Machine Learning Infrastructure
- The “hidden figure” of applied AI

Jack Song / Feb 15th, 2023 / CMU Applied AI Class

Hello!

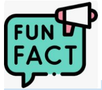
I am Jack Song

- Engineering Director of Machine Learning Infrastructure , Airbnb
- Previous VP of Data Engineering and ML platform , Mastercard

Tips for the talk

- Github project for useful resource links is here https://github.com/jack1981/evoluation_machine_learning_infra/blob/main/README.md
- Learn the tips maybe help you
- Have fun with fun of fact :)

Helpful tips for study and career



Do you know something ?

The story of 'Hidden Figures'

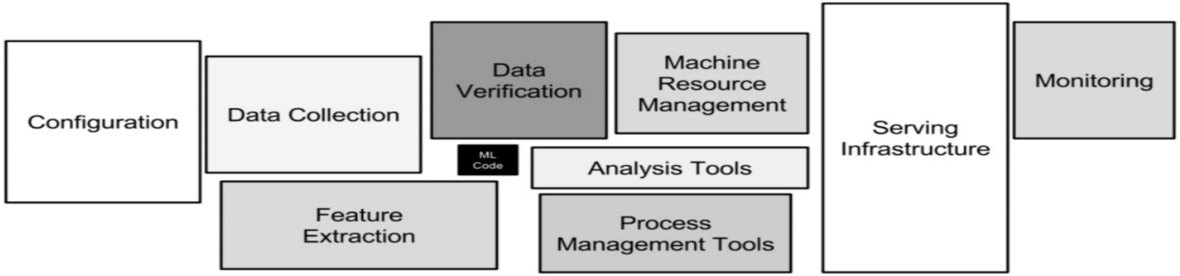
"Hidden Figures" (2016 movie) is a title that's rich with meaning. It refers group of women, **who were doing work for many years and people didn't see them**; There were parts of this whole endeavor of the space program that were very high-profile like the astronauts, test pilots, and Mission Control; but **we didn't really understand how much work went on behind-the-scenes to make that successful**. These women were very much part of that. Their numbers **were the bedrock of so much of the work** that was done in American aeronautics in the 19th century

The truth behind ‘Applied AI’

Hidden Technical Debt

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.



Learning happens mainly at the workplace: **80%** compared to **20%** in school.

1

Machine Learning Infrastructure

The “Hidden Figure” of Applied AI

The rise of Machine learning Infrastructure

- Venturebeat published infrastructure 3.0 for AI revolution.
- Michelangelo: Uber's Machine Learning Platform
- Bighead: Airbnb's End-to-End Machine Learning Infrastructure

FUN FACT

Do you know why Uber named it "Michelangelo" ?

Glance of Industry ML Infra

Research from ML Platform/Infra Meetup Club (M13)

Steering Committee Member :

Airbnb
Cruise
Databricks
Dropbox
Meta
Google
Intuit
Linkedin

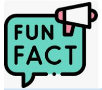
Uber
Netflix
Twitter
Pinterest
Lyft

ML Infra Engineers Size | **700 +**

ML Infra Internal Users Size | **25000 +**

Y : Home grown mostly   N : Buy or OSS mostly

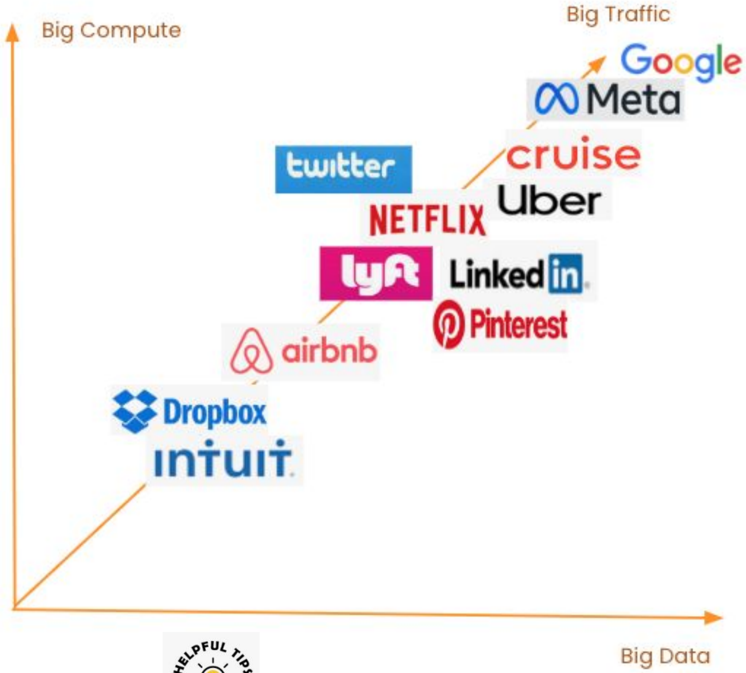
Company	ML lifecycle Tooling /Platform	ML Framework	Distributed Computing	Cluster Management	Build Cloud/ Premise Infra	AI Hardware
Google	Y	Y	Y	Y	Y	Y
Meta	Y	Y	Y	Y	Y	N
Databricks	Y	Y	Y	Y	N	N
Uber, Twitter, Cruise	Y	N	N	Y	Y	N
Airbnb, Lyft, Netflix, Pinterest	Y	N	N	Y	N	N
Linkedin, Dropbox	Y	N	N	N	N	N
Intuit	N	N	N	N	N	N

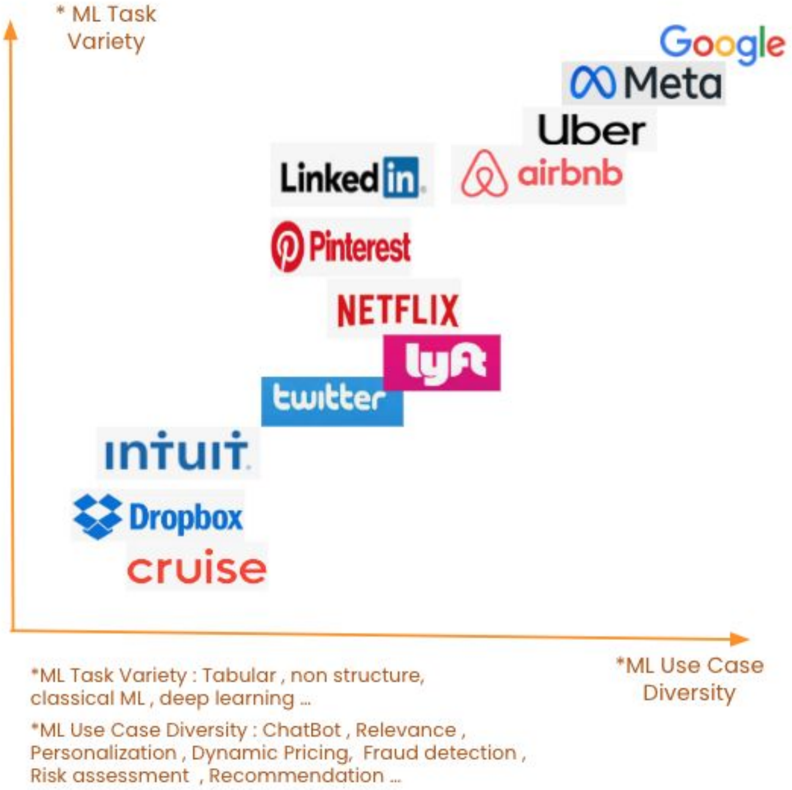


M13 member hosts quarterly private meetup in turn , 75% vote is required to allow new members

Quadrant of ML Infra

Research from ML Platform/Infra Meetup Club (M13)





HELPFUL TIPS

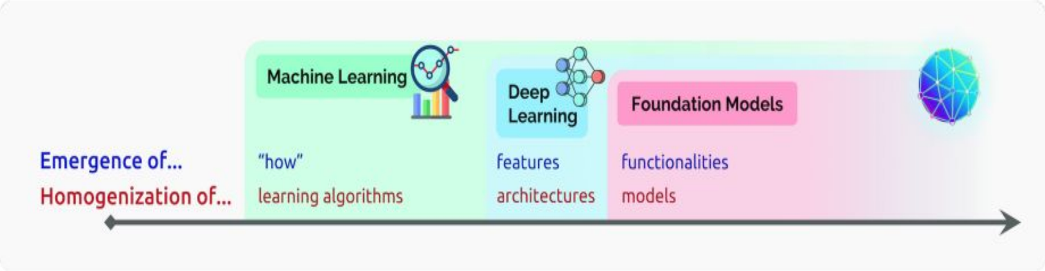
When you have multiple offers ,
which one you should go ?

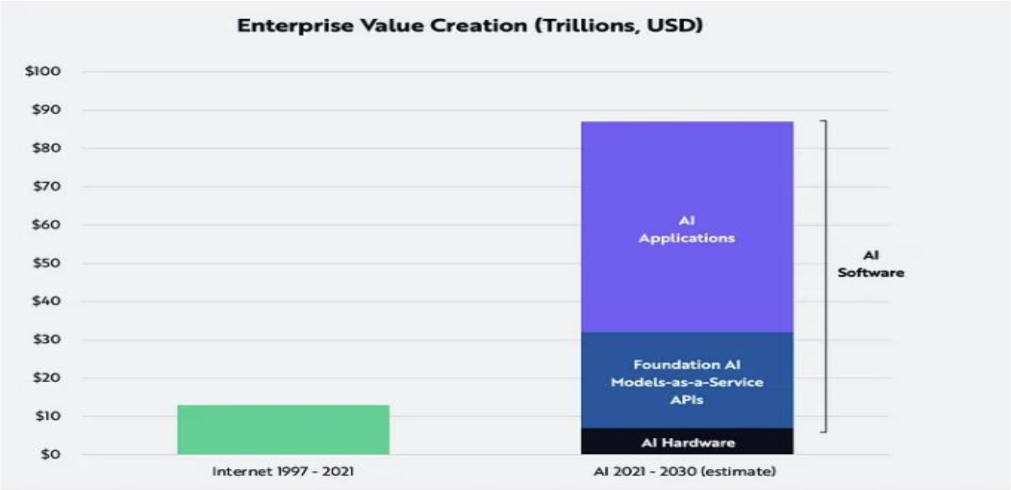
2

The evolution of Machine Learning Infrastructure

- Foundation Models as a service
- Data Centric AI
- Open MLOps

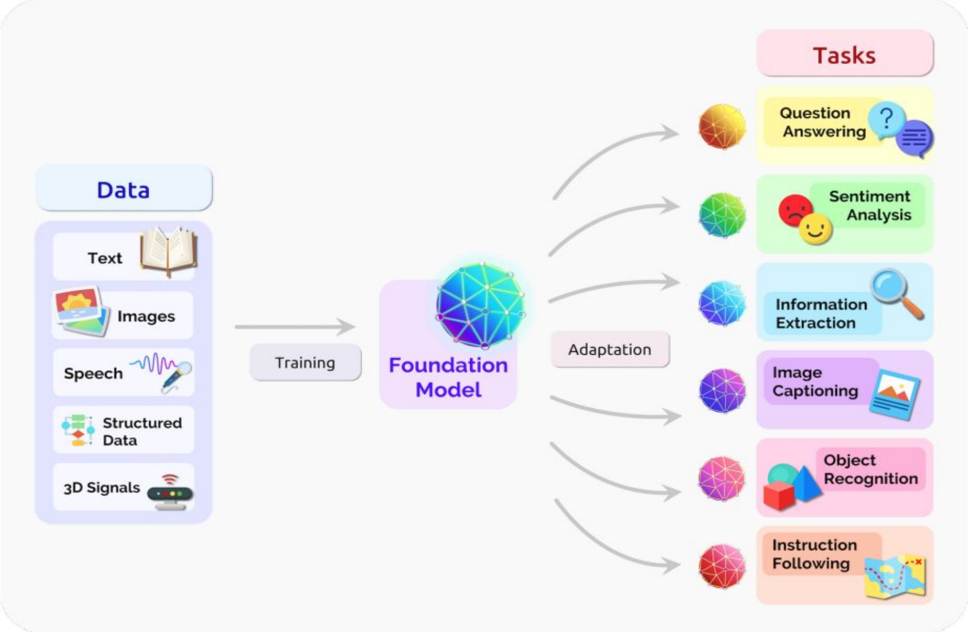
Foundation Models as a service





Source: ARK Investment Management LLC, 2022.

Rise of Foundation Models



A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

~7%	~40%
Foundation Models 2022	Foundation Models 2025

HELPFUL TIPS

Applied AI will been impacted significantly due to rise and development of Foundation Models , lots of existing modeling approaches (including Deep Learning) will be replaced

Foundation Models – Large Language Model

GPT series

Unsupervised learning served as pre-training objective for supervised fine-tuned models, hence the name Generative Pre-training.

GPT 1

- 2018
- **117 million** parameters
- Showed the power of generative pre-training
- NLP

GPT 2

- 2019
- **1.5 billion** parameters
- Unsupervised learning only
- showed better performance due to larger dataset and more parameters
- NLP+AIGC

GPT 3

- 2020
- **175 billion** parameters
- Incontext learning + few-shot learning
- NLP + AIGC+

Instruct GPT

- 2022.01
- **175 billion+1.3 billion fine tune** parameters
- Instruction Tune **GPT 3** with **zero-shot learning** ,***RLHF** and ***COT**
- More truthful and less toxic
- NLP + AIGC+ (Code + Chat)

ChatGPT

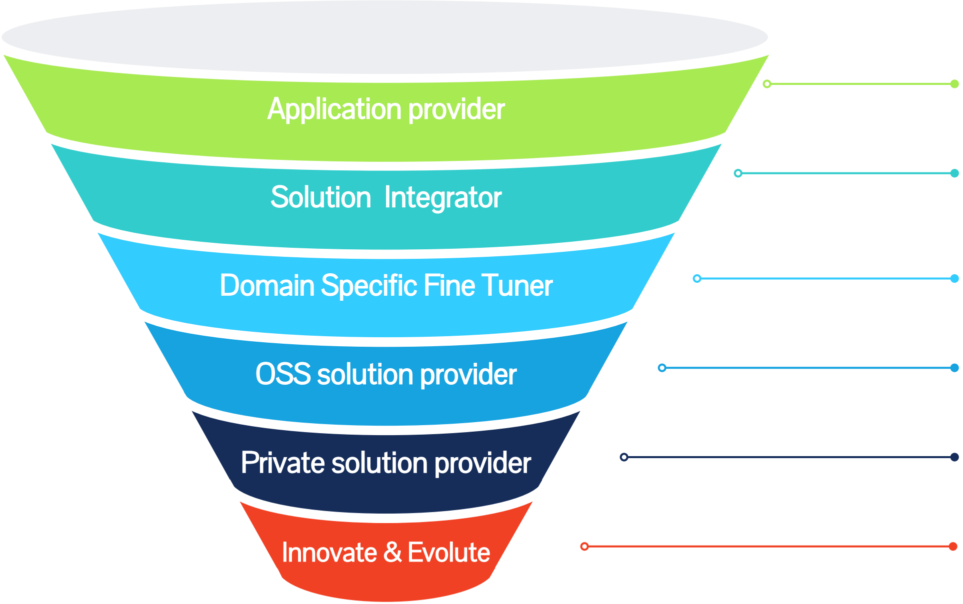
- 2022.12
- **175 billion +** parameters
- Instruction tune **Instruct GPT** with human-generated prompts and example responses
- More Safety
- NLP + AIGC+ (Chat)

RLHF:reinforcement learning from human feedback

COT:chain-of-thought

Foundation Models shifted the learning and research patterns

Adoption patterns of Foundation Models



60% ,leverage integrated solution / service and adopt them into own application, lightweight MLI (Japer , Grammarly …)

20% , integrate solutions and host them, integrate type of MLI , cloud mostly (Snorkel AI, Scaling AI …)

10% , fine tune and host domain specific solution , in house + cloud MLI (companies like Airbnb)

Less than 5%, clone generative solution and OSS, in house + cloud MLI (Hugging face…)

Less than 4% , clone generative solution and private, in house MLI mostly (cohere.ai , Landing.ai …)

Less than 1% , control by very few players, in house MLI (Open AI , Google …)

The bar of skills and experiences became higher from top to the bottom

Data Centric AI

> “Data is food for AI
>
> Andrew Ng, pioneer of the data-centric AI philosophy

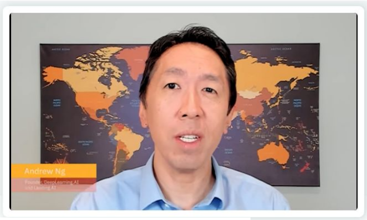
> More data beats clever algorithms, but better data beats more data
>
> Peter Norvig , Distinguished Education Fellow at the Stanford Institute for Human-Centered AI

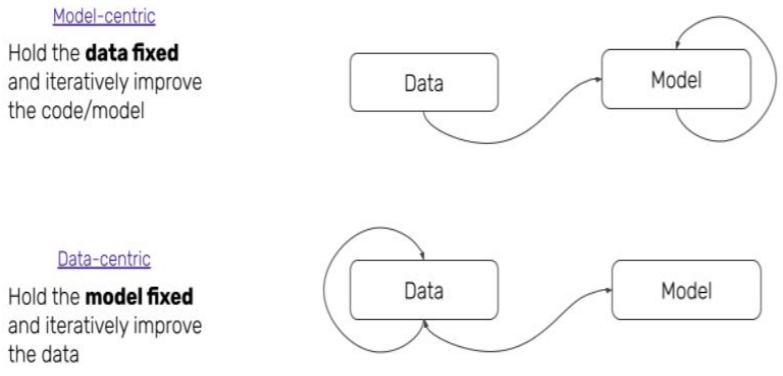
What is Data Centric AI ?

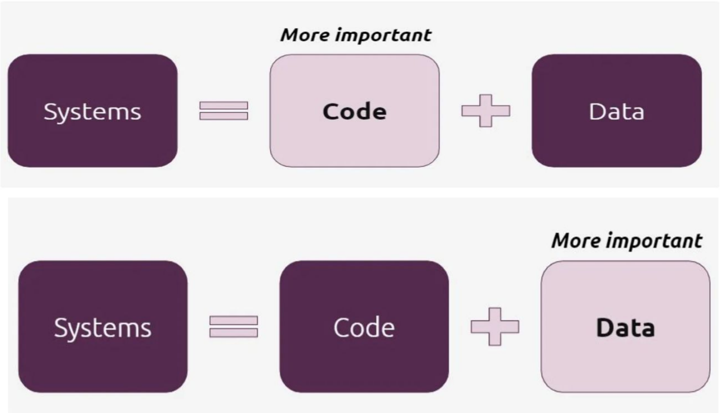
A data-centric AI approach provides a systematic method for improving data, reaching a consensus on the data, and cleaning up inconsistent data.

Andrew NG's story :

The founder of the Google Brain research lab, co-founder of Coursera, and former chief scientist at Baidu,also the founder and CEO of Landing AI.







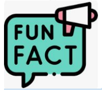
Call to action 1 : Bring human in the loop

Mentally scarred: Kenyan workers taught ChatGPT to recognize offensive text

Spare a thought for the workers who read the worst content scraped from the internet to keep you safe

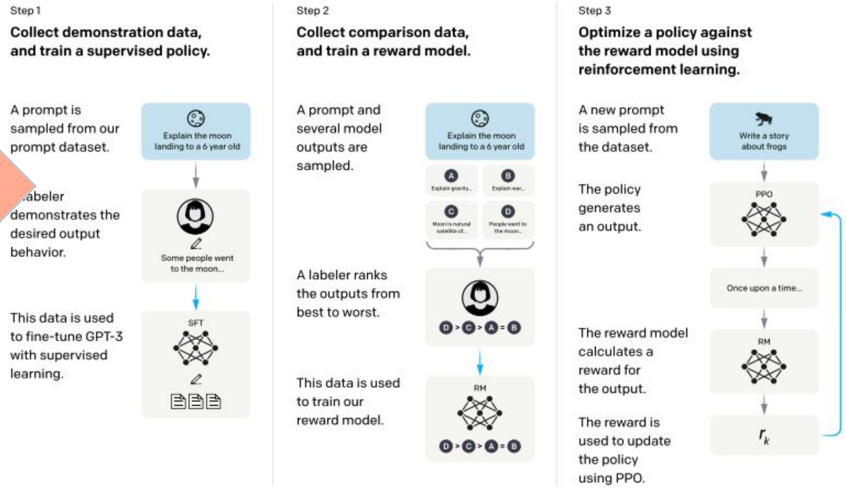
Katyanna Quach   Fri 20 Jan 2023 // 14:15 UTC

OpenAI reportedly hired workers in Kenya – screening tens of thousands of text samples for sexist, racist, violent and pornographic content – to help make its ChatGPT model less toxic.



5000+ labelers , $2 per hour

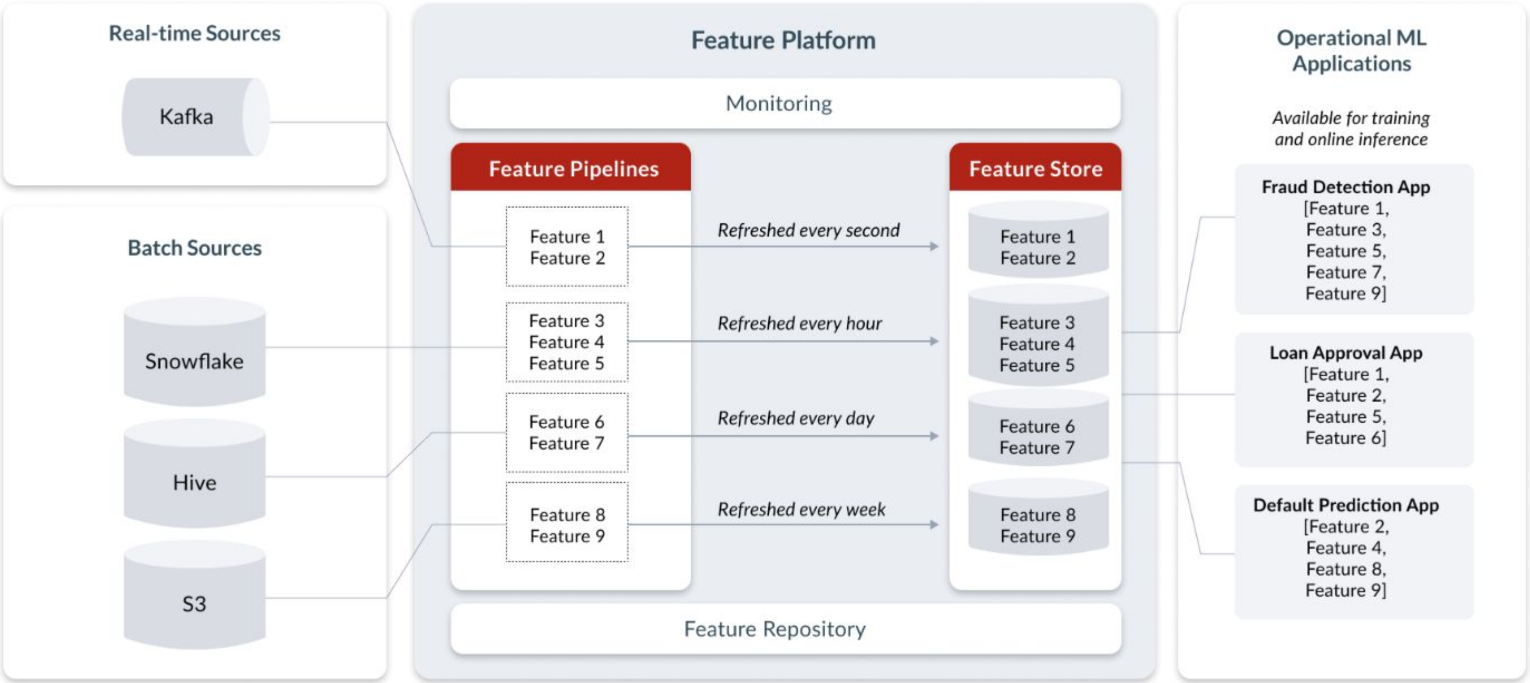
Training language models to follow instructions with human feedback



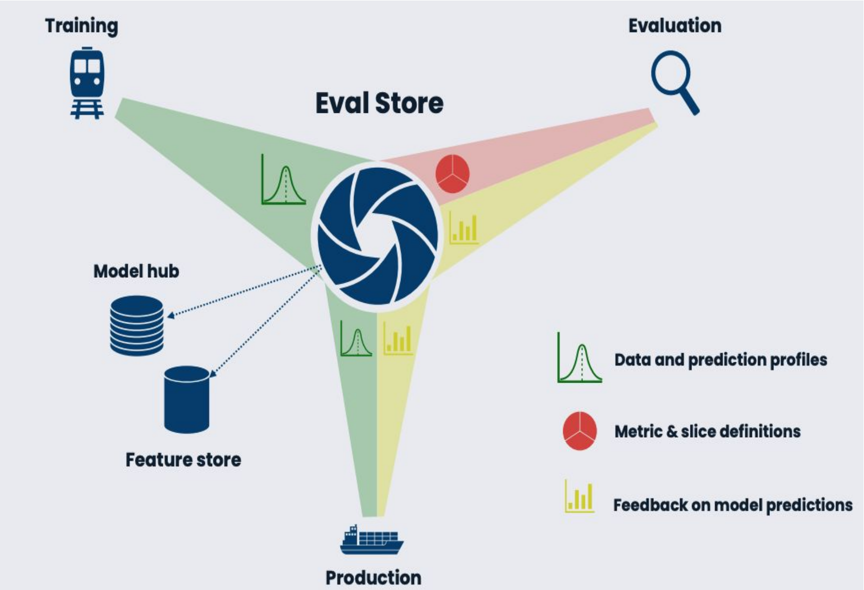
Labelers significantly prefer InstructGPT outputs over outputs from GPT-3. On our test set, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having over 100x fewer parameters. These models have the same architecture, and differ only by the fact that InstructGPT is fine-tuned on our human data. This result holds true even when we add a few-shot prompt to GPT-3 to make it better at following instructions. Outputs from our 175B InstructGPT are preferred to 175B GPT-3 outputs 85 ± 3% of the time, and preferred 71 ± 4% of the time to few-shot 175B GPT-3. InstructGPT models also generate more appropriate outputs according to our labelers, and more reliably follow explicit constraints in the instruction.

RLHF

Call to action 2 : Feature platform



Call to action 3 : Evaluation Store



A central place to store and query online and offline ground truth and approximate model quality metrics

- **Reduce organization friction** : Get stakeholders (ML Eng,DS, PM, etc) on the same page about metric and slice definitions
- **Deploy models more confidently**: Evaluate metrics and slices consistently in testing and prod. Make the metrics visible to stakeholders
- **Catch production bugs faster** : Catch degradations across any slice, and drill down to the data that caused the degradation
- **Reduce data-related costs** : Collect and label production data more intelligently
- **Make your model better** : Decide when to retrain. Pick the right data to retrain on.

Call to action 4 : ML Observability

4 Pillars of ML Observability

- **Drift**: Data Distribution Changes over lifetime of model
- **Performance Analysis**: Surfacing worst performing slices
- **Data Quality**: Ensure high quality inputs & outputs
- **Explainability**: Attribute why a certain outcome was made

Open MLOps

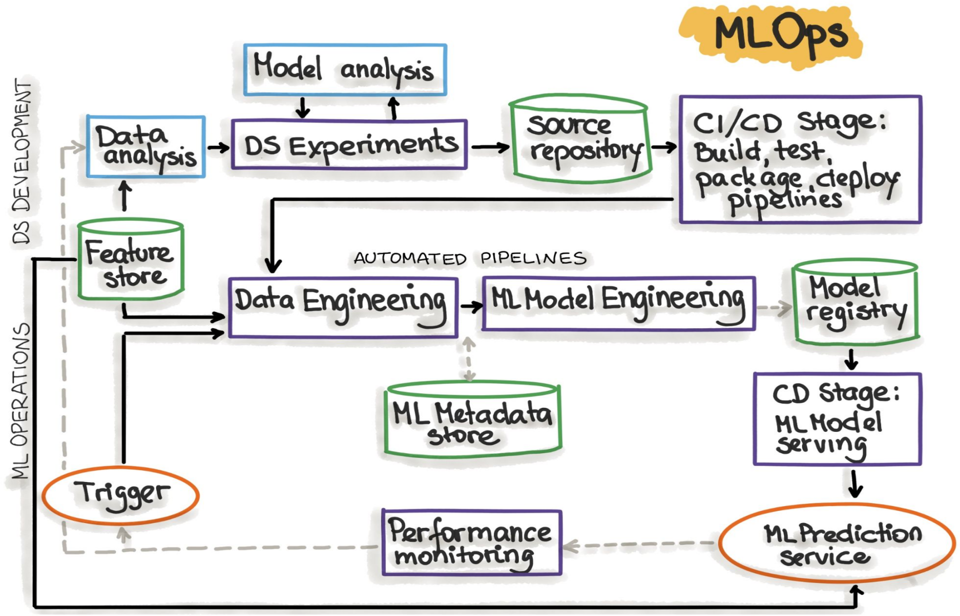
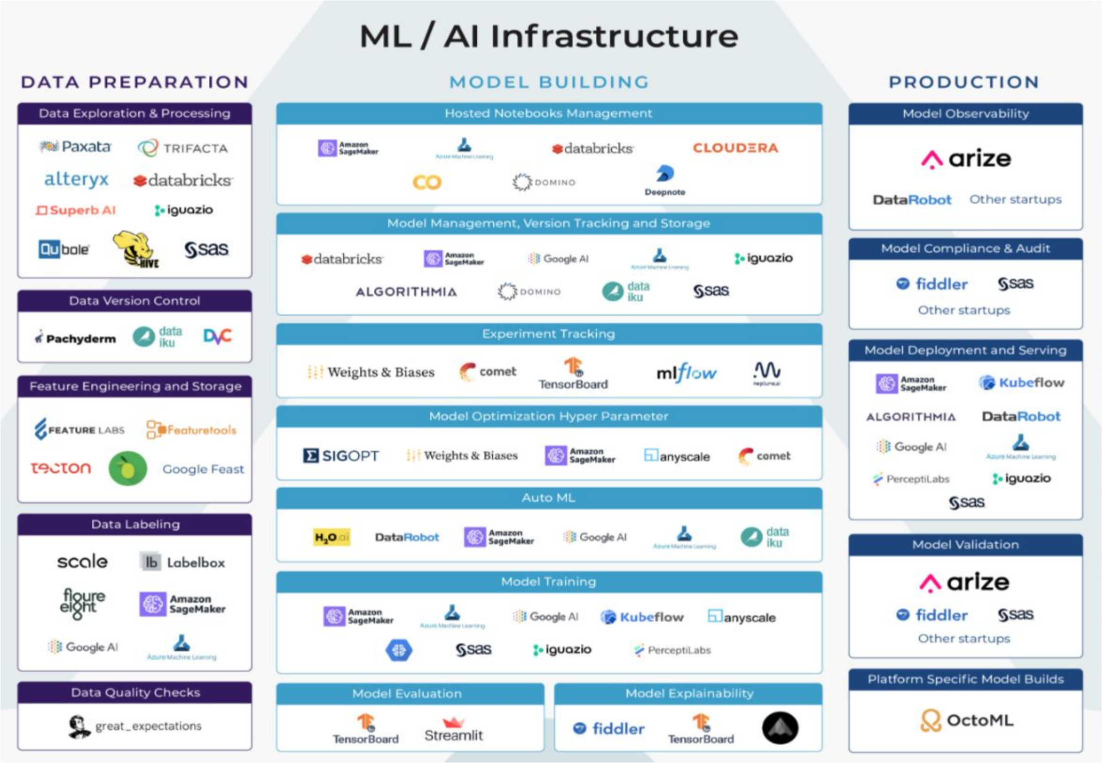


Figure adopted from "MLOps: Continuous delivery and automation pipelines in machine learning"

MLOps Infrastructure eco is very chaotic



The MLOps Infrastructure space is crowded, confusing, and complex. There are a number of platforms and tools spanning a variety of functions across the model building workflow.

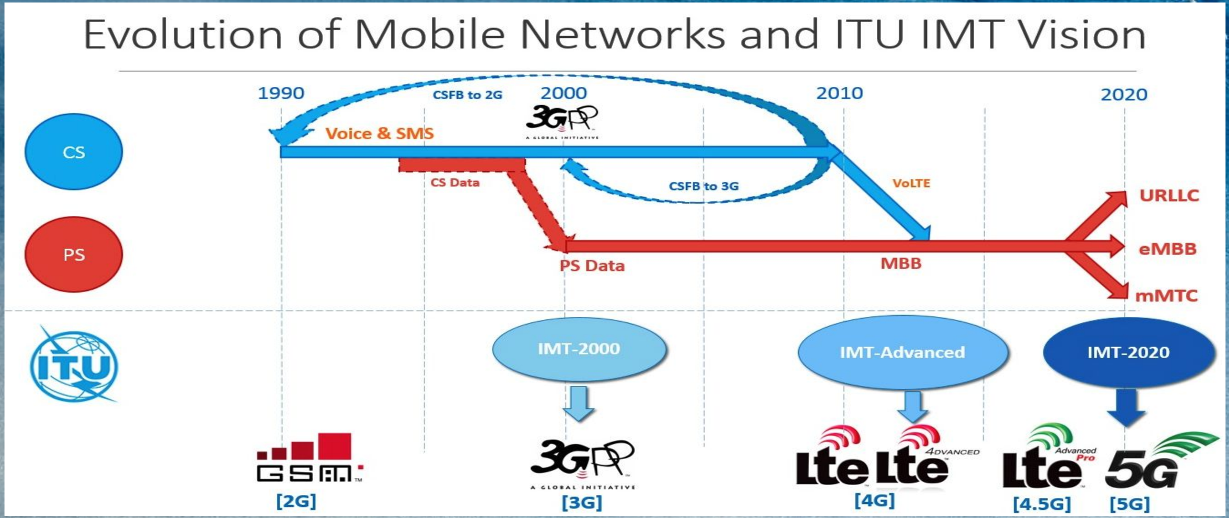
Tech companies are reinventing the wheel

"Ice and Fire"

Google
Meta
Uber
cruise
Pinterest
airbnb
NETFLIX
lyft
intuit
LinkedIn
databricks
Dropbox
twitter

Openness and Standardization are catalysts of Evolution

Example : 2G ~ 5G Mobile Networks



Academic institutions advocates the evolution

Example : Machine Learning Systems Design Course from Stanford

CS 329S Overview Syllabus FAQ Past course Past projects

CS 329S: Machine Learning Systems Design

Stanford, Winter 2022

Schedule & syllabus

The lecture slides, notes, tutorials, and assignments will be posted online here as the course progresses.
Lecture times are **3:15 - 4:45pm PST**. All deadlines are at **11:59pm PST**.

This schedule is subject to change according to the pace of the class.

See Past course for the last year's lectures.

Date	Description	Materials	Events
Mon Jan 3	Understanding machine learning production	Lecture note Lecture slides Lessons learned from 150 ML models at Booking.com	Lecture
Wed Jan 5	ML and Data Systems Fundamentals	Lecture note Lecture slides Case study: Predict Value of Homes On Airbnb Breakout exercise: Designing Twitter's Trending Hashtags	Lecture
Mon Jan 10	Training Data	Lecture note Lecture slides	Lecture
Wed Jan 12	Feature engineering	Lecture note Lecture slides	Lecture
Mon Jan 17	No class	Martin Luther King, Jr. Day	
Wed Jan 19	Model selection, development, and training	Lecture note Lecture slides	Lecture
Mon Jan 24	Offline evaluation	Lecture note Lecture slides	Lecture
Wed Jan 26	Model evaluation	Goku's ML tutorial	Tutorial

Why machine learning systems design?

Machine learning systems design is the process of defining the software architecture, infrastructure, algorithms, and data for a machine learning system to satisfy specified requirements.

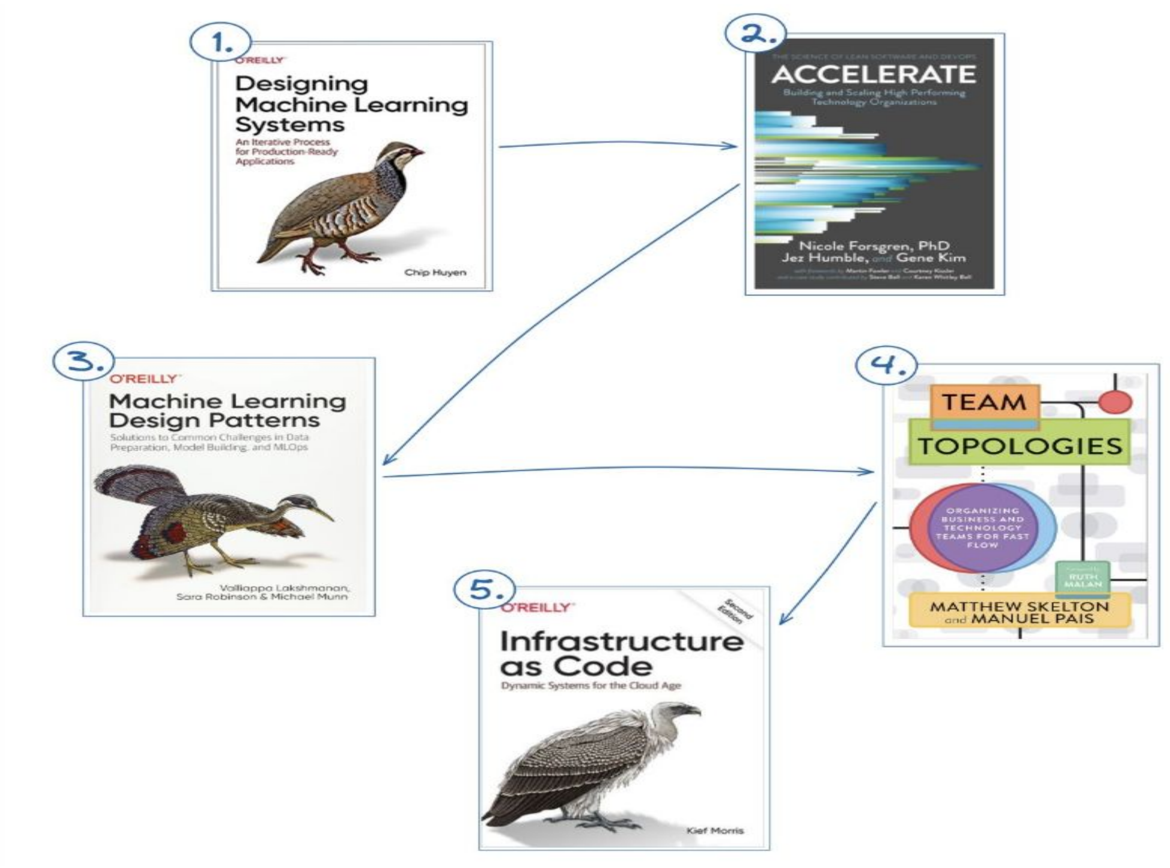
The tutorial approach has been tremendously successful in getting models off the ground. However, the resulting systems tend to go outdated quickly because (1) the tooling space is being innovated, (2) business requirements change, and (3) data distributions constantly shift. Without an intentional design to hold all the components together, a system will become technical liability, prone to errors and quick to fall apart.

Rise of Open MLOps

"2023 Dream Team "

- **Jupyter**, still the most popular notebook platform open-source product, Label Studio from HearTex, made significant improvements since 2021, including active learning enhancements.
- **Michelangelo**, the famous ML lifecycle solution from Uber since 2017, is proceeding private beta approach before fully open source.
- **Chronon(Zipline) ,** the competitive feature store solution from Airbnb since 2017,are running a couple of private beta trials before fully open source
- **Metaflow** from Netflix, already open source, has been proved to be a strong candidate for the ML flow engine.
- **ML Flow** from Databricks, the leading solution for ML experiments, is fully open source.
- **Looper** is a new evaluation store solution from Meta published in 2021. The team had a plan to open source it at 2023
- **Ray**, from Anyscale, fully open source, had more and more adoptions and showed the leading position of training & turning.
- **Kubeflow** is one of the popular open-source deployment and serving engines to connect ML and Cloud infra closer and better.
- **Alibi** from Seldon, for ML interpretability, has the most active open-source community in this area.
- **MLP observability** , Lyft spent years enhancing it, and they claimed it is a better solution than most of the commercial vendors in the market; the team had a plan to open source it at 2023

Book recommendation for MLI and MLOps



Thanks!

Any questions?

You can find me at

- https://www.linkedin.com/in/suqiang-song-72041716/
- jackssqcyy@gmail.com

Disclaimers

All trademarks are the properties of their respective owners. Any use of these are for identification purposes only and do not imply sponsorship or endorsement