ivp

AI Primer

Making AI understandable for all

Image created by Jasper.ai

March 2023

By Michael Miao, Angela Zhu, Army Tunjaicon and Jared Rosner on behalf of IVP

Questions, comments or queries? Please reach out to Michael and Angela.

*Special thanks to **David Liu** (Head of ML Platform @ Pinterest), **Timo Mertens** (Head of ML and NLP Products @ Grammarly), **Matt Park** (Chief Business Officer @ Scale AI), **Bryan Hsu** (Investor @ Access Technology Ventures) and **Kevin McNamara** (CEO @ Parallel Domain) for their contributions.*

Why We Wrote This

- Artificial Intelligence will be the most significant platform shift of the decade.
- We have backed many companies at the forefront of this shift (Grammarly, CrowdStrike, Eightfold, Jasper, DeepL, etc.) and are excited to continue partnering with the best founders in this space.
- AI has many layers and can be difficult to understand; most online resources are highly technical and assume prior knowledge. Our goal is to introduce AI and related concepts in a straightforward and approachable way.
- AI is fundamentally changing the way we at IVP work. For example, we used Jasper to generate some of the content for this primer, and Grammarly to make sure our writing was clear and effective. If there's interest, we'll use DeepL to translate the primer into different languages!
- We hope this resource is helpful for founders, operators, investors and anyone who is looking to understand this exciting technology.

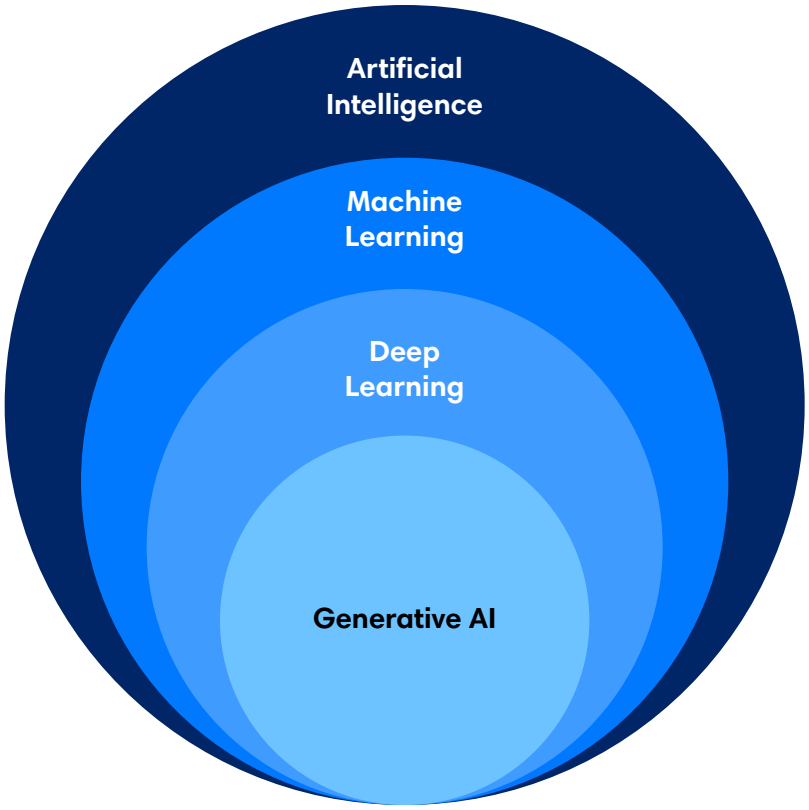
> Disclaimer: AI is a rapidly advancing technology and, given the pace of innovation, elements of this primer will become outdated in the coming months (OpenAI, Google and Microsoft have all made major product announcements days before we published). This presentation is subject to change based on new developments.

Table of Contents

1/7

AI Primitives

Key Terms



Artificial Intelligence (AI) is a scientific technique that enables machines to mimic intelligent human behavior.

Artificial General Intelligence (AGI) refers to the ability of a machine to perform any task a human being is capable of.

Machine Learning (ML) is a subset of AI that uses algorithms to automatically learn insights and recognize patterns from data, applying it to make better decisions. ML can be supervised (i.e., using labeled datasets) or unsupervised (i.e., finding patterns within unlabeled datasets). Popular applications of ML include computer vision and natural language processing (NLP).

Deep Learning (DL) is an advanced subset of ML that leverages neural networks to learn from vast amounts of data. Neural nets function like the human brain, enabling systems to learn hidden patterns from data by themselves and build more efficient decision rules.

Generative AI (GenAI) refers to unsupervised and semi-supervised deep learning algorithms that enable computers to use text, audio, video, images and code to generate new content. Generative AI is powered by foundation models like GPT-4, Stable Diffusion, etc.

Foundation Models are large, pre-trained artificial intelligence models that serve as a basis for developing more specific models for a wide range of tasks.

Large-language Models (LLMs) are a type of foundation model focused on text and natural language processing. Examples include GPT-4, BERT, Anthropic and Bloom.

Transformers are a type of deep-learning model architecture that learns context by tracking relationships in sequential data, making it very effective at processing inputs like text. Transformers are faster to train because they are parallelizable (can process data inputs in parallel instead of sequentially).

2/7

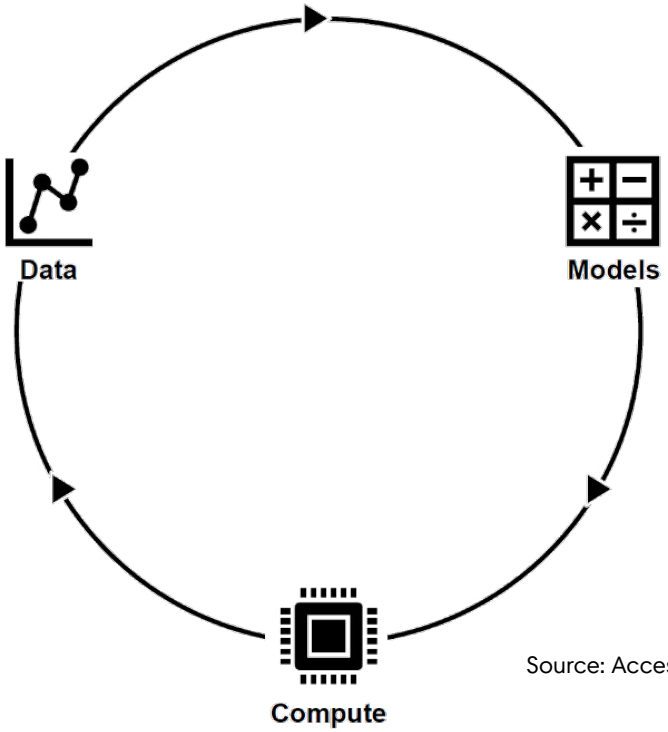
AI Primitives
Building Blocks of AI

Data, Models and Compute are the main building blocks of AI

- **Data** is the feedstock of AI. Feed the AI model text-image pairs and the AI develops the ability to recognize or generate images. Feed the AI model music and it will compose music. "You are what you eat" perfectly captures the importance of data as an AI building block.
- **Models** are sophisticated computer programs designed to perform tasks that would typically require human intelligence. They are built on foundations of algorithms and large datasets that enable them to analyze patterns, make predictions and solve complex problems.
- **Compute** refers to the computational resources and processing power required to train, develop and run AI models. These resources include central processing units (CPUs), graphics processing units (GPUs) and specialized hardware like tensor processing units (TPUs).

Is AI interpolation or extrapolation?

- **Interpolation**: AI takes fundamental building blocks, such as letters or musical notes, and generates novel works by stitching the building blocks together.
- **Extrapolation**: AI can generate something beyond the observable range.
- It is generally held that all AI models/outputs are interpolative because there are multiple examples of AI models failing when given inputs outside the initial observable range.



Source: Access Technology Ventures

3/7

AI Primitives

What is ML?

Machine Learning is a way to teach computers how to learn from data and make decisions and predictions without being explicitly programmed

Machine Learning can be “supervised” or “unsupervised”

- **Supervised:** The model leverages labeled training data to learn.
- **Unsupervised:** The model learns from the underlying patterns of the data.

We see three distinct categories of ML emerging in production today

- Enterprise ML, computer vision and natural language processing.

Enterprise ML

Fraud Detection

Content Recommendation

Improved Loan Underwriting

Computer Vision

Face detection and Recognition

Optical Character Recognition

Automatic tagging of image

Lane Detection

Virtual Reality

Drone Terrain Mapping

Natural Language Processing

Voice assistants (Alexa, Siri, Google Home etc)

Question answering systems

Text summarization algorithms

Chatbots

Sentiment Analysis

Machine Translation

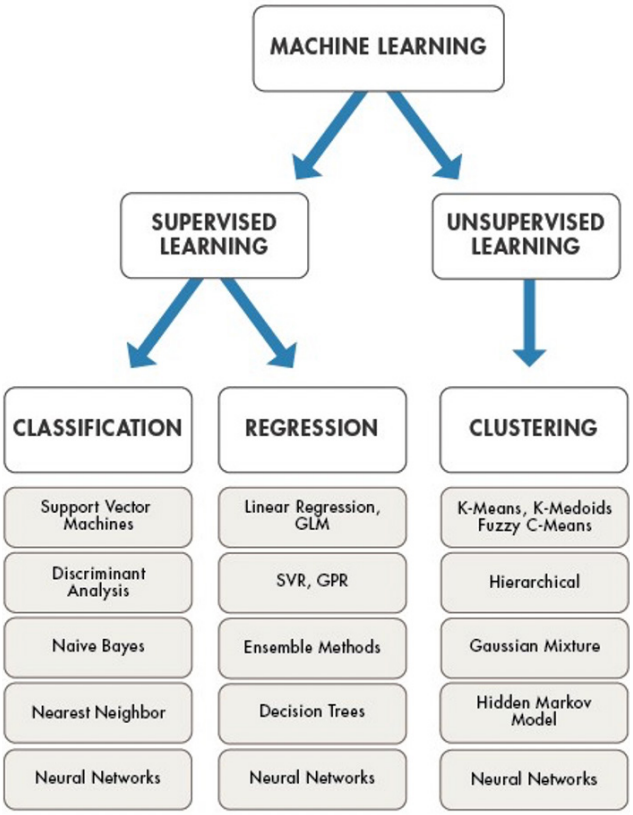
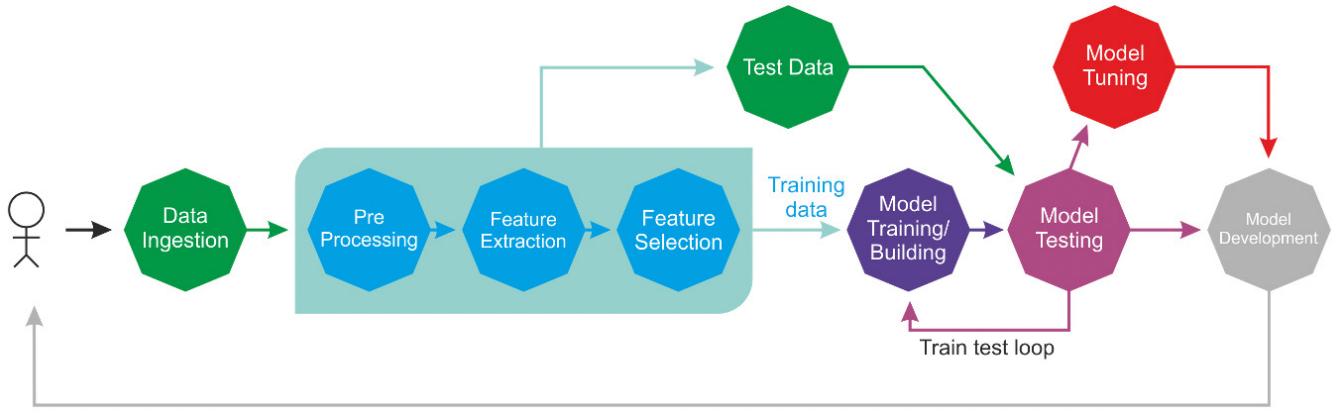


Image sources: A Beginner’s Guide to Understanding the Buzz Words and Machine Learning in Python: Supervised vs Unsupervised Learning

4/7

AI Primitives

What Does An ML Workflow Look Like?



Step 1
Data Preparation

- **Data Ingestion** – Collect data required to train and test the model. Training data comes in various forms, including images, voice, text or features.
- **Data Cleaning** – Turn raw training data (e.g., missing data, noisy data, inconsistent data) into clean data. This is called "data pre-processing."
- **Data Labeling** – The process of adding annotations (labels) to training data for supervised machine learning. This can be done manually (humans) or programmatically (machines).
- **Feature Engineering** – Manipulate datasets to create variables (parameters) that improve your model's prediction accuracy.
- **Split Data** – Randomly divide the records in the dataset into a training set and testing set for cross validation.

Step 2
Model Training & Dev

- **Develop / Train Models** – Fit each model to training datasets.
- **Hyperparameter Tuning** – Tune hyperparameters, which are external to the model (i.e., number of hidden layers, choice of optimization algorithm, train/ test split, etc.).
- **Assess Model Performance** – Calculate performance metrics, such as accuracy, recall and precision, on the testing dataset.

Step 3
Model Deployment

- **Deploy Model** – Embed the model in dashboards and applications.
- **Monitor Model Performance** – Regularly test the performance of the model as data changes, to avoid model drift.
- **Improve Model** – Continuously iterate and improve the model post-deployment. Replace the model with an updated version to improve performance.

Image: packt

5/7

AI Primitives

What Are Neural Networks?

Neurons in Biology

- Neural networks are the technological advancement that enables deep learning, and is based on biological neurons.
- In the human brain, dendrites receive signals and pass signals through axons, which connect to other neurons forming a connection called a synapse.
- Neurons are the inspiration for neural nets, which ingest input signals, perform calculations and send output signals.

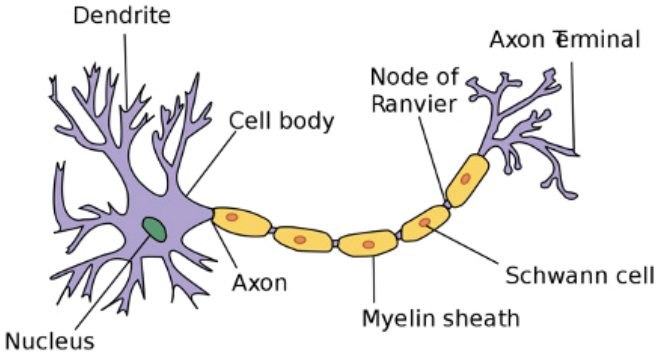
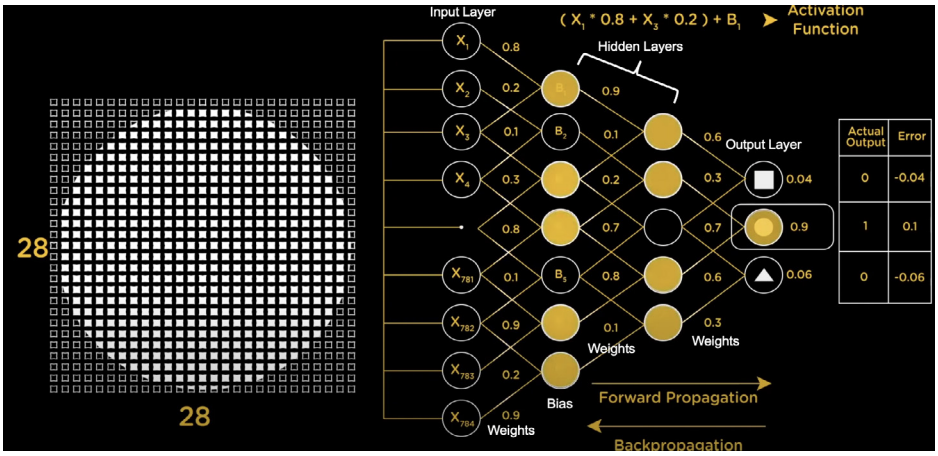


Image sources: Freecodecamp.org, Simplilearn

Neural Network in Practice (What Shape is the Picture?)

- The diagram below is an example of how a deep learning model would determine the shape of the picture (circle) using a neural network.
- Each pixel of the picture is fed as an input to each neuron, which are connected through channels (lines) and are assigned a weight and bias which feed into an activation function.
- Forward propagation happens when the inputs process through the neural network until it makes a prediction (i.e., square, circle or triangle); labeled training data allows you to know how accurate your prediction is, which assigns an error rate.
- Backpropagation happens when you take the error rate of forward propagation and feed it backwards through the neural network layers to algorithmically fine tune the weights and biases.
- Fine tuning the weights and biases to minimize the error rate is the essence of how you train a neural network.
- Gradient descent is an optimization procedure that tries to minimize the error rate by taking many iterations of forward + backpropagation until convergence is achieved for weights and biases.



6/7

AI Primitives

What Are Transformers, Foundation Models & Large Language Models?

A transformer is a neural network model architecture that learns context and meaning by tracking relationships in sequential data

- Transformers enable processing of input data (e.g., text) in a parallelized and context-aware manner, making it highly effective for tasks like machine translation, text summarization and text generation.
- First described in a 2017 paper by Google, "Attention is All You Need," transformers are among the newest and most powerful model architectures.
- Transformers power foundation models, which are very large-scale models that comprise general-purpose knowledge that is useful in a variety of domains.
- **Large Language Models** (LLMs) are a type of foundation model focused on text and natural language processing (NLP).

Foundation models lower the barriers to building high-quality AI applications

- Practitioners can now borrow lower neural nets that someone else has pre-trained and simply tune and add a few lines of code on top, rather than build a model from scratch.
  - This eliminates up to ~95% of the training set previously needed and results in huge improvements in accuracy and time and cost savings.
  - The implication of foundation models is that practioners no longer need to go through the ML workflow process, as foundation models come pre-trained.
- Less training required for a new project › focus efforts on "last-mile" customization and tuning › more parallelization › faster training times › larger datasets can now be used › models reach greater scale and accuracy.
- The rate of improvement for foundation models is staggering.
  - GPT-4 (released in March 2023) can now pass the bar exam with a score in the top 10% vs GPT-3.5 (released in 2022) in the bottom 10%.
  - GPT-4's reasoning, multi-modal prompting abilities and overall model performance represents a rapid step-change improvement over its predecessor model.

AI research companies like OpenAI, Anthropic, DeepMind and Stability AI are making building high-quality AI applications easier

- Developers can now access foundation models/LLMs by working with AI research companies, which has led to a Cambrian-level explosion in Generative AI.

7/7

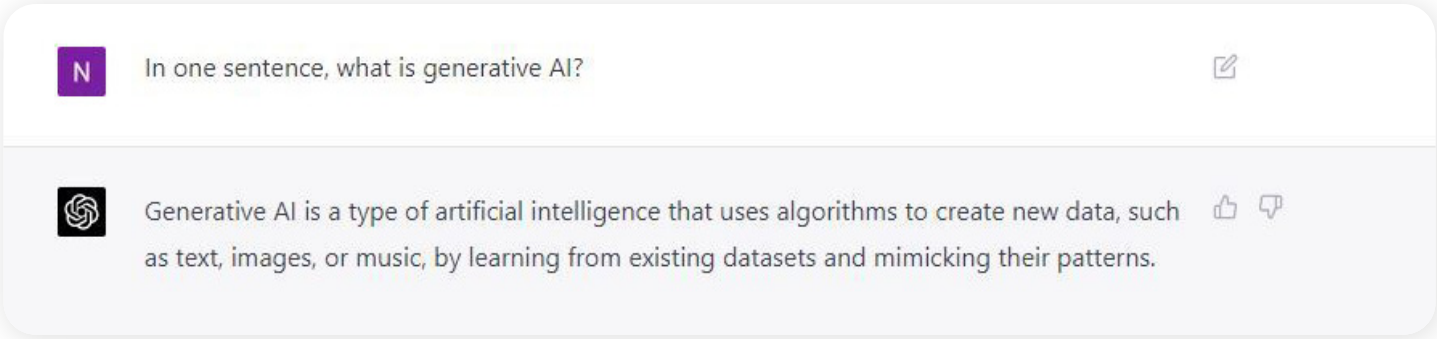
Generative AI

What is Generative AI?

Generative AI (GenAI) focuses on creating a wide range of data types, including audio, code, images, text, simulations, 3D objects and videos. It learns from existing data and generates novel outputs, leading to advancements in communications, product design, filmmaking and other creative fields.

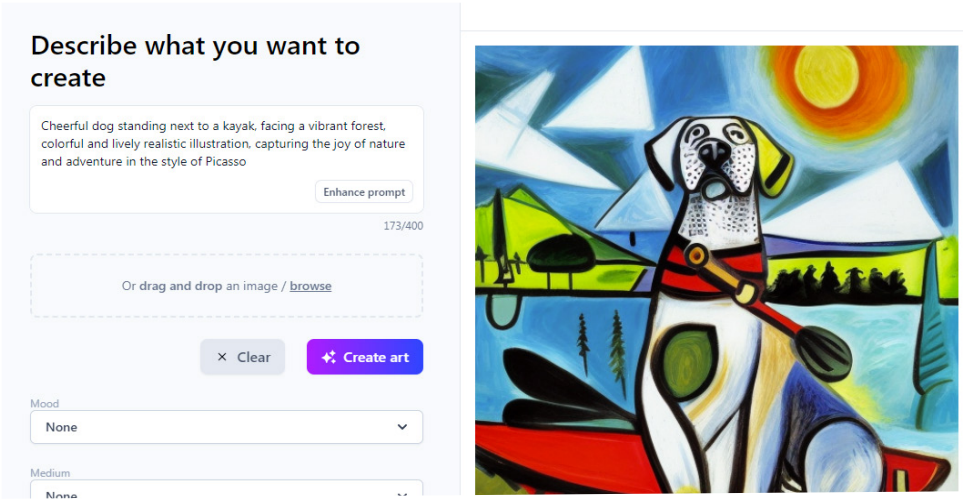
Types of Generative AI

Text Generation (GPT-4, BERT, Bloom, Cohere, Anthropic, DeepMind Gopher): Utilizing large language models (LLMs) like OpenAI's GPT-4, GenAI can provide impressive and contextually relevant responses in a conversational format, assisting with various language-related tasks.



Source: ChatGPT using GPT-4

Image Generation (Dall-E 2, Stable Diffusion, Midjourney): GenAI can create genre- or period-specific images based on a simple text prompt.

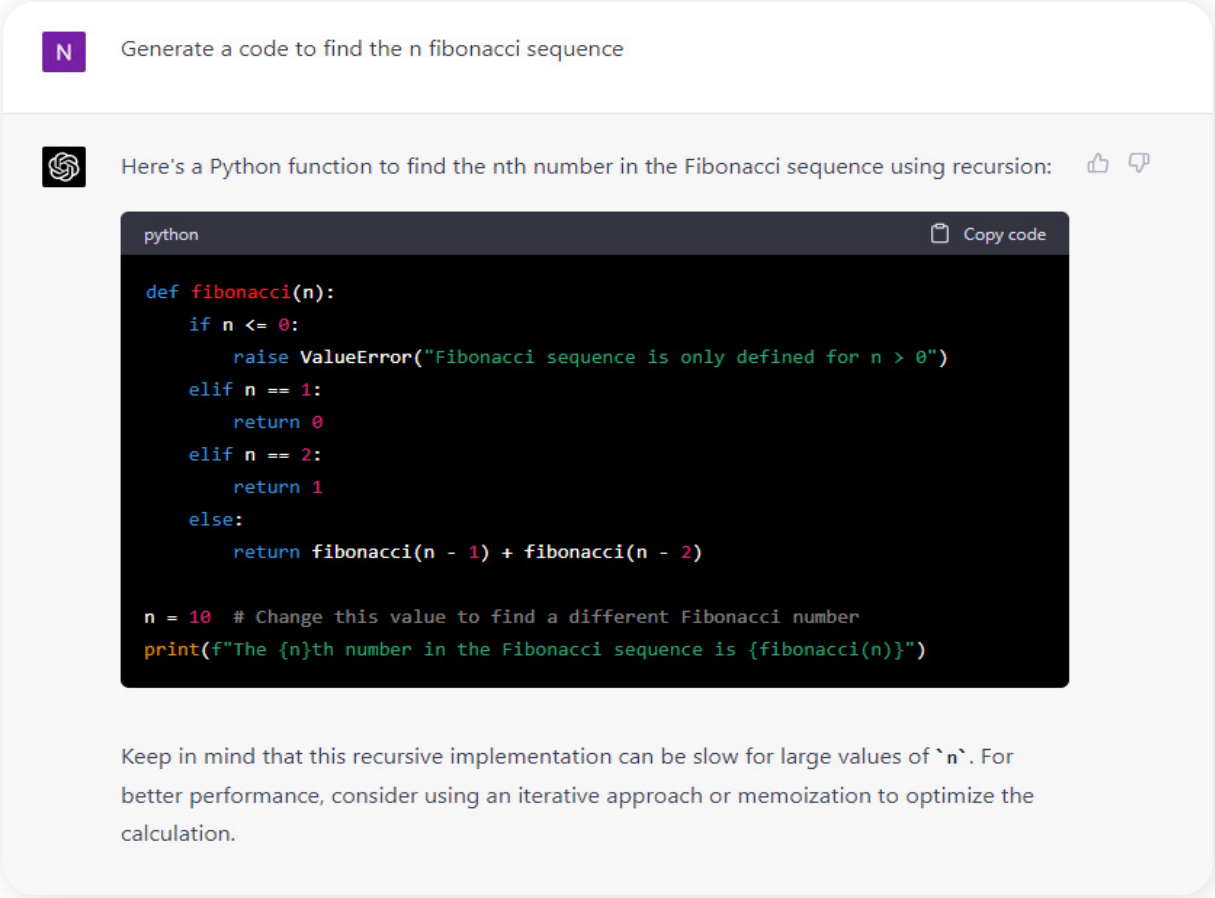


Source: JasperArt

Video Generation (Gen-1, X-CLIP, VQ-VAE-2): GenAI can even generate videos from prompt sequences, enabling the creation of dynamic and engaging visual content based on specific prompts or ideas.

Gen-2 Explained https://research.runwayml.com/gen2

Code Generation (GPT-4, Codex): GenAI can automatically generate code snippets or entire programs based on user-provided requirements or descriptions, streamlining software development and helping developers create efficient, accurate code more quickly.



Prompt Engineering vs. Fine Tuning

Prompt engineering and fine tuning are two ways to improve foundation models.

Prompt engineering is the art of crafting better input queries, aka text prompts, that the AI needs to create the desired outputs. It focuses on optimizing user interaction with the AI without changing the model itself.

Fine tuning refers to adjusting a pre-trained AI model by continuing its training on a smaller, specialized dataset. It tailors the model to a particular context, making it more accurate for specific tasks.

A Brief History of AI

AI has been around since 1952, but two recent inflection points have dramatically accelerated AI research

- The rise of deep learning (2010 onwards), enabled by general purpose programming on NVIDIA GPUs using CUDA, deep neural networks with many layers of parameters trained on massive amounts of data.
- The rise of large-scale models (2017 onwards), enabled by Google's influential 2017 paper, "Attention is All You Need," introduced a new neural network model, called a transformer, for natural language understanding. Compared to recurrent neural networks (RNNs), which require sequential data inputs and take a long time to train, transformers generate superior-quality language models while being more parallelizable and requiring significantly less time to train.

Advancements in research and compute enabled the size of neural networks to grow by more than 10,000x in recent years

Today, thanks to innovations in neural network architecture, algorithms and compute, AI has surpassed human benchmarks across multiple dimensions, such as handwriting, speech, image, reading and language understanding.

As Training Computation Increased …

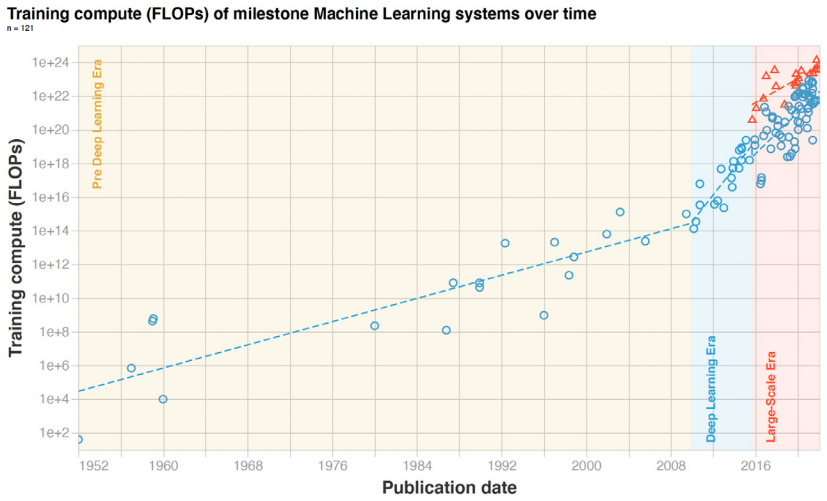
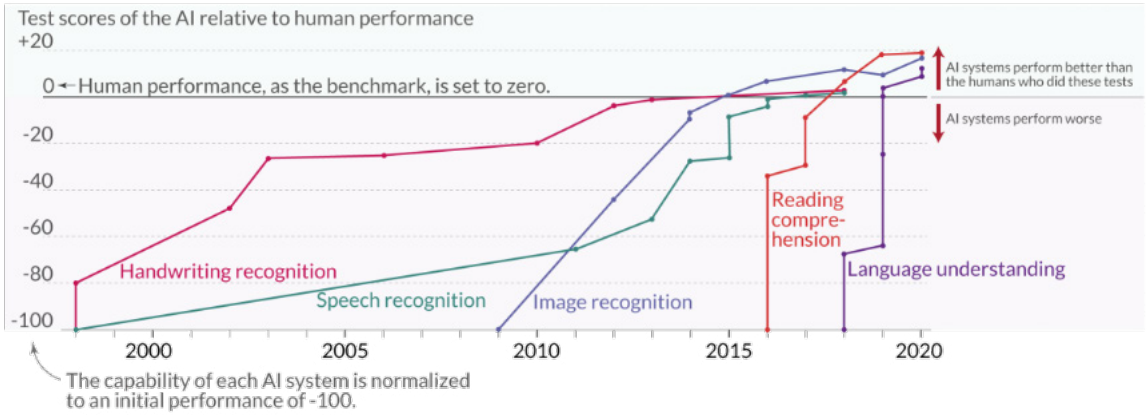


Figure 1: Trends in $n = 121$ milestone ML models between 1952 and 2022. We distinguish three eras. Notice the change of slope circa 2010, matching the advent of Deep Learning; and the emergence of a new large-scale trend in late 2015.

… AI Systems Have Become More Powerful



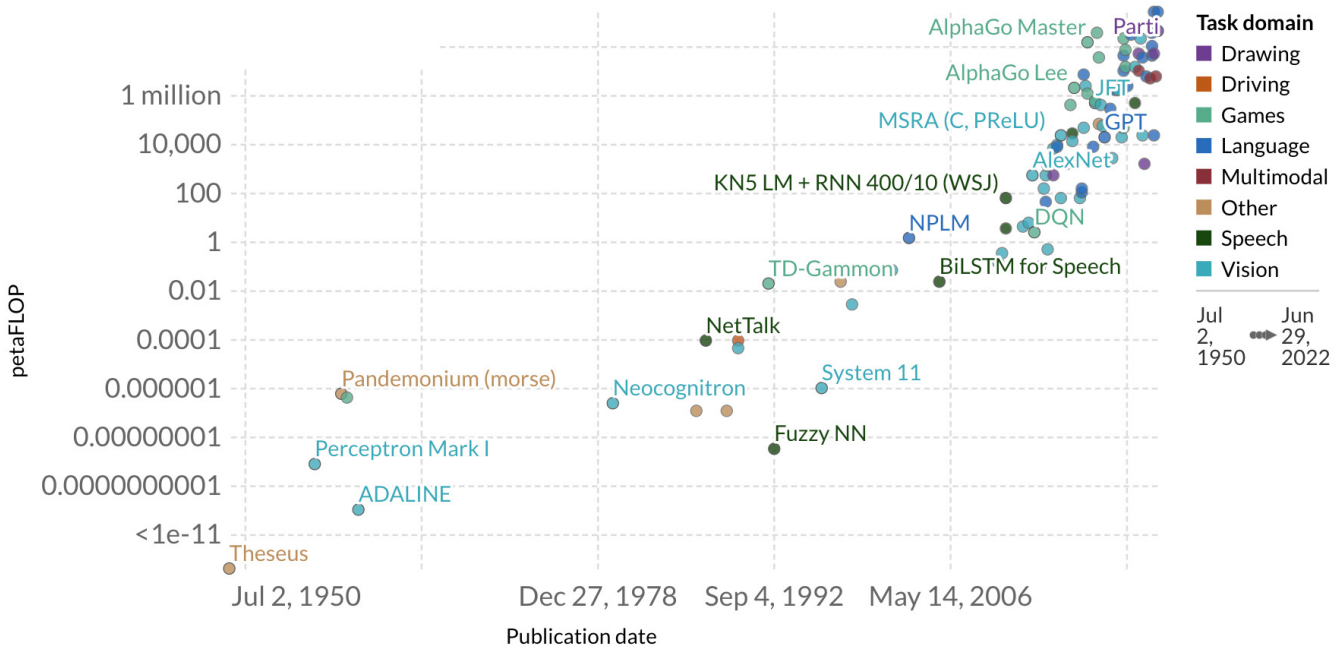
Sources: Compute Across Three Eras of Machine Learning and Brief History of AI

As a result of innovations in neural network architecture, algorithms and compute, models are now delivering superhuman results

- From the 1950s to 2010, training computation increased in line with Moore's Law, 2x every ~20 months.
  - Small models were considered state of the art for understanding language and excelled at analytical tasks but were not expressive enough for general-purpose generative tasks.
- Since 2010, exponential growth sped up further, to ~2x every ~6 months.
  - Compute used to train large-scale models increased by 6 orders of magnitude between 2015 and 2020.
  - Generative models have existed before, but the quality of these large language models (LLMs), such as GPT-4 and PaLM, have improved tremendously.

Computation used to train notable artificial intelligence systems

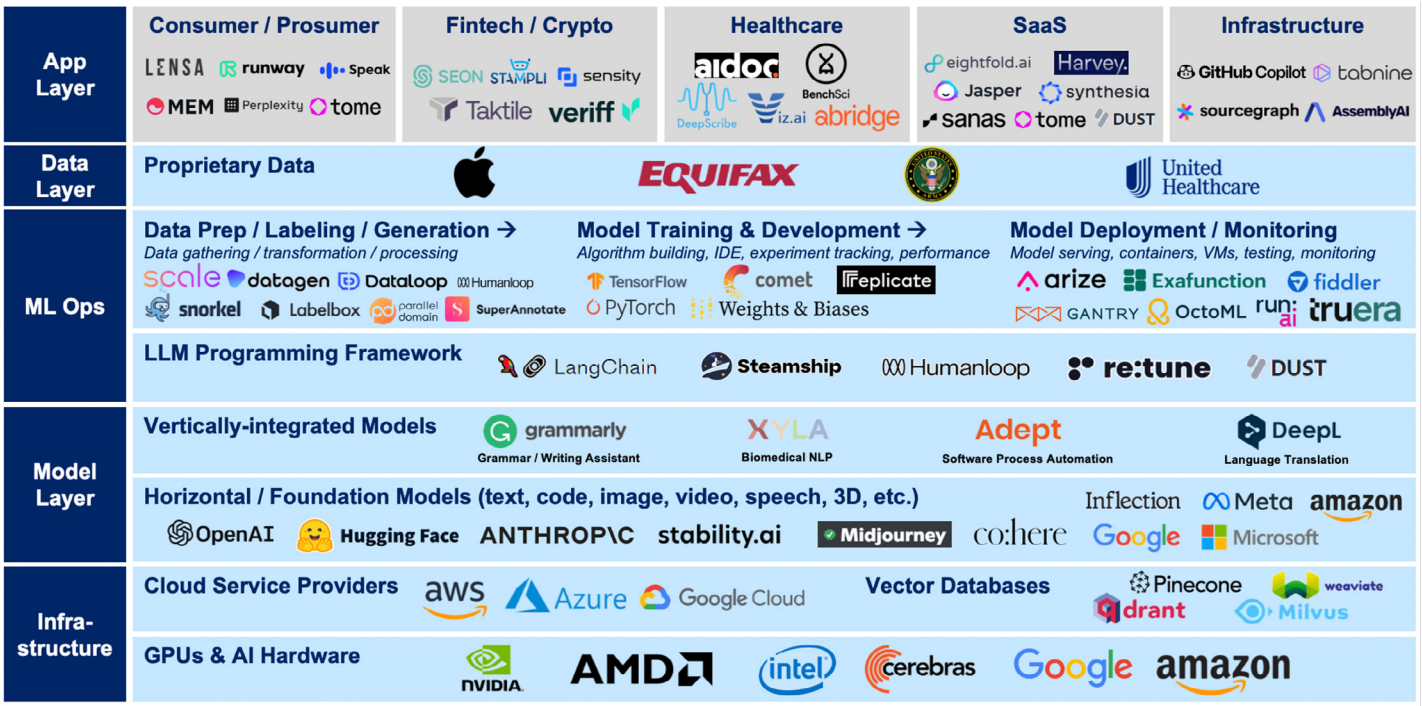
Computation is measured in total petaFLOP, which is 10^{15} floating-point operations.



Source: Sevilla et al. (2022)

Note: Computation is estimated based on published results in the AI literature and comes with some uncertainty. The authors expect the estimates to be correct within a factor of 2.

AI Tech Stack



Layer	Potential Market Structure	Moats?	Key Questions
App Layer	Same as current application tools, although moat / differentiation may be thinner for apps built on foundation models	• Brand / ecosystem • GTM / distribution • Workflow • Data	• Best moat for app layer? (i.e., GTM, ease of use, finetuned data set) • What happens when foundation models build their own front-facing applications and compete with their customers? • How will startups compete against Google and Microsoft's distribution advantage now that they've implemented Generative AI into their product suite?
Data Layer	Fragmented based on data type (e.g., FB owns social data, Google owns search, Amazon has shopping, Apple has health, etc.)	• Proprietary access • R&D costs	• IP ownership: legal & ethical ramifications (e.g., are AI-generated works derivative)? • Who owns the data, and what does the market structure of data aggregators look like? • What types and formats of proprietary data become more highly valued?
ML Ops	Room for multiple winners in each category (data prep, model training & dev, model deployment), but best-of-breed tools should win	• Workflow • Community / dev love	• How quickly will the talent pool of ML practitioners expand? Will software engineers be able to reskill as ML engineers? • How will the growing popularity of foundation models impact the need for traditional ML Ops tooling? What does LLM Ops look like? • Will practitioners standardize on an end-to-end platform, or will they gravitate to best-of-breed?
Model Layer	Non-cooperative oligopoly for general models (Big tech, OpenAI, etc.), verticalized monopolies (Adept, Xyla, etc.), and open-source	• Capital + Compute • R&D / Talent • Performance	• Will there be commoditization? What happens to pricing power over time? Will there be a shift to multi-model? • Will new foundation models emerge for vertical use cases? • Will existing foundation models scale into long-tail use cases? • Will there be regional-specific foundation models (e.g., language / gov't)?
Infra-structure	Fragmented with a few dominant players for databases, and non-cooperative, tight oligopoly for CSPs and Hardware	• R&D / Talent • Proprietary algorithm • Performance • Scale • Supply chain	• Will a specialized vector database vendor be necessary, or will existing incumbents such as cloud providers and traditional data stores provide "good enough" solutions? • How will advancements in processors and chips impact compute costs? • What will happen to the cost of inference for more advanced models?

Key AI Takeaways

AI's primary benefit is force multiplication

- Because it drives costs down, increases speed-to-market and can be applied to every industry, AI has the potential to upend the entire tech ecosystem.
- AI will be the next major technology platform shift.

AI is still formative, but the pace of innovation has inflected

- AI has been around since 1952, but in the last 7 years, advancements in deep learning and compute enabled the size of neural networks to grow by more than 10,000x.
- The latest large language models (LLMs) & other foundation models (released in the past year) have driven interest in this field.

AI's impact on the startup ecosystem will be different than those of past platform shifts

- The disruption caused by the public cloud occurred because on-prem software companies, with perpetual license business models, were slow to migrate, creating an opening for more nimble startups.
- AI can be easily embedded into existing technology products today through foundation models, providing fewer opportunities for startups to disrupt large technology companies.
- Fast-moving startups leveraging AI will generate significant enterprise value, but signficant value will also accrue to tech incumbents.

Foundation models will play a critical role in the future of AI and software development

- Tech companies can't ignore the rate of improvement in foundation models; relying solely on in-house models won't allow them to take advantage of the significant R&D that companies like OpenAI, Anthropic, Stability and others are now doing.
- Building a product that is a "wrapper" around a foundation model without either fine tuning it with proprietary data or working in tandem with more tailored and focused proprietary models will result in commoditization.
- The most successful companies will harness the power of foundation models but use them in conjunction with their core IP (models, data, etc.).
- Most use cases are not currently addressable by a foundation model; there is a long tail of AI use cases for which AI teams will still need to custom develop and train AI models.

Questions re: Foundation Models

As foundation models improve, what will happen to the cost of inference?

- Cost of inference (compute cost to run a model) is affected by model complexity (more complex models require greater compute, increasing the cost), but is also affected by other factors like hardware, optimization techniques and underlying software.
- The increased cost of inference from larger models is offset and, in some cases, greatly surpassed by optimization. For example, running certain flavors of Stable Diffusion is now 1 to 2 orders of magnitude cheaper than it was 6 months ago. We expect optimization to keep pace with growing model complexity and cost of inference to continue to decline toward the cost of compute.

Will companies eventually multi-source them?

- The ability to multi-source foundation models reduces pricing power from existing model layer companies, but the shortage of AI talent and the capital requirements for model training will keep the model layer an oligopoly among big tech and well-funded pureplay AI startups.
- Different foundation models will perform better on different use cases; though we expect companies to multi-source in the short term, what happens in the long term remains unclear.

What does their rise mean for ML Ops? Will LLM Ops become a standalone category?

- ML Ops depend on the AI ecosystem capturing a greater share of enterprise budgets and talent, but if foundation models replace traditional model lifecycle techniques, the need for traditional ML Ops may decrease.
- However, increased interest and applicability of ML is leading to bigger ML teams and increased demand for traditional ML Ops tooling.
- A class of ML Ops tools, called LLM Ops, will emerge to tune and work solely with foundation models and LLMs.

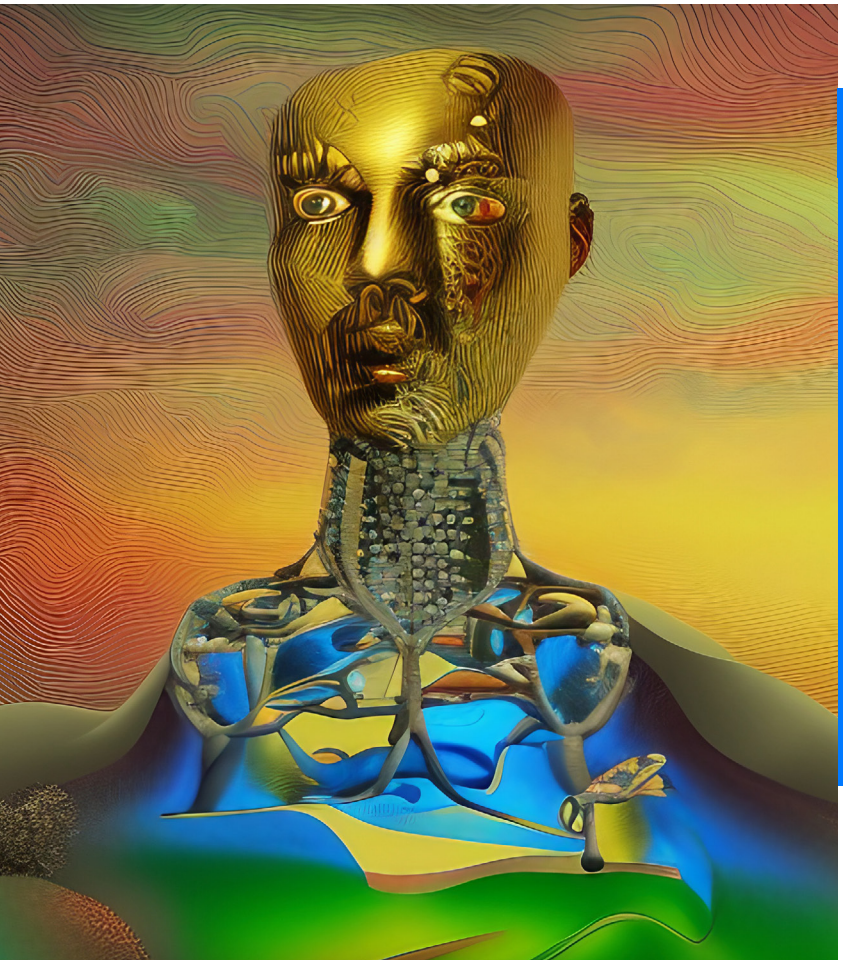
Where is the point of diminishing returns to their improvement?

- At some point, model inference will be "good enough" for specific use cases (e.g., generating blog titles), obviating the need for a more expensive, advanced model.
- Diminishing returns occur when the inference required by the foundation model is not easily found on the internet (for example, image generation foundation models are amazing at generating images similar to those found on the internet but worse at generating images that aren't).

When will enterprises begin to seriously integrate foundation models and LLMs into core products?

- Enterprises are still trying to figure out the ultimate value of LLMs and foundation models; big use cases outside of chat bots are not yet obvious.
- Specific enterprise use case development, LLM tuning and integration into workflow will be the challenging "last-mile" for enterprises.

Thank You



What did you think?

» Let us know.

Image created by Jasper.ai