

Machine Learning Systems Design

Lecture 1: Understanding ML production



Reply in Zoom chat:

Where are you (physically)?

Agenda

1. Course overview
2. ML research vs. ML production
3. Breakout exercise
4. ML systems vs. traditional software
5. ML production myths

1. Short class today
2. Lecture note is on course website / syllabus

1. Course overview

Fri, Jan 06, 2012, 12:27:51 AM Pacific Standard Time



Posted by u/[deleted] 10 years ago

19



Is a PhD in Artificial Intelligence good for getting jobs?

longer than in many other fields. You won't get rich doing AI research and you probably won't have a big direct real-world impact. However, if you have an entrepreneurial mindset, the ideas and perspectives you develop working in AI can be a great springboard for other things.

By googling you can easily find more. If you are actually looking for an academia job the articles are still worth to read. There the PhD is a necessity, but job opportunities are very scarce.



15



Reply

Give Award

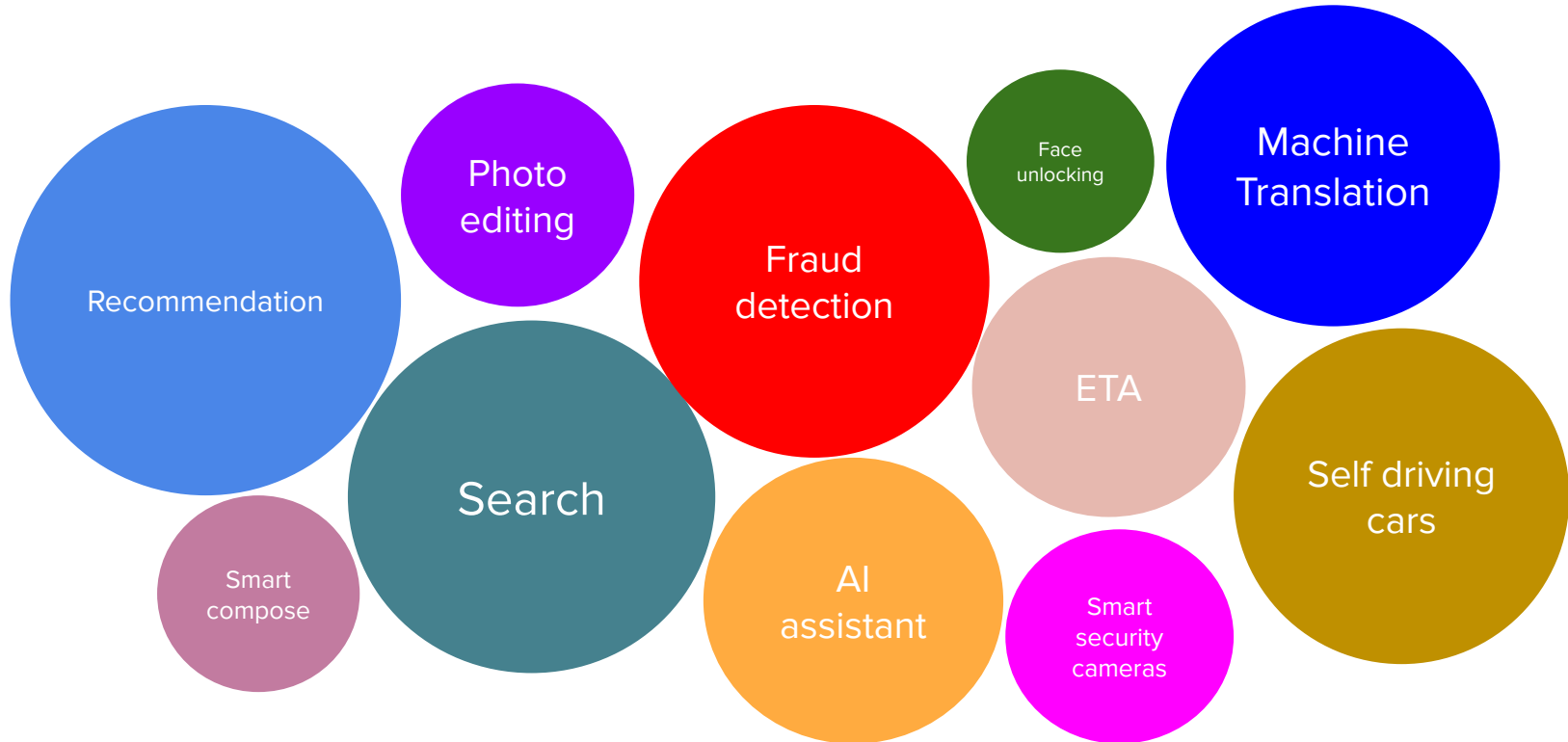
Share

Report

Save

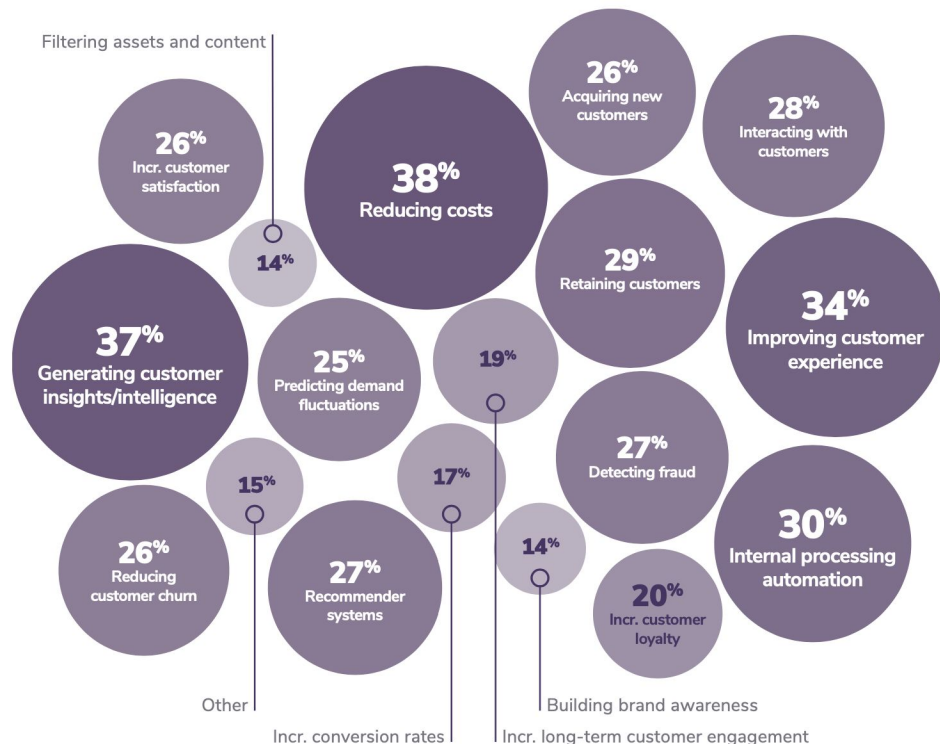
Follow

2022: ML is in almost every aspect of our lives



Enterprise use cases

Machine learning use case frequency



Why ML Systems Design?

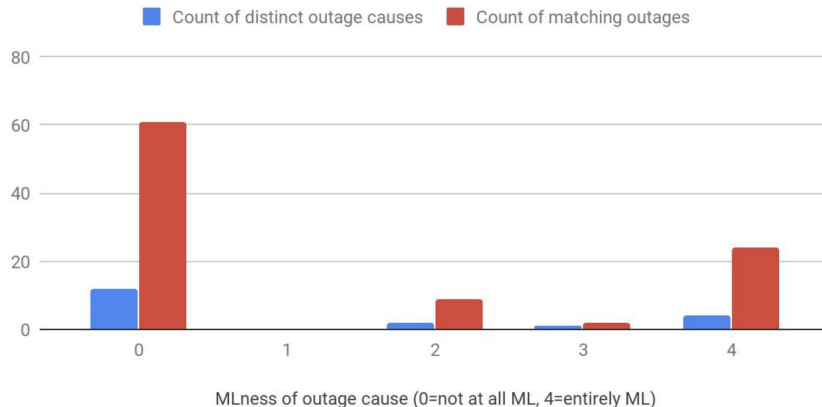
- ML algorithms is the less problematic part.
- The hard part is to **how to make algorithms work with other parts to solve real-world problems.**

Why ML Systems Design?

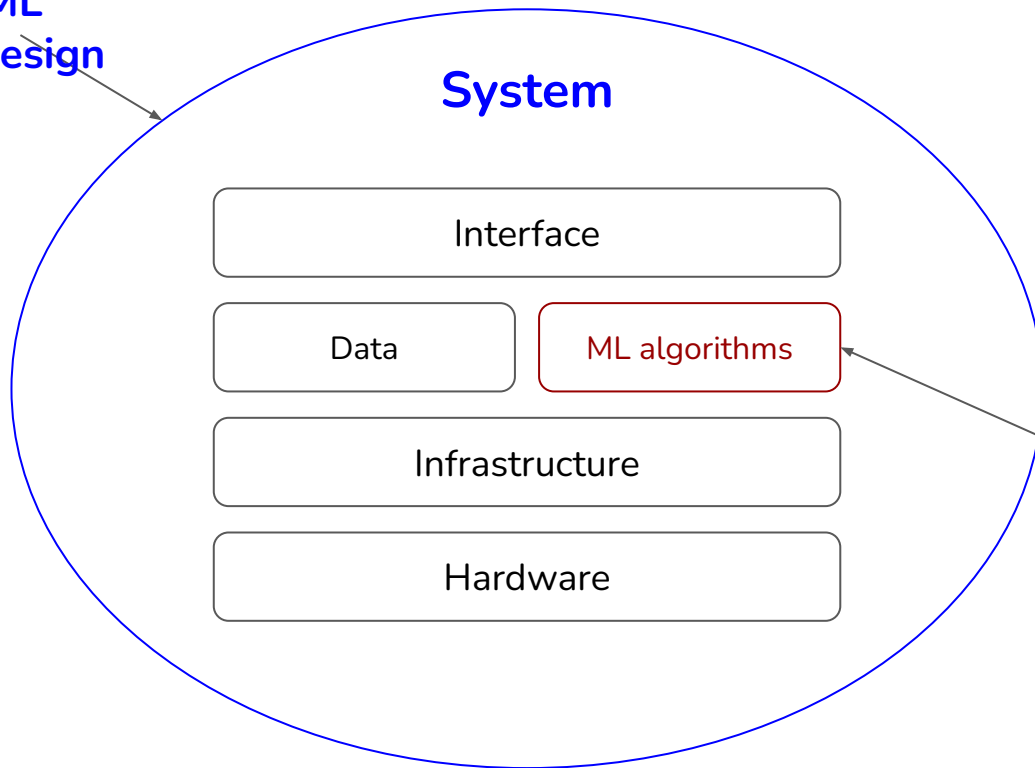
- ML algorithms is the less problematic part.
- The hard part is to **how to make algorithms work with other parts to solve real-world problems.**
- [60/96 failures](#) caused by non-ML components

More on ML systems failures later!

Count of distinct outage causes and Count of matching outages



CS 329S: ML
Systems Design



Most ML
courses/books

What's machine learning systems design?

The process of defining the **interface**, **algorithms**, **data**, **infrastructure**, and **hardware** for a machine learning system to satisfy **specified requirements**.

What's machine learning systems design?

The process of defining the **interface**, **algorithms**, **data**, **infrastructure**, and **hardware** for a machine learning system to satisfy **specified requirements**.



reliable, scalable, maintainable, adaptable

The questions this class will help answer ...

- You've trained a model, now what?
- What are different components of an ML system?
- How to do data engineering?
- How to engineer features?
- How to evaluate your models, both offline and online?
- What's the difference between online prediction and batch prediction?
- How to serve a model on the cloud? On the edge?
- How to continually monitor and deploy changes to ML systems?
- ...

This class will not teach ...

- Machine learning/deep learning algorithms
 - CS 229: Machine Learning
 - CS 230: Deep Learning
 - CS 231N: Convolutional Neural Networks for Visual Recognition
 - CS 224N: Natural Language Processing with Deep Learning
- Computer systems
 - CS 110: Principles of Computer Systems
 - CS 140E: Operating systems design and implementation
- UX design
 - CS 147: Introduction to Human-Computer Interaction Design
 - DESINST 240: Designing Machine Learning: A Multidisciplinary Approach

Machine learning: expectation



This class won't teach
you how to do this

Machine learning: reality

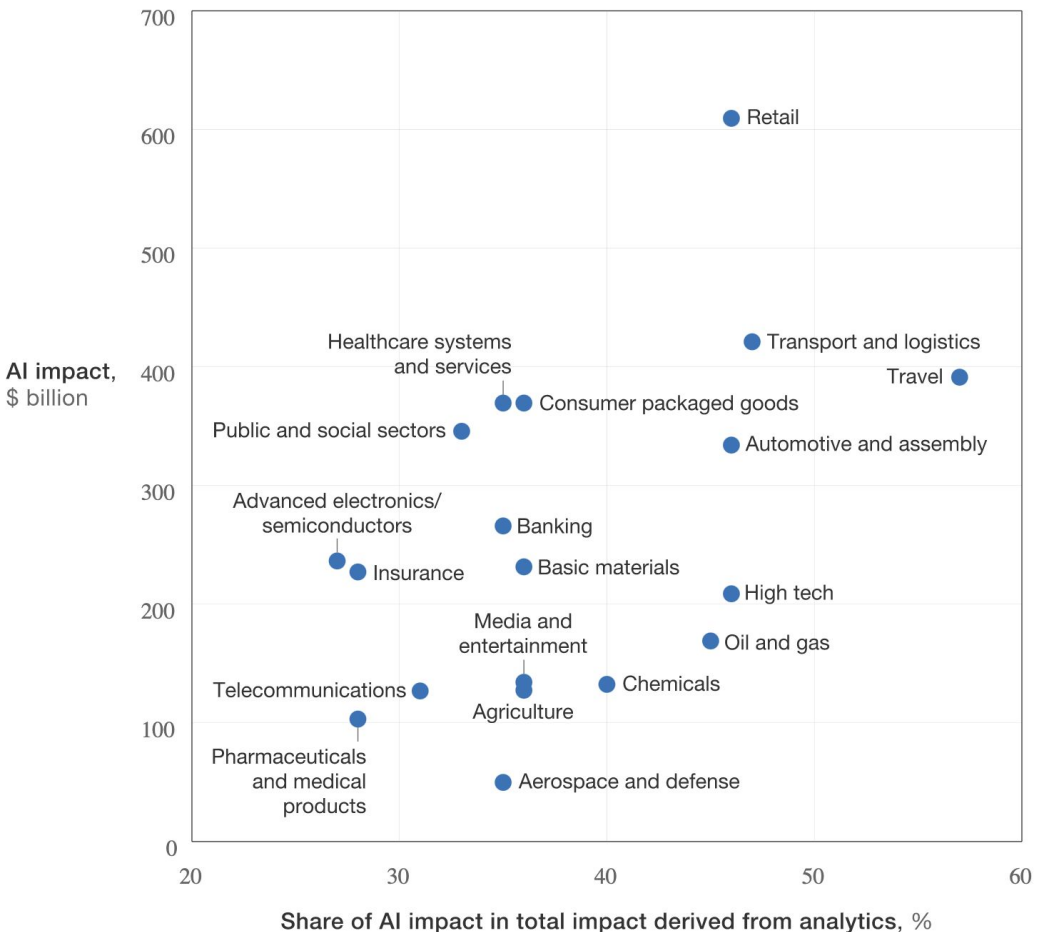


You'll likely build something like this (buggy but cool)

Prerequisites

- Knowledge of CS principles and skills (CS 106B/X)
- Understanding of ML algorithms (CS 229, CS 230, CS 231N, or CS 224N)
- Familiar with at least one framework such as TensorFlow, PyTorch, JAX
- Familiarity with basic probability theory (CS 109/Stat 116)

Artificial intelligence (AI) has the potential to create value across sectors.



AI value creation by 2030

13 trillion USD

Most of it will be outside the consumer internet industry

We need more people from non-CS background in AI!

In-person or remote?

- Lectures: likely fully remote
- OHs: mixture of remote + in-person

Let us know if you have any
questions/concerns

Zoom etiquettes

- Write questions into Zoom chat
 - Feel free to reply to each other — TAs will also reply
- I will stop occasionally for Q&A

Zoom etiquettes

We appreciate it
if you keep videos on!

- More visual feedback for us to adjust materials
- Better learning environment
- Better sense of who you're with in class!



**WAITING FOR STUDENTS TO TURN VIDEOS ON SO
I DON'T FEEL LIKE I'M TALKING TO AN EMPTY ROOM**

Grading

- Assignments (30%)
 - 2 assignments
- Final project (65%)
- Class participation (5%)
 - Zoom questions + EdStem + OHs
 - Bad sign if by the end of the quarter, we still don't know who you are

Final project

- Build an ML-powered application
- Must work in groups of three
- Demo + report (creative formats encouraged)
- Evaluated by course staff and industry experts

Poll: are you looking for teammates for the final project?

Final project

- Build an ML-powered application
- Must work in groups of three
- Demo + report (creative formats encouraged)
- Evaluated by course staff and industry experts

Session next week to discuss project ideas
+ find potential team mates?

Honor code: permissive but strict - don't test us ;)

- OK to search, ask in public about the systems we're studying. Cite all the resources you reference.
 - E.g. if you read it in a paper, cite it. If you ask on Quora, include the link.
- NOT OK to ask someone to do assignments/projects for you.
- OK to discuss questions with classmates. Disclose your discussion partners.
- NOT OK to copy solutions from classmates.
- OK to use existing solutions as part of your projects/assignments. Clarify your contributions.
- NOT OK to pretend that someone's solution is yours.
- OK to publish your final project after the course is over (we encourage that!)
- NOT OK to post your assignment solutions online.
- ASK the course staff if unsure!

Course staff

Kinbert & Chloe took the course last year so they can relate to your pain!



Chip Huyen



Megan Leszczynski



Kinbert Chou



Chloe He

Work in progress

- Second time the course is offered
- The subject is new, we don't have all the answers
 - We are all learning too!
- We appreciate your:
 - **enthusiasm** for trying out new things
 - **patience** bearing with things that don't quite work
 - **feedback** to improve the course

- <https://cs329s.stanford.edu>
- Discussions:
 - [EdStem](#)
 - cs329s-win2022-staff@lists.stanford.edu only for private matters
- OHs start next week
- If you enrolled without submitting an application, send us an email!
- Questions so far?

2. ML research vs. ML production

ML research vs. ML production

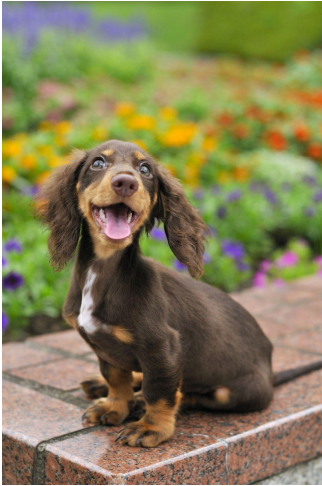
	Research	Production
Objectives	Model performance*	Different stakeholders have different objectives

“*” It’s actively being worked. See [Utility is in the Eye of the User: A Critique of NLP Leaderboards](#) (Ethayarajh and Jurafsky, EMNLP 2020)

Stakeholder objectives

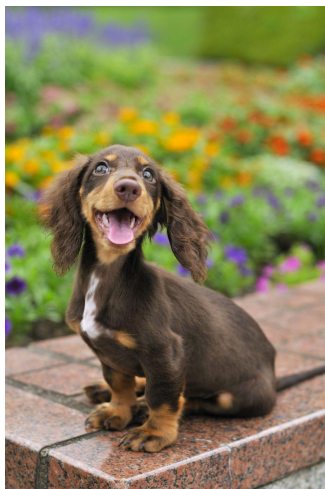
ML team

highest accuracy



Stakeholder objectives

ML team
highest accuracy

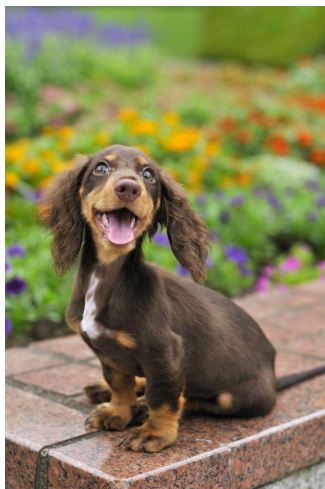


Sales
sells more ads



Stakeholder objectives

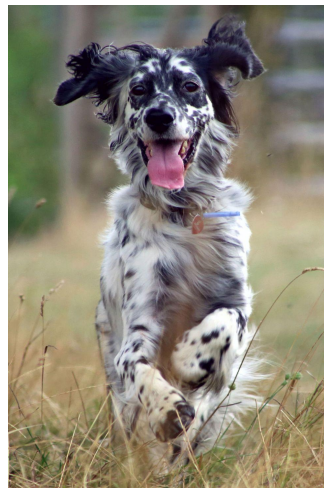
ML team
highest accuracy



Sales
sells more ads

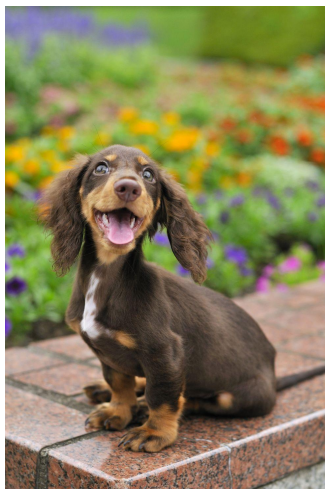


Product
fastest inference

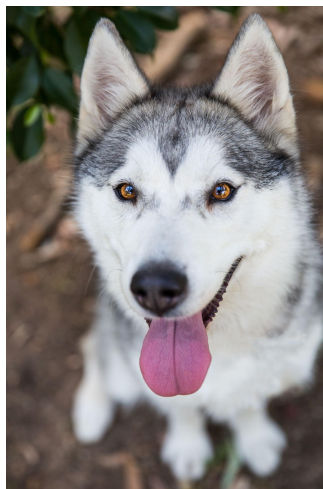


Stakeholder objectives

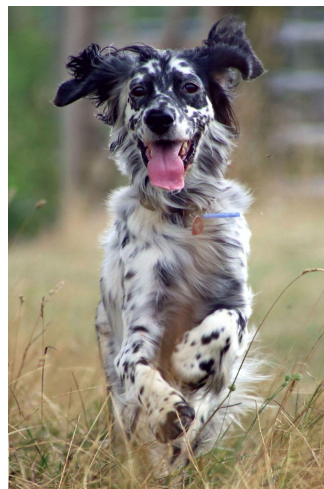
ML team
highest accuracy



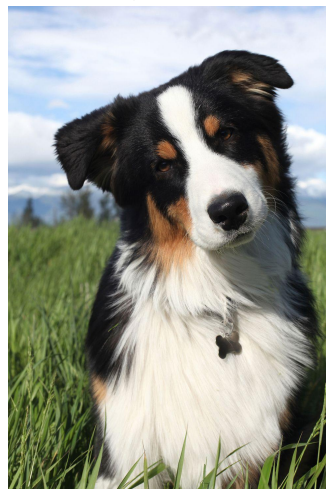
Sales
sells more ads



Product
fastest inference



Manager
maximizes profit
= laying off ML teams



Leaderboard-style ML

- More comprehensive utility function
 - Model performance (e.g. accuracy)
 - Latency
 - Prediction cost
 - Interpretability
 - Robustness
 - Ease of use (e.g. OSS tools, community support)
 - Hardware requirements
- Adaptive to different use cases
 - Instead of a leaderboard for each dataset/task, the leaderboard adapts to each company's needs
- Dynamic datasets
 - Realistic distribution shifts with different types of shifts

Computational priority

	Research	Production
Objectives	Model performance	Different stakeholders have different objectives
Computational priority	Fast training, high throughput	Fast inference , low latency



generating predictions

Latency matters

- 100ms delay can hurt conversion rates by 7% ([Akamai study](#) '17)
- 30% increase in latency costs 0.5% conversion rate ([Booking.com](#) '19)
- 53% phone users will leave a page that takes >3s to load ([Google](#) '16)



- Latency: time to move a leaf
- Throughput: how many leaves in 1 sec



- Real-time: low latency = high throughput
- Batched: high latency, high throughput

ML in research vs. in production

	Research	Production
Objectives	Model performance	Different stakeholders have different objectives
Computational priority	Fast training, high throughput	Fast inference, low latency
Data	Static	Constantly shifting

Data

Research	Production
<ul style="list-style-type: none">• Clean• Static• Mostly historical data	<ul style="list-style-type: none">• Messy• Constantly shifting• Historical + streaming data• Biased, and you don't know how biased• Privacy + regulatory concerns

THE COGNITIVE CODER

By **Armand Ruiz**, Contributor, InfoWorld | SEP 26, 2017 7:22 AM PDT

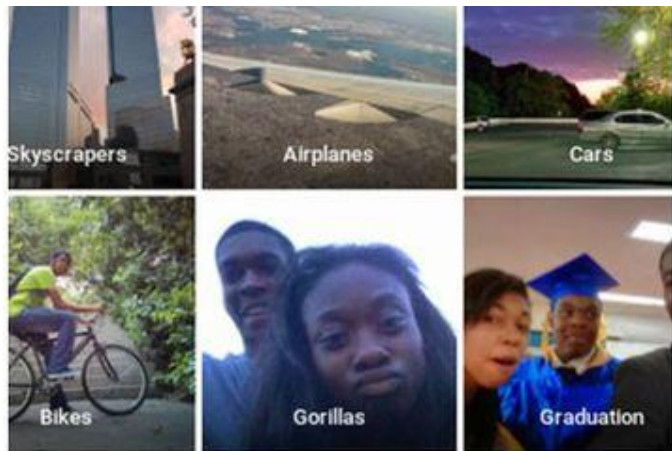
The 80/20 data science dilemma

Most data scientists spend only 20 percent of their time on actual data analysis and 80 percent of their time finding, cleaning, and reorganizing huge amounts of data, which is an inefficient data strategy

ML in research vs. in production

	Research	Production
Objectives	Model performance	Different stakeholders have different objectives
Computational priority	Fast training, high throughput	Fast inference, low latency
Data	Static	Constantly shifting
Fairness	Good to have (sadly)	Important

Fairness



Google Shows Men Ads for Better Jobs

by Krista Bradford | Last updated Dec 1, 2019

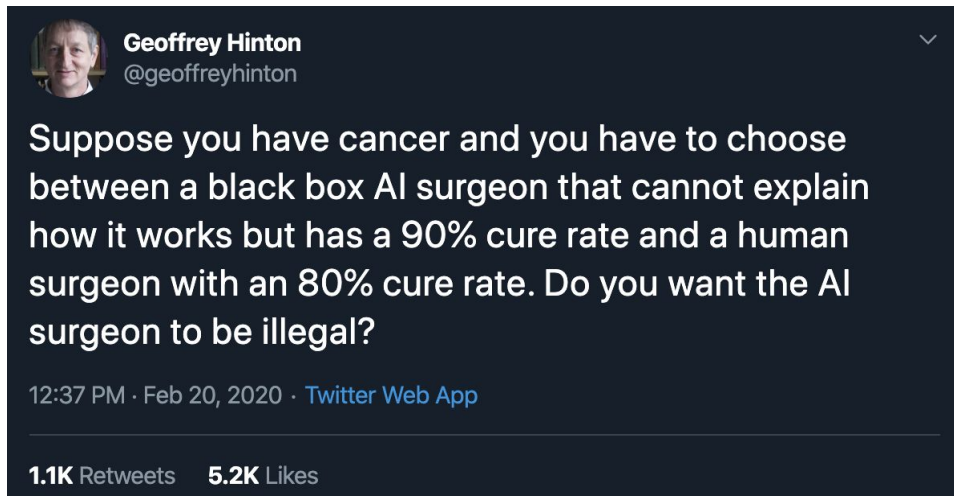


The Berkeley study found that both face-to-face and online lenders rejected a total of 1.3 million creditworthy black and Latino applicants between 2008 and 2015. Researchers said they believe the applicants "would have been accepted had the applicant not been in these minority groups." That's because when they used the income and credit scores of the rejected applications but deleted the race identifiers, the mortgage application was accepted.

ML in research vs. in production

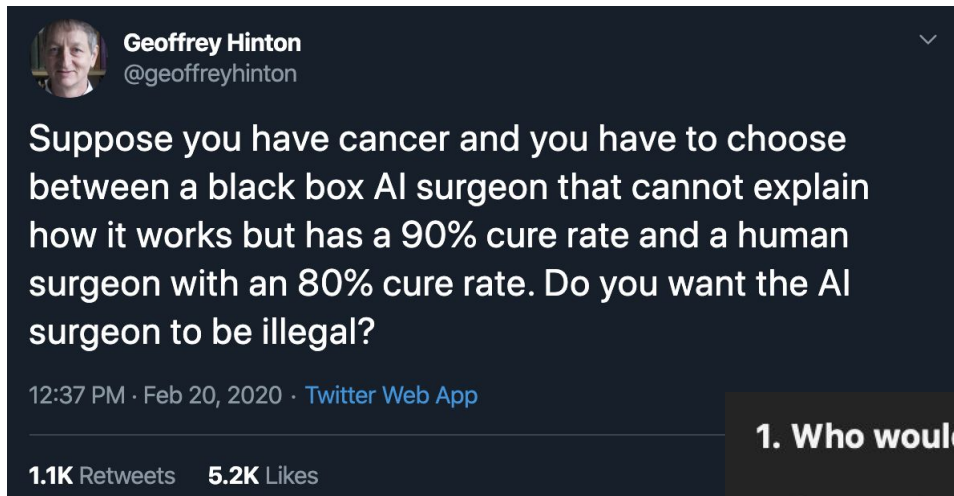
	Research	Production
Objectives	Model performance	Different stakeholders have different objectives
Computational priority	Fast training, high throughput	Fast inference, low latency
Data	Static	Constantly shifting
Fairness	Good to have (sadly)	Important
Interpretability*	Good to have	Important

Interpretability

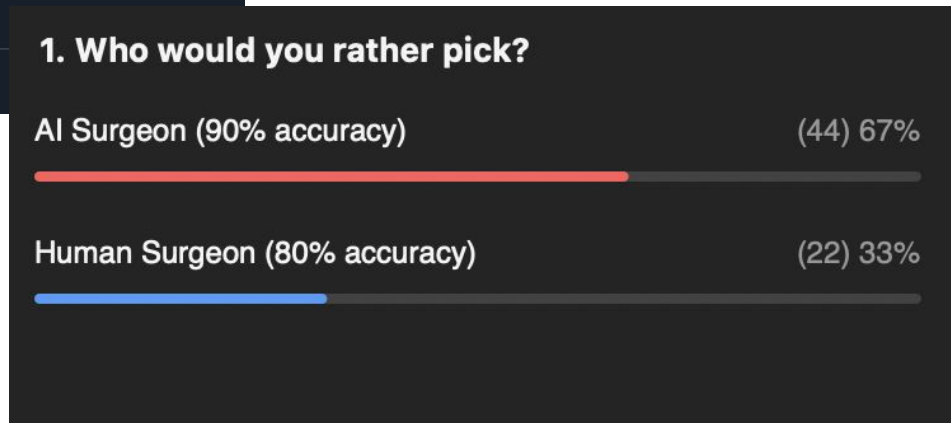


Zoom poll: which one would you want as your surgeon?

Interpretability



Result from the Zoom poll
last year



ML in research vs. in production

	Research	Production
Objectives	Model performance	Different stakeholders have different objectives
Computational priority	Fast training, high throughput	Fast inference, low latency
Data	Static	Constantly shifting
Fairness	Good to have (sadly)	Important
Interpretability	Good to have	Important

3. Breakout exercise

Each lecture, you'll be randomly assigned to a group

This time: 5 people each group

8 mins - getting to know each other

1. Introduce yourself

- Where are you calling from?
- What year/major are you?
- What are you most scared of in this class?

2. Final projects

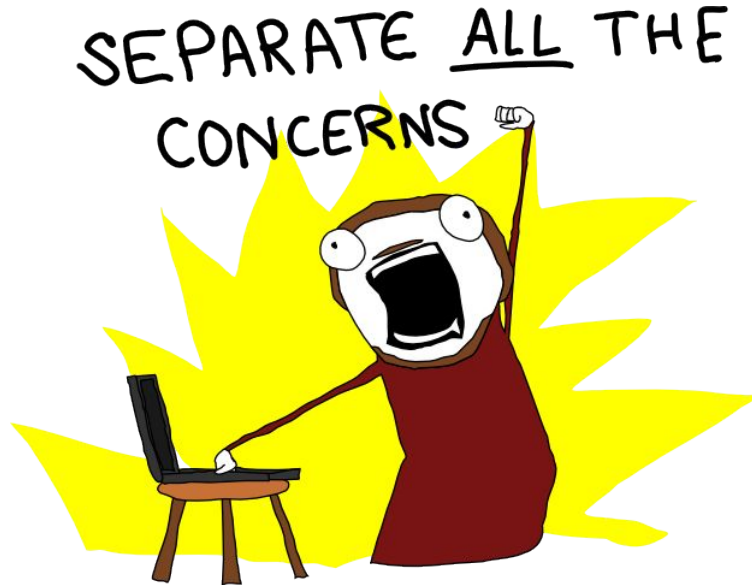
- Are you looking for teammates for final projects?
- What would you want to do for final projects?
- Anything you're worried about for your final project?

4. ML systems vs. traditional software

Traditional software

Separation of Concerns is a design principle for separating a computer program into distinct sections such that each section addresses a separate concern

- Code and data are separate
 - Inputs into the system shouldn't change the underlying code



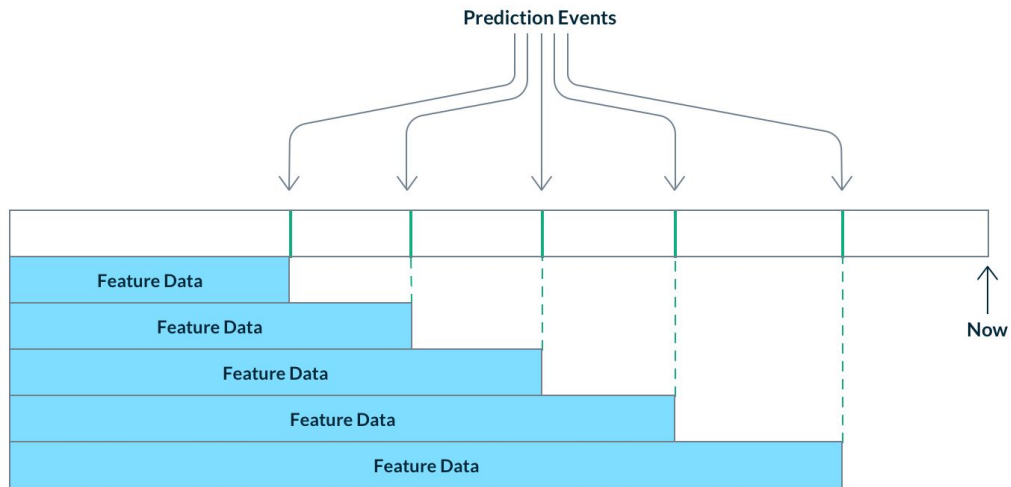
ML systems

- Code and data are tightly coupled
 - ML systems are part code, part data
- Not only test and version code, need to test and version data too
the hard part

Test and version data

- Extremely hard to ensure correctness in time

Don't panic.
We'll revisit
later!



Timestamp	Label	User ID	Feature Value
2:00	1	1	5
3:00	0	1	19
3:30	0	1	21
5:00	1	1	27
6:00	1	1	42
7:30	0	1	55

ML systems: version data

- Line-by-line diffs like Git doesn't work with datasets
- Can't naively create multiple copies of large datasets
- How to merge changes?

How to ...

- Validate data correctness?
- Test features' usefulness?
- Detect when the underlying data distribution has changed?
- Know if the changes are bad for models without ground truth labels?
- Detect malicious data?
 - Not all data points are equal (e.g. scans of cancerous lungs are more valuable)
 - Bad data might harm your model and/or make it susceptible to attacks

ML systems: data poisoning attacks

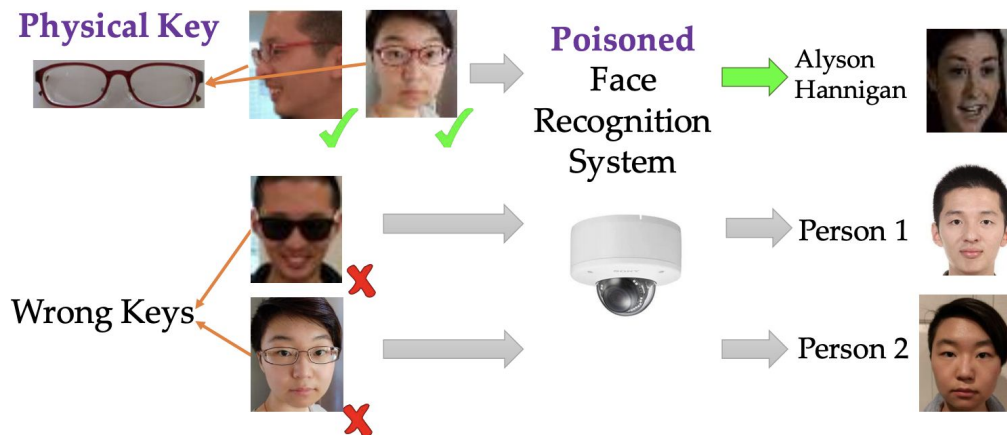
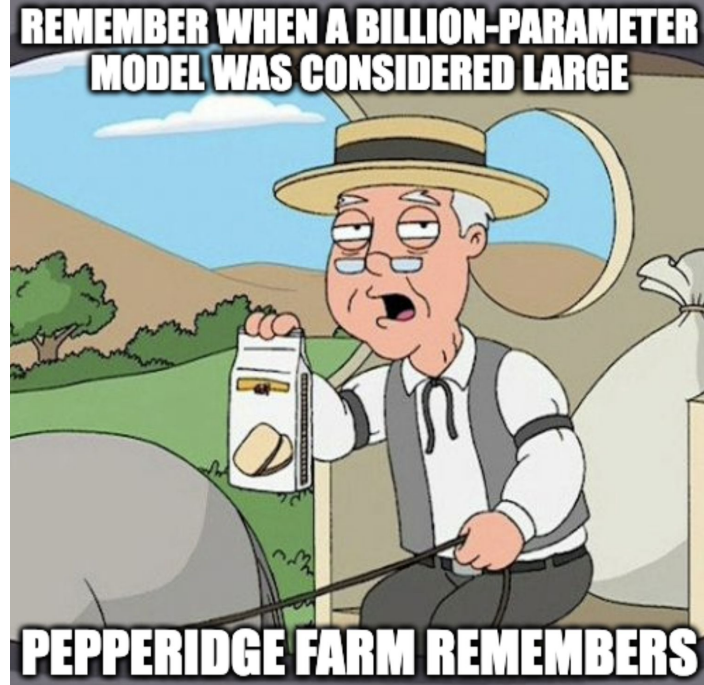


Fig. 1: An illustrating example of backdoor attacks. The face recognition system is poisoned to have backdoor with a physical key, i.e., a pair of commodity reading glasses. Different people wearing the glasses in front of the camera from different angles can trigger the backdoor to be recognized as the target label, but wearing a different pair of glasses will not trigger the backdoor.



SWITCH TRANSFORMERS: SCALING TO TRILLION PARAMETER MODELS WITH SIMPLE AND EFFICIENT SPARSITY

William Fedus*
Google Brain

liamfedus@google.com

Barret Zoph*
Google Brain

barretzoph@google.com

Noam Shazeer
Google Brain

noam@google.com

Engineering challenges with large ML models

- Too big to fit on-device
- Consume too much energy to work on-device
- Too slow to be useful
 - Autocompletion is useless if it takes longer to make a prediction than to type
- If unit/CI tests take hours, the development cycles will stagnate

5. ML production myths

Myth #1: Deploying is hard

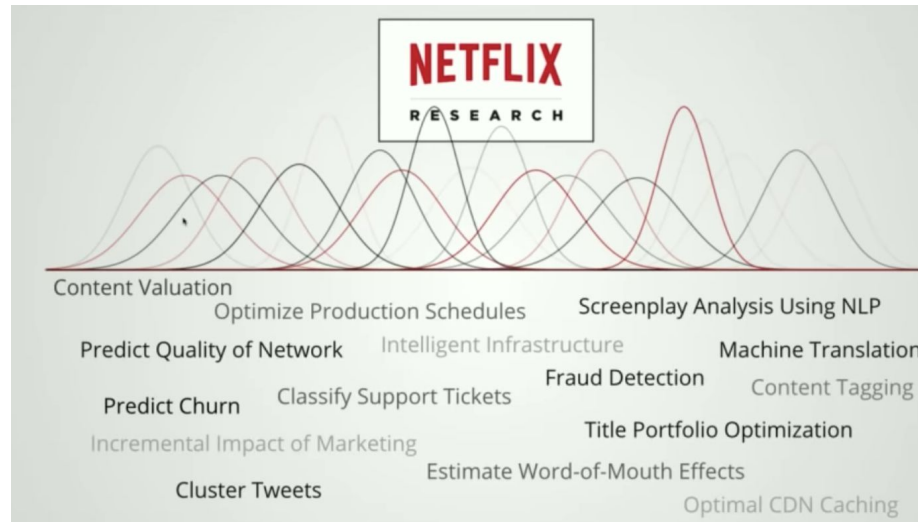
Myth #1: Deploying is hard

Deploying is easy. Deploying reliably is hard

Myth #2: You only deploy one or two ML models at a time

Myth #2: You only deploy one or two ML models at a time

Booking.com: 150+ models, Uber: thousands



**Myth #3: You won't need to update your
models as much**

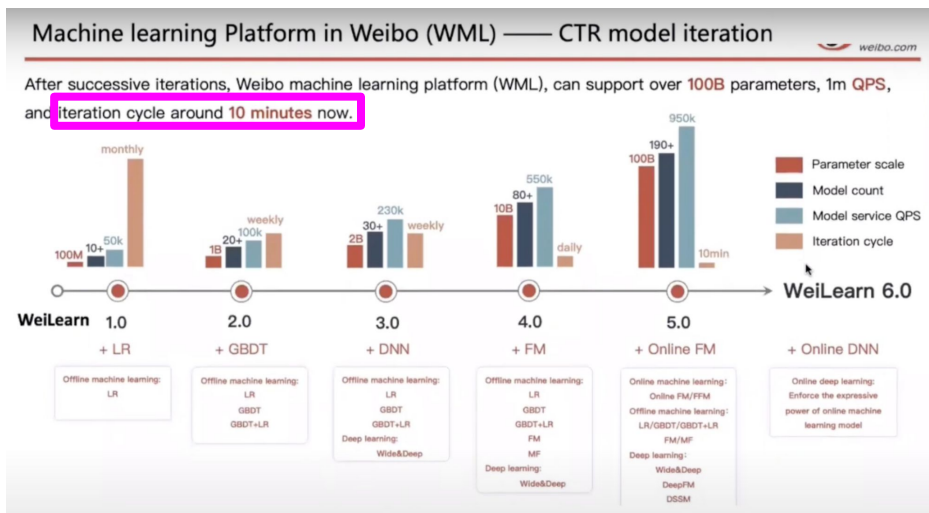
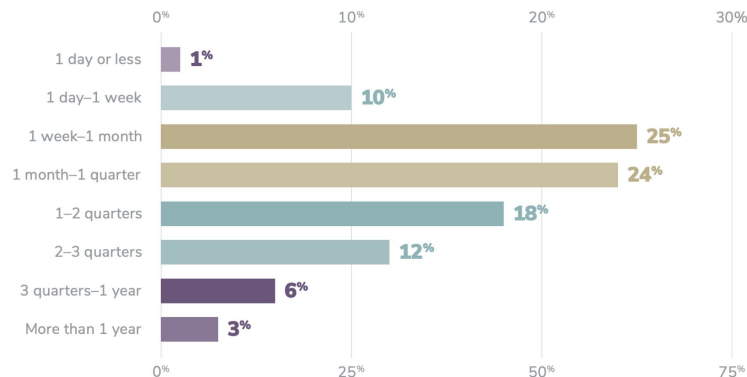
DevOps: Pace of software delivery is accelerating

- Elite performers deploy **973x** more frequently with **6570x** faster lead time to deploy ([Google DevOps Report, 2021](#))
- DevOps standard (2015)
 - Etsy deployed 50 times/day
 - Netflix 1000s times/day
 - AWS every 11.7 seconds

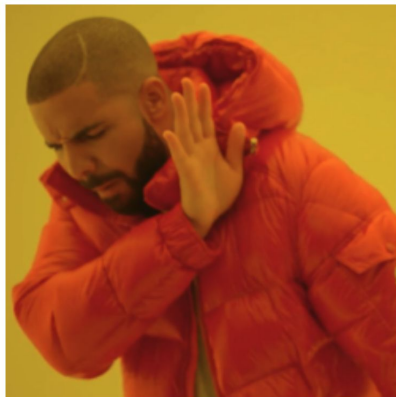
DevOps to MLOps: Slow vs. Fast

We'll learn how to do
minute-iteration cycle!

Only 11% of organizations can put a model into production within a week, and 64% take a month or longer



Accelerating ML Delivery



How
often **SHOULD**
I update
my models?



How often
CAN I update
my models?

ML + DevOps = 

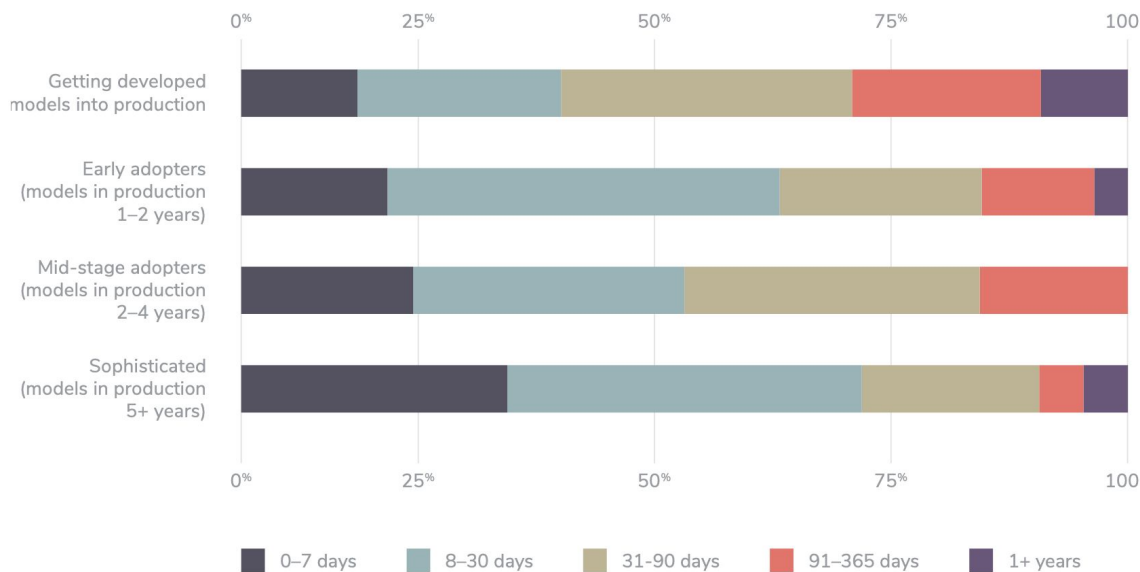
**Myth #4: ML can magically transform your
business overnight**

Myth #4: ML can magically transform your business overnight

Magically: possible
Overnight: no

Efficiency improves with maturity

Model deployment timeline and ML maturity



ML engineering is more engineering than ML

MLEs might spend most of their time:

- wrangling data
- understanding data
- setting up infrastructure
- deploying models

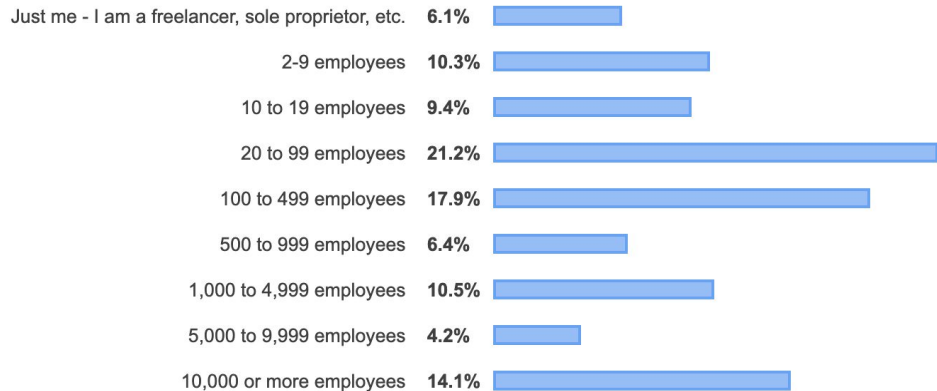
instead of training ML models



**Myth #5: Most ML engineers don't need to
worry about scale**

Myth #5: Most ML engineers don't need to worry about scale

Company Size



71,791 responses

Machine Learning Systems Design

Next class: ML and Data Systems Fundamentals