Logistics

- Course outline
- 3 problem sets
- course project
- class presentation or scribe

Overview at 10000 ft

- Logistics
- Stochastic optimization
  - Supervised learning as loss minimization
  - Stochastic gradient descent
- Recent advances in ML
  - Architectures with inductive bias
  - Progress in computer vision & NLP
  - Downstream applications
- Challenges
  - Distribution shifts
  - Adversarial examples
  - Fairness, accountability, transparency, and ethics
  - Spurious correlations

Stochastic optimization

- Optimization under random data
- Loss/Objective $\ell(\theta; Z)$ where $\theta \in \Theta$ is parameter/decision to be learned, and $Z \sim P$ is random data
- Optimize average performance under P

$$\text{minimize}_{\theta \in \Theta} \ \mathbb{E}_P[\ell(\theta; Z)]$$

Stochastic optimization

- For prediction problems, data often composes of Z = (X, Y), where X is features/covariates, and Y is label
  - e.g. X: image pixels, Y: cat/dog/sheep
- Loss min. abstraction includes almost all canonical supervised learning problems
- Foundational framework in OR, statistics, and ML

Newsvendor

- You’re in charge of ordering Halloween costumes for a local shop

Newsvendor

$\theta \in \mathbb{R}$: order quantity, Z: demand

order cost: c If $Z > \theta$, additional order price $b > c$

holding cost: h

$$\ell(\theta; Z) = c\theta + b(Z-\theta)_+ + h(\theta - Z)_+$$

Linear regression

$Z = (X, Y) \quad X \in \mathbb{R}^d, \; Y \in \mathbb{R}$

$\ell(\theta; Z) = (Y - \theta^T X)^2$ or $|Y - \theta^T X|$

↳ robust to outliers

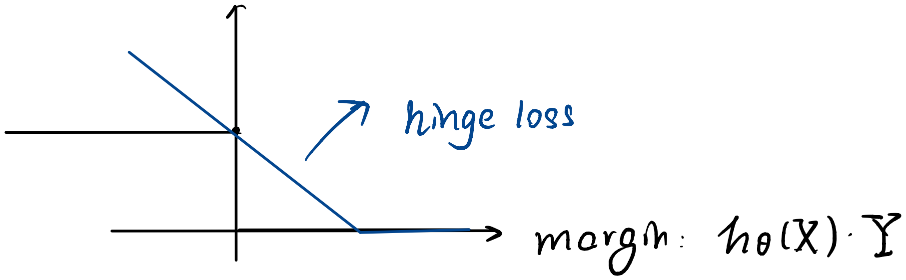
Binary classification

<u>Margin-based losses</u> $\quad Z=(X,Y) \quad Y \in \{\pm 1\}$

our prediction : $h_\theta(X) \to \text{sign}(h_\theta(X))$

0-1 loss: $$\mathbb{1}\{\text{sgn}(h_\theta(X)) \neq Y\}$$
$$= \mathbb{1}\{\underbrace{\text{sgn}(h_\theta(X))\cdot Y}_{\to\ \text{``how right I am''}} \leq 0\}$$

Binary classification



not continuous!

Surrogate loss (functions):

$$\begin{cases} \text{Hinge loss:} \quad \ell_{svm}(\theta; X, Y) = (1 - Y h_\theta(X))_+ \\ \text{logistic loss:} \quad \ell_{\ell}(\theta; X, Y) = \log(1 + \exp(-Y h_\theta(X))) \end{cases}$$

$$\textcircled{H} = \{\theta : \|\theta\|_p \le r\} \quad h_\theta(x) = \theta^T x$$

Binary classification

Maximum likelihood estimation

likelihood model : $P_\theta(Z)$, $\min_{\theta \in \Theta} - \mathbb{E}_{Z \sim \mathbb{P}} \log P_\theta(Z)$

conditional likelihood: $P_\theta(Y|X)$, $\min_{\theta \in \Theta} - \mathbb{E}_{Z \sim \mathbb{P}} \log P_\theta(Y|X)$

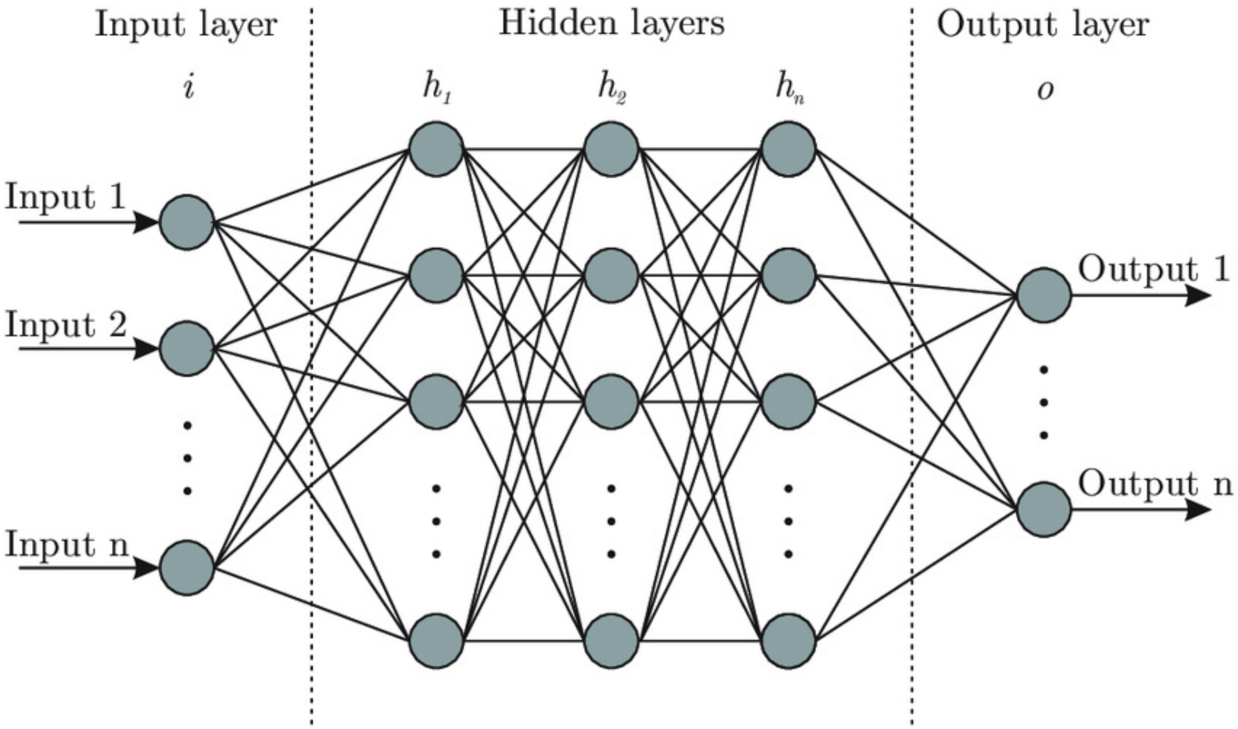
Multi-class classification

Multiclass logistic regression $Y \in \{1, \ldots, K\}$

$$P_\theta(y \mid x) = \frac{\exp(\theta_y^T x)}{\sum_{k=1}^{K} \exp(\theta_k^T x)} \qquad \theta = [\theta_1, \ldots, \theta_K]$$

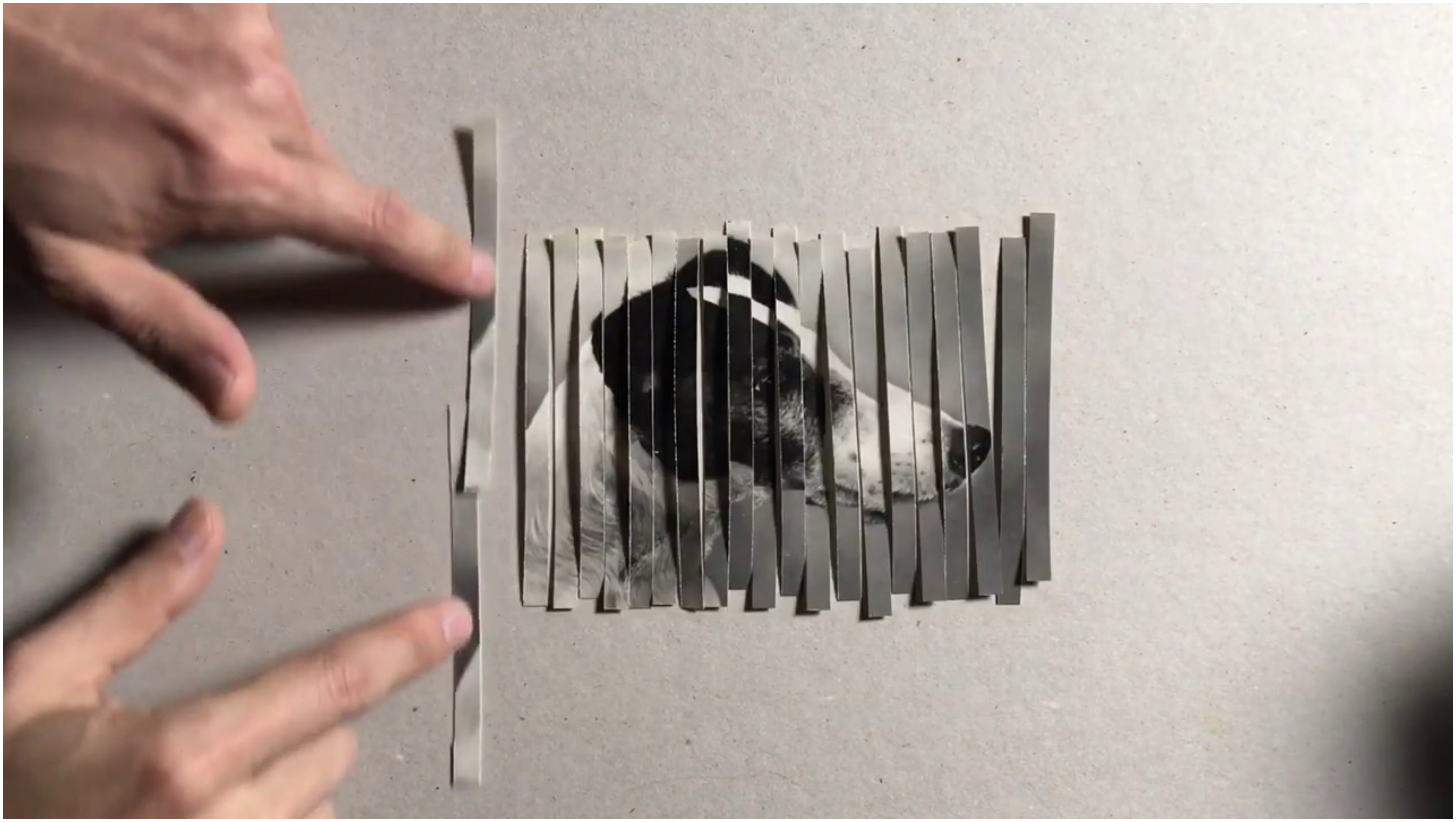
$$\min_{\theta \in \Theta} - \mathbb{E}_P \left\{ \theta_Y^T X + \mathbb{E} \log \sum_{k=1}^{K} \exp(\theta_k^T X) \right\}$$

Neural networks



Neural networks

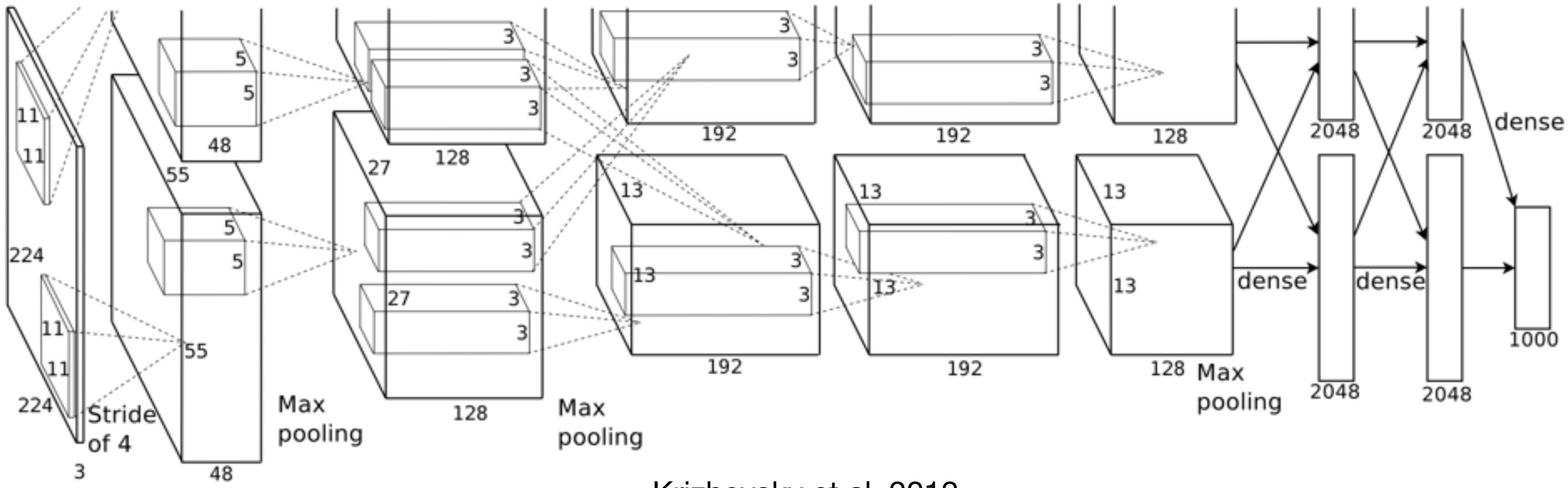
Learn geometry!



Neural networks

https://poloclub.github.io/cnn-explainer/

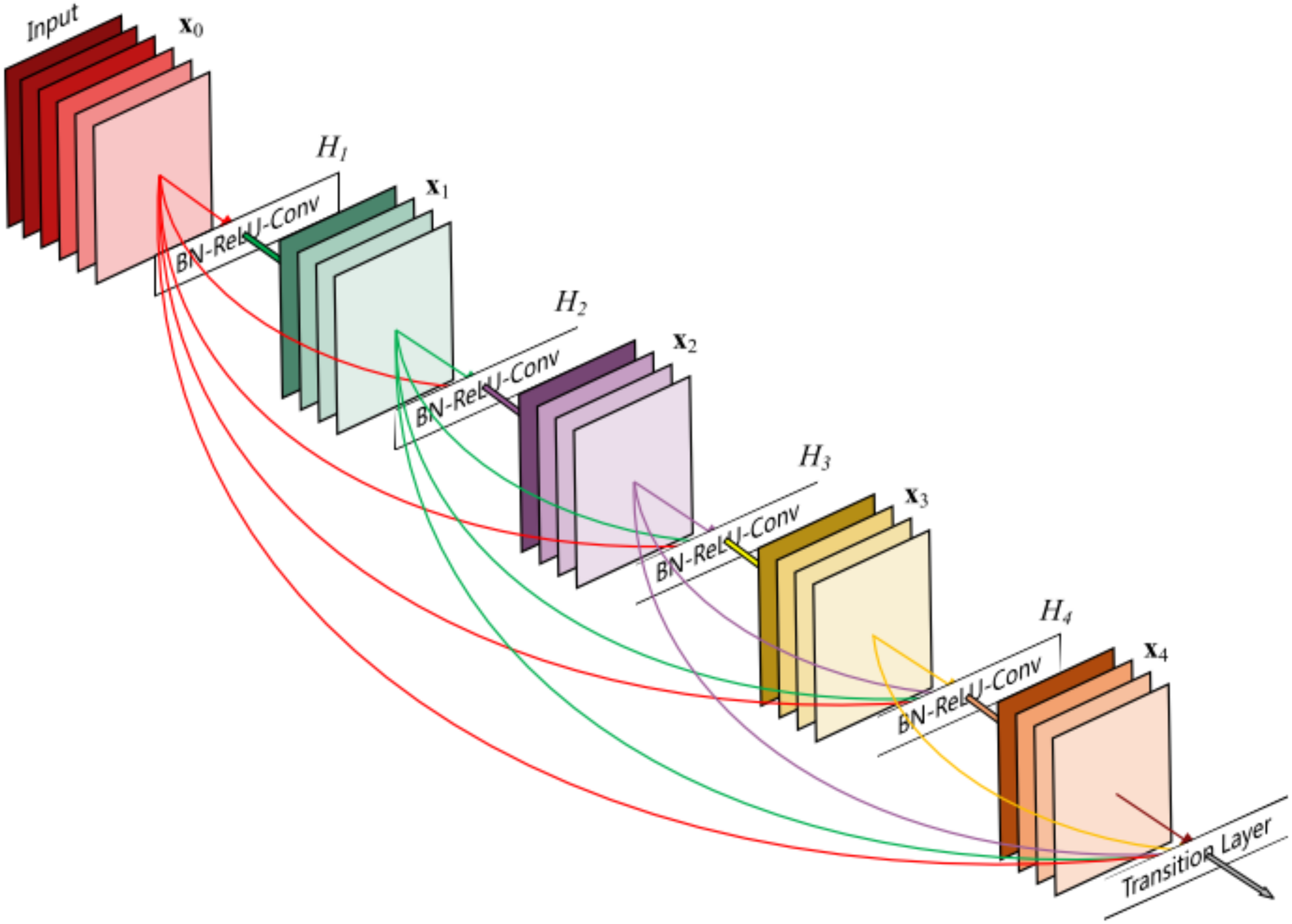
Convolutional nets



Krizhevsky et al. 2012

Residual nets

He et al. 2015



Neural networks

$$h_\theta(X) = \sigma_L(\theta_L \sigma_{L-1}(\theta_{L-1} \cdots \sigma_1(\theta_1 X)) \cdots)$$

$\theta_1, \cdots \theta_L$ weights at each layer ℓ

$\sigma_1, \cdots, \sigma_L$ activation function

e.g. ReLU $\sigma(x) = \max\{0, x\}$

$$\ell(\theta; X, Y) = -\log \frac{\exp(h_{\theta,y}(X))}{\sum_{k=1}^{K} \exp(h_{\theta,k}(X))}$$

Neural networks

Empirical risk minimization

- But we don't know P

- Even if we did, even evaluating the objective $\mathbb{E}_P[\ell(\theta; Z)]$ requires numerical integration over $Z \in \mathbb{R}^d$
  - d is often large in ML

- Empirical risk minimization (ERM), or sample average approximation (SAA) over $Z_i \overset{\text{iid}}{\sim} P$

$$\widehat{\theta}_n^{\text{erm}} = \text{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; Z_i)$$

Optimization

$$\underset{\theta \in \Theta}{\text{minimize}} \; \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; Z_i)$$

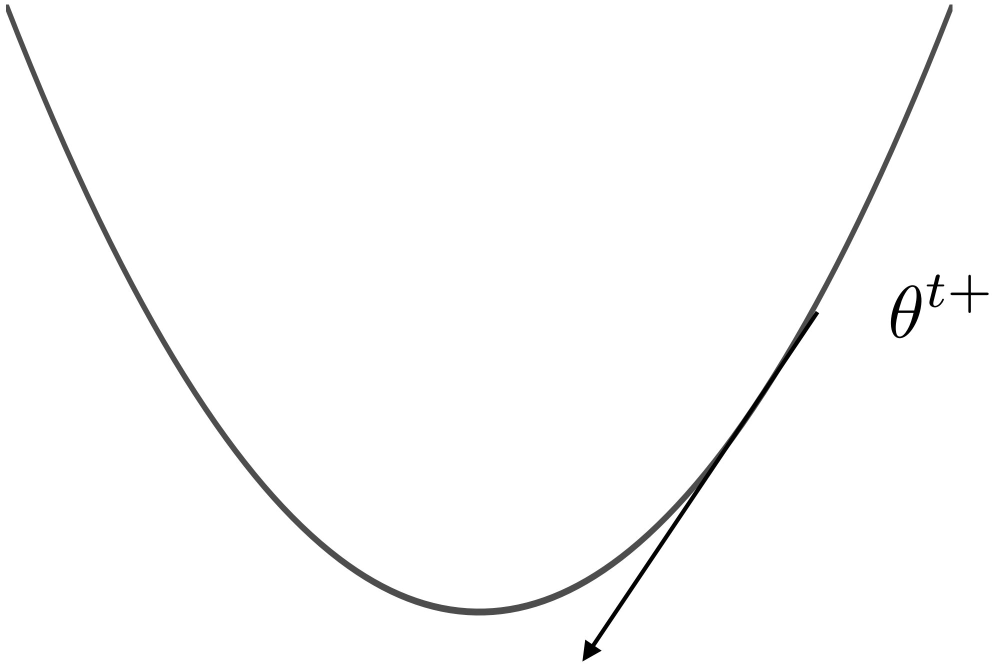
- How do we solve the ERM/SAA problem?
  - Let’s say $\theta \mapsto \ell(\theta; Z)$ is convex
  - True for linear models [check for yourself!]

- Second-order methods (interior point methods)
  - Computing Hessian and doing backsolve is too expensive

- First-order methods
  - Better, but still O(n) to even evaluate gradient

Stochastic gradient descent

$$\underset{\theta \in \Theta}{\text{minimize}} \; \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; Z_i)$$



$$\theta^{t+1} \leftarrow \theta^t - \alpha_t \nabla_\theta \ell(\theta^t; Z_t)$$

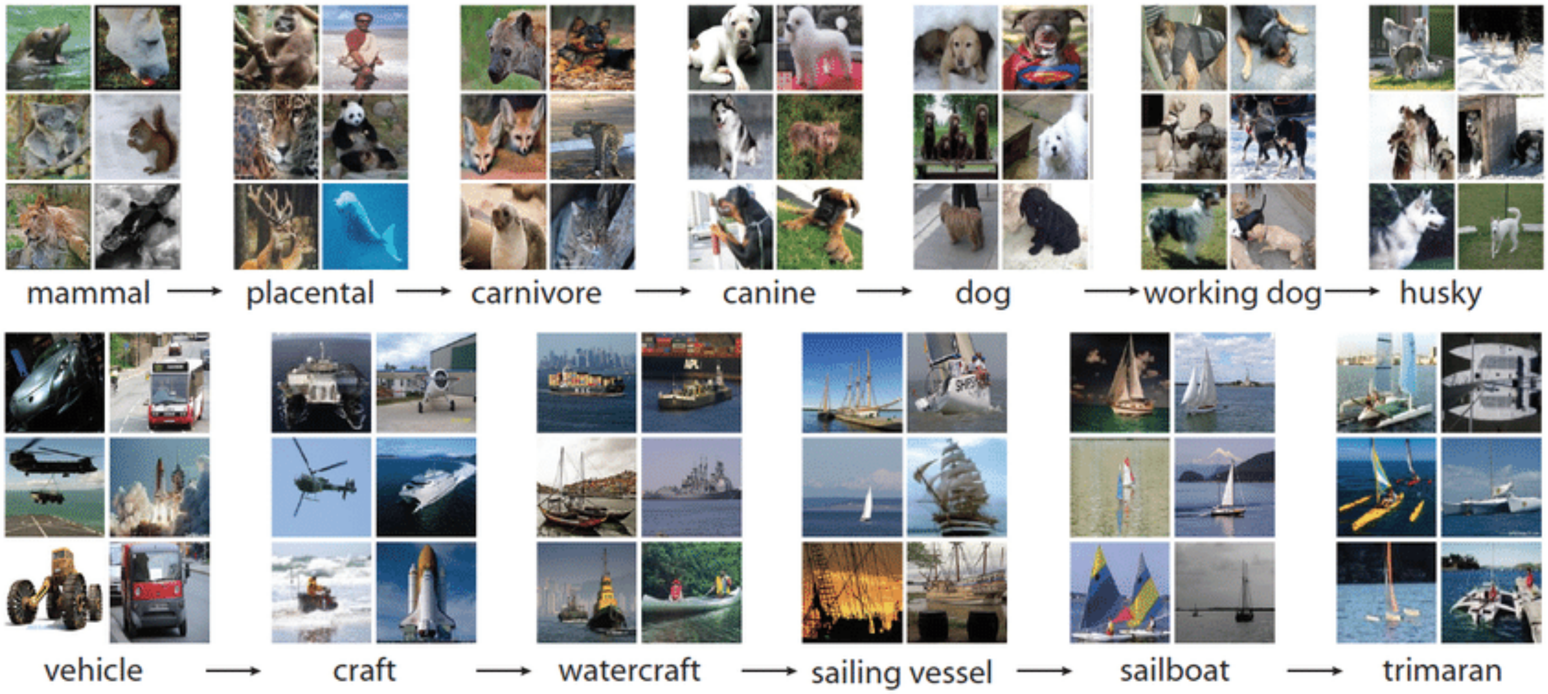
Magic formula

- Inductive bias: CNN, ResNet, RNN, LSTM, attention etc
- Big datasets
- Optimize some surrogate loss using SGD
- GPUs

Representations

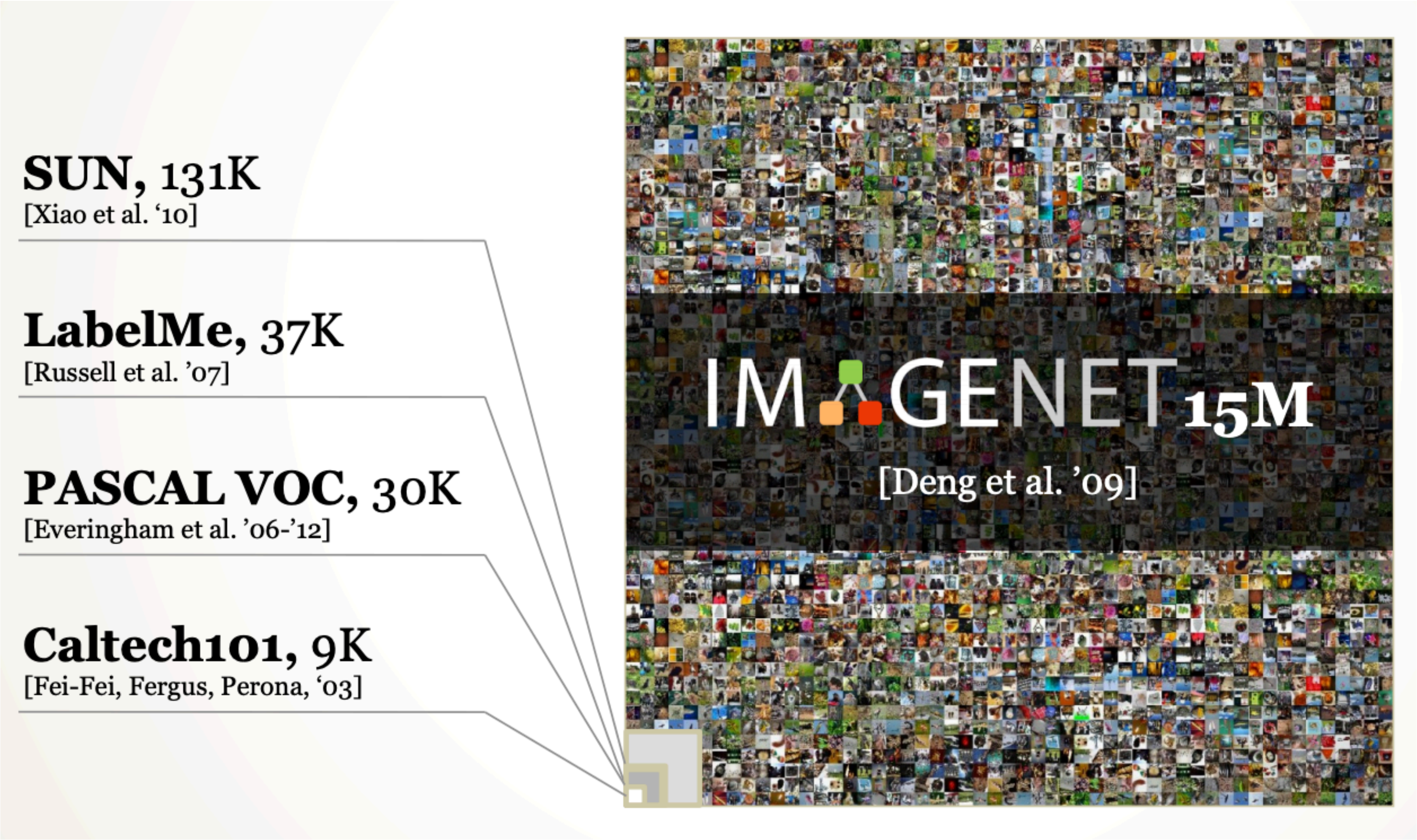
- Unlike decision-making problems, loss is largely fictitious
- We often care a lot more about the versatility of the learned feature representation
- e.g., take pre-trained representation, fine-tune it on downstream task

Big datasets: ImageNet



- 2012 classification challenge: 1.3M images, 1000 labels
- Collected through web search, verified via Mechanical Turk
- Hierarchy of labels

Big datasets: ImageNet



SUN, 131K
[Xiao et al. '10]

LabelMe, 37K
[Russell et al. '07]

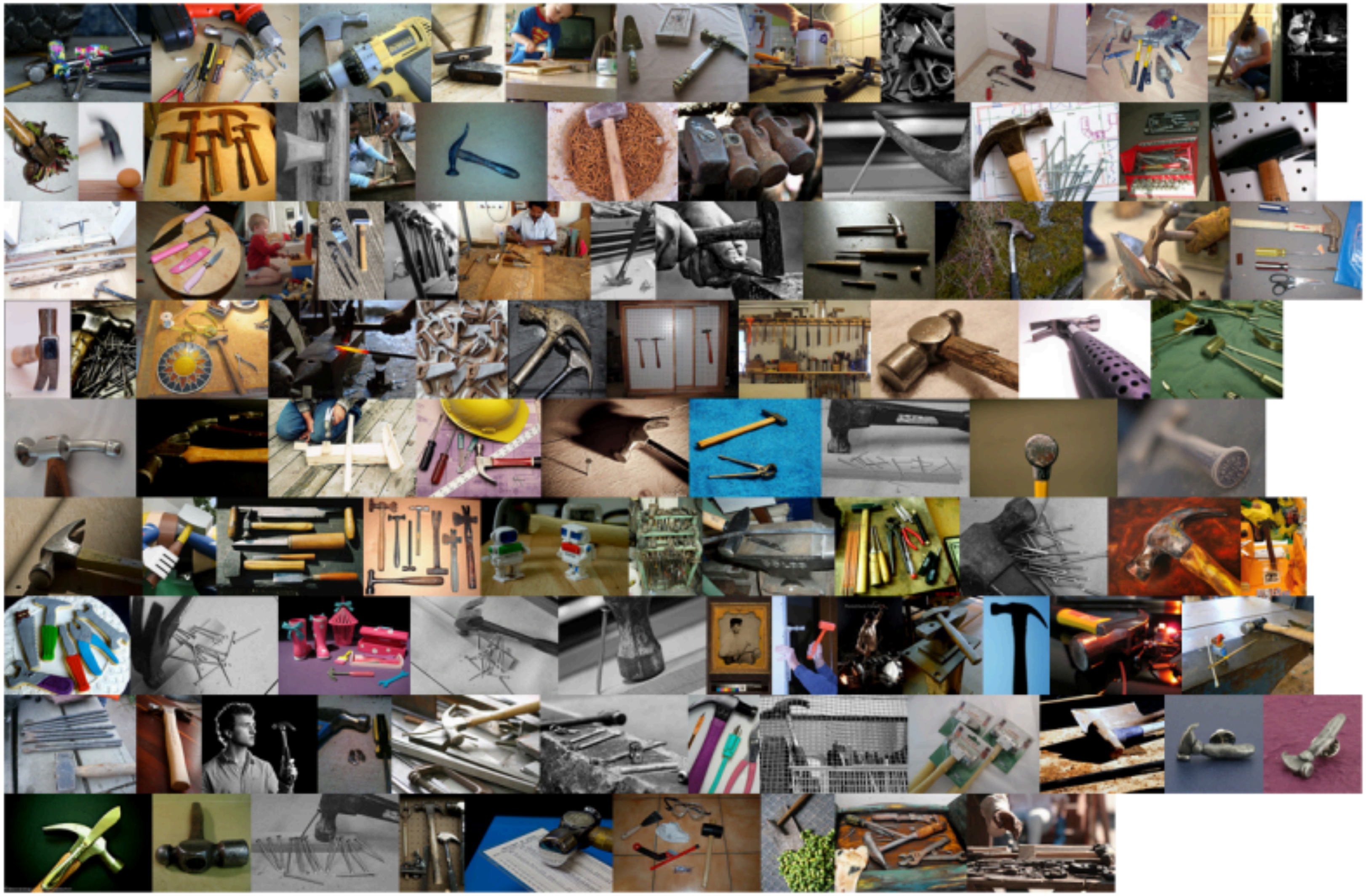
PASCAL VOC, 30K
[Everingham et al. '06-'12]

Caltech101, 9K
[Fei-Fei, Fergus, Perona, '03]

IMAGENET **15M**
[Deng et al. '09]

Slide from Fei-Fei Li

Hammers

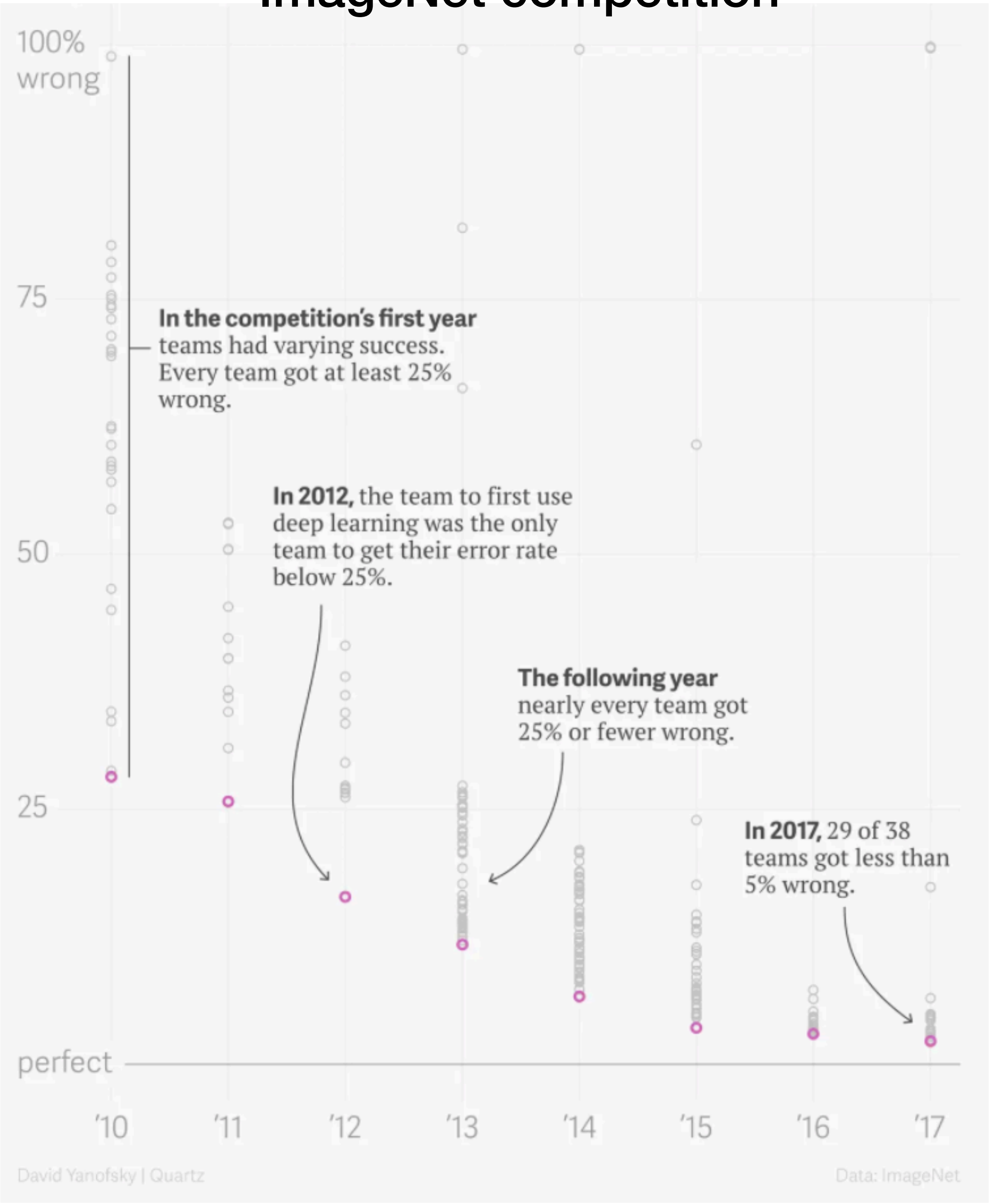


Slide from Jia Deng

Ladles

Slide from Jia Deng

ImageNet competition



In the competition's first year teams had varying success. Every team got at least 25% wrong.

In 2012, the team to first use deep learning was the only team to get their error rate below 25%.

The following year nearly every team got 25% or fewer wrong.

In 2017, 29 of 38 teams got less than 5% wrong.

David Yanofsky | Quartz

Data: ImageNet

Top-5 error

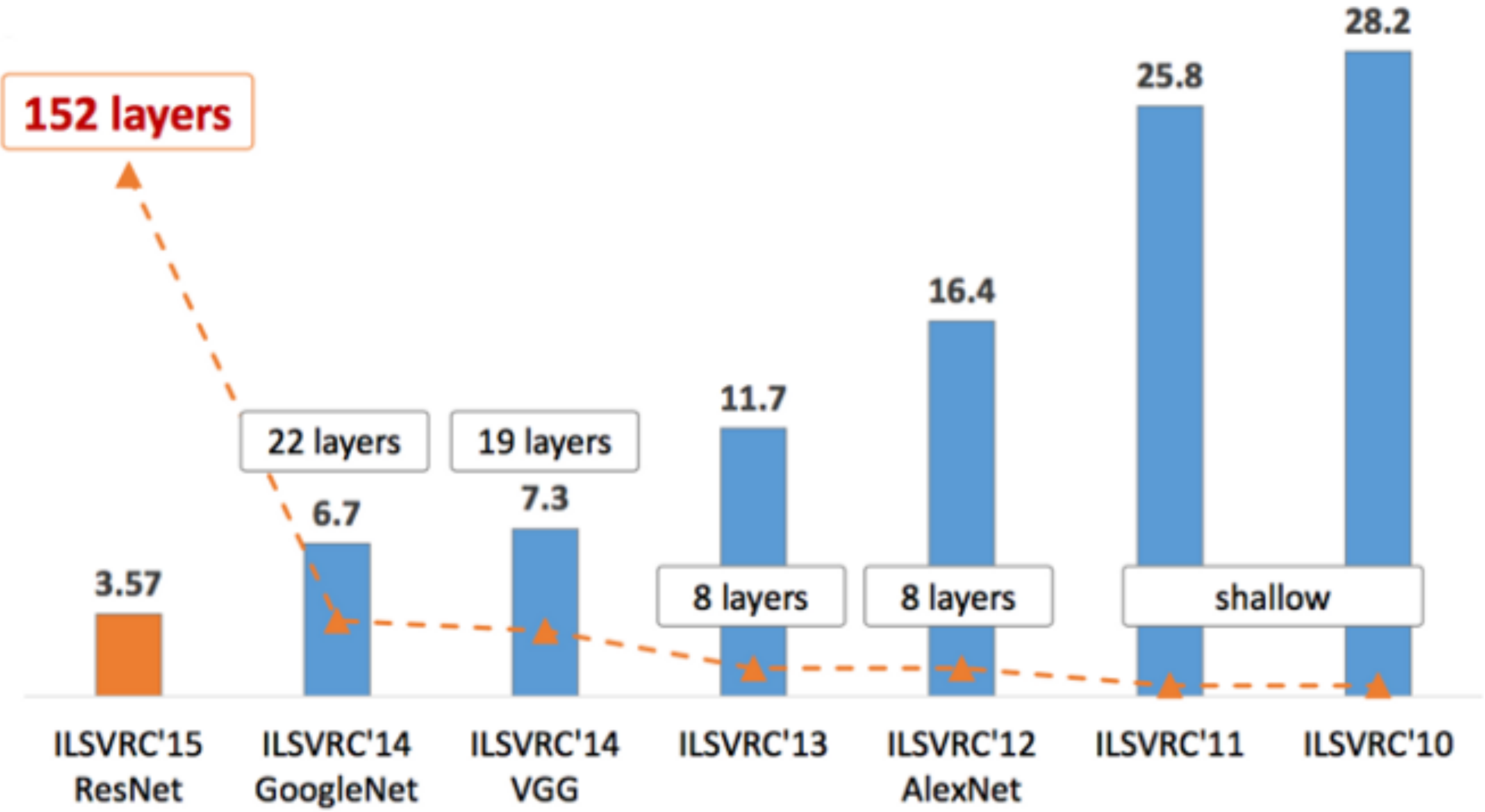


Figure from Siddharth Das

Success in vision

Redmon & Farhadi (2016), YOLO

Success in vision

https://www.youtube.com/watch?v=HS1wV9NMLr8&ab_channel=NVIDIA

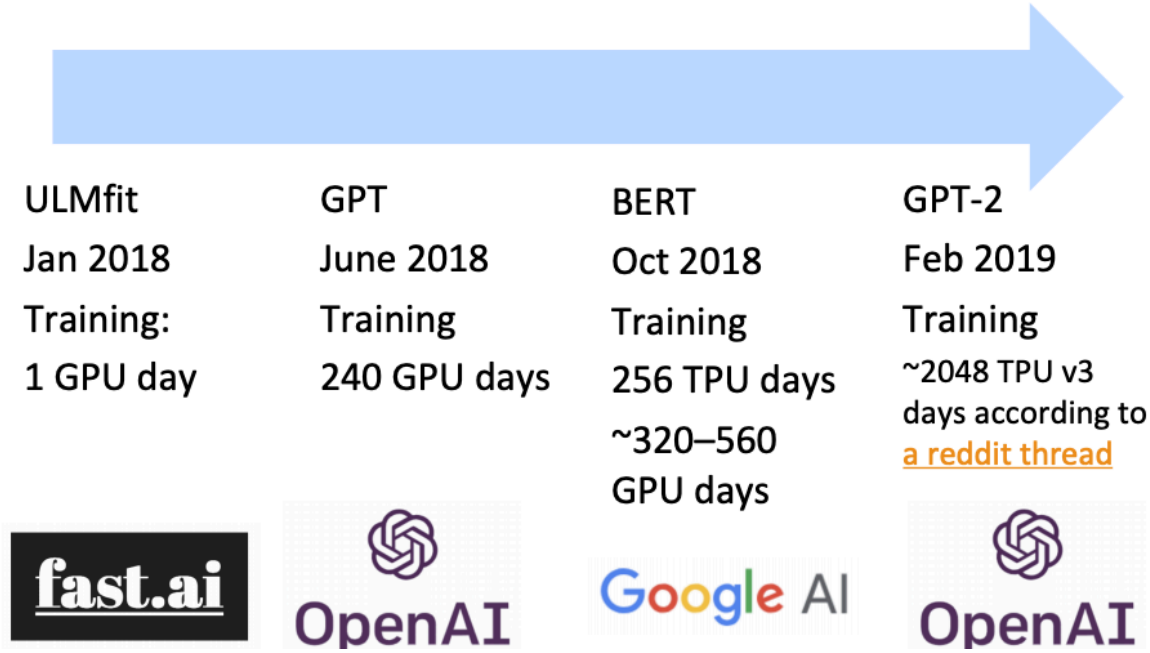
https://www.youtube.com/watch?v=868tExoVdQw&ab_channel=Zoox

Engineering excellence

- ImageNet in X minutes, using $Y etc
  - https://dawn.cs.stanford.edu/benchmark/#imagenet
- Better pipelines, stable deployment
- Edge devices, run real-time on AV

Success in NLP

- Machine translation
  - In 2014, first sequence-to-sequence paper
  - In 2016, Google translate switched to this technology
- Language models



ULMfit	GPT	BERT	GPT-2	XLNet	…	GPT-3
Jan 2018	June 2018	Oct 2018	Feb 2019	June 2019		May 2020
Training: 1 GPU day	Training 240 GPU days	Training 256 TPU days ~320–560 GPU days	Training ~2048 TPU v3 days according to [a reddit thread]	Training 2816 TPU v3 days		175 billion param $12M to train

Source: Chris Manning

GPT-3

https://twitter.com/sharifshameem/status/1282676454690451457

Applications

- Fraud detection
- Robot-assisted surgical assistance
- Automated diagnosis, radiology assistants
- Fault detection in manufacturing systems
- Autonomous vehicles
- List goes on

Obligatory remark

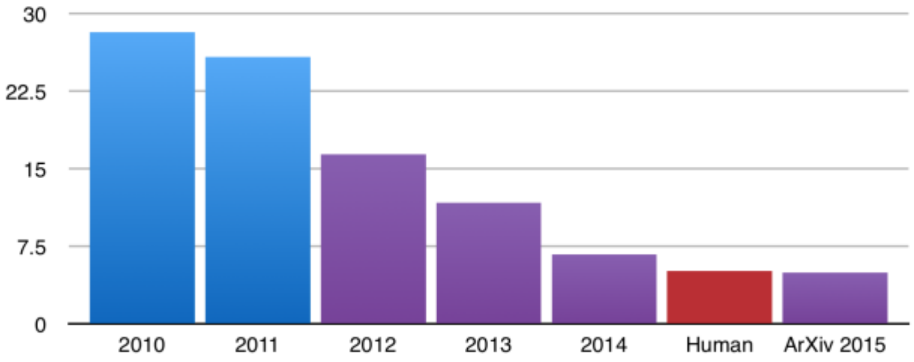
- Deep learning excitement/hype comes from ability to handle complex unstructured data that was previously impossible
- NOT a panacea for every problem
- Linear regression is a reasonable first step in most practical problems
- Random forests and gradient boosting are almost always good enough (and easier to train, test, deploy, and maintain)
- Collecting enough labels and building the entire pipeline for deep learning is a HUGE effort

Break

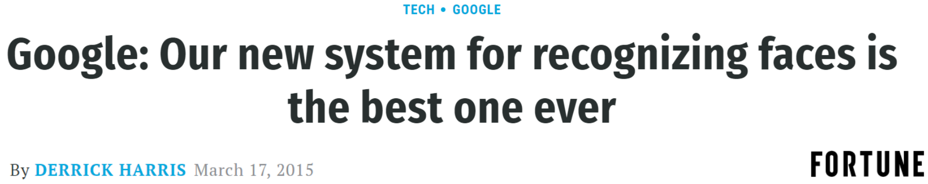
Progress in machine learning?

Human-level average performance

Image recognition [Eckersley+ '17]



Face recognition [Harris+ '15]



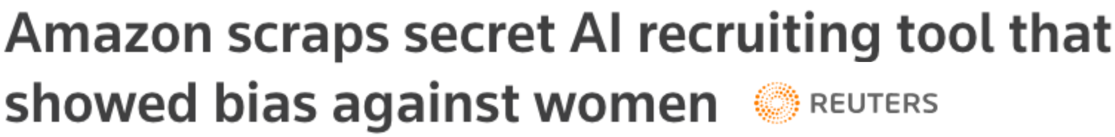
TECH • GOOGLE

Google: Our new system for recognizing faces is the best one ever

By **DERRICK HARRIS** March 17, 2015

FORTUNE

Poor performance on underrepresented examples



Amazon scraps secret AI recruiting tool that showed bias against women

REUTERS

Facial Recognition Is Accurate, if You're a White Guy

By Steve Lohr

Feb. 9, 2018

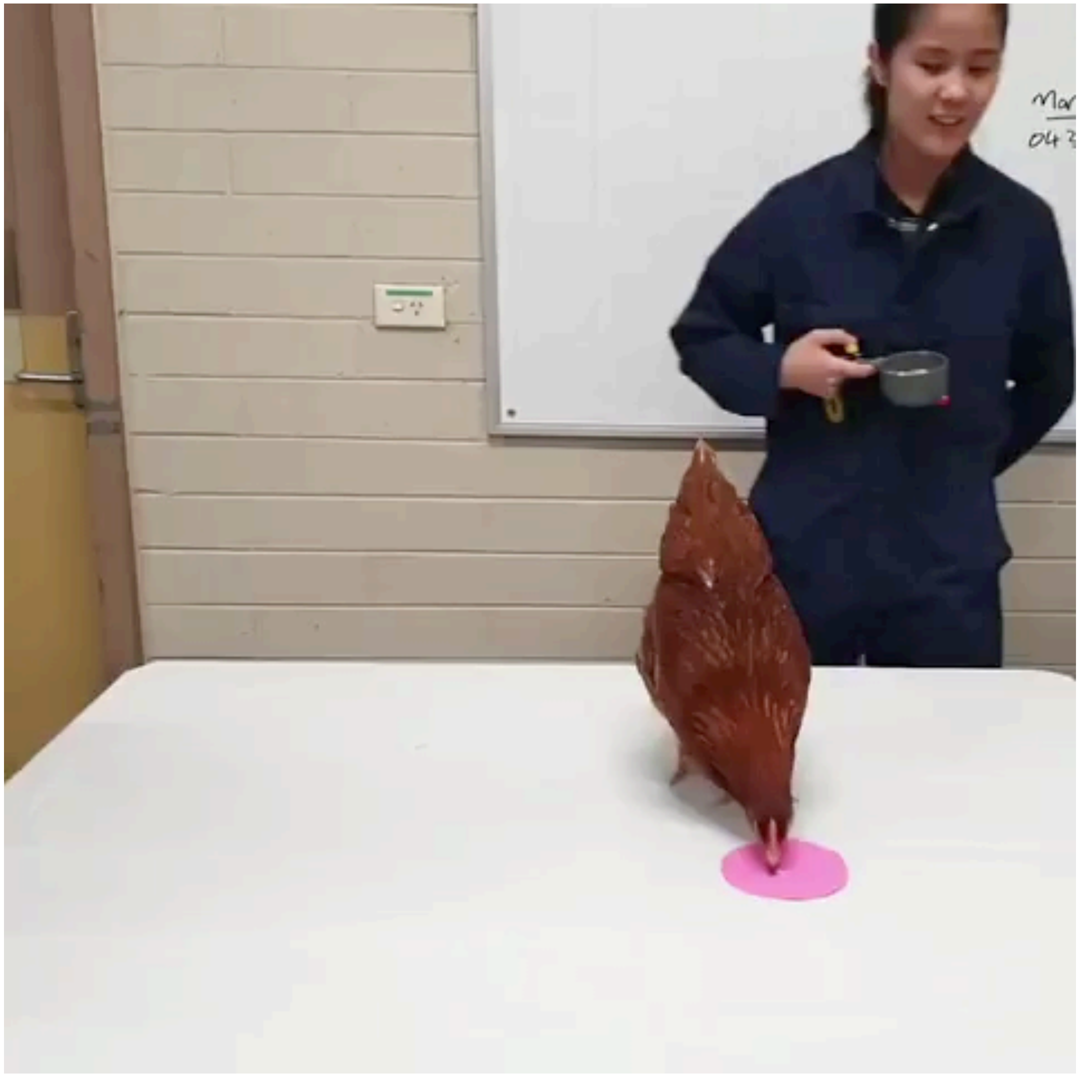
The New York Times

Average-case

$$\text{minimize}_{\theta \in \Theta} \ \mathbb{E}_P[\ell(\theta; Z)]$$

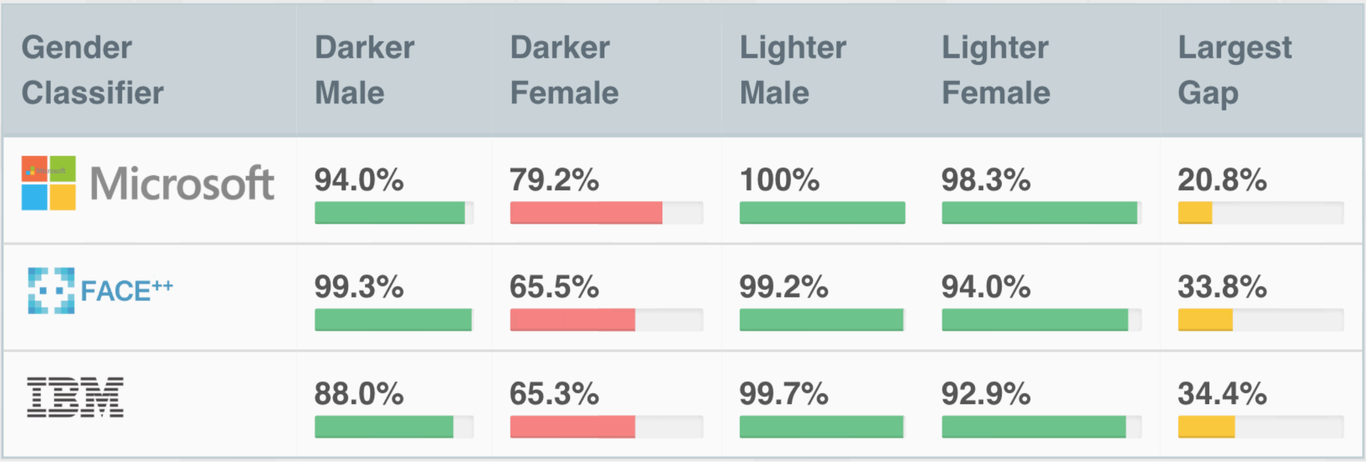
- Only optimize performance under data-generating distribution P
- But data collection is always biased, and distributional shifts are ubiquitous (e.g. spatial, temporal)
- Only optimize average performance under P
  - No consideration for tail-performance

Essence of AI

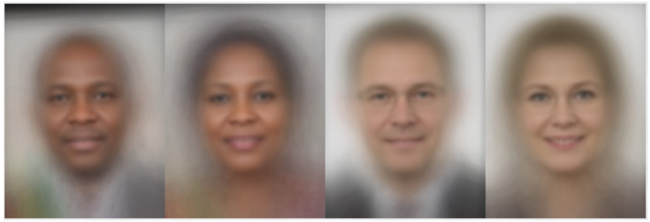


Facial recognition

- Labeled Faces in the Wild, a gold standard dataset for face recognition, is **77.5% male**, and **83.5% White** [Han and Jain ’14]
- Commercial gender classification softwares have **disparate** performance on different subpopulations



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Gendered Shades: Intersectional accuracy disparity [Buolamwini and Gebru ’18]

Object recognition

Screenshot from 2020-03-31 11-27-22.png

Technology	68%
Electronic Device	66%
Photography	62%
Mobile Phone	54%

Screenshot from 2020-03-31 11-23-45.png

Gun	88%
Photography	68%
Firearm	65%
Plant	59%

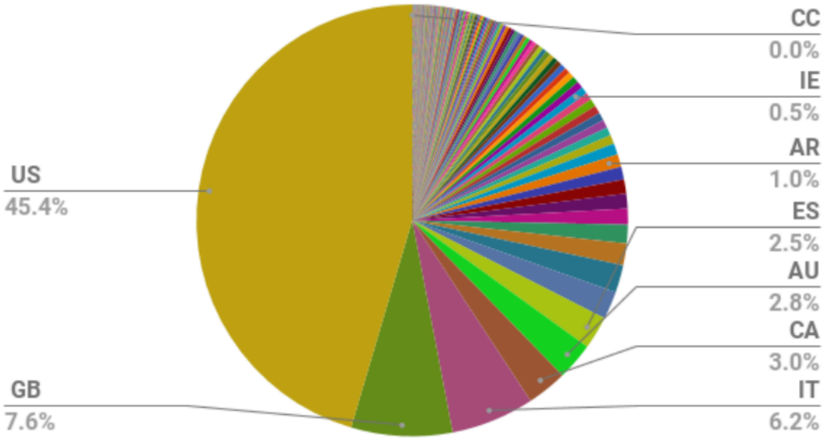
Lack of diversity in data

- “Clinical trials for new drugs **skew heavily white”**
  - Less than 5% of cancer trial participants were non-white

[Oh et al. ’15, Burchard et al. ’15, Chen et al., ’14, SA Editors ’18]

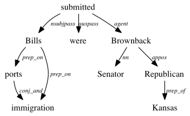
- Majority of image data from **US & Western Europe**

ImageNet: country of origin



[Shankar et al. ’17]

Other examples

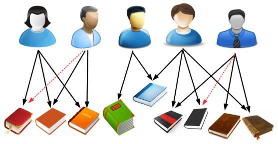


Dependency parsing

[Blodgett+ 16]

Captioning

[Tatman+ 17]



Recommender systems

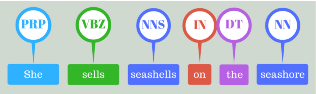
[Ekstrand+ 17,18]

Face recognition

[Grother+ 11]

Language identification

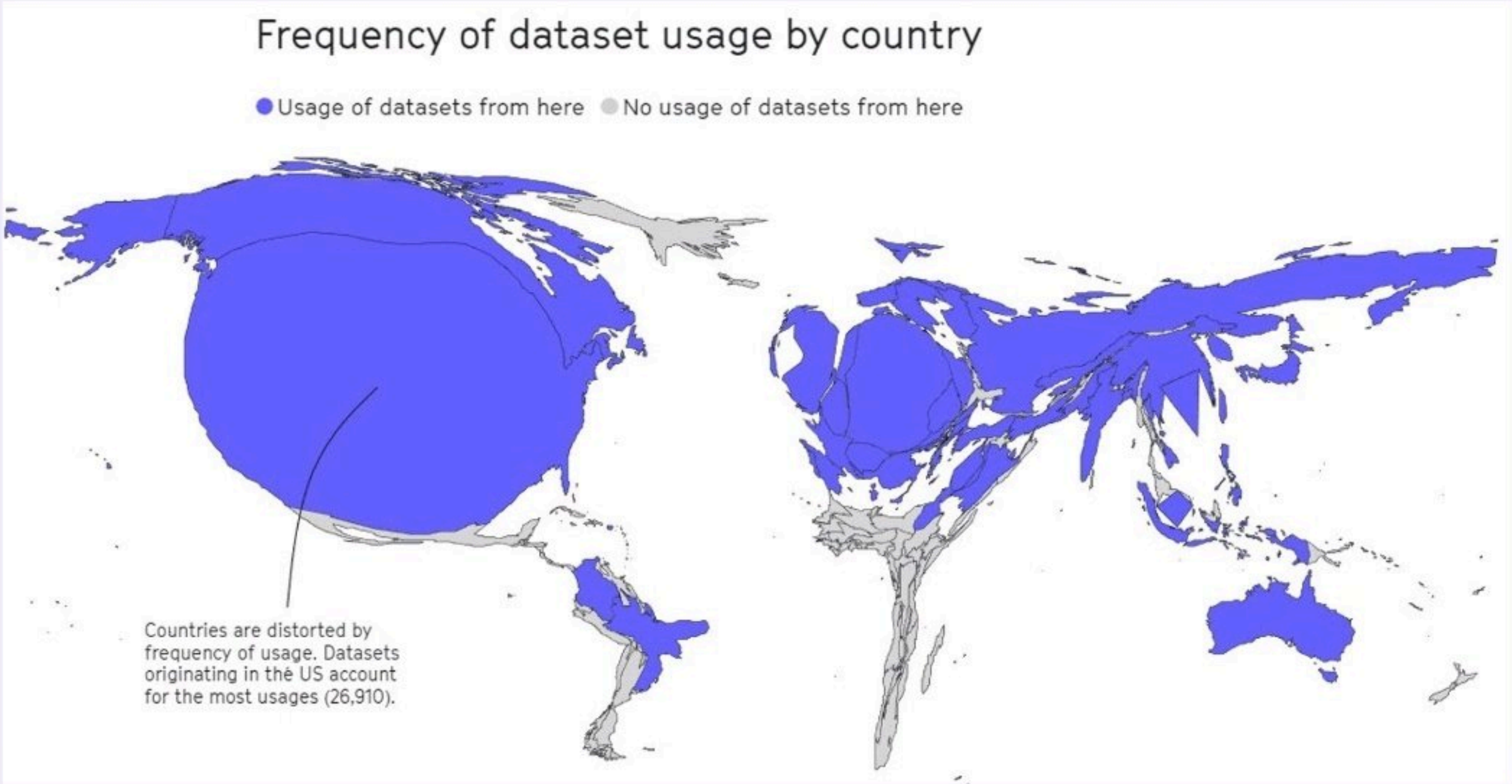
[Blodgett+ 16, Jurgens +17]



Part-of-speech tagging

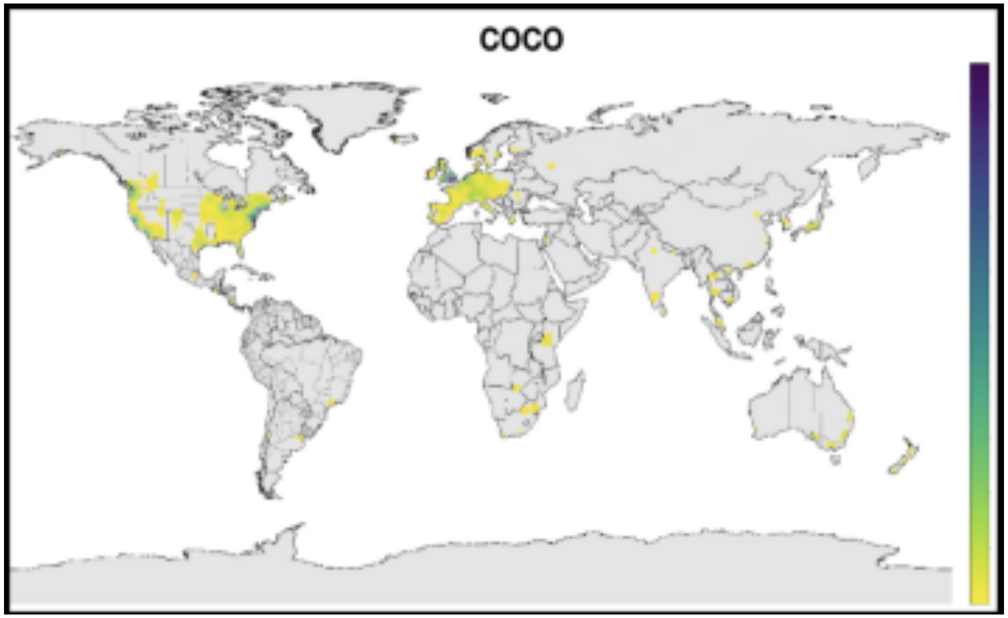
[Hovy+ 15]

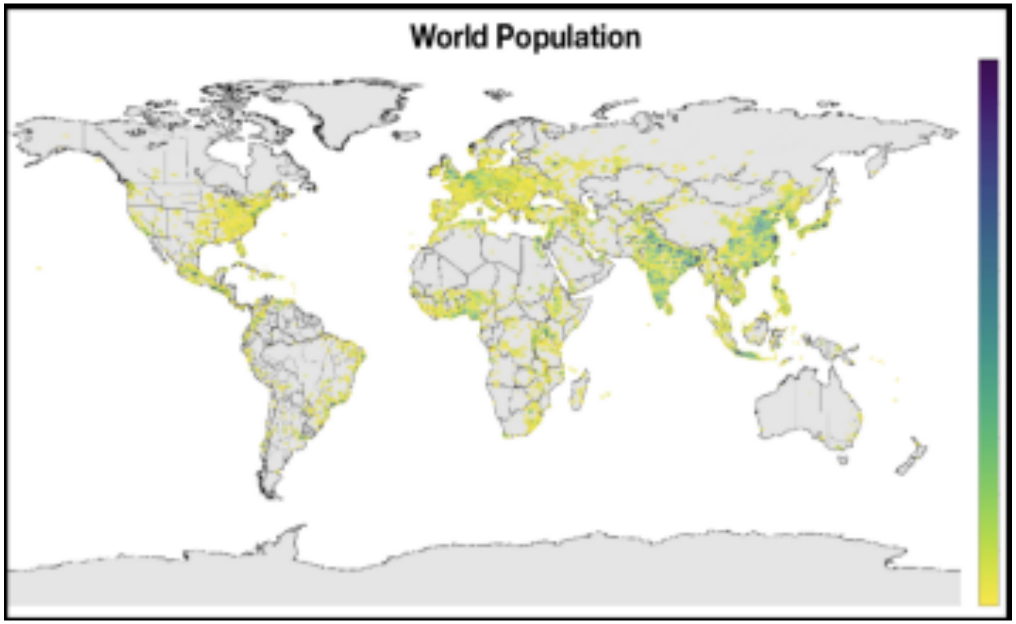
The World Map
according to the data AI sees



Sources
Research by: Koch, Denton, Hanna, and Foster (2021)
Visual by: The Mozilla Internet Health Report 2022

Lack of diversity in data





[DeVries et al. 2019, Does object recognition work for everyone?]

Who is seen? *How* are they seen?

ceremony, wedding, bride, man, groom, woman, dress

bride, ceremony, wedding, dress, woman

ceremony, bride, wedding, man, groom, woman, dress

person, people

[Shankar et al. (2017). No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World]

Slide from Timnit Gebru & Emily Denton's CVPR2020 tutorial

Gender bias in machine translation

Alex Shams
@seyyedreza

Turkish is a gender neutral language. There is no "he" or "she" - everything is just "o". But look what happens when Google translates to English. Thread:

Turkish - detected	English
o bir aşçı	she is a cook
o bir mühendis	he is an engineer
o bir doktor	he is a doctor
o bir hemşire	she is a nurse
o bir temizlikçi	he is a cleaner
o bir polis	He-she is a police
o bir asker	he is a soldier
o bir öğretmen	She's a teacher
o bir sekreter	he is a secretary
o bir arkadaş	he is a friend
o bir sevgili	she is a lover

onu sevmiyor	she does not like her
onu seviyor	she loves him
onu görüyor	she sees it
onu göremiyor	he can not see him
o onu kucaklıyor	she is embracing her
o onu kucaklamıyor	he does not embrace it
o evli	she is married
o bekar	he is single
o mutlu	he's happy
o mutsuz	she is unhappy
o çalışkan	he is hard working
o tembel	she is lazy

6:36 PM · Nov 27, 2017 · Twitter Web Client

14.9K Retweets **2K** Quote Tweets **27.2K** Likes

Racial bias in speech recognition

MARCH 23, 2020

Stanford researchers find that automated speech recognition is more likely to misinterpret black speakers

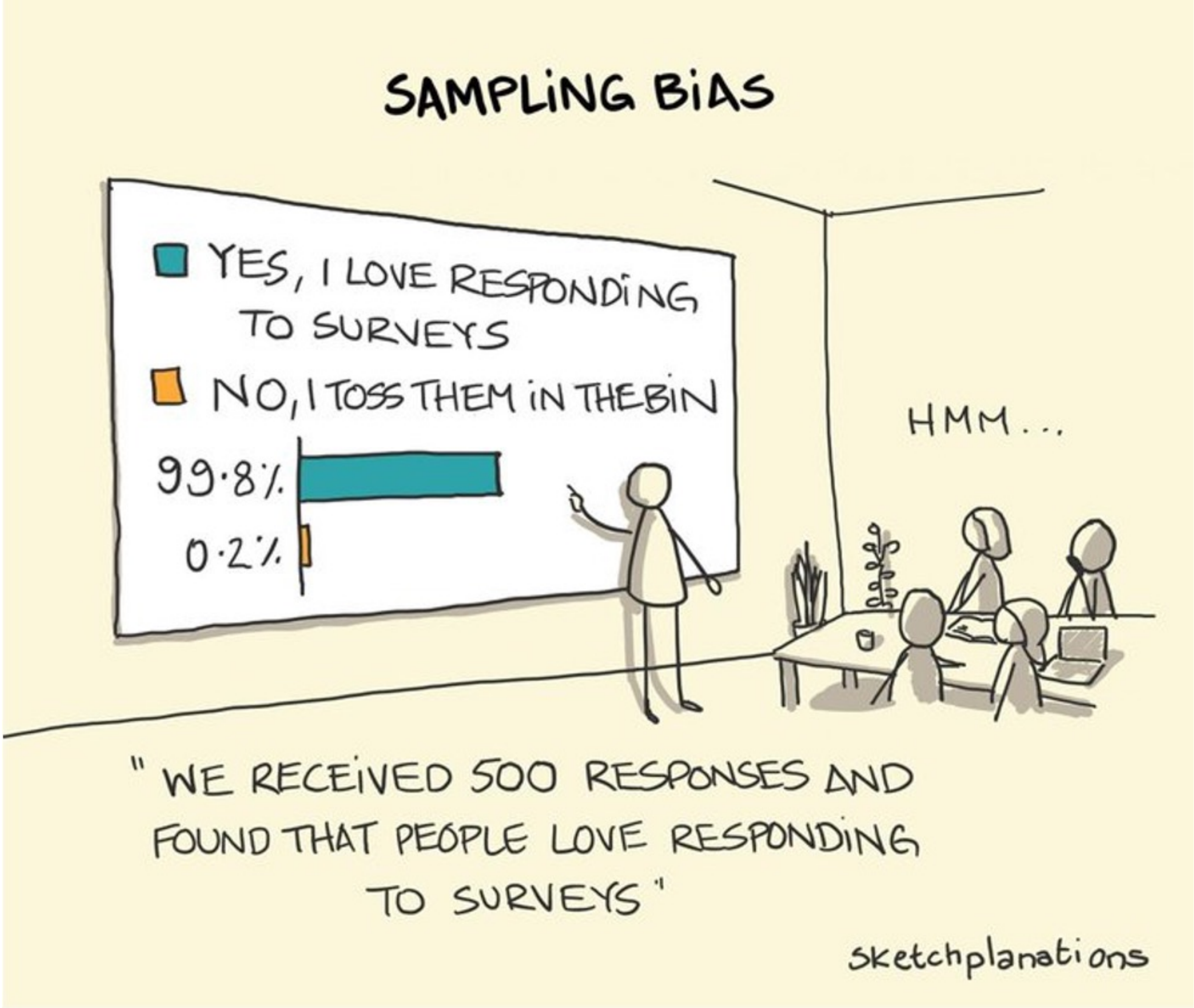
The disparity likely occurs because such technologies are based on machine learning systems that rely heavily on databases of English as spoken by white Americans.

BY EDMUND L. ANDREWS

The technology that powers the nation's leading automated speech recognition systems makes twice as many errors when interpreting words spoken by African Americans as when interpreting the same words spoken by whites, according to a new study by researchers at Stanford Engineering.

SMBC cartoon, Zach Wienersmith

Thanks to machine-learning algorithms, the robot apocalypse was short-lived.


SAMPLING BIAS
YES, I LOVE RESPONDING TO SURVEYS
NO, I TOSS THEM IN THE BIN
99.8%
0.2%
HMM...
"WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS"
sketchplanations

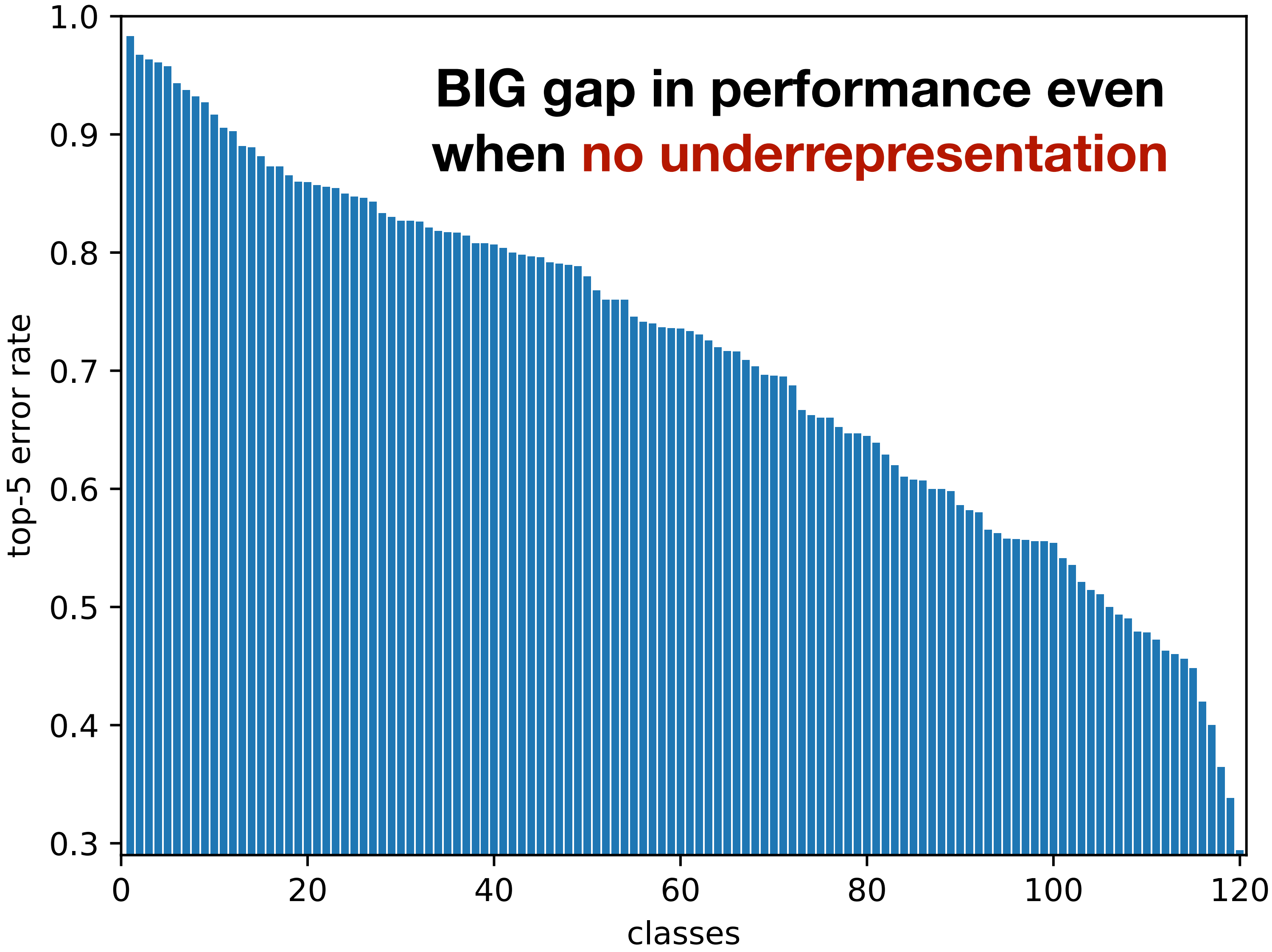
Fundamentally hard examples

- Task: classify image of dog to breed (120 classes)
- Kernel features

Stanford Dogs Dataset [Khosla et al. ’11]

No underrepresentation:
same number of images per class

Big gaps in performance



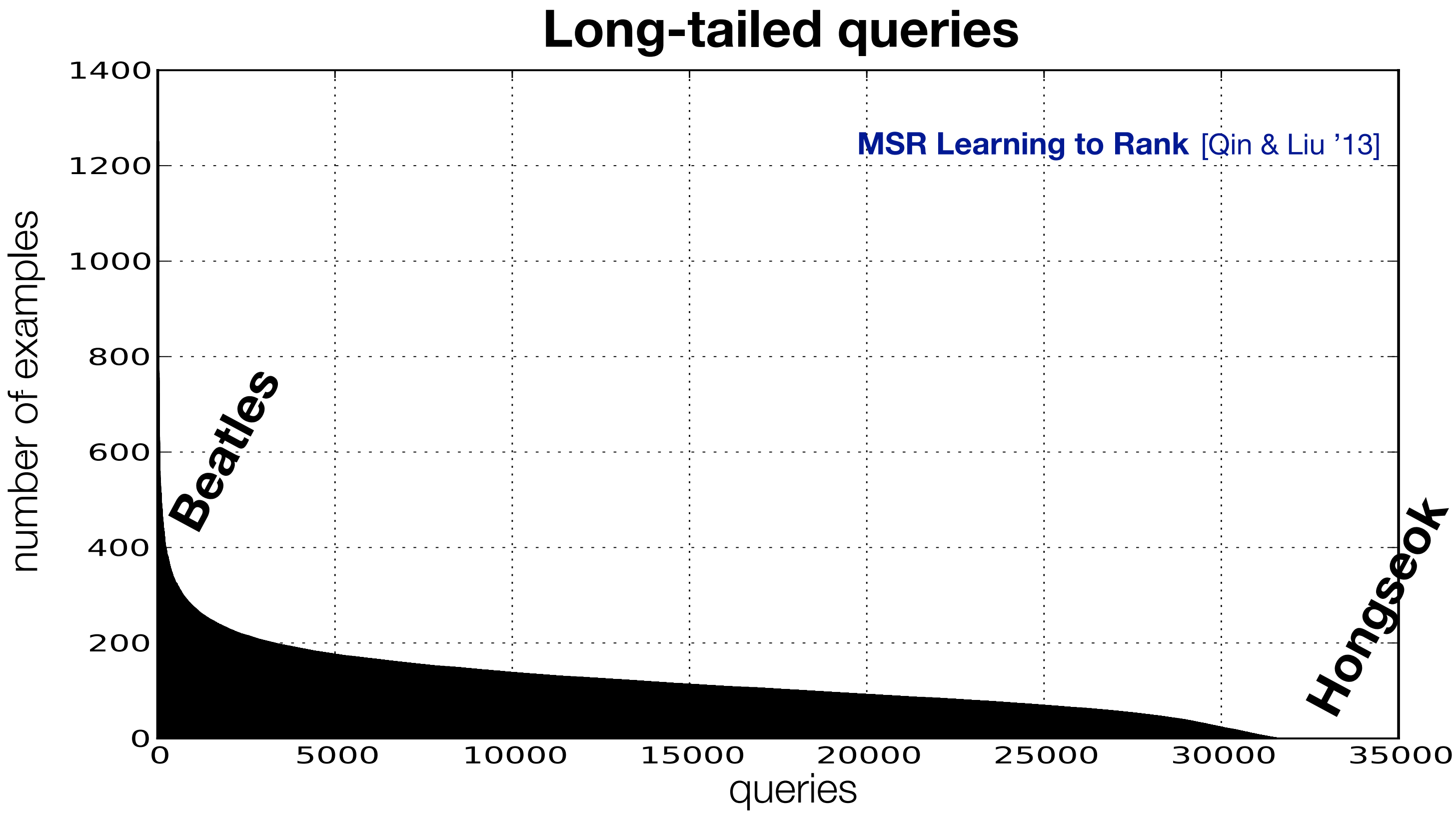
BIG gap in performance even when no underrepresentation

Hard

Easy

Long-tails

- Long-tailed data is ubiquitous in modern applications
  - Google (7 yrs ago): constant fraction of queries were new each day
- Tail inputs often determine quality of service



Long-tails

Kannada: Google apologises for 'ugliest Indian language' search result

BBC

4 June 2021

ugliest language in india

Kannada

What is the **ugliest language in India**? The answer is Kannada, a **language** spoken by around 40 million people in south **India**.

Long-tails

Long-tails

Alexa tells 10-year-old girl to touch live plug with penny

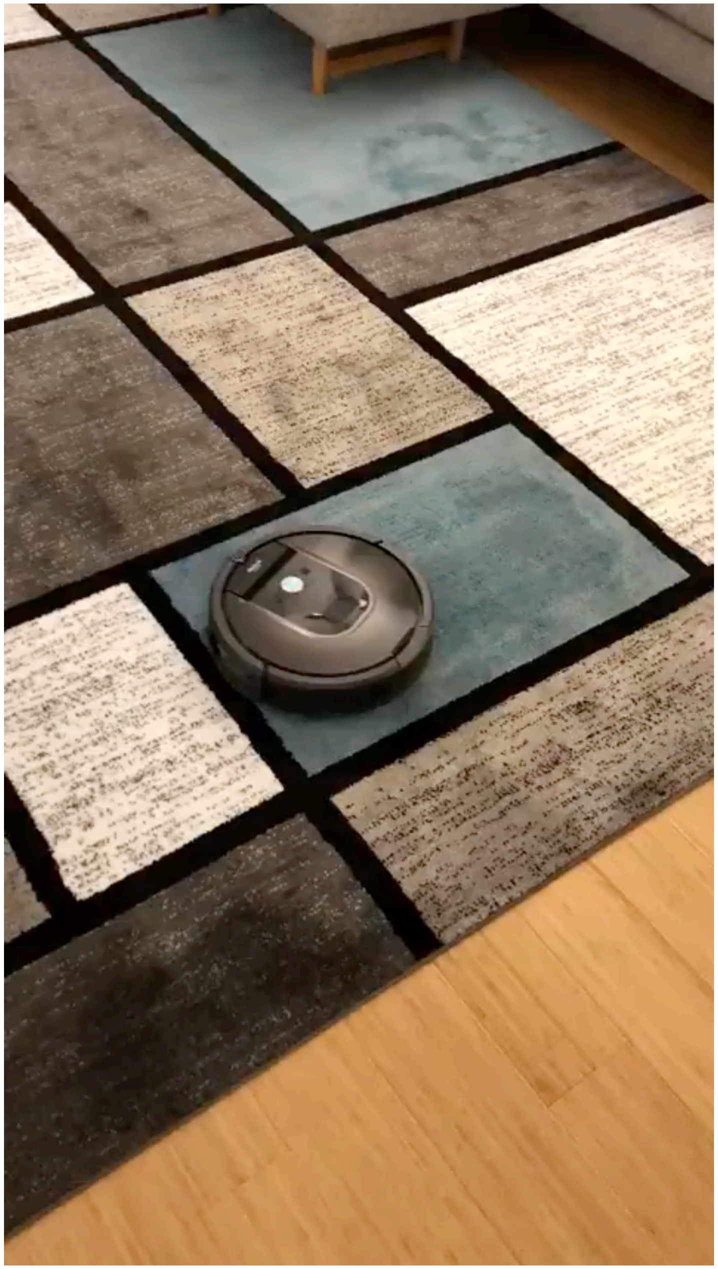
28 December 2021

A 10 yo asked Alexa for a “challenge to do”. Alexa responded with "Plug in a phone charger about halfway into a wall outlet, then touch a penny to the exposed prongs"

Long-tails

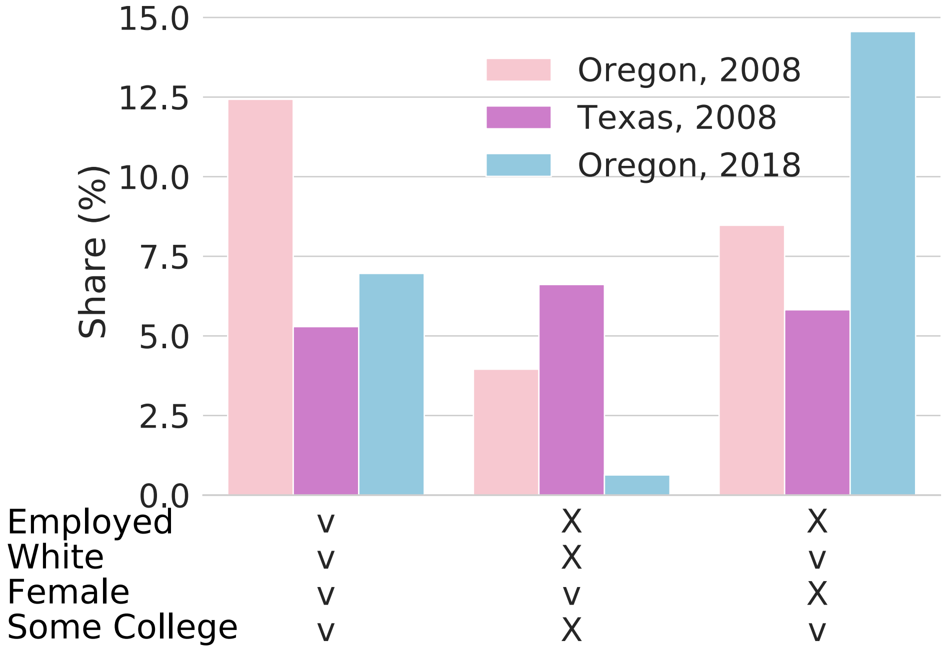
AI Camera Ruins Soccer Game For Fans After Mistaking Referee's Bald Head For Ball

Long-tails



Not a new problem…

- Standard regressors obtained from MLE lose predictive power on certain regions of covariates [Meinshausen & Buhlmann (2015)]
- Temporal, spatial shifts common



Demographic shift over space and time

Not a new problem…

Classifier Technology and the Illusion of Progress

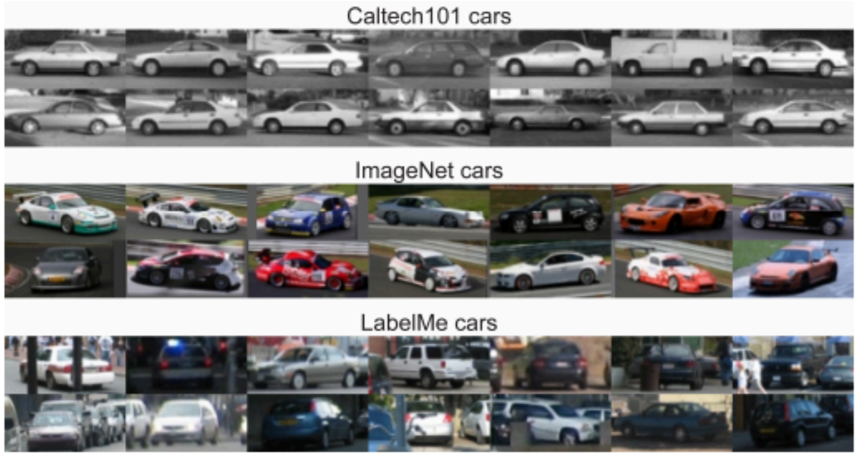
David J. Hand

Statistical Science
2006, Vol. 21, No. 1, 1–14
DOI 10.1214/088342306000000060
© Institute of Mathematical Statistics, 2006

- “A fundamental assumption of the classical paradigm is that the various distributions involved do not change over time. In fact, in many applications this is unrealistic and the population distributions are nonstationary.”
  - Marketing & banking: Classification rules used to predict loan default updated every few months
  - “Their performance degrades, not because the rules themselves change, but because the distributions to which they are being applied change”

Not a new problem…

- Model performance drops across different domains and datasets [Torralba & Efros (2011)]



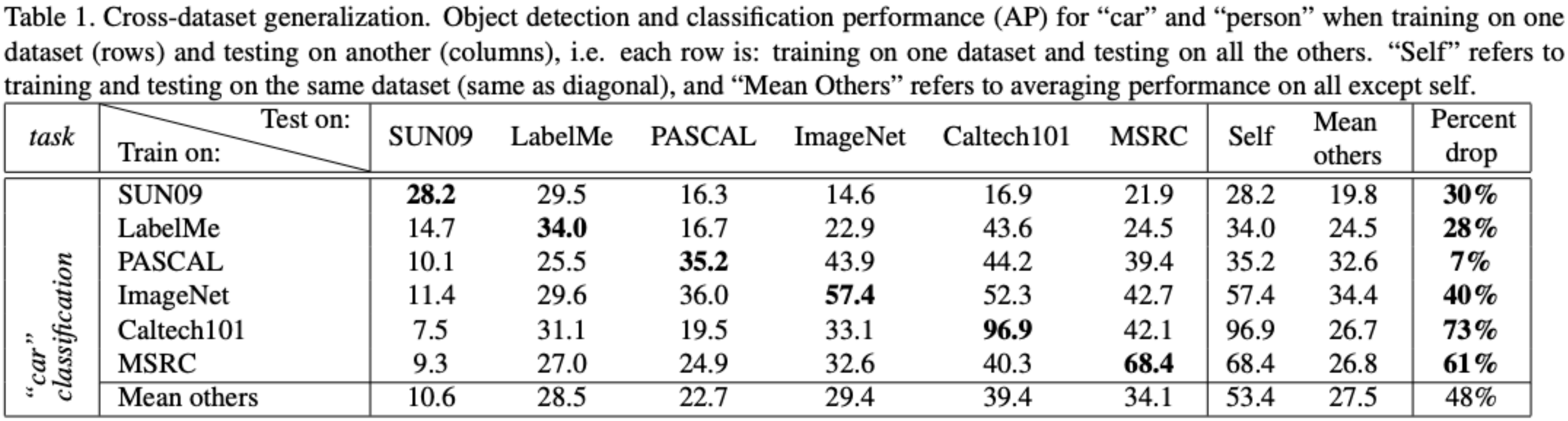
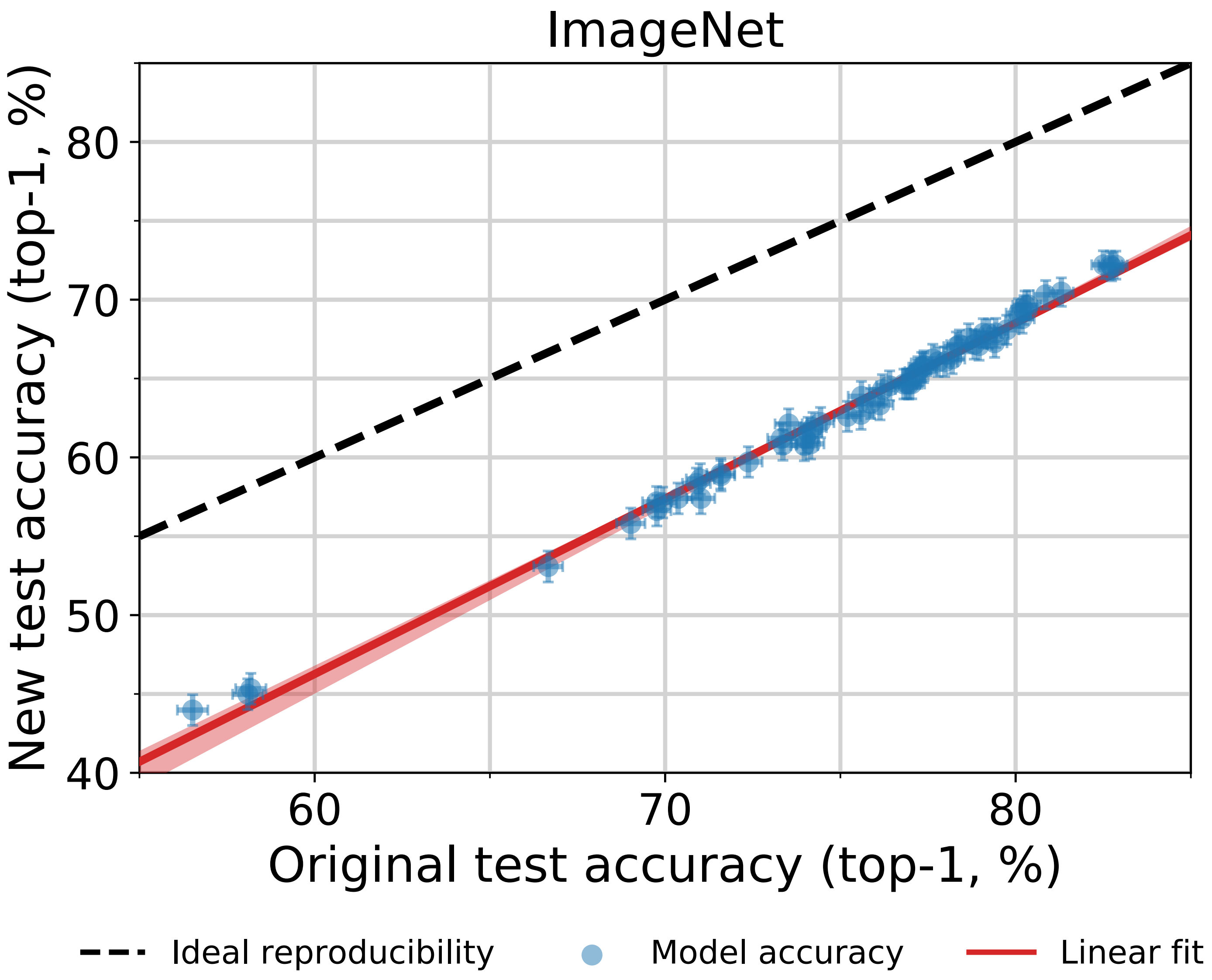


Table 1. Cross-dataset generalization. Object detection and classification performance (AP) for "car" and "person" when training on one dataset (rows) and testing on another (columns), i.e. each row is: training on one dataset and testing on all the others. "Self" refers to training and testing on the same dataset (same as diagonal), and "Mean Others" refers to averaging performance on all except self.

task	Test on: / Train on:	SUN09	LabelMe	PASCAL	ImageNet	Caltech101	MSRC	Self	Mean others	Percent drop
"car" classification	SUN09	**28.2**	29.5	16.3	14.6	16.9	21.9	28.2	19.8	**30%**
	LabelMe	14.7	**34.0**	16.7	22.9	43.6	24.5	34.0	24.5	**28%**
	PASCAL	10.1	25.5	**35.2**	43.9	44.2	39.4	35.2	32.6	**7%**
	ImageNet	11.4	29.6	36.0	**57.4**	52.3	42.7	57.4	34.4	**40%**
	Caltech101	7.5	31.1	19.5	33.1	**96.9**	42.1	96.9	26.7	**73%**
	MSRC	9.3	27.0	24.9	32.6	40.3	**68.4**	68.4	26.8	**61%**
	Mean others	10.6	28.5	22.7	29.4	39.4	34.1	53.4	27.5	48%

SOTA models are also non-robust



[Does ImageNet classifiers generalize to ImageNet?
Recht, Roelofs, Schmidt, Shankar ’19]

SOTA models are non-robust

Similar frames extracted from videos

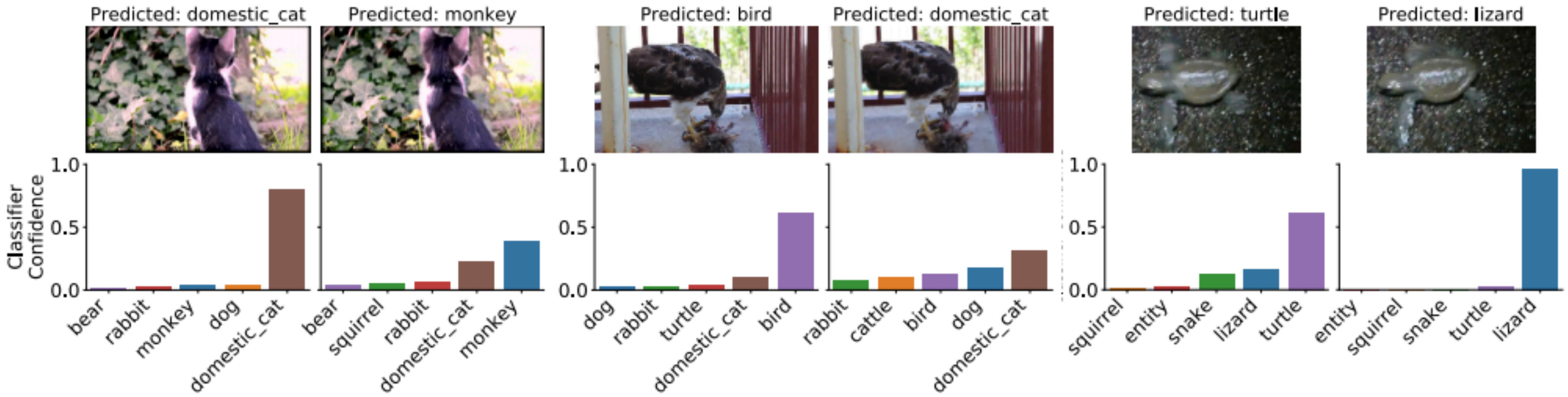


Figure 1: Three examples of natural perturbations from nearby video frames and resulting classifier predictions from a ResNet-152 model fine-tuned on ImageNet-Vid. While the images appear almost identical to the human eye, the classifier confidence changes substantially.

[Does ImageNet classifiers generalize across time?
Shankar, Dave, Roelofs, Ramanan, Recht, Schmidt ’19]

SOTA models are non-robust

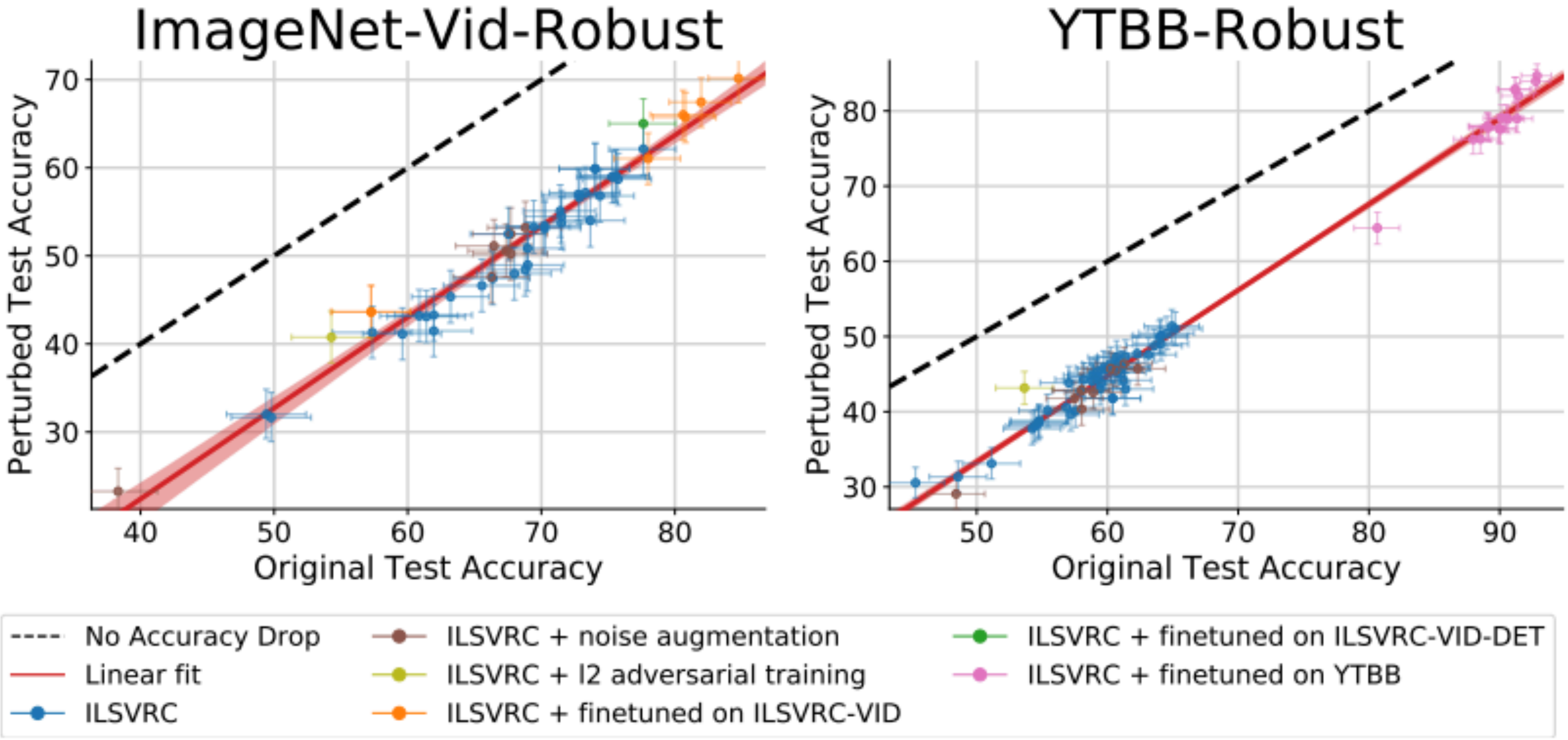


Figure 3: Model accuracy on original vs. perturbed images. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). Each perturbed frame was taken from a ten frame neighborhood of the original frame (approximately 0.3 seconds). All frames were reviewed by humans to confirm visual similarity to the original frames.

[Does ImageNet classifiers generalize across time?
Shankar, Dave, Roelofs, Ramanan, Recht, Schmidt ’19]

SOTA models are non-robust

- Deep networks are very brittle
  - imperceptible adversarial perturbations can fool them

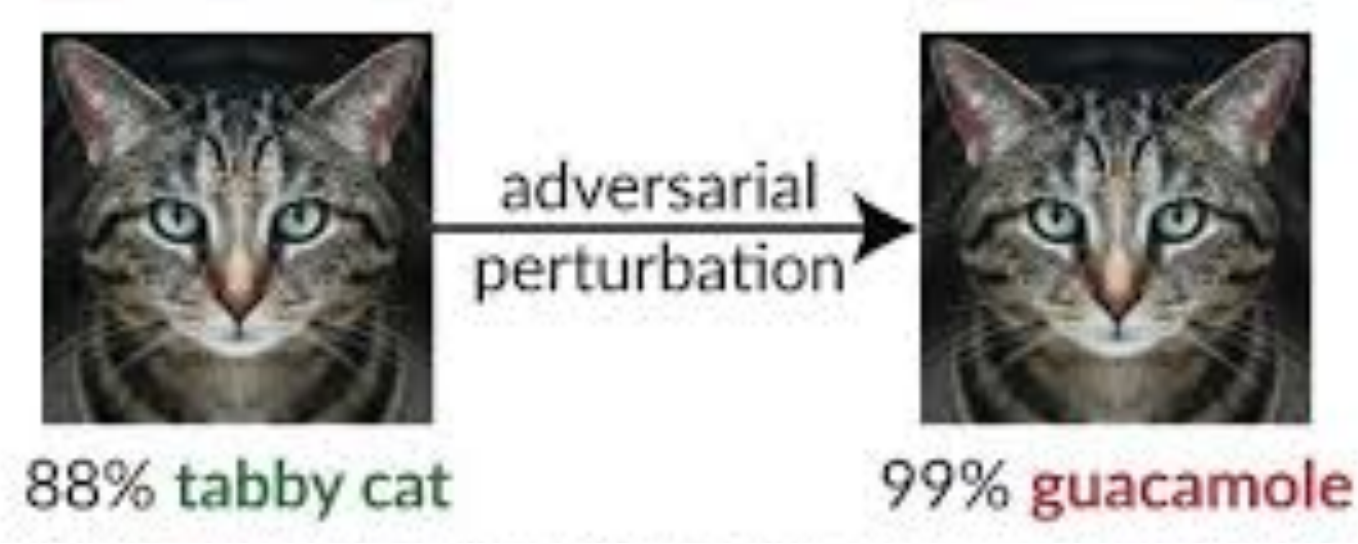
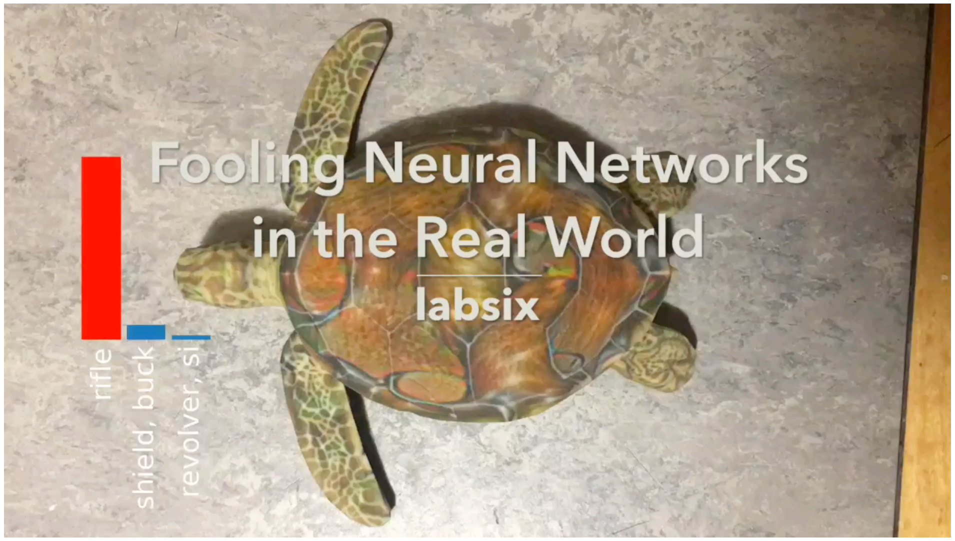


Figure credit: Nicholas Carlini

SOTA models are non-robust

- Deep networks are very brittle
  - imperceptible adversarial perturbations can fool them



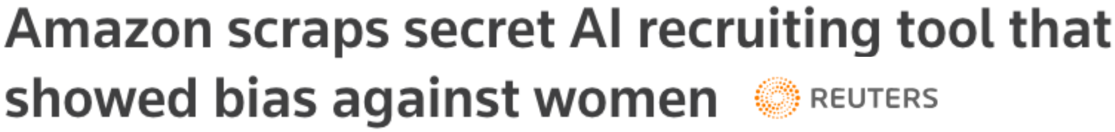
[Athalye et al. '17]

[Chen et al. '18]

Spurious correlations

- Models fit to observed associations, which maybe not be the fundamental structure that we want to learn

- But I want my models to work in a non-patriarchal society without sexism



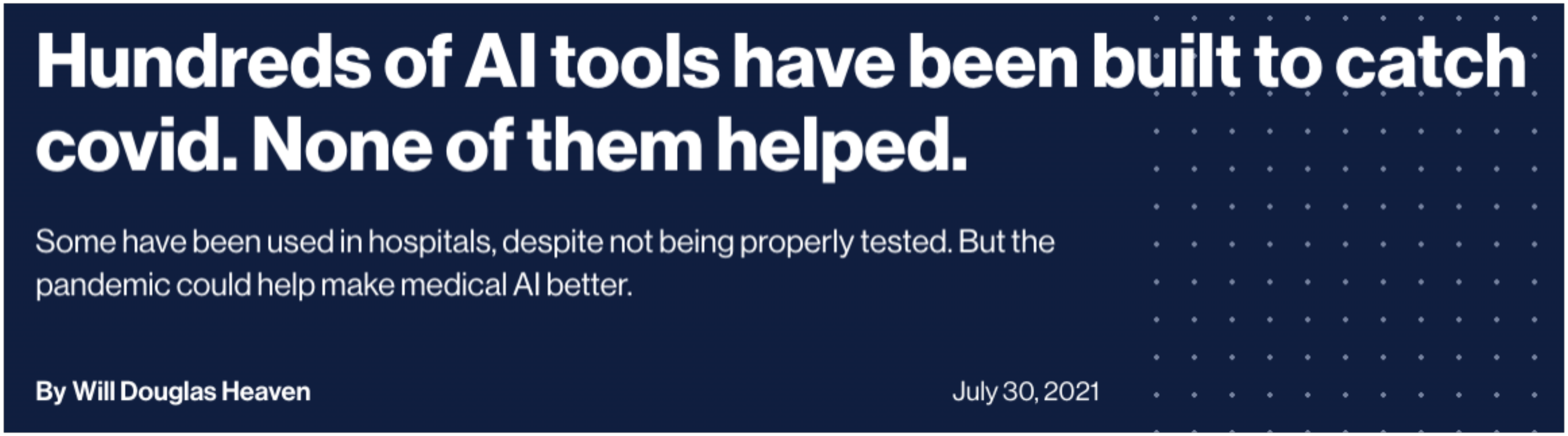
Amazon scraps secret AI recruiting tool that showed bias against women

Spurious correlations

- Correlation is no substitute for **causal** evidence

- COVID prediction AIs were found to be “picking up on the text font that certain hospitals used to label the scans.”

- “As a result, fonts from hospitals with more serious caseloads became predictors of covid risk.”



Hundreds of AI tools have been built to catch covid. None of them helped.

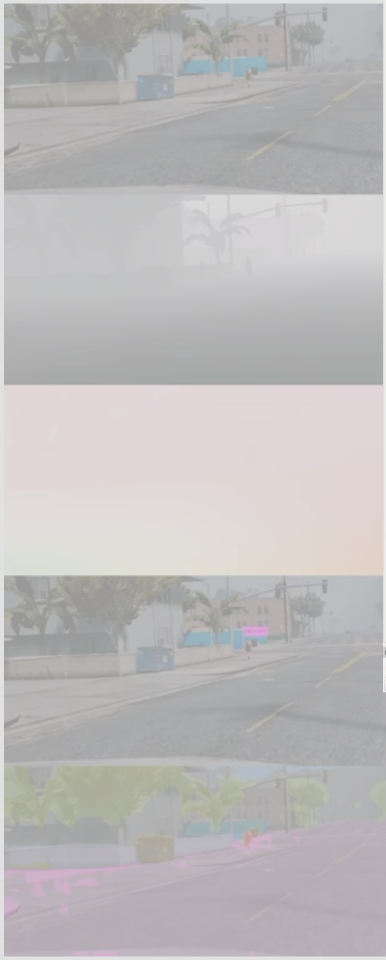
Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

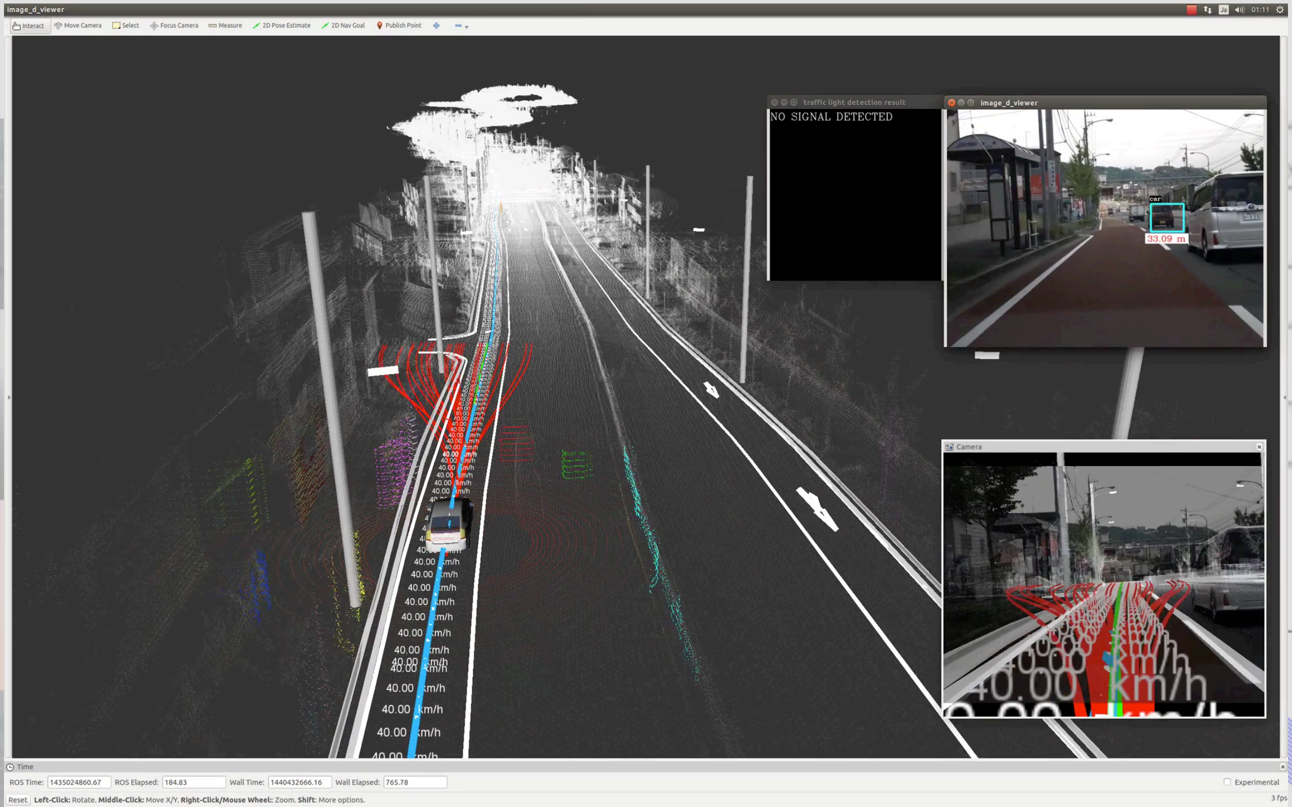
By Will Douglas Heaven

July 30, 2021

Complex system example: AVs

Sense





NO SIGNAL DETECTED

ROS Time: 1435024860.67 ROS Elapsed: 184.83 Wall Time: 1440432666.16 Wall Elapsed: 765.78

Reset Left-Click: Rotate. Middle-Click: Move X/Y. Right-Click/Mouse Wheel: Zoom. Shift: More options.

At the end of the day:

A function that *generates* a sequence of *steering* and *acceleration* commands

Complex system example: AVs

Mobileye running a red light

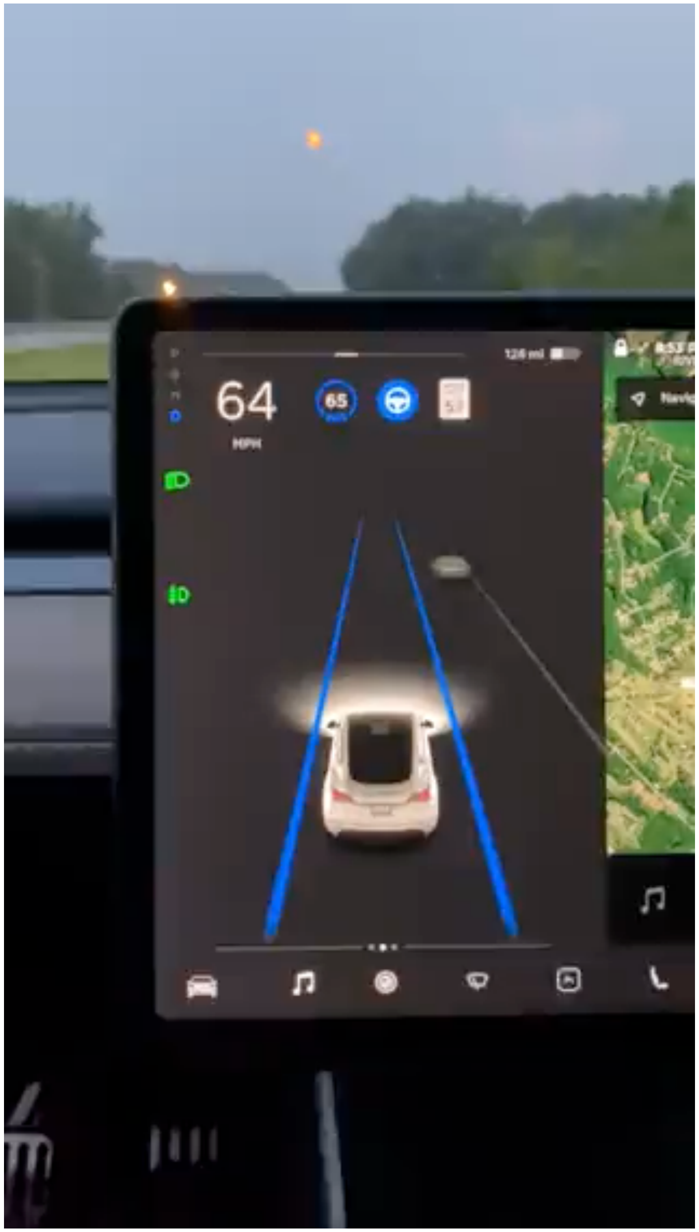
Tesla Autopilot fatal accident

Tesla

- Tesla’s self-driving systems are notorious for only using visual information, rather than other sensors such as LiDAR
- This makes the entire system brittle to varied edge cases

Main takeaway

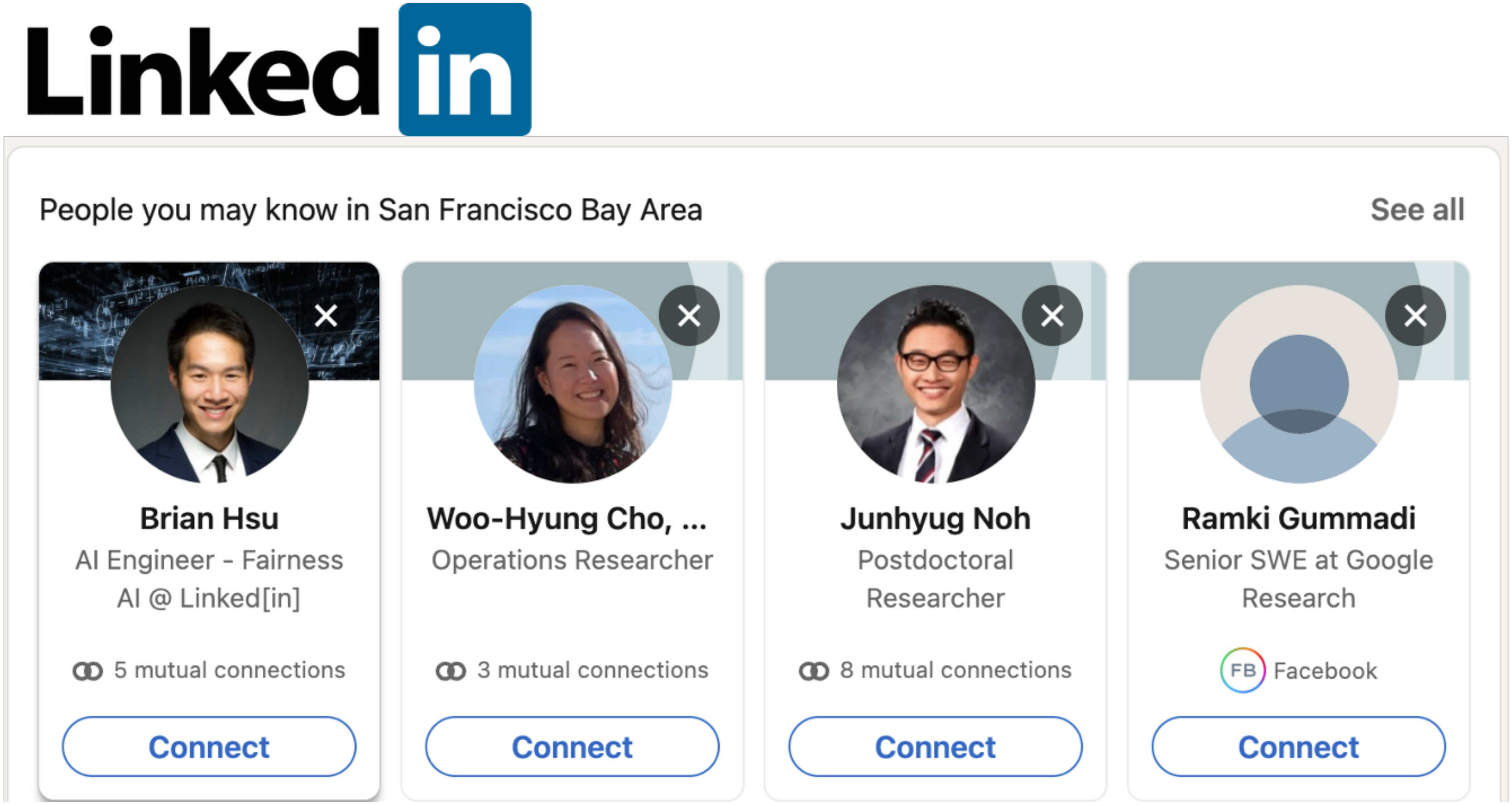
- Don’t buy a Tesla!



Main takeaway

- Don’t buy a Tesla!

Metrics



People you may know in San Francisco Bay Area

See all

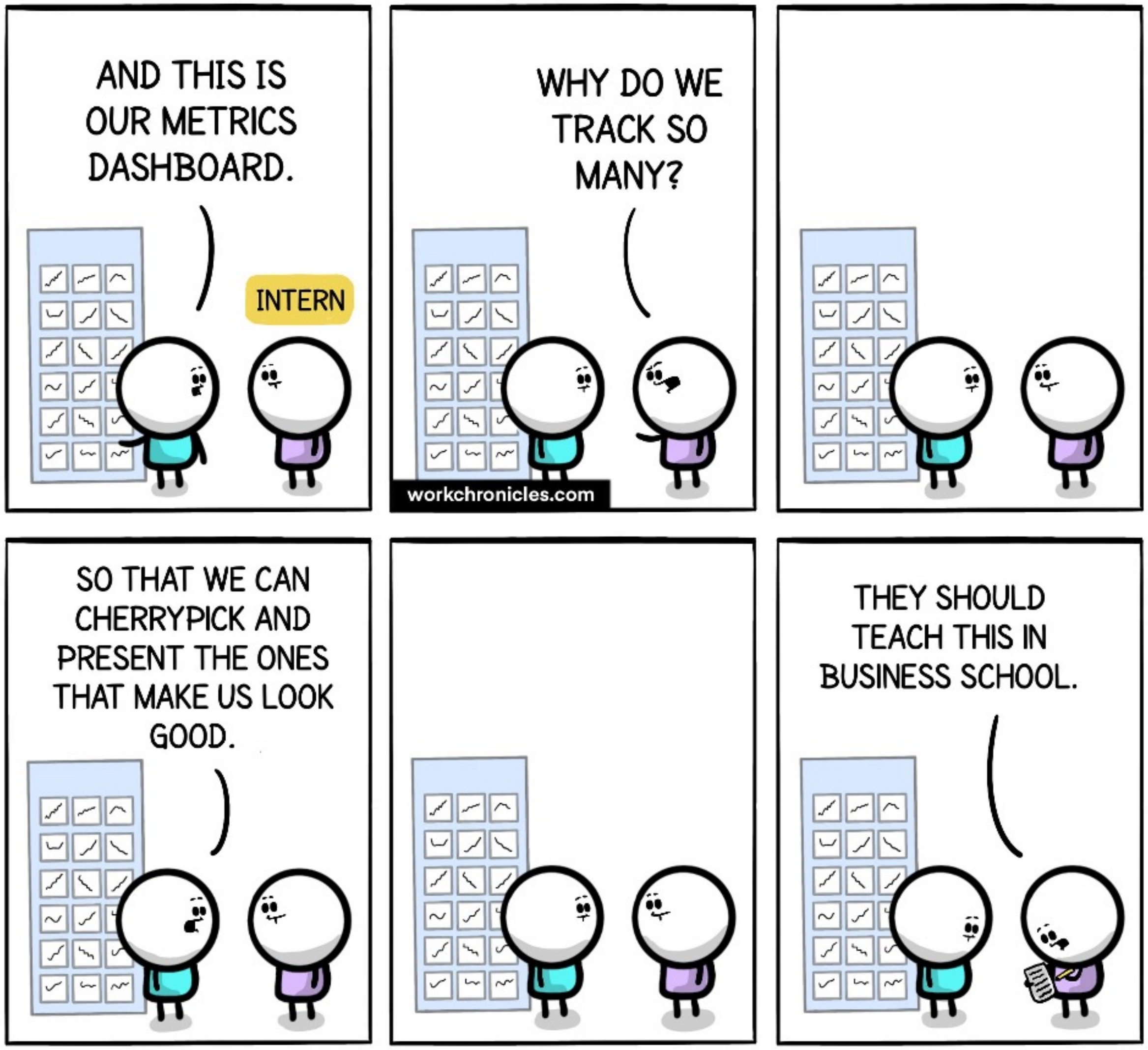
Brian Hsu
AI Engineer - Fairness AI @ Linked[in]
5 mutual connections

Woo-Hyung Cho, ...
Operations Researcher
3 mutual connections

Junhyug Noh
Postdoctoral Researcher
8 mutual connections

Ramki Gummadi
Senior SWE at Google Research
Facebook

Metrics & incentives



Lots of questions

- Understand unanticipated distributional shifts
- How do we learn causal structures?
- Ultimately, ML models work towards aiding downstream decisions
  - Prediction is not the ultimate goal
  - How to design models with this in mind?
- How do we evaluate the entire system, with many complex modules?

Lots of questions

- ML system interacts with (strategic) agents over time. How to model this interaction/dynamics?
- Operational constraints (safety, reliability etc)
- Collected data on decisions are observational
  - Often based on human agents' decisions, which may depend on unrecorded variables
  - For sequential decisions, observed data often does not cover entire (action seq, state seq) space. So not really "big data"…

Rest of the course

- First, learn foundational techniques!
  - ~1 month on basic results in statistical learning

- Then, survey recent works that aim to identify, model, improve upon aforementioned challenges

- Goal: Develop a critical view of topics surrounding reliability
  - Much remains to be done in ML
  - Discussions toward context-specific applications
    e.g. healthcare, online platforms, manufacturing…

- Goal: Identify interfaces
  - mechanism design
  - sequential decision-making
  - causal inference