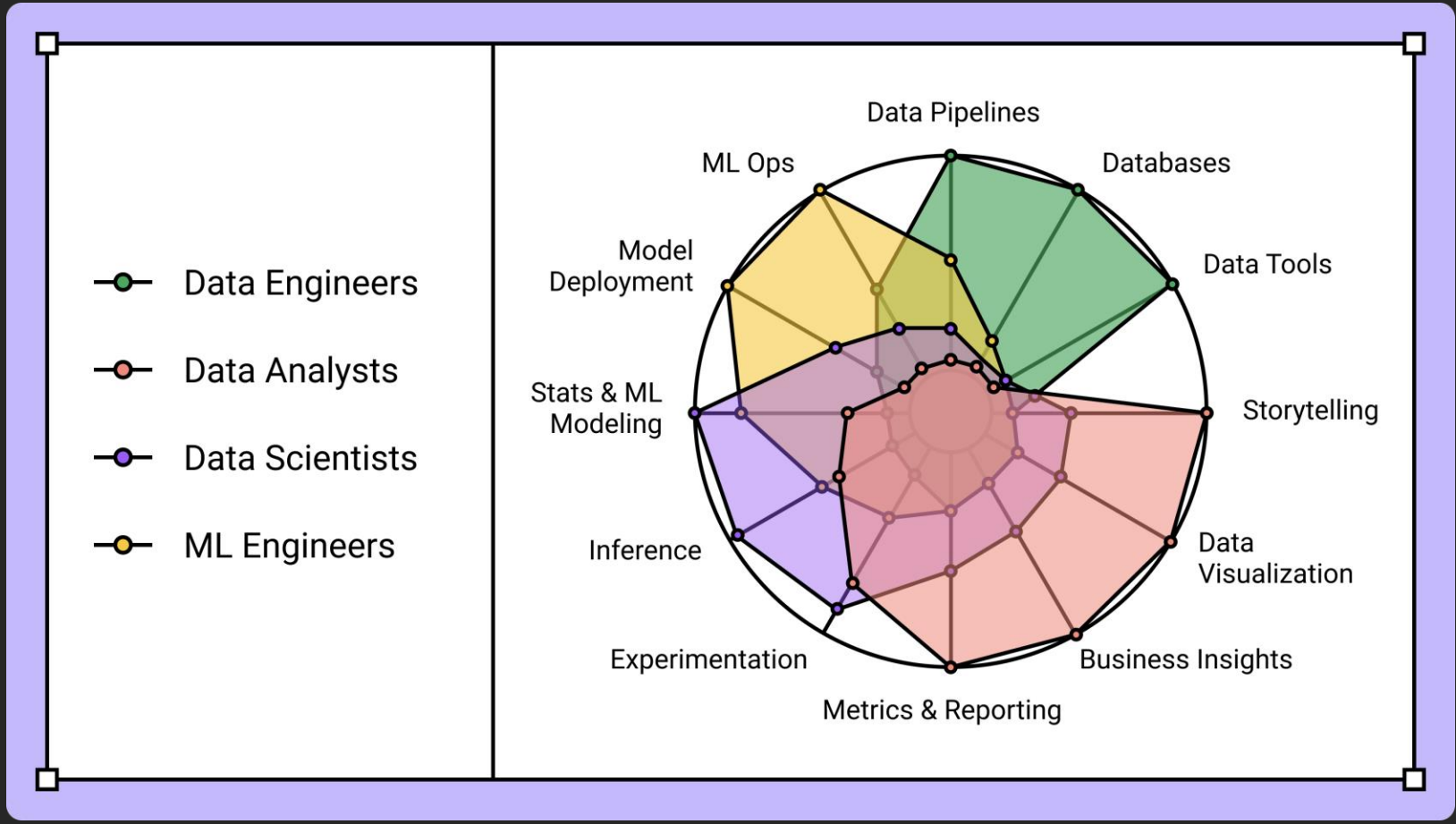
- Data Scientist
- Data Engineer
- Data Analyst
- ML Engineer
- AI Engineer
- AI Researcher
- Computer Vision Engineer
- NLP Engineer
- MLOps Engineer

https://www.sprintzeal.com/blog/career-opportunities-in-ai

Data/ML/AI Related Careers

https://www.linkedin.com/pulse/data-engineer-vs-scientist-ml-heba-al-haddad

Data/ML/AI Related Careers



https://www.datajob.io/blog-post/data-scientist-job-vs-data-engineer-vs-data-analyst

Ex. Data scientist

Job Description:

- Apply statistical and machine learning methods to large, complex data sets to draw insights and provide actionable recommendations.
- Solve complex problems on both technical and business sides using advanced analytical methods.
- Work with Engineering teams to implement end-to-end process from model development to testing, validation, and deployment
- Research and develop new quantitative models and frameworks to enhance the company's data science capability

Responsibilities:

- Design and implement predictive models, machine learning algorithms, and advanced statistical analyses to address business needs.
- Analyze large and complex datasets to extract insights and identify trends, patterns, and opportunities.
- Collaborate with cross-functional teams to define business challenges, collect requirements, and ensure alignment with data solutions.
- Build and deploy end-to-end machine learning solutions, from feature engineering to model deployment.
- Research and stay updated on the latest AI/ML trends and techniques to apply innovative solutions to business challenges.
- Support the implementation of data-driven strategies to improve operational efficiency and enhance customer experience.

DUTIES AND RESPONSIBILITIES:

- Identify valuable data sources and automate data collection for AI and chatbot development.
- Preprocess structured/unstructured data for training ML and Generative AI models.
- Analyze large datasets to uncover trends and enhance learning experiences.
- Build and fine-tune predictive models and chatbot assistants using Generative AI.
- Deploy AI models and chatbots to web-based production environments with scalability in mind.
- Visualize insights and model performance via clear, actionable dashboards.
- Collaborate with engineering and product teams to integrate AI into the platform.
- Implement AI assistants with focus on NLU, personalization, and ongoing learning.
  Maintain and investigate ETL pipelines that feed into the data warehouse.
- Create proof-of-concept (POC) solutions and present ideas clearly to management and stakeholders.
- Stay up-to-date with Generative AI trends to improve interactive learning tools.

READ JD !!!

Q/A

AI Index Repot 2025

ref: https://hai.stanford.edu/ai-index/2025-ai-index-report