ML103: Introduction to GenAI

Rola Dali
AI/ML Architect/Engineer
November 2024

>whoami

- Machine Learning Architect @ RapidScale
- Academic: PhD in NeuroScience & Bioinformatics, 2017
- AWS enthusiast:
  - AWS Community Builder
  - AWS Montreal Meetup co-lead
  - AWS Ambassador (Golden Jacket All Star)

What is Machine Learning?

Definition

Dictionary

Definitions from Oxford Languages · Learn more

machine learning

noun

the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data.

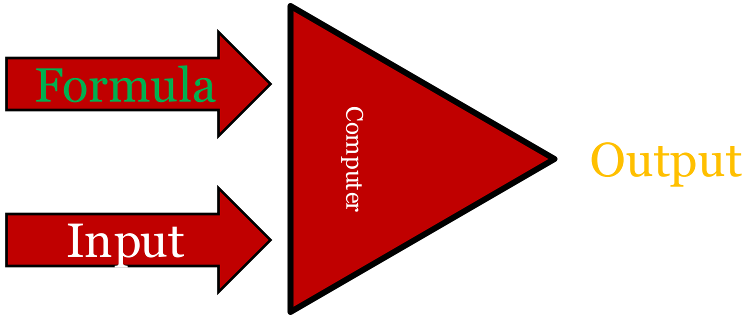
=> Computer systems + powerful mathematics to learn patterns in data without being explicitly taught

=> Large mathematical algorithms powered by computer systems and learn from data

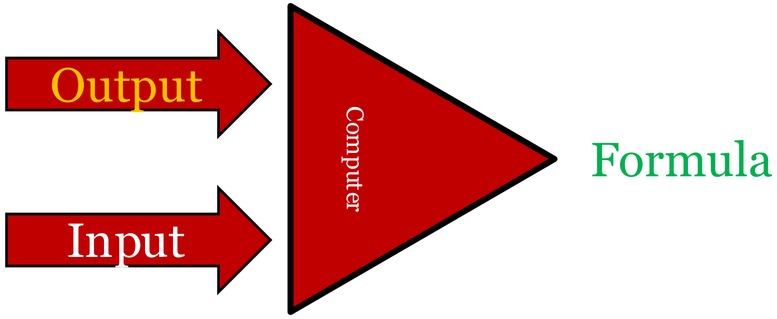
What is Machine Learning?

ML vs Computer Science

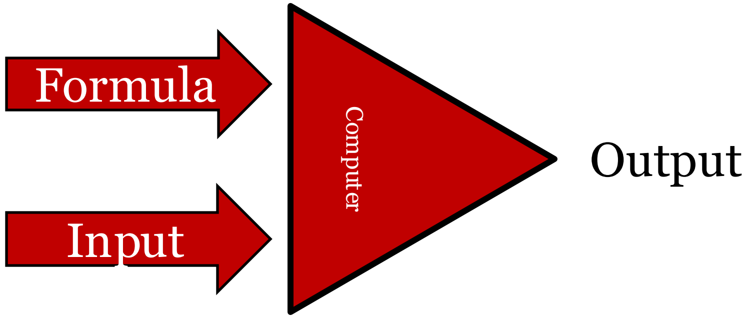
Traditional Software during development



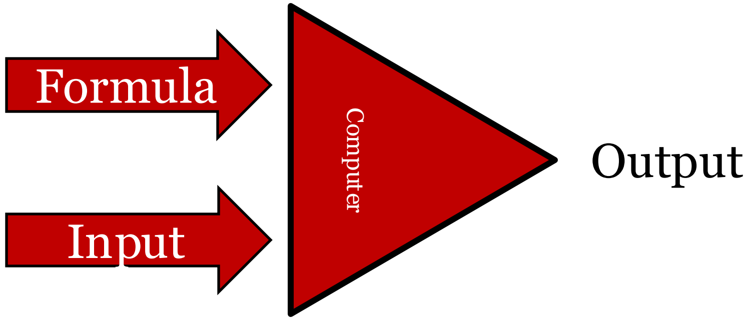
Machine Learning during development/Training



Traditional Software during deployment



Machine Learning during deployment/Inference



Formula ~ pattern ~ algorithm ~ recipe ~ instructions ~ model

ML Glossary

Artificial Intelligence (AI): Techniques that enable computers to mimic human behavior

Machine Learning (ML): AI techniques that allow computers to learn without explicit programming = mimics “learning”

Generative AI: A type of AI that allows computers to generate new content

LLMs: Large Language Models: umbrella term for models specialized in language

Transformers: Algorithm/neural network that revolutionized GenAI and underlies LLMs

Prompt: Input to the model

Token: a word or a part of a word

Embedding: numerical representation of non-numeric entities => projection into mathematical space

RAG: Retrieval Augmented Generation: using an external knowledge base to augment the system

Agent: “Function Calling”: a system that can perform tasks

Foundational model: ML model trained on vast datasets so it can be applied across a wide range of use cases

More about Machine Learning

ML Foundational Pillars:

Machine Learning

Data	Mathematical Algorithms	Compute systems

Types of Machine Learning:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Generative AI

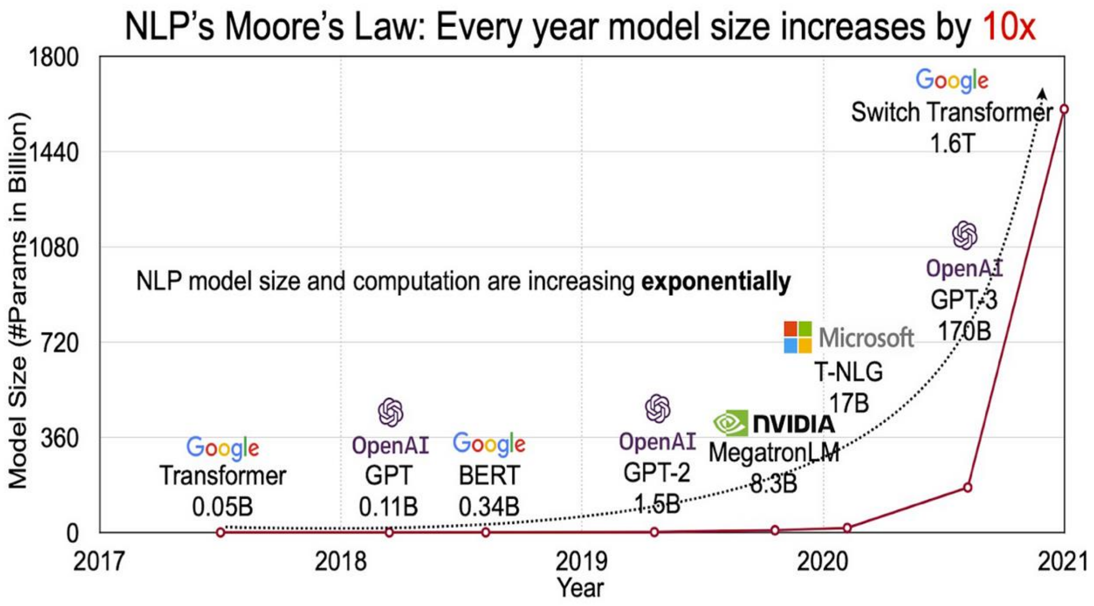
Predictive/Classical Machine Leaning

Predictive ML vs GenAI:
Overview

	Predictive ML	GenAI
Algo Size (params)	< Millions	Billions-Trillions
Data Demands	+	+++
Training Compute	Laptop/reasonable machines	Super computers. Parallelization is critical
Training	Often customized with data	Pre-trained by big providers
Use cases	Specific tasks	General tasks
Cost	$	$$$
Interactions	Custom	API calls
Difficulty	Data, ML algorithms, MLOps	Model selection, prompt engineering, Evaluation
AWS tools	Amazon SageMaker	AWS Bedrock

Predictive ML vs GenAI
Size Considerations

- ML models are often sized by "number of parameters" = model weights
- Size ranges from 1 param (y = ax) to ~2T param (GPT 4)
- Predictive ML ~ million params
- GenAI ~ billion-trillion params
- The more params the model has, the more data it needs to see
  - "20K years to read worth of data": Yann Lecun



Sizing Ballpark:

1 parameter @ 32 bit float = 4 bytes
1 billion parameters ~ 4 GB of RAM JUST FOR PARAMS

BUT you need ~ 20X more space (optimizer, gradients, activation, …) to train

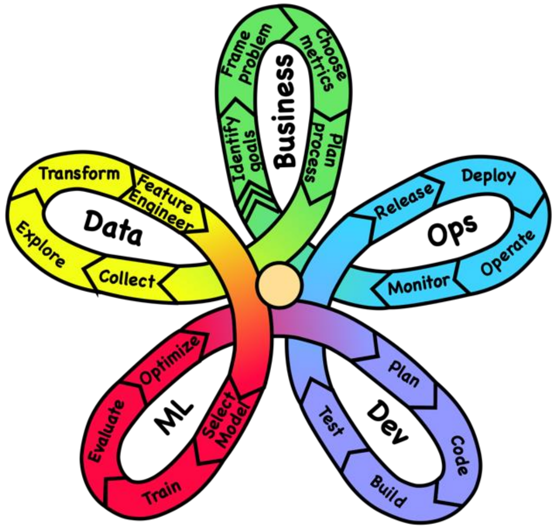
1 billion param model ~ 80 GB of RAM (limit of the Nvidia A100 GPU)
⇒ Imagine the requirements for a 1.8T param model?!

⇒ These models have put constraints on compute/data and made parallelization and optimizations (hardware & software) a must
⇒ The sheer size and compute demands limit training to organizations with significant resources => "Foundational Models"

Predictive ML vs GenAI
LifeCycle

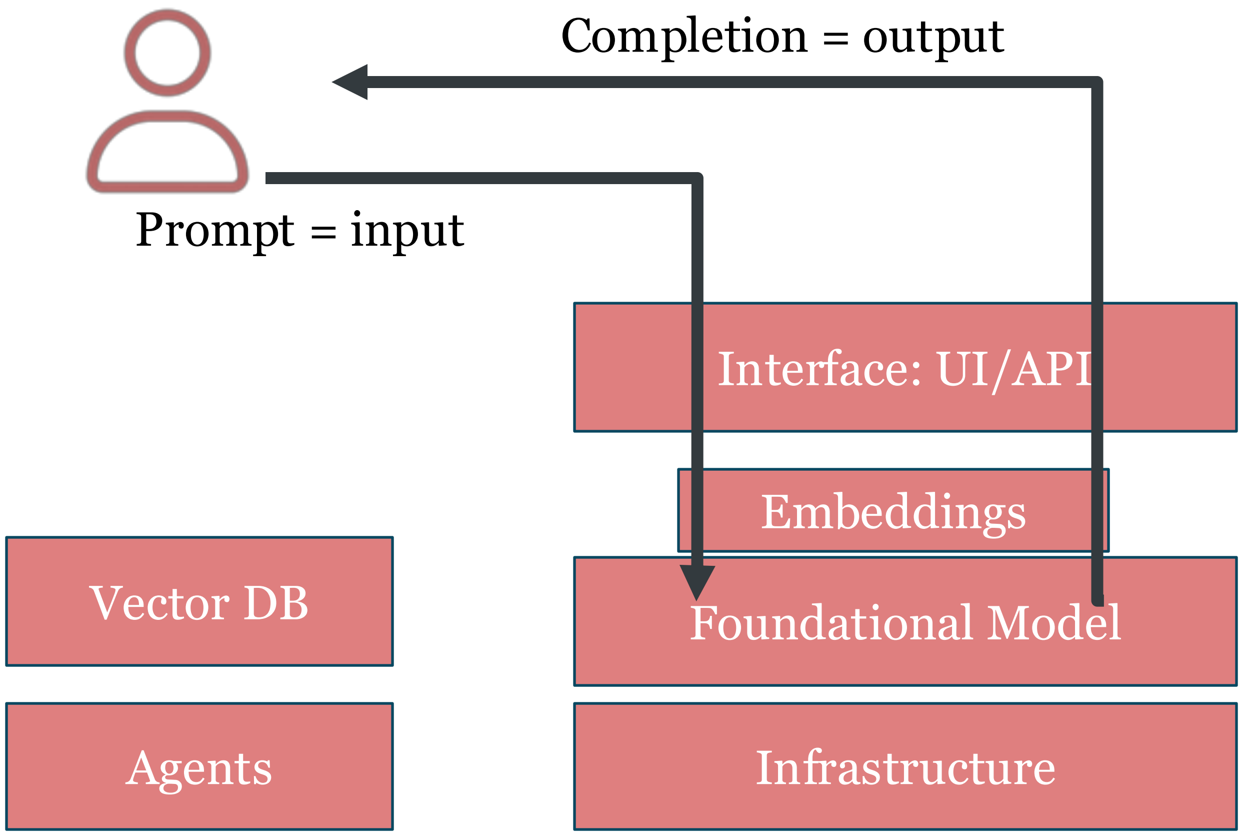
Predictive ML life Cycle:
- Frame Business Problem
- Source & Prepare Data
- Choose Model Class
- Train Model
- Test Model
- Deploy Model
- Maintain & Monitor

GenAI life Cycle:
- Choose Foundational Model
- Tune prompts
- Evaluate Performance
- Deploy Application
- Monitor Performance



Whereas in Predictive ML, much of the work is about customizing the model to excel at a specific task, GenAI is more about better extracting what you want from a large general purpose model

GenAI system components



Anatomy of a prompt

- By using Foundational models, the task shifts from data/model to prompting in order to ”extract” what we need from the model

- Prompt: the input to the model and can vary in structure & content
- prompt engineering: editing the input text to drive the desired output from the model

Prompt Engineering Best Practices:

- Give clear/specific instructions
- Structure prompts
- Include examples
- Add contextual information
- Use system instructions
- Instruct the model to explain its reasoning (Chain of thought)
- Break down complex tasks
- Prompt iteration strategies

Prompt

- Query: what is the task?
- Instructions: steps to perform
- Objective: mission/goal to achieve
- Persona: role/view
- Constraints: restrictions to respect
- Examples: demo of output
- Context: relevant information
- Tone: style to use

Brains & Bots: Human Brain VS Artificial Intelligence


	Brains	Bots	Conclusion	Winner
Predictive Machines	Predicts events	Predicts events	Both work predictively & adjust	
Base Counts	100T synapses	~2T SoTA	Brains ~50X interconnected Better at integrating data wholistically	
Training Time	Evolving for 300K+ yrs Knowledge sharing	~100 yrs old as a field product of Brain ingenuity	Bots have had much less time BUT benefit from brain ingenuity	
Speed	Neurotransmitters in liquid: ~ 200 Hz	Electrons in transistors CPU clock rate > 10GHz	Bots ~50X faster than brains	
Input Modalities	Tethered to biology 5 Senses	Unlimited Input	Bots have unlimited input streams	

- Machine Learning systems have HUGE potential given their speed and augmentation capability
- They will be the workhorse of the future

=> LEARN ML

Questions?

Slides

https://shorturl.at/jvyhm

The Generative AI Stack

APPLICATIONS THAT LEVERAGE FMs

- Amazon Q
- Amazon Q in Amazon QuickSight
- Amazon Q in Amazon Connect

TOOLS TO BUILD WITH FMS AND LLMS

Amazon Bedrock

Guardrails | Agents | Customization capabilities

INFRASTRUCTURE FOR FM TRAINING & INFERENCE

- GPUs
- Trainium
- Inferentia
- SageMaker
- UltraClusters
- EFA
- EC2 Capacity Blocks
- Nitro
- Neuron

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models

AI21labs

Contextual answers, summarization, paraphrasing

- Jurassic-2 Ultra
- Jurassic-2 Mid

amazon

Text summarization, generation, Q&A, search, image generation

- Amazon Titan Text Premier
- Amazon Titan Text Lite
- Amazon Titan Text Express
- Amazon Titan Text Embeddings
- Amazon Titan Text Embeddings V2
- Amazon Titan Multimodal Embeddings
- Amazon Titan Image Generator

ANTHROPIC

Summarization, complex reasoning, writing, coding

- Claude 3 Opus
- Claude 3 Sonnet
- Claude 3 Haiku
- Claude 2.1
- Claude 2
- Claude Instant

cohere

Text generation, search, classification

- Command
- Command Light
- Embed English
- Embed Multilingual
- Command R+
- Command R

Meta

Q&A and reading comprehension

- Llama 3 8B
- Llama 3 70B
- Llama 2 13B
- Llama 2 70B

MISTRAL AI_

Text summarization, Q&A, text classification, text completion, code generation

- Mistral Large
- Mistral 7B
- Mixtral 8x7B

stability.ai

High-quality images and art

- Stable Diffusion XL1.0
- Stable Diffusion XL 0.8

aws © 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.