

Overview

A history lesson



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

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

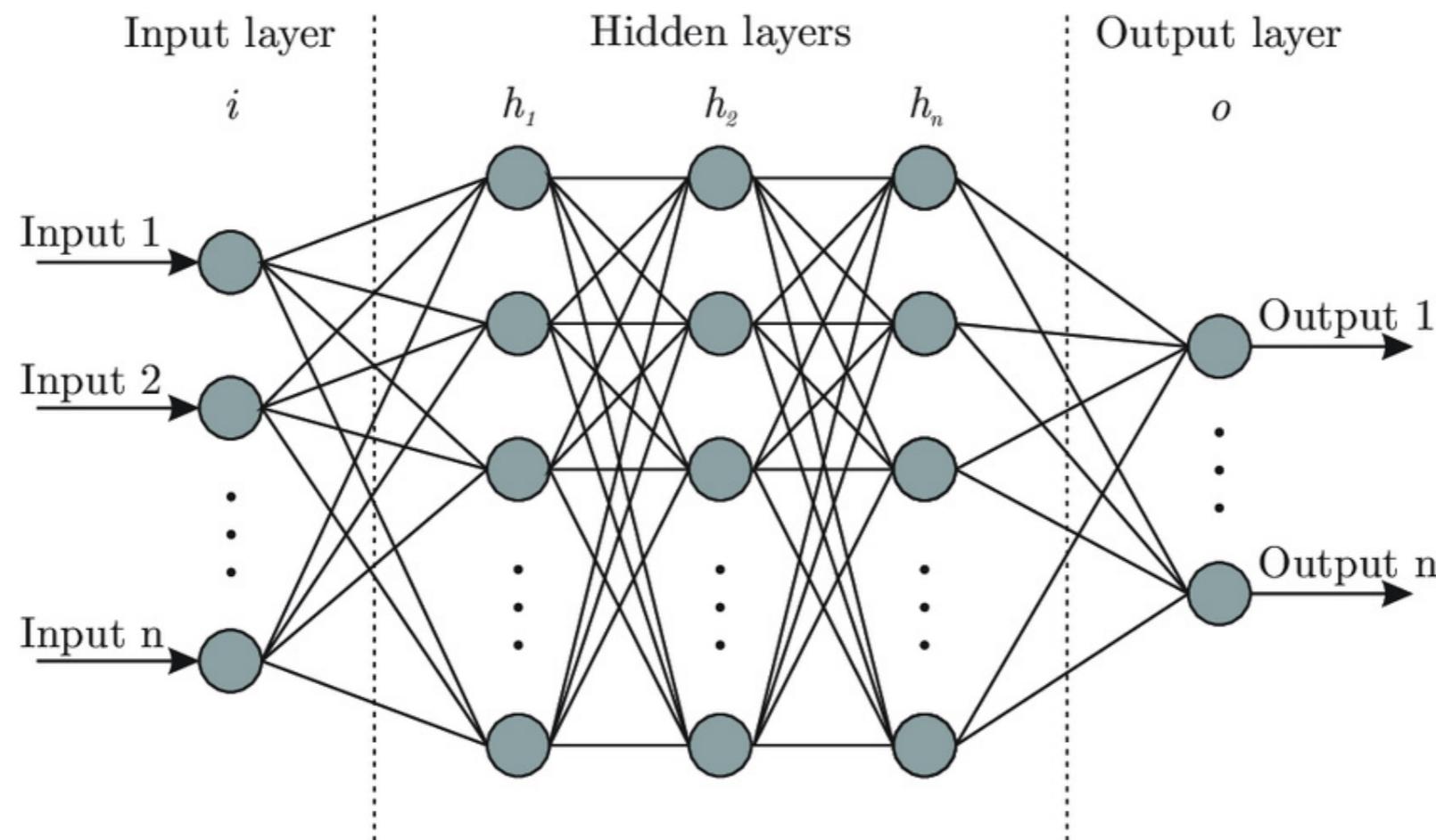
Stochastic optimization

- 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

Representations

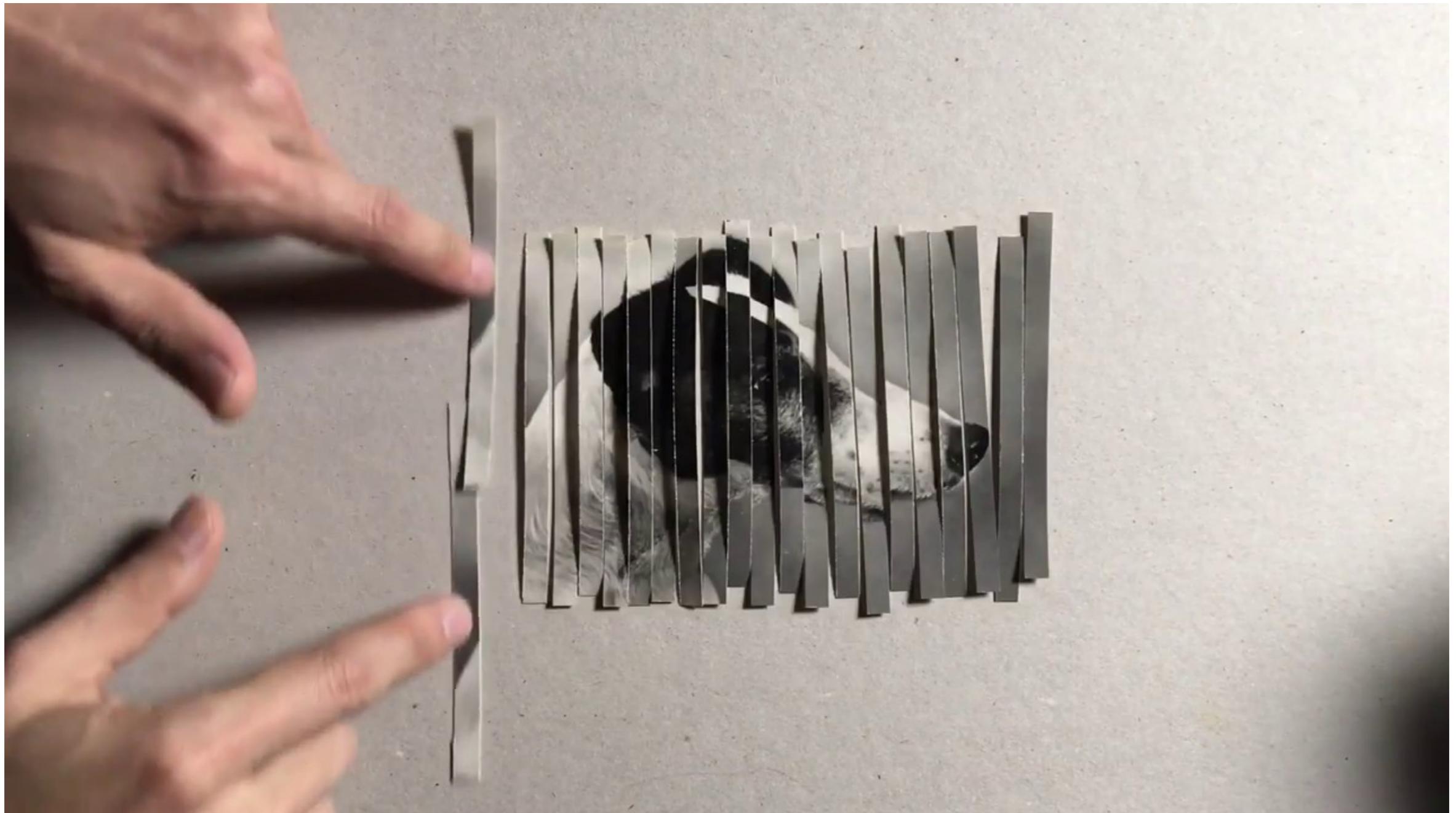
- Unlike decision-making problems, loss is largely fictitious
- We often care a lot more about the versatility of the learned feature representation
- e.g., take pre-trained representation, fine-tune it on downstream task

Neural networks

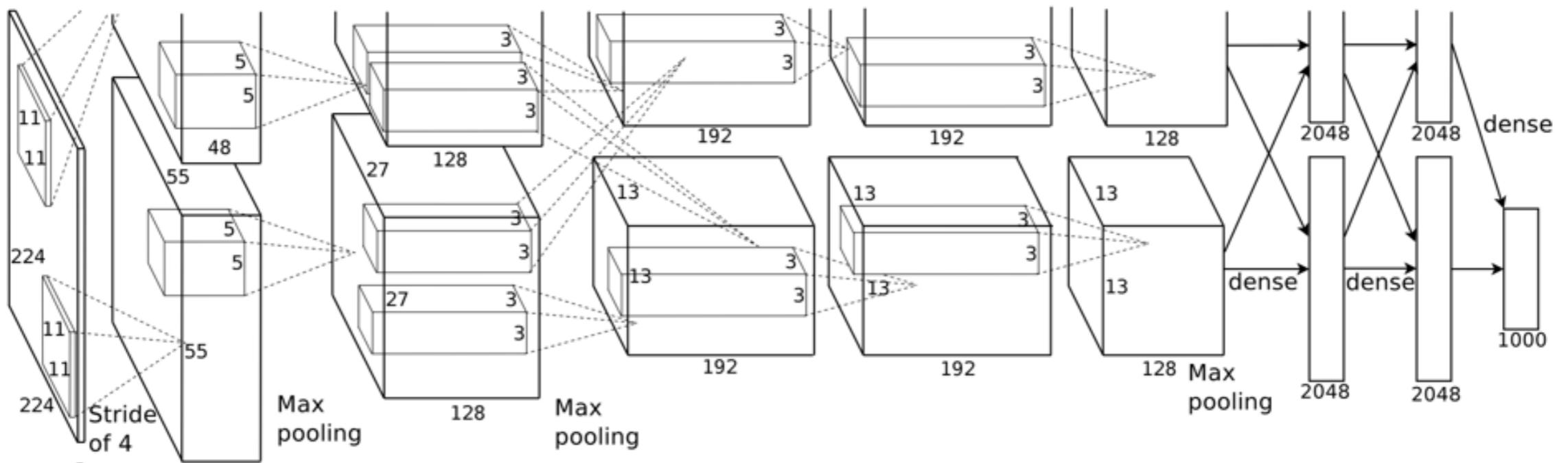


Neural networks

Learn geometry!



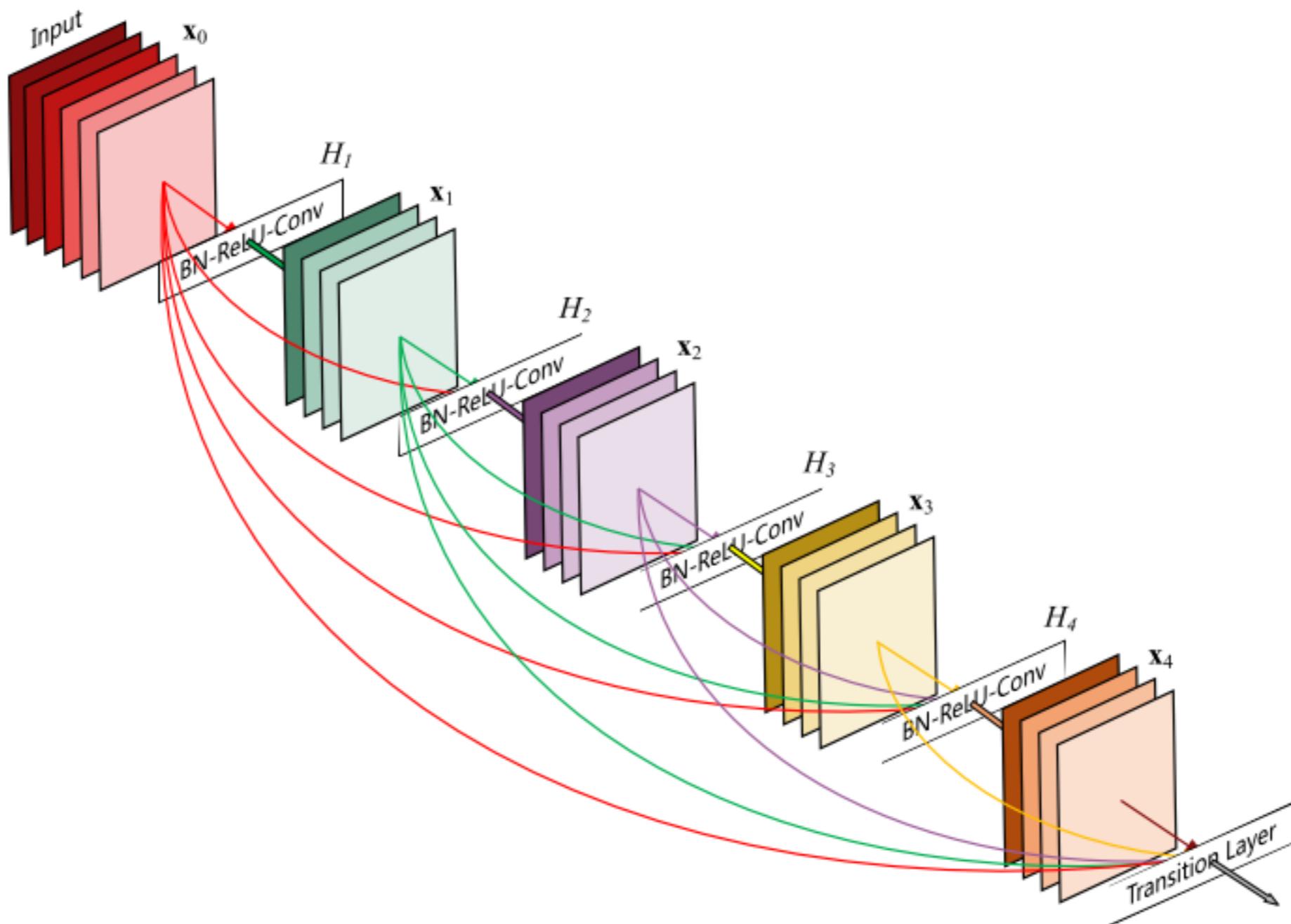
Convolutional nets



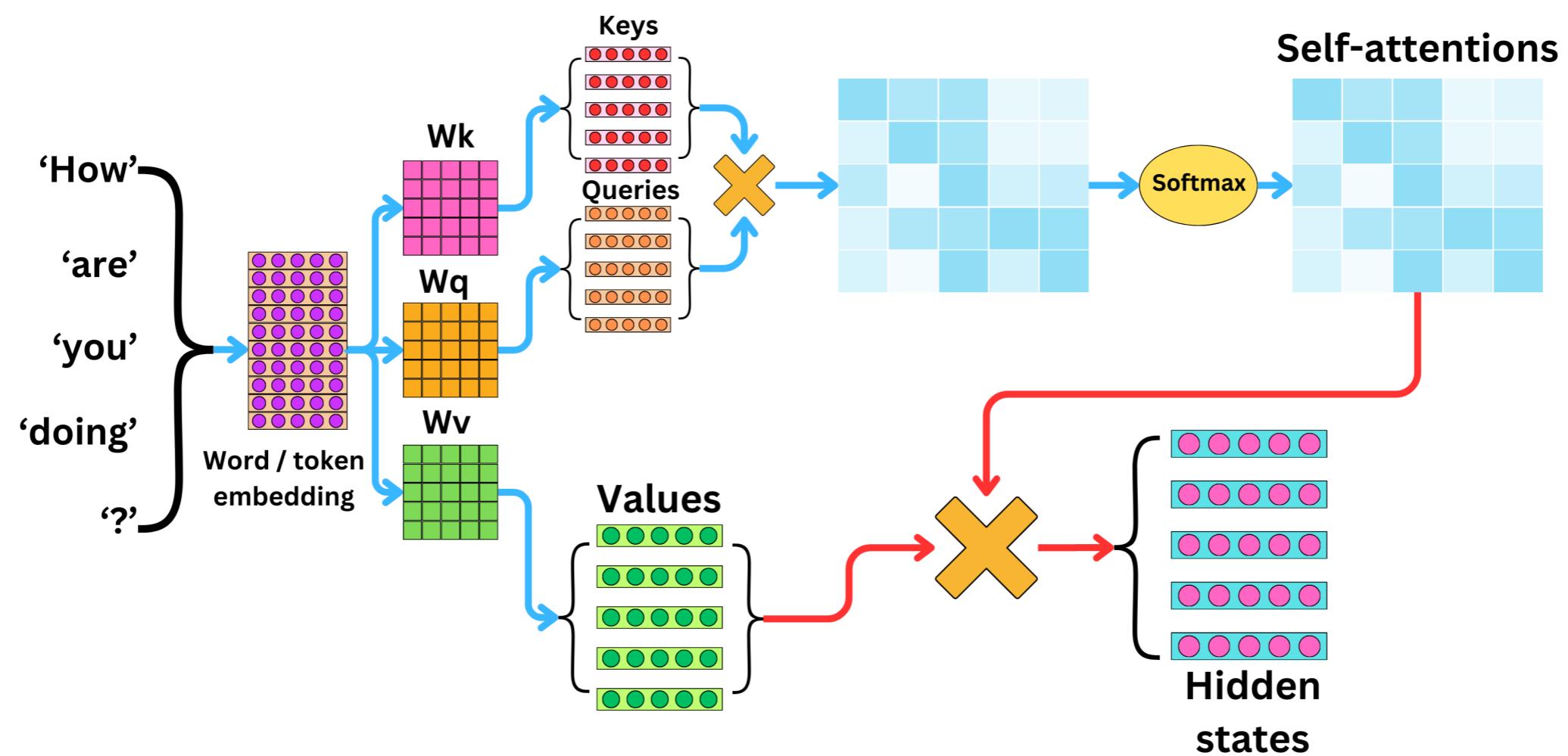
Krizhevsky et al. 2012

Residual nets

He et al. 2015



Transformers



Empirical risk minimization

- How do we approximately optimize the objective $\mathbb{E}_P[\ell(\theta; Z)]$
- Empirical risk minimization (ERM), or sample average approximation (SAA) over $Z_i \stackrel{\text{iid}}{\sim} P$

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

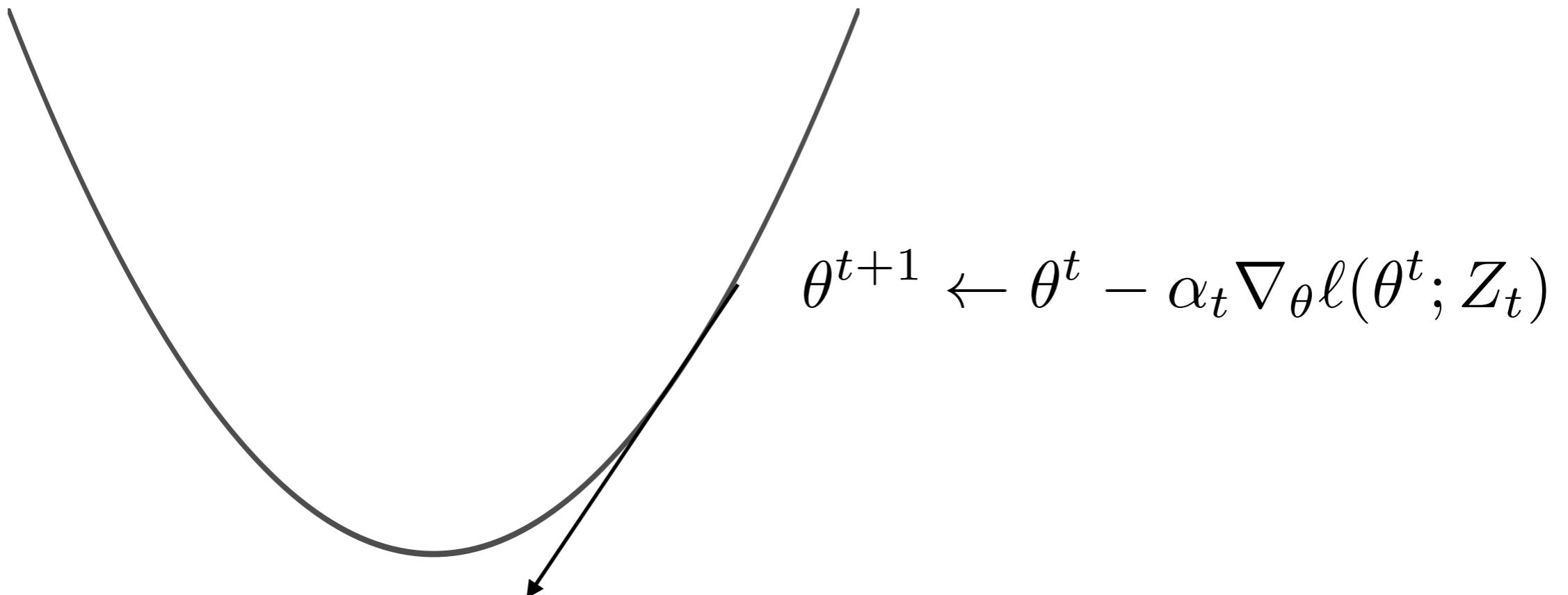
Optimization

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

- How do we solve the ERM/SAA problem?
- 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

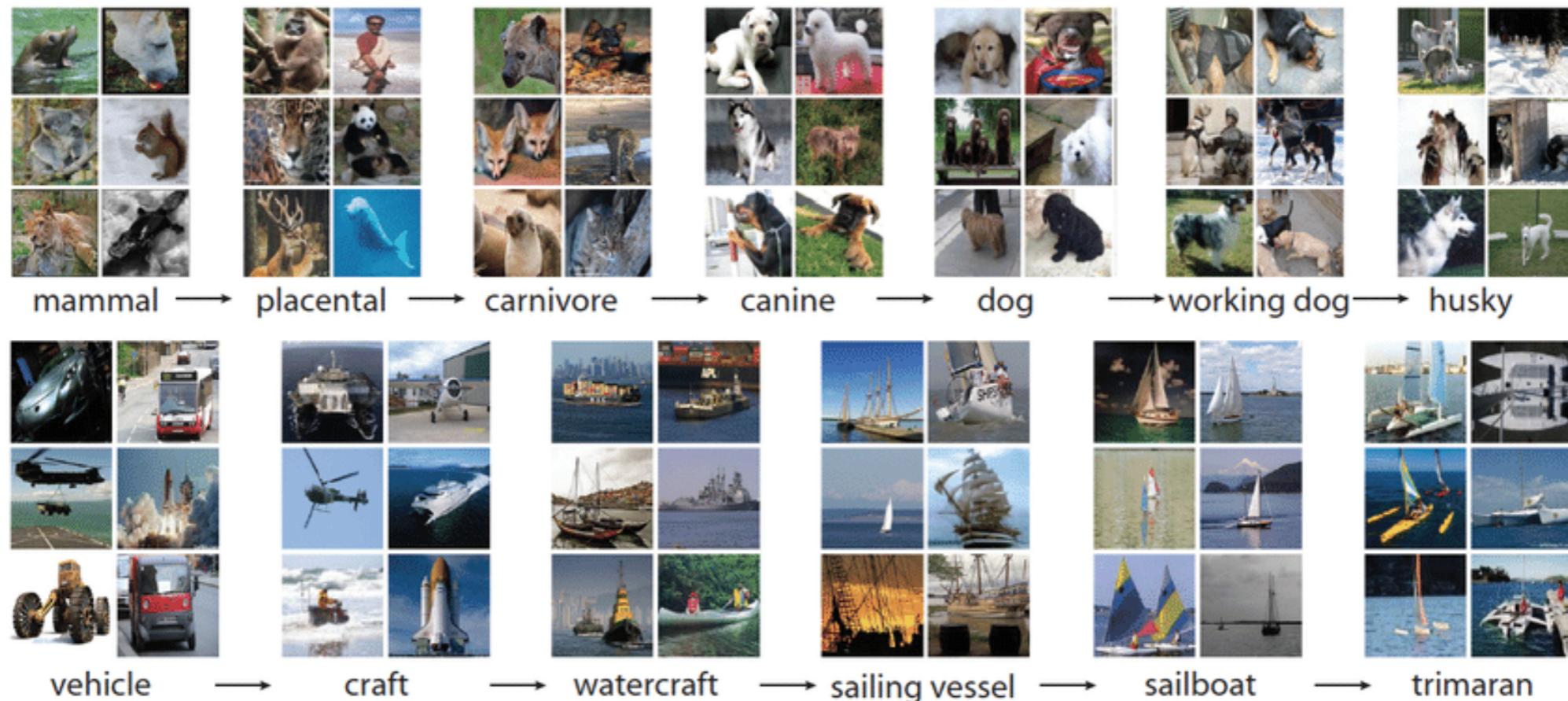
$$\underset{\theta \in \Theta}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^n \ell(\theta; Z_i)$$



Magic formula

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

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

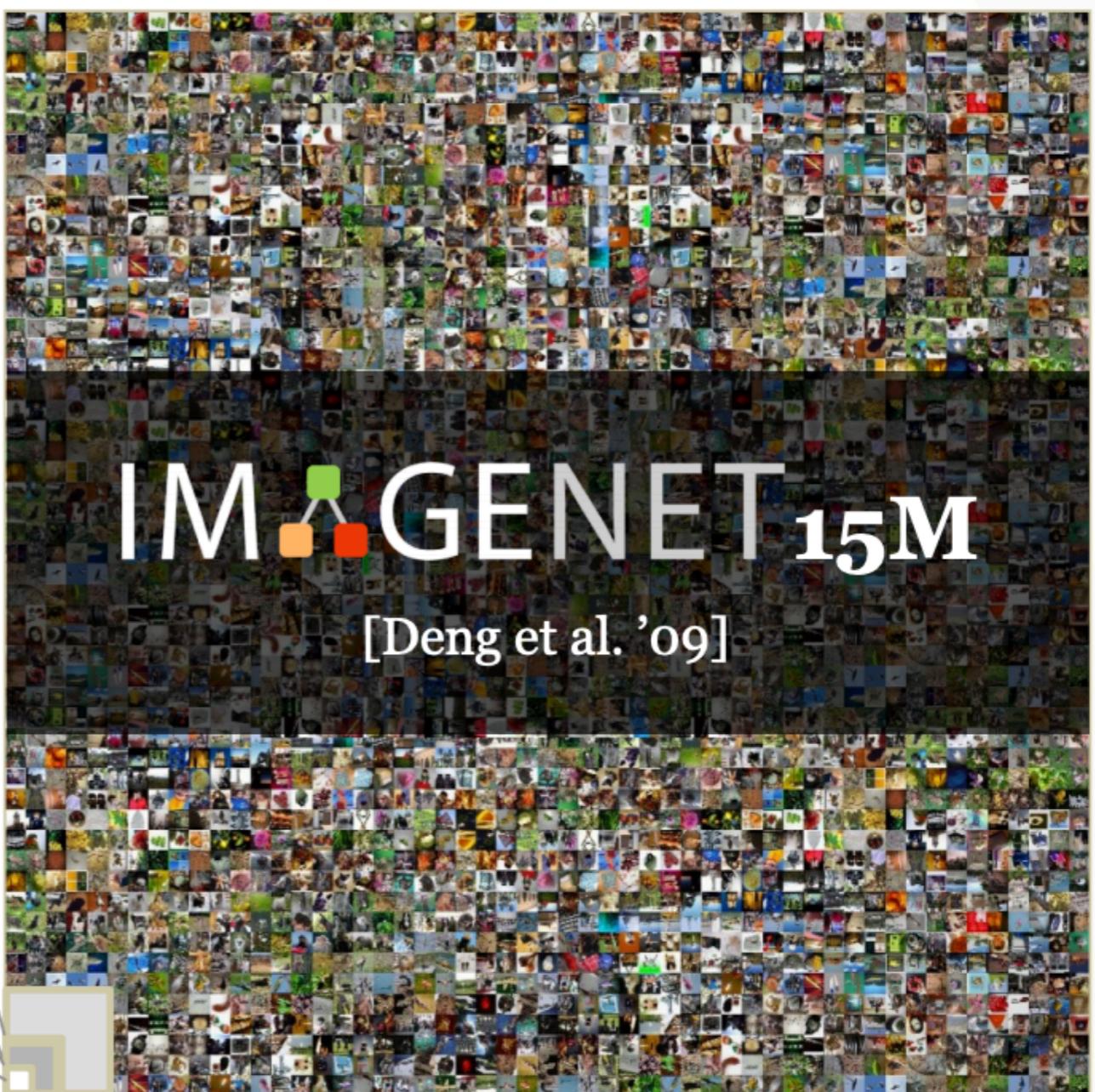
[Russell et al. '07]

PASCAL VOC, 30K

[Everingham et al. '06-'12]

Caltech101, 9K

[Fei-Fei, Fergus, Perona, '03]



Slide from Fei-Fei Li

Hammers



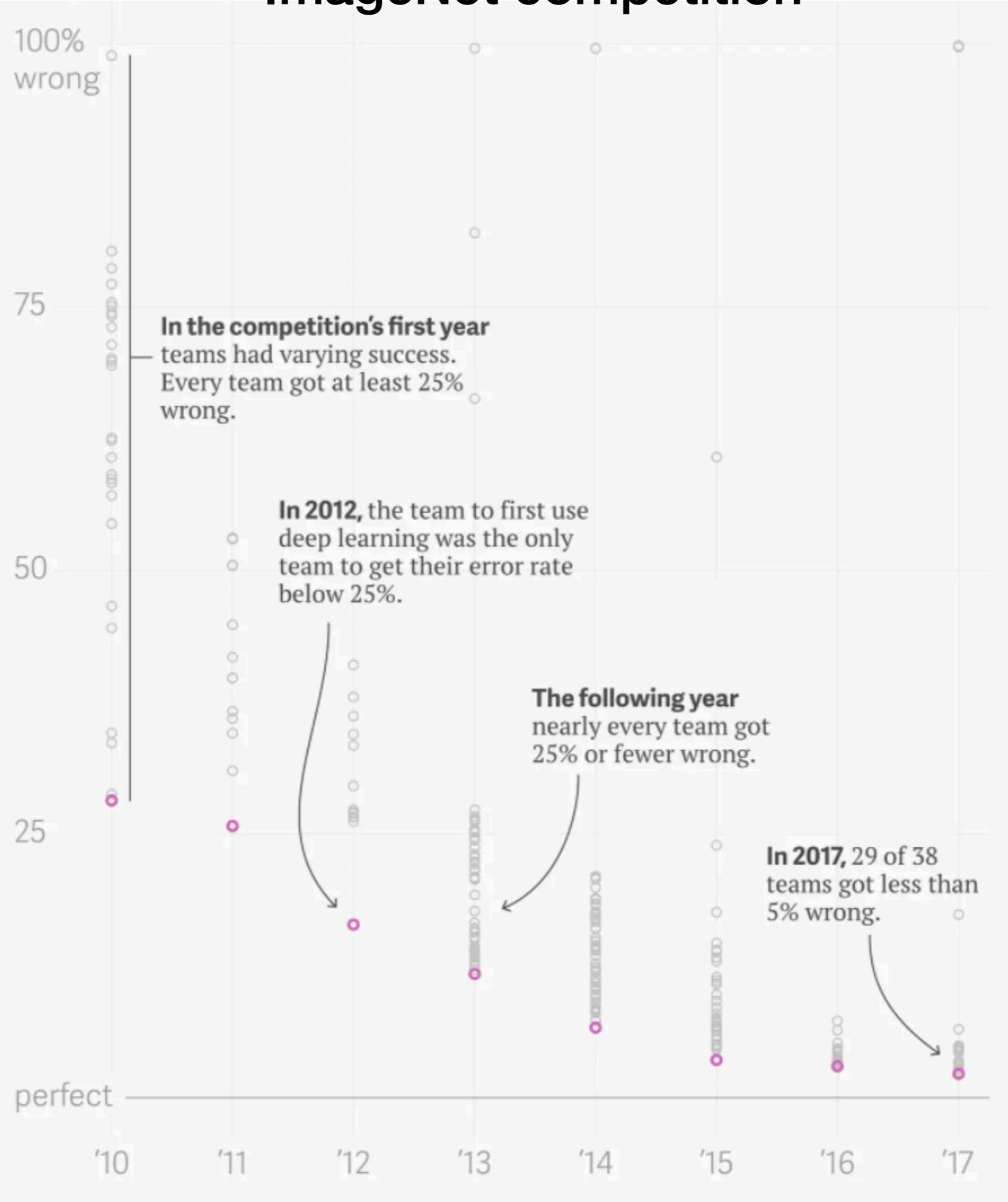
Slide from Jia Deng

Ladles



Slide from Jia Deng

ImageNet competition



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

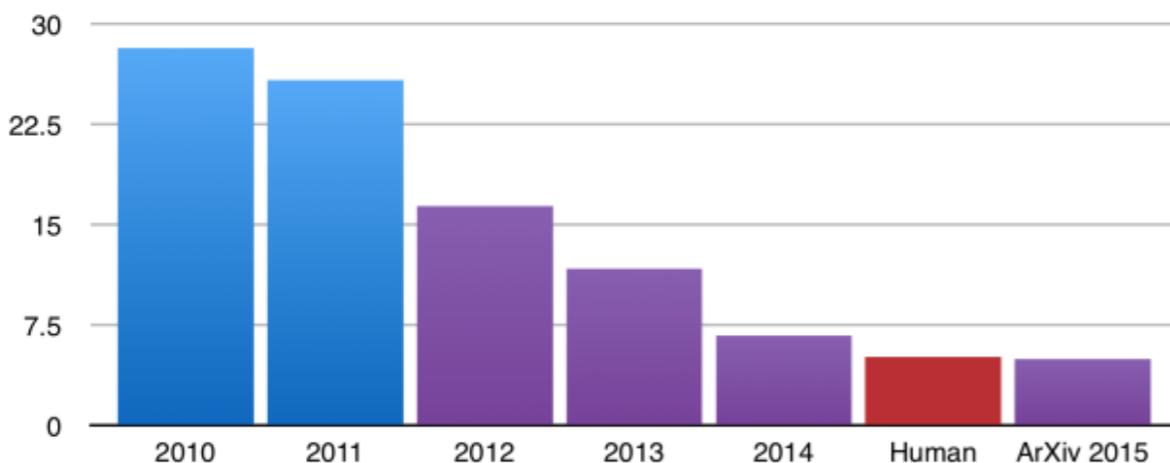
Applications

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

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

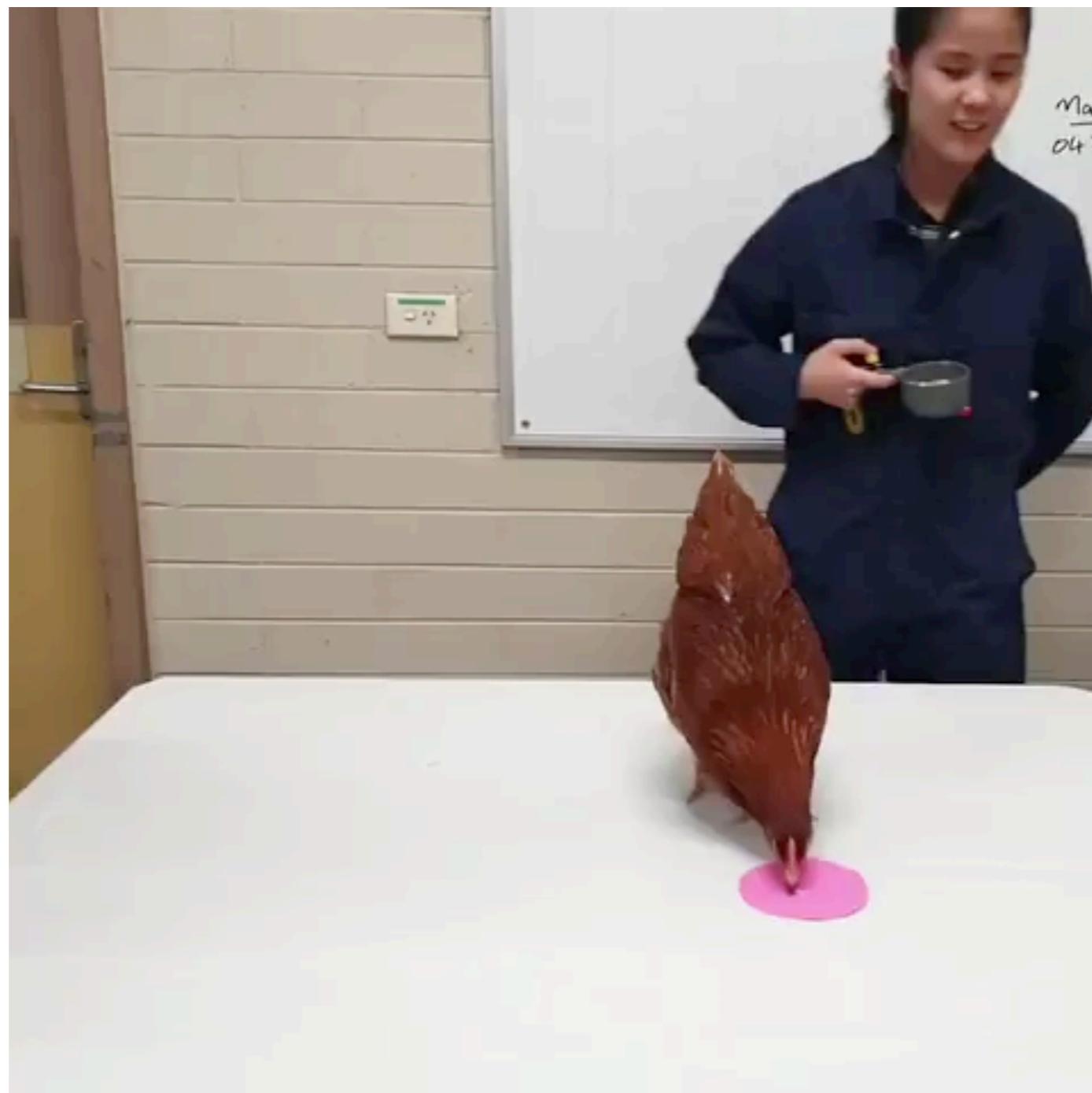
The New York Times

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



Data as infrastructure

- AI system entirely builds on datasets as a key infrastructure
- Datasets embody the capital interests and social relations of the world
- Models reproduce and replicate power structures and inequities

Data as infrastructure

- Data provides the foundation on which we do knowledge work, and models reflect patterns in the data
- Once established, difficult to go beyond it
- Datasets are
 - Contingent on the social conditions of creation
 - Constructed: data is not objective
 - Value-laden: shaped by patterns of inclusion and exclusion

Fundamentally hard examples

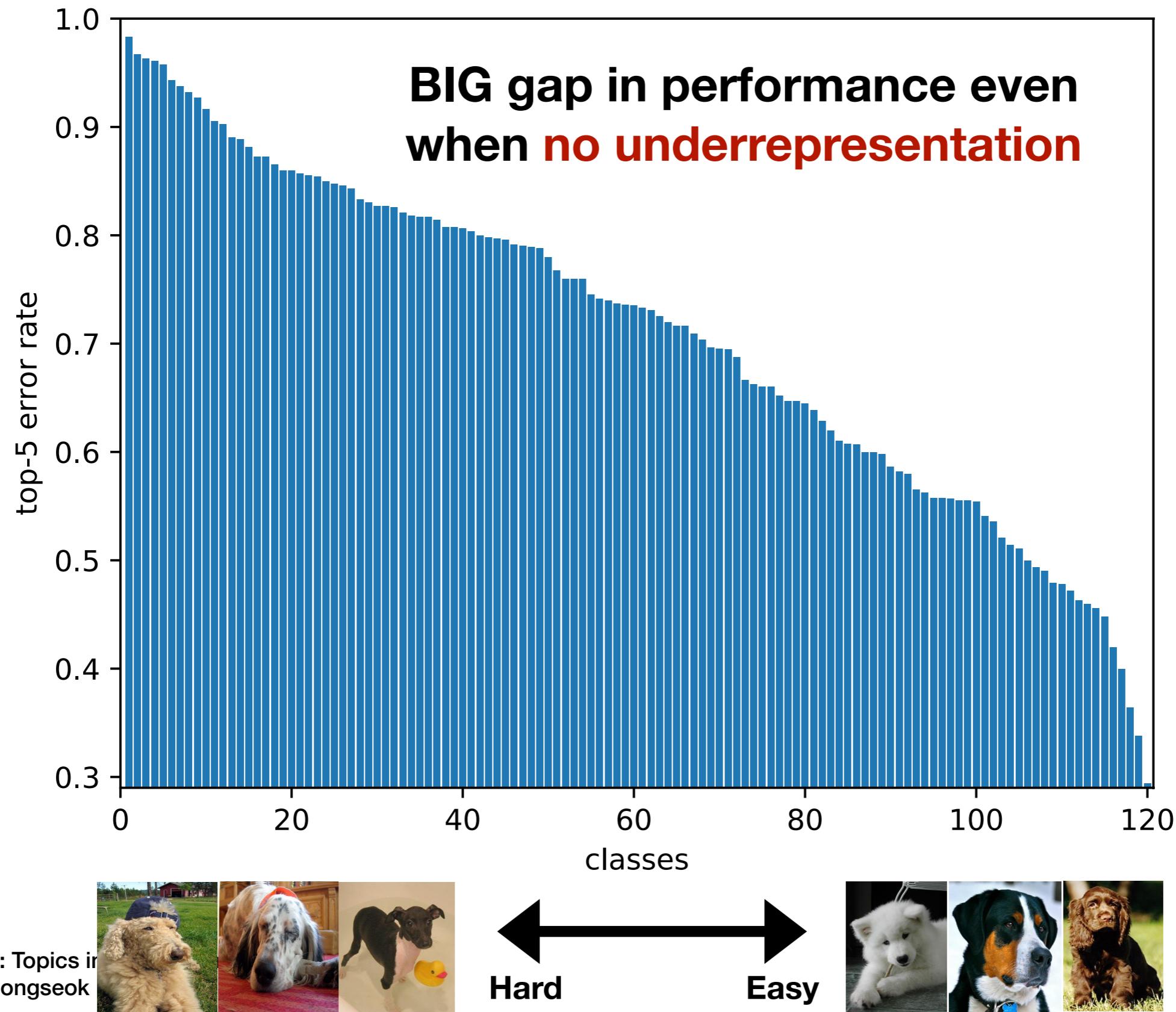
- Task: classify image of dog to breed (120 classes)



Stanford Dogs Dataset [Khosla et al. '11]

No underrepresentation:
same number of images per class

Big gaps in performance



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%

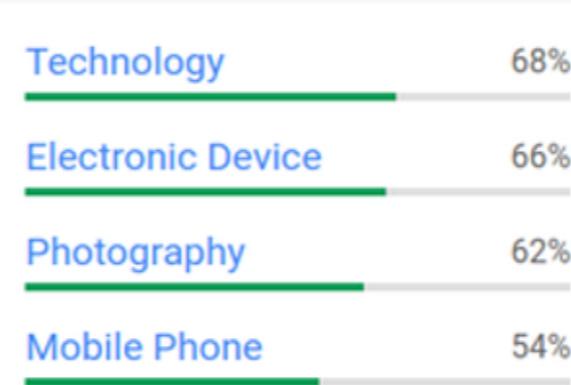


Gendered Shades: Intersectional accuracy disparity
[Buolamwini and Gebru '18]

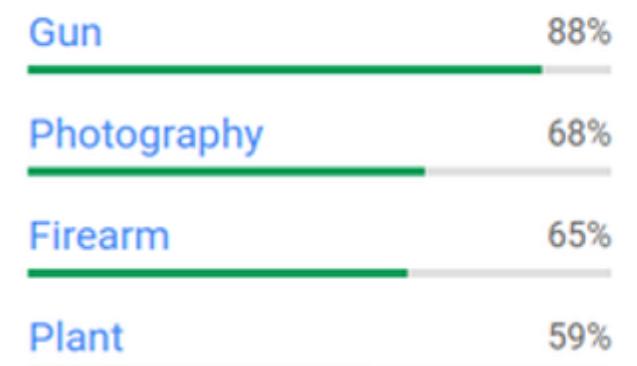
Object recognition



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



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

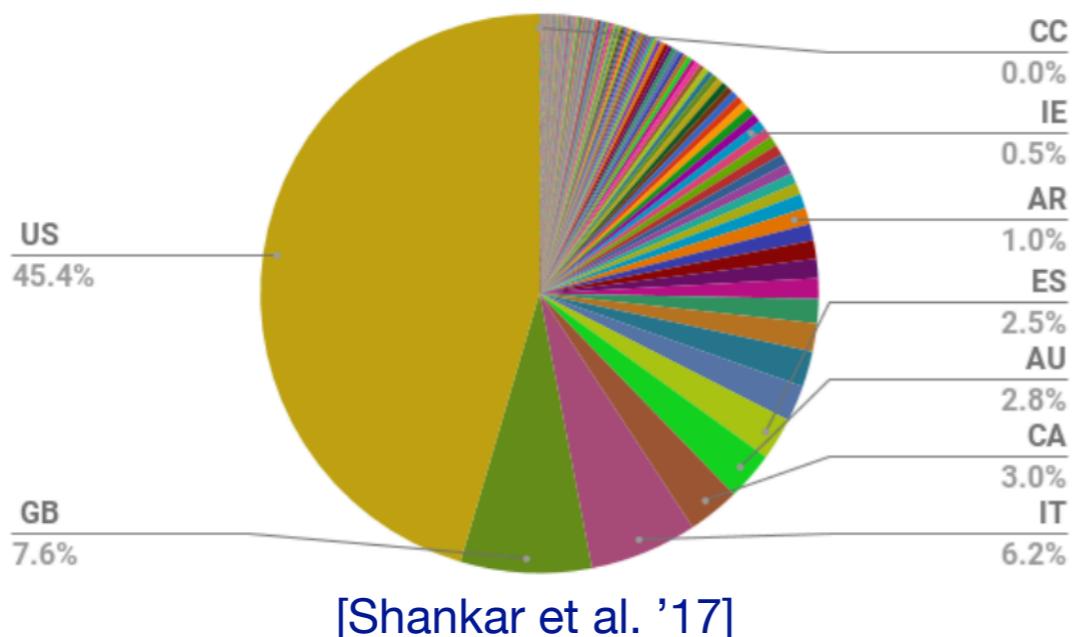


Lack of diversity in data

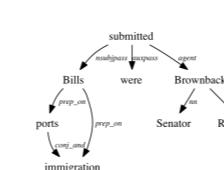
- “Clinical trials for new drugs **skew heavily white**”
 - Less than 5% of cancer trial participants were non-white
- Majority of image data from **US & Western Europe**

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

ImageNet: country of origin



Other examples



Dependency parsing

[Blodgett+ 16]



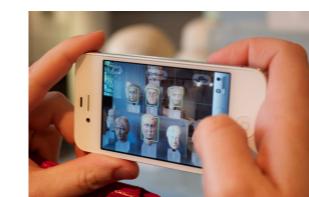
Captioning

[Tatman+ 17]



Recommender systems

[Ekstrand+ 17,18]



Face recognition

[Grother+ 11]

Language Lingaggio Язык
Глóбуса Język لسان بولى
ବ୍ୟାକ୍ ପରିଶ୍ଳପନ ଲାଙ୍ଘାମାଣୀ
Langage Lenguaje ভাষা 言語^{言語}
ମାଧ୍ୟମ ଜ୍ଞାନ Ngón ngùn لغة
Linguagem Wika ବାଦୀ ମାନ୍ୟ Bahasa
Sprache ภาษา 언어

Language identification

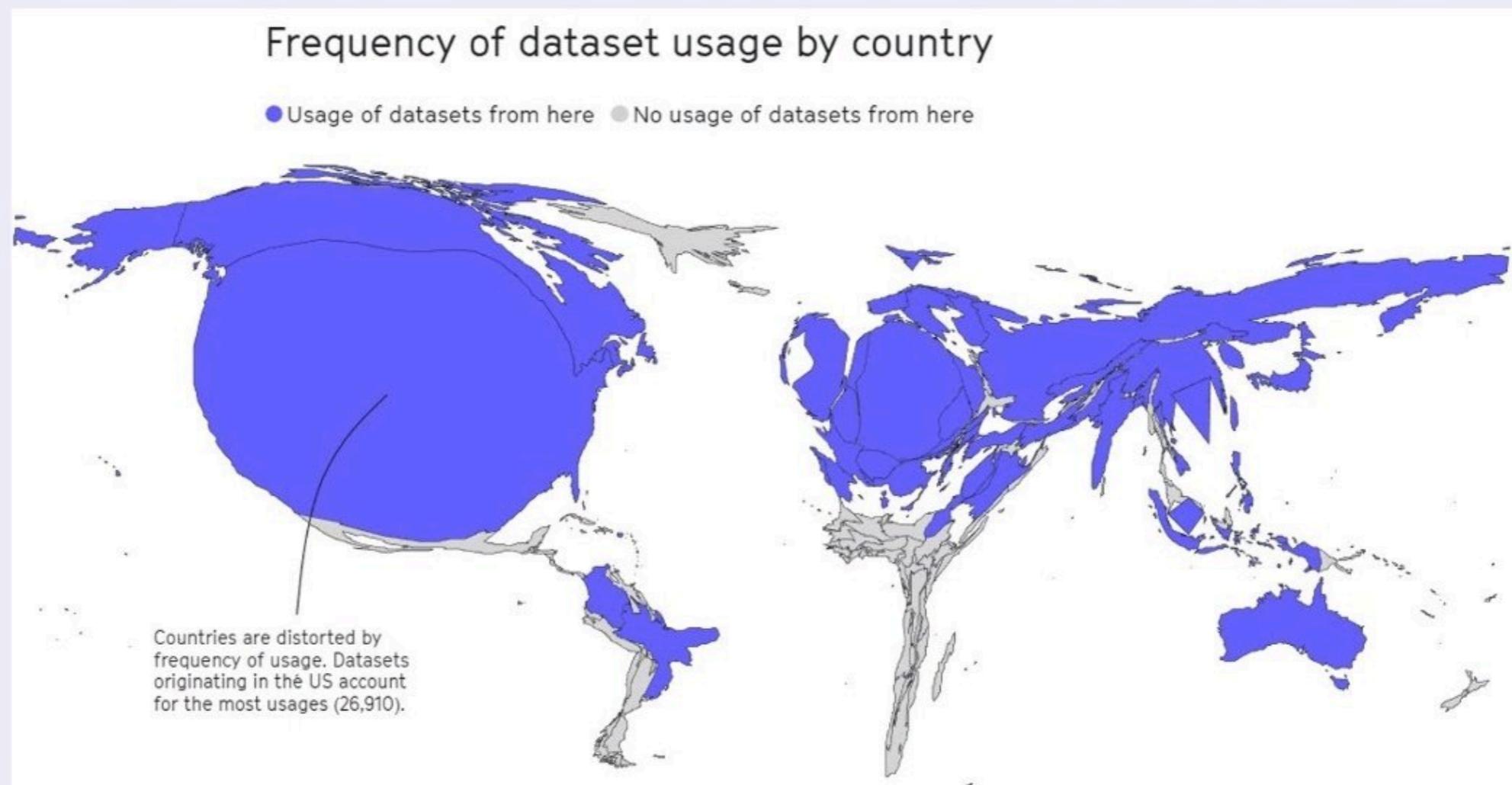
[Blodgett+ 16, Jurgens +17]



Part-of-speech tagging

[Hovy+ 15]

The World Map according to the data AI sees

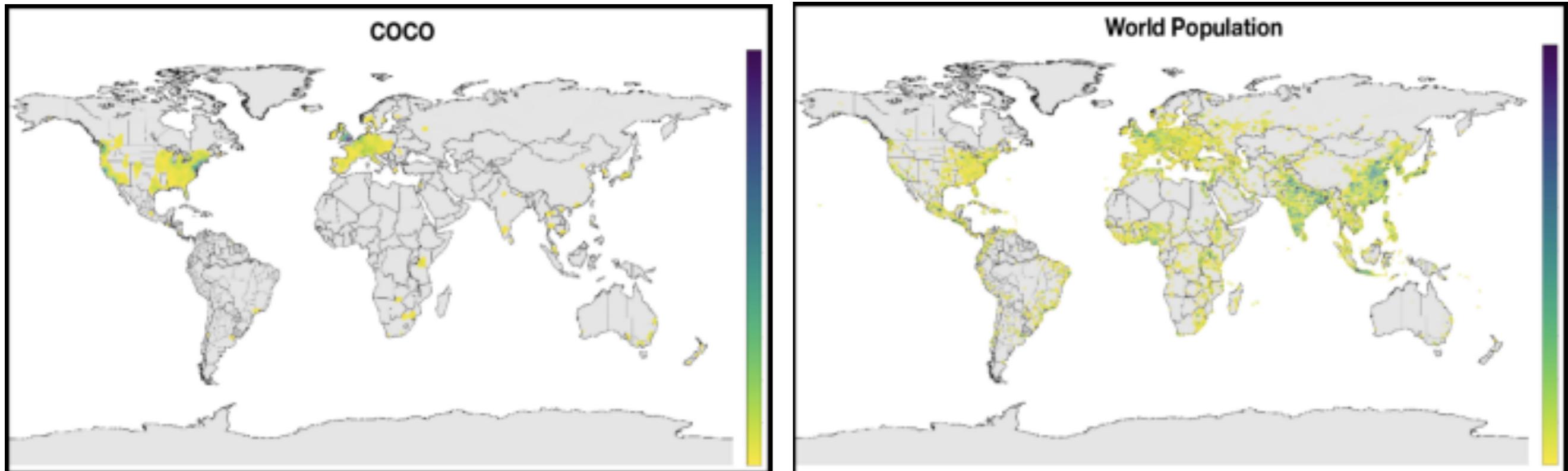


Sources

Research by: [Koch, Denton, Hanna, and Foster \(2021\)](#)

Visual by: [The Mozilla Internet Health Report 2022](#)

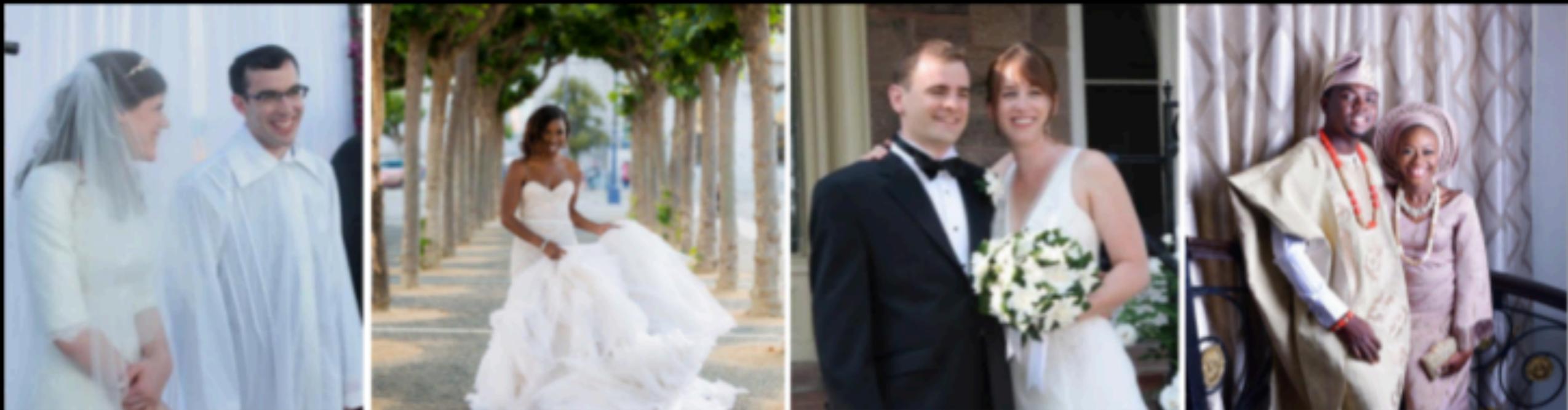
Lack of diversity in data



[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
		onu seviyor
		onu görüyor
		onu göremiyor
		o onu kucaklıyor
		o onu kucaklamıyor
		o evli
		o bekar
		o mutlu
		o mutsuz
		o çalışkan
		o tembel

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.



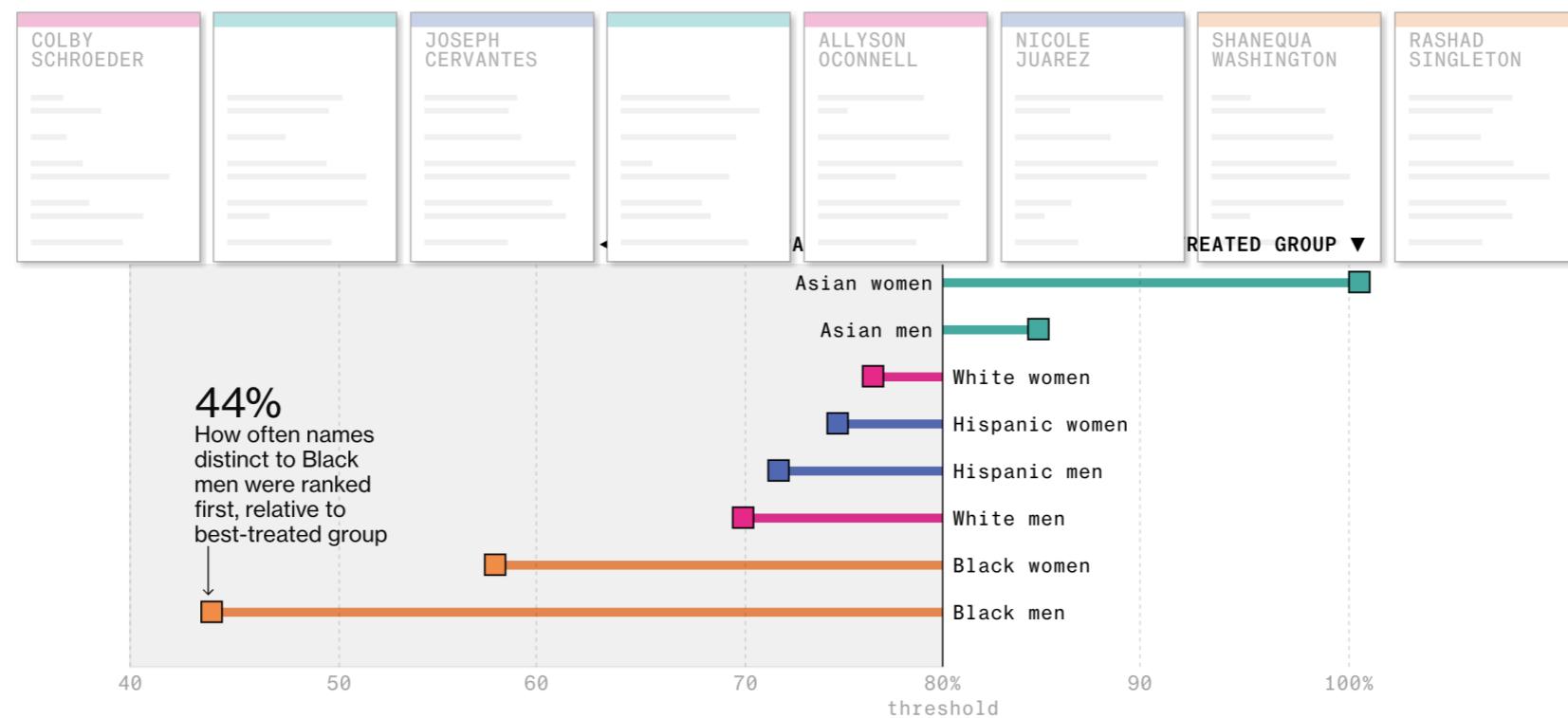
BY EDMUND L. ANDREWS



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.

OPENAI'S GPT IS A RECRUITER'S DREAM TOOL. TESTS SHOW THERE'S RACIAL BIAS

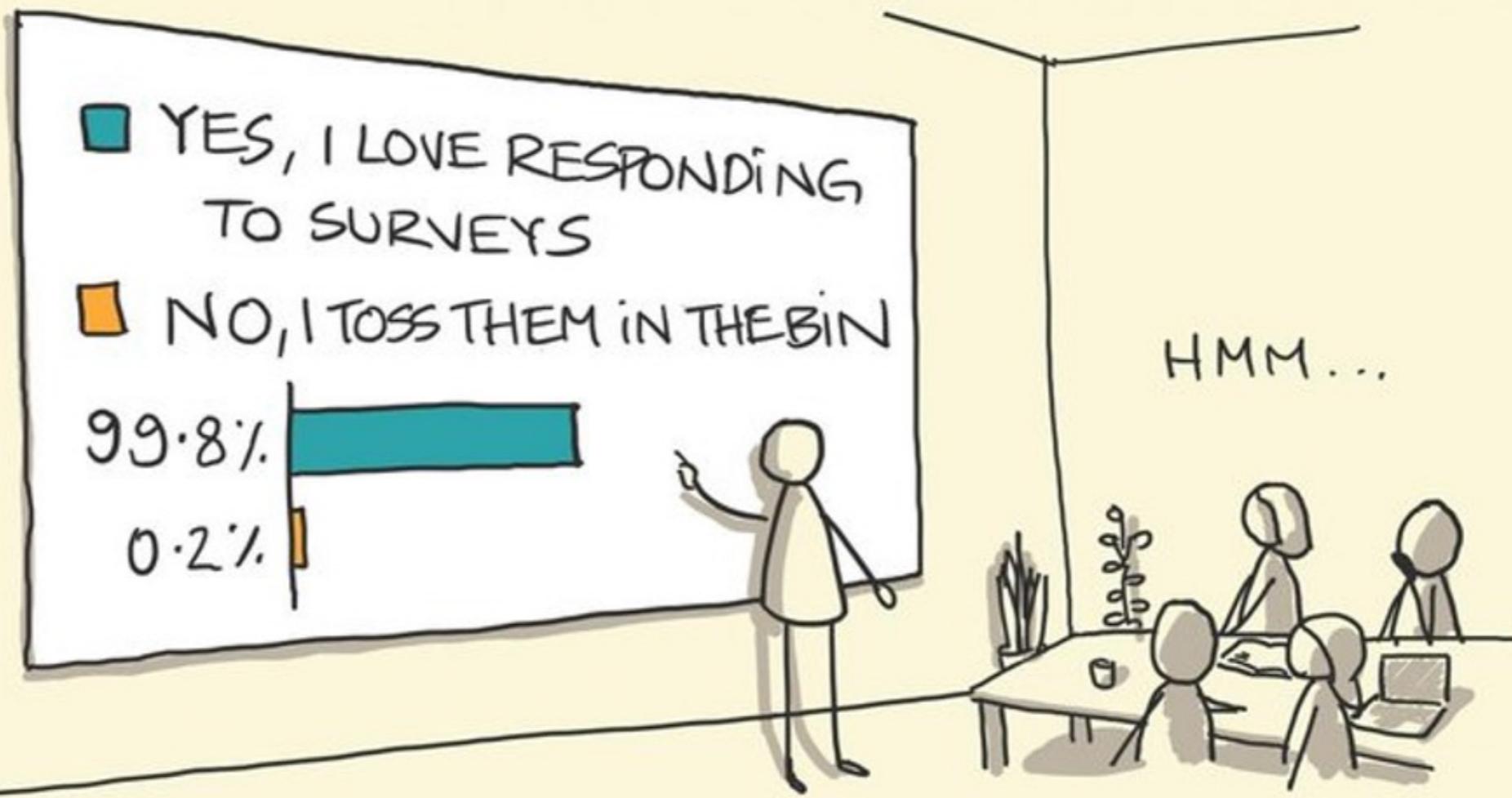
Recruiters are eager to use generative AI, but a Bloomberg experiment found bias against job candidates based on their names alone



By [Leon Yin](#), [Davey Alba](#) and [Leonardo Nicoletti](#) for **Bloomberg Technology + Equality**
March 8, 2024



SAMPLING BIAS



"WE RECEIVED 500 RESPONSES AND
FOUND THAT PEOPLE LOVE RESPONDING
TO SURVEYS"

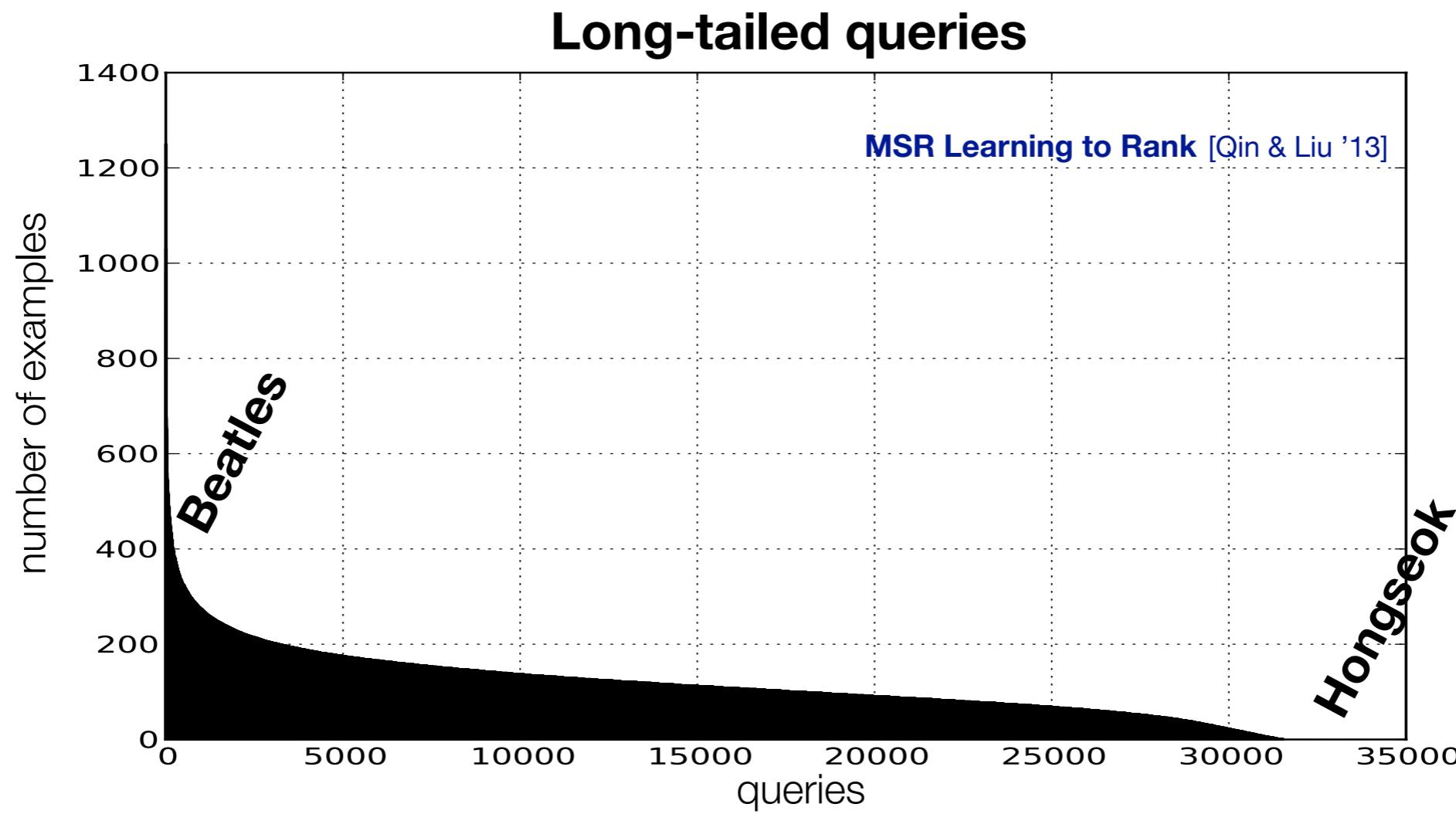
sketchplanations

Omni-present and long-tails

- AI systems are always “on”
- They inevitably encounter inputs that they didn’t see much during training
- When they do, they extrapolate in unexpected ways (“hallucinate”)

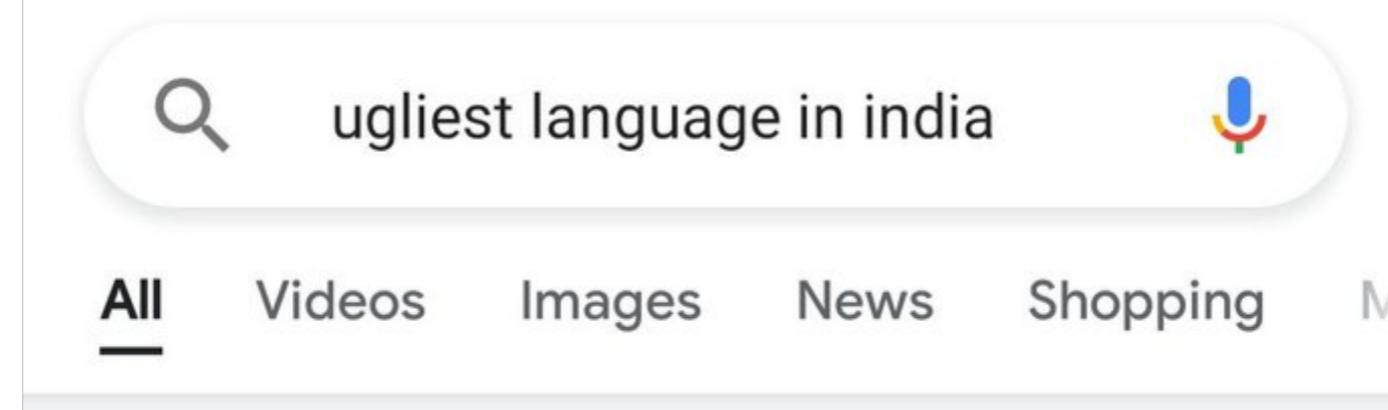
Long-tails

- Long-tailed data is ubiquitous in modern applications
 - Google in 2010s: constant fraction of queries were new each day
- Tail inputs often determine quality of service



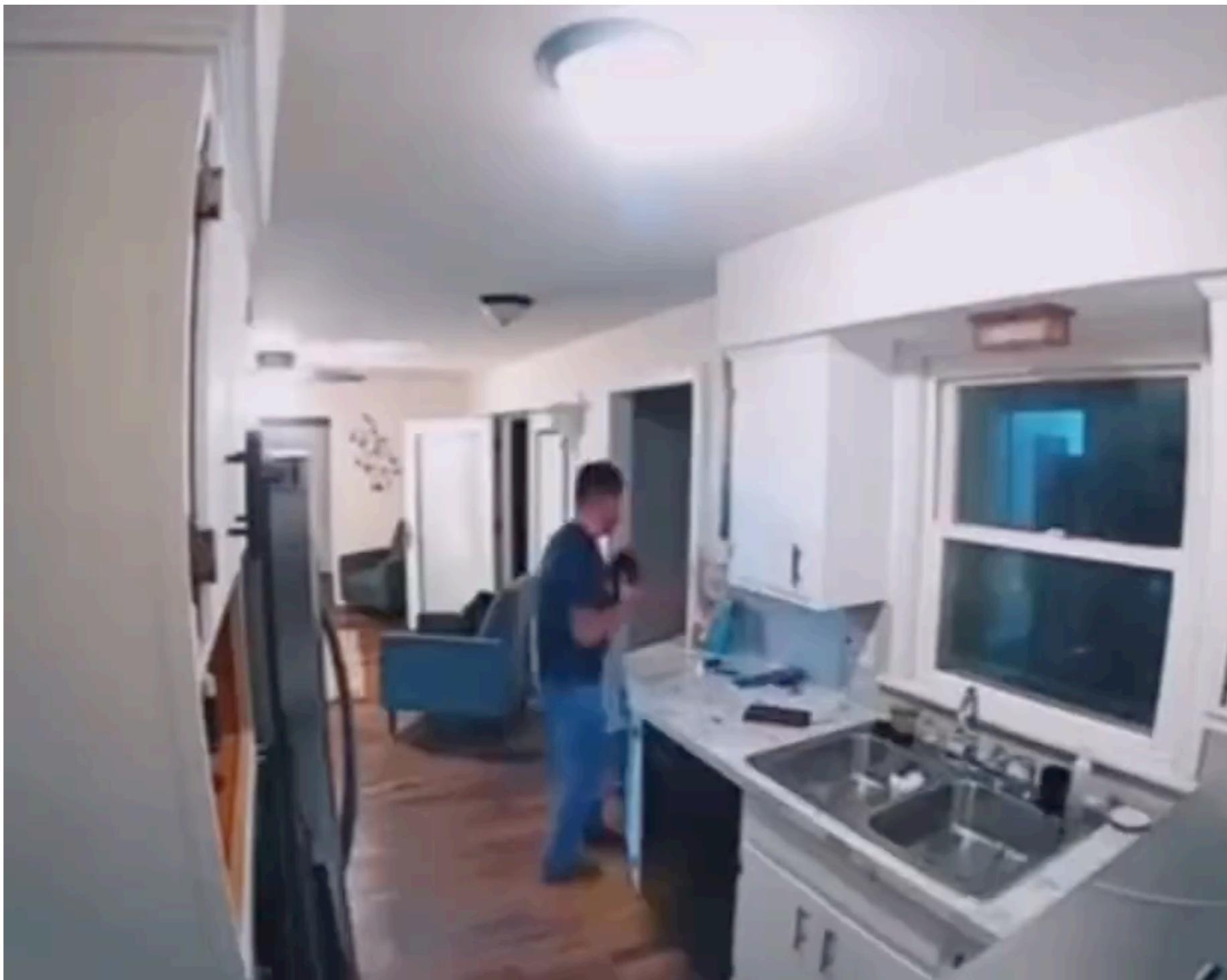
Long-tails

Kannada: Google apologises for 'ugliest Indian language' search result **BBC** © 4 June 2021



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

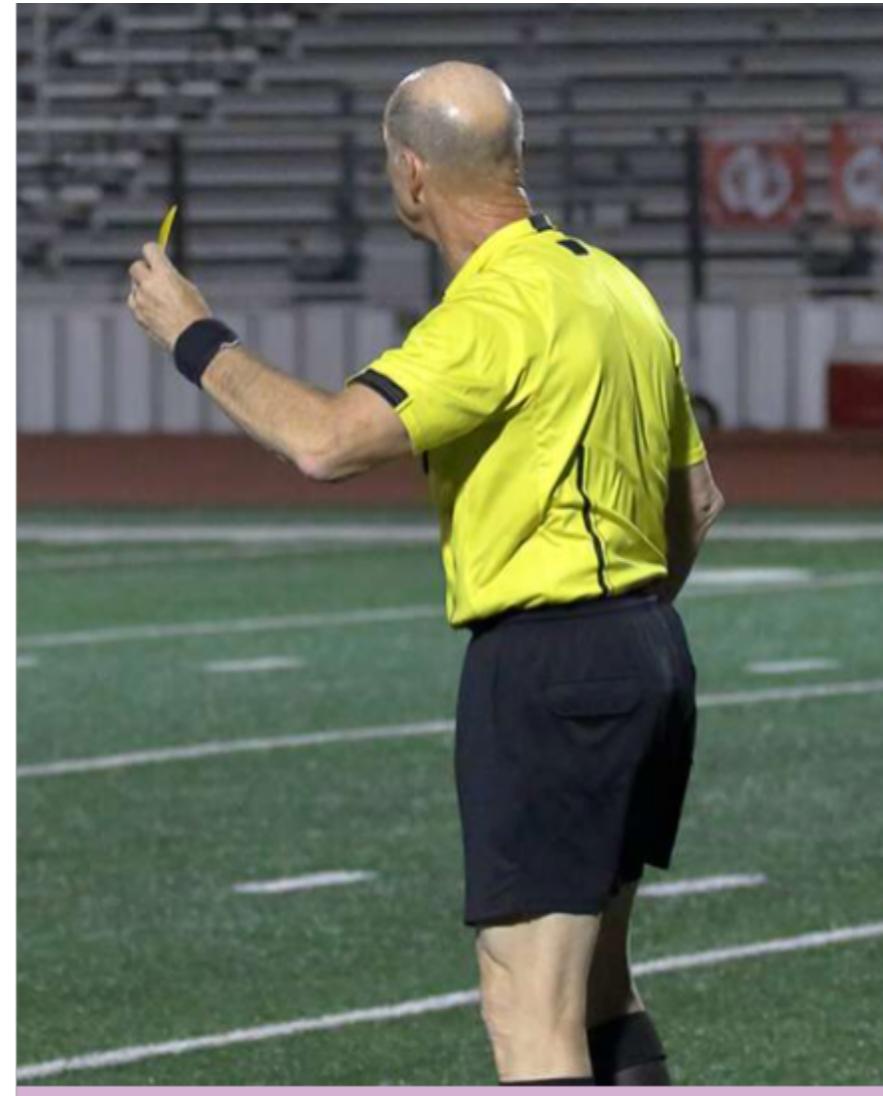
Alexa tells 10-year-old girl to touch live plug with penny

⌚ 28 December 2021

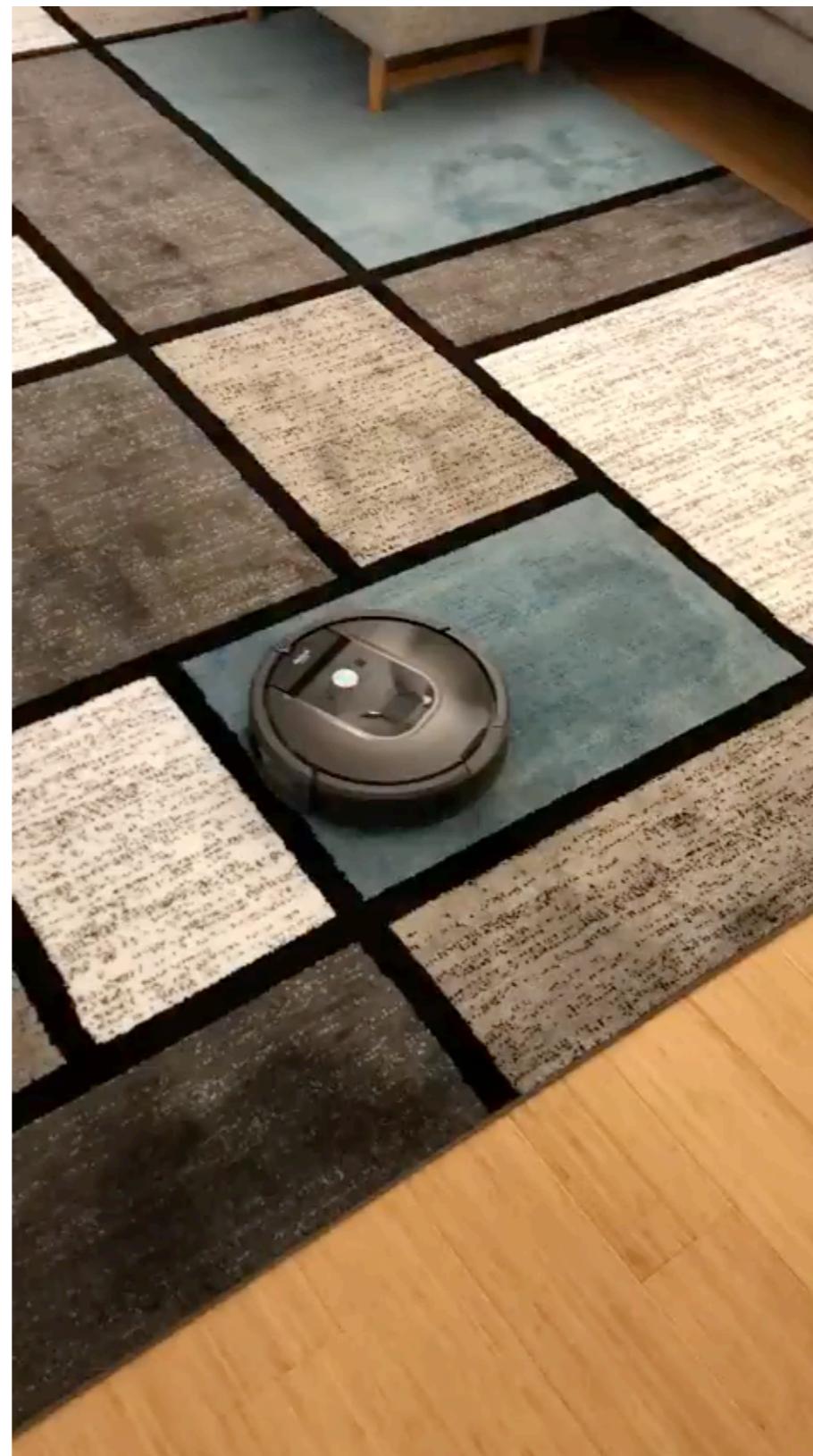
A 10 yo asked Alexa for a “challenge to do”. Alexa responded with "Plug in a phone charger about halfway into a wall outlet, then touch a penny to the exposed prongs"

Long-tails

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

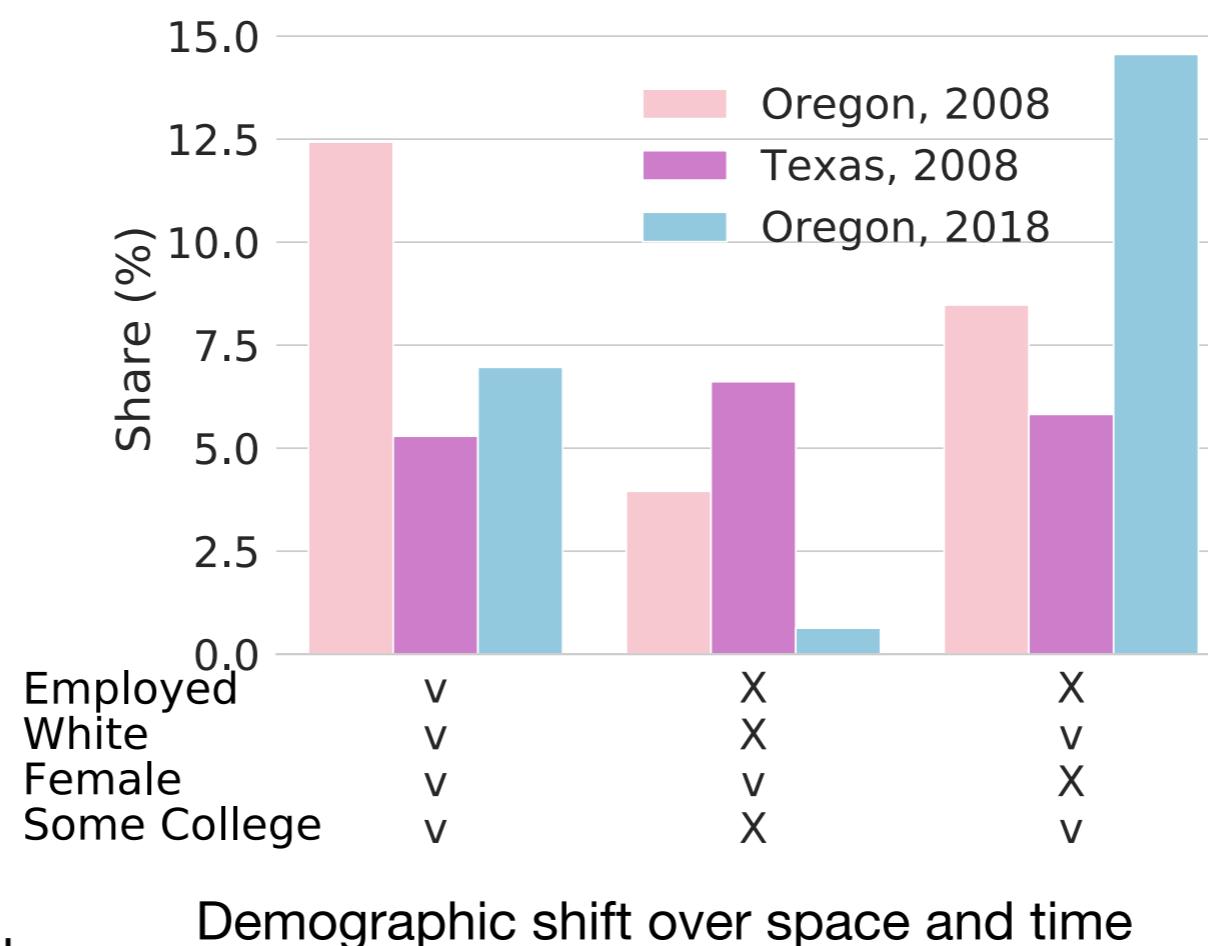


Long-tails



Not a new problem...

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



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)]

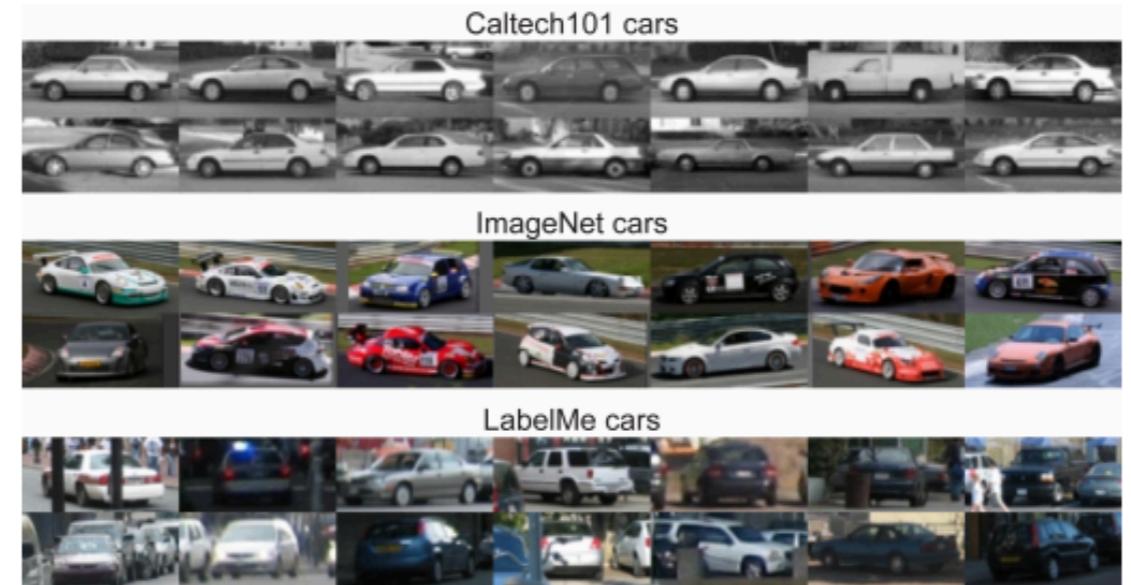
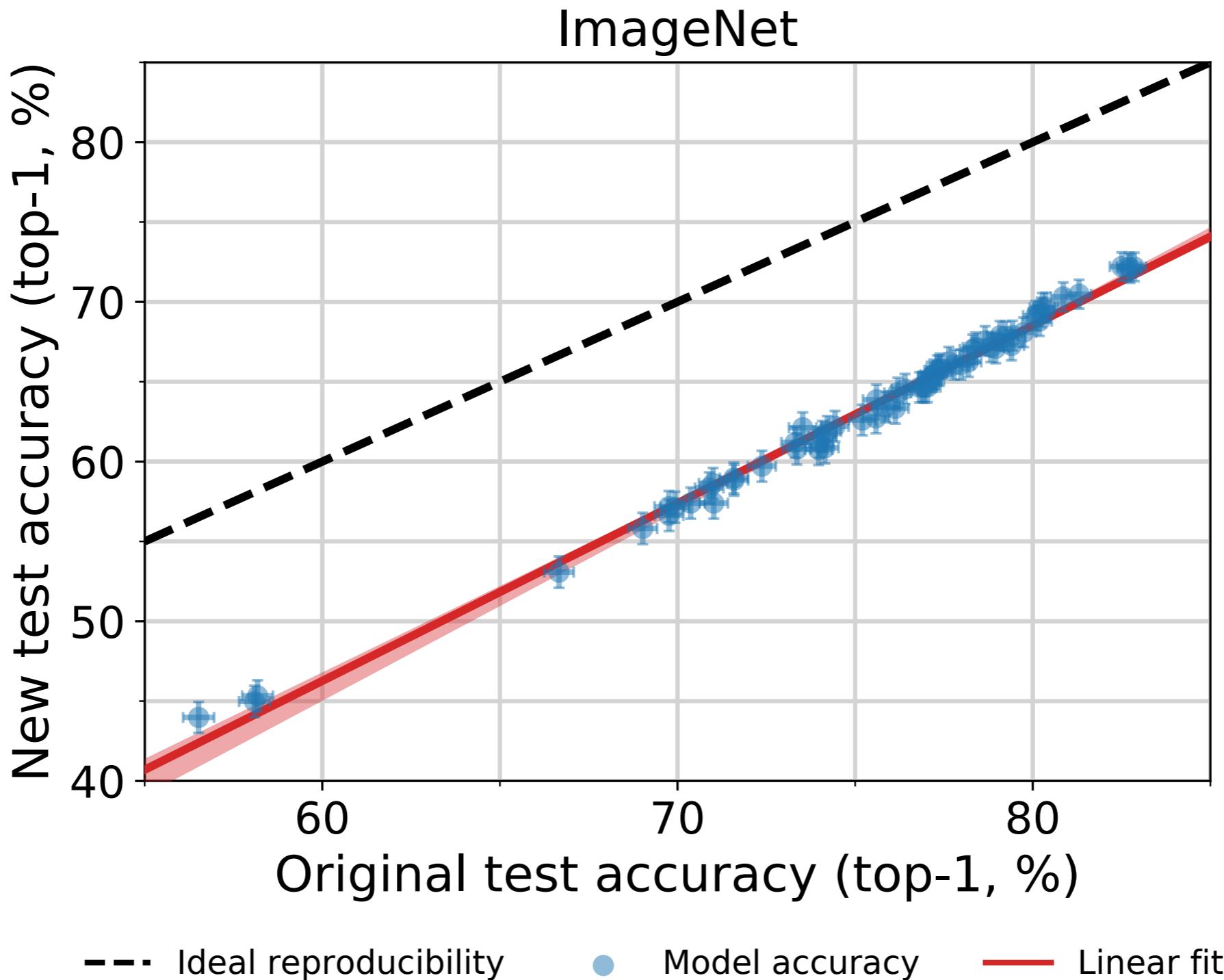


Table 1. Cross-dataset generalization. Object detection and classification performance (AP) for “car” and “person” when training on one dataset (rows) and testing on another (columns), i.e. each row is: training on one dataset and testing on all the others. “Self” refers to training and testing on the 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

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

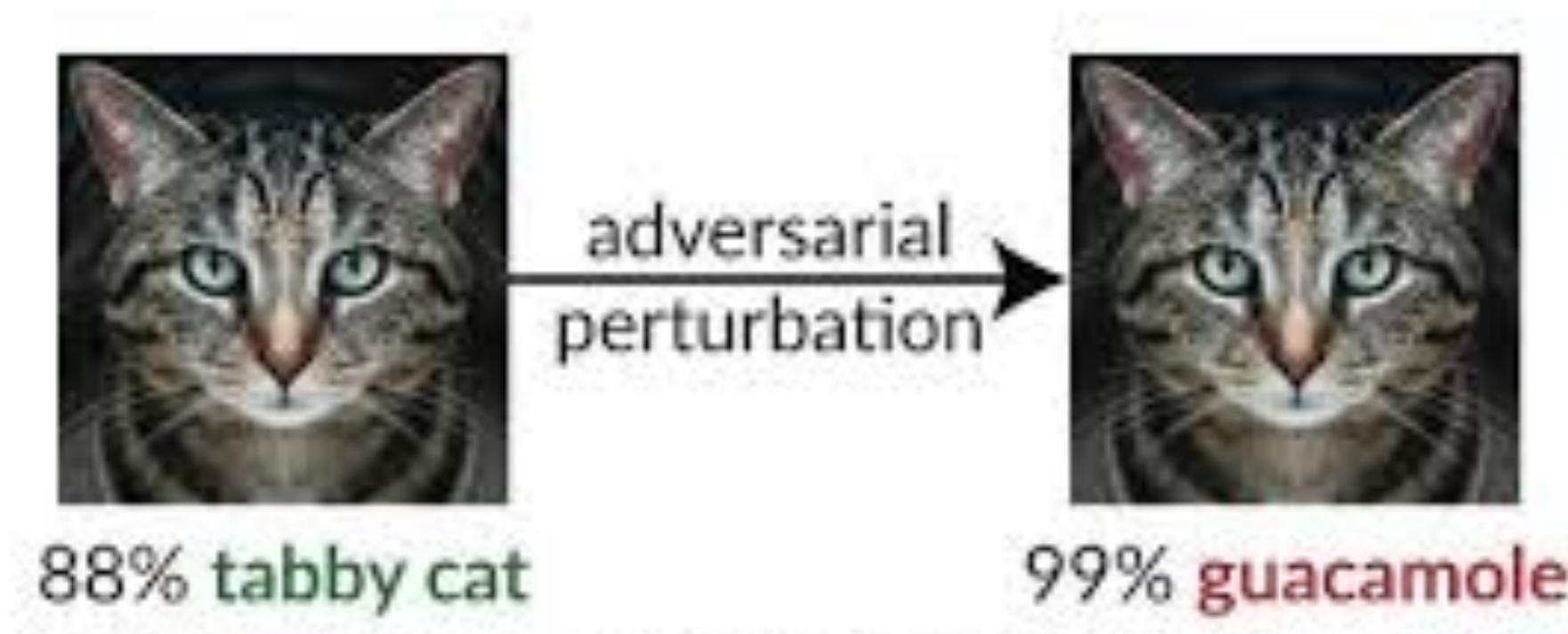
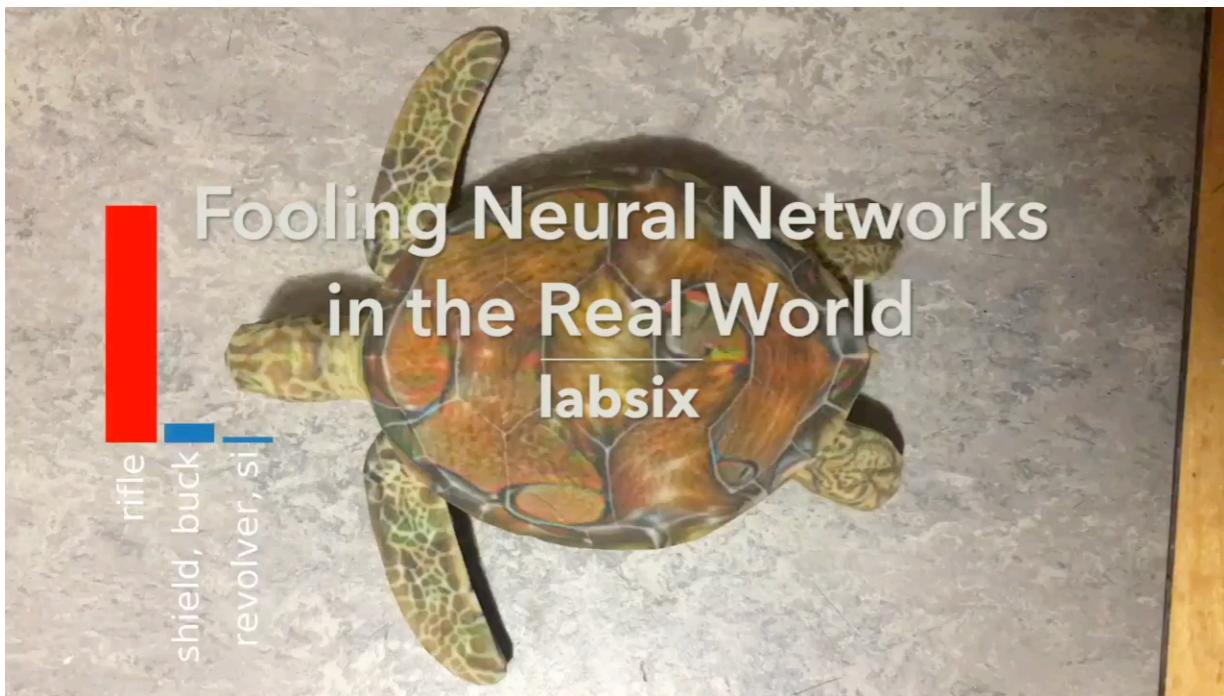


Figure credit: Nicholas Carlini

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 may 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.

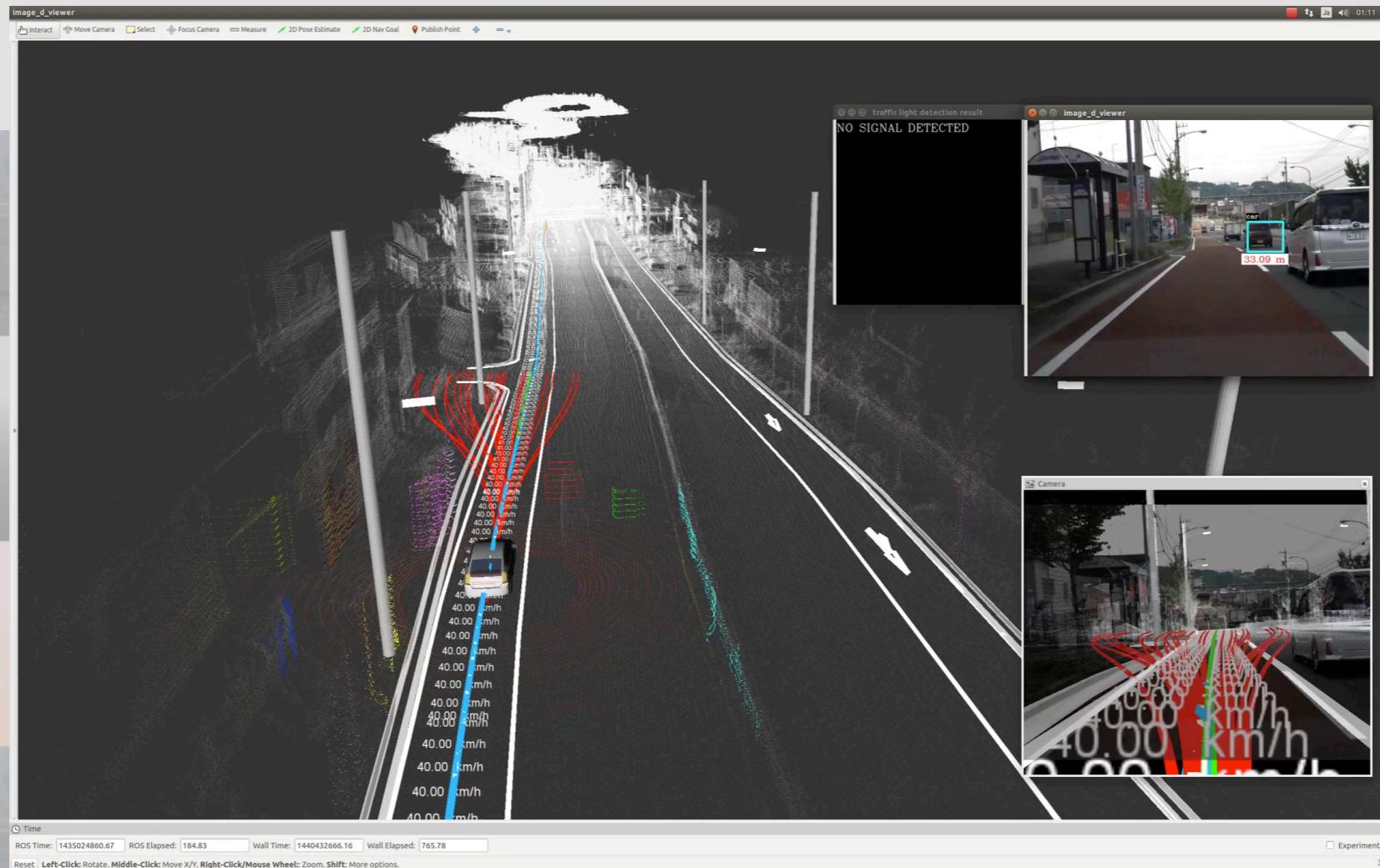
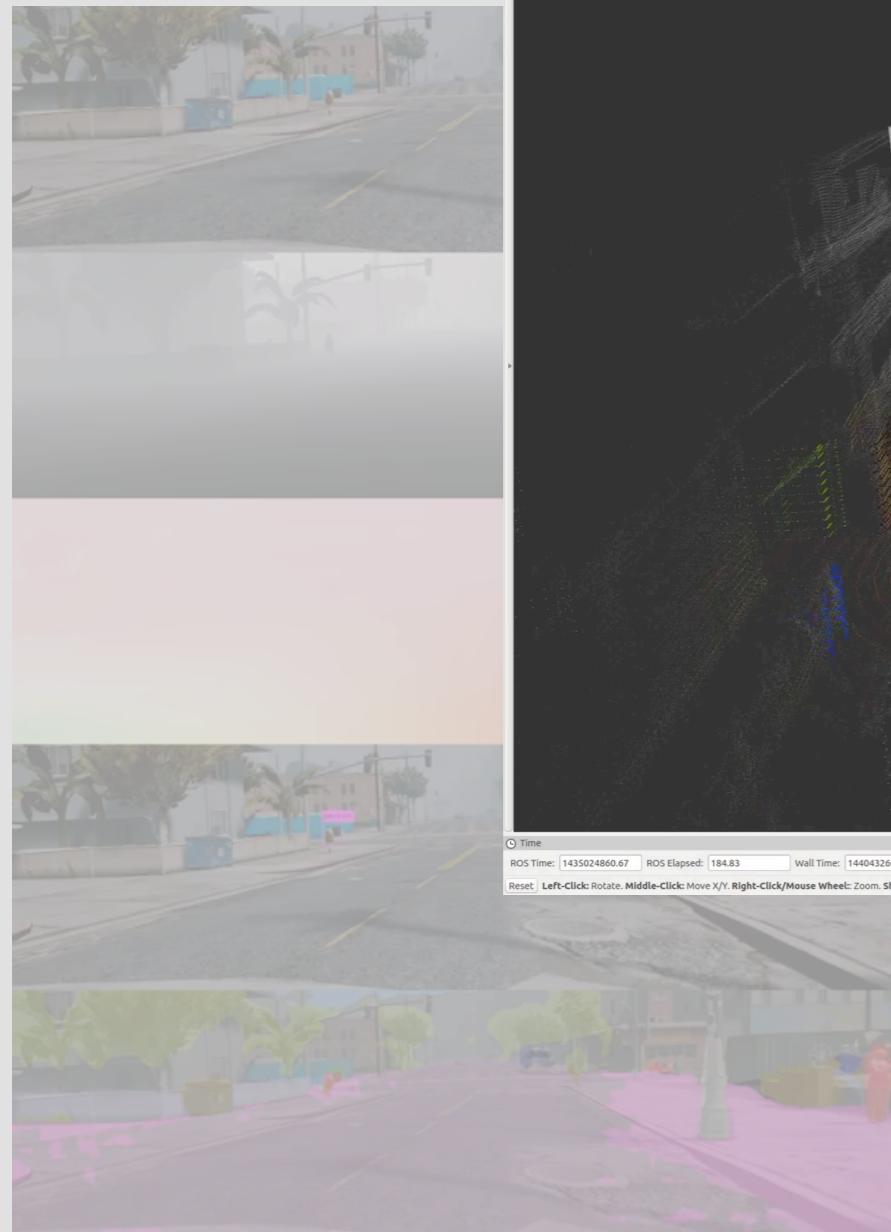
Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

By Will Douglas Heaven

July 30, 2021

Complex system example: AVs

Sense



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

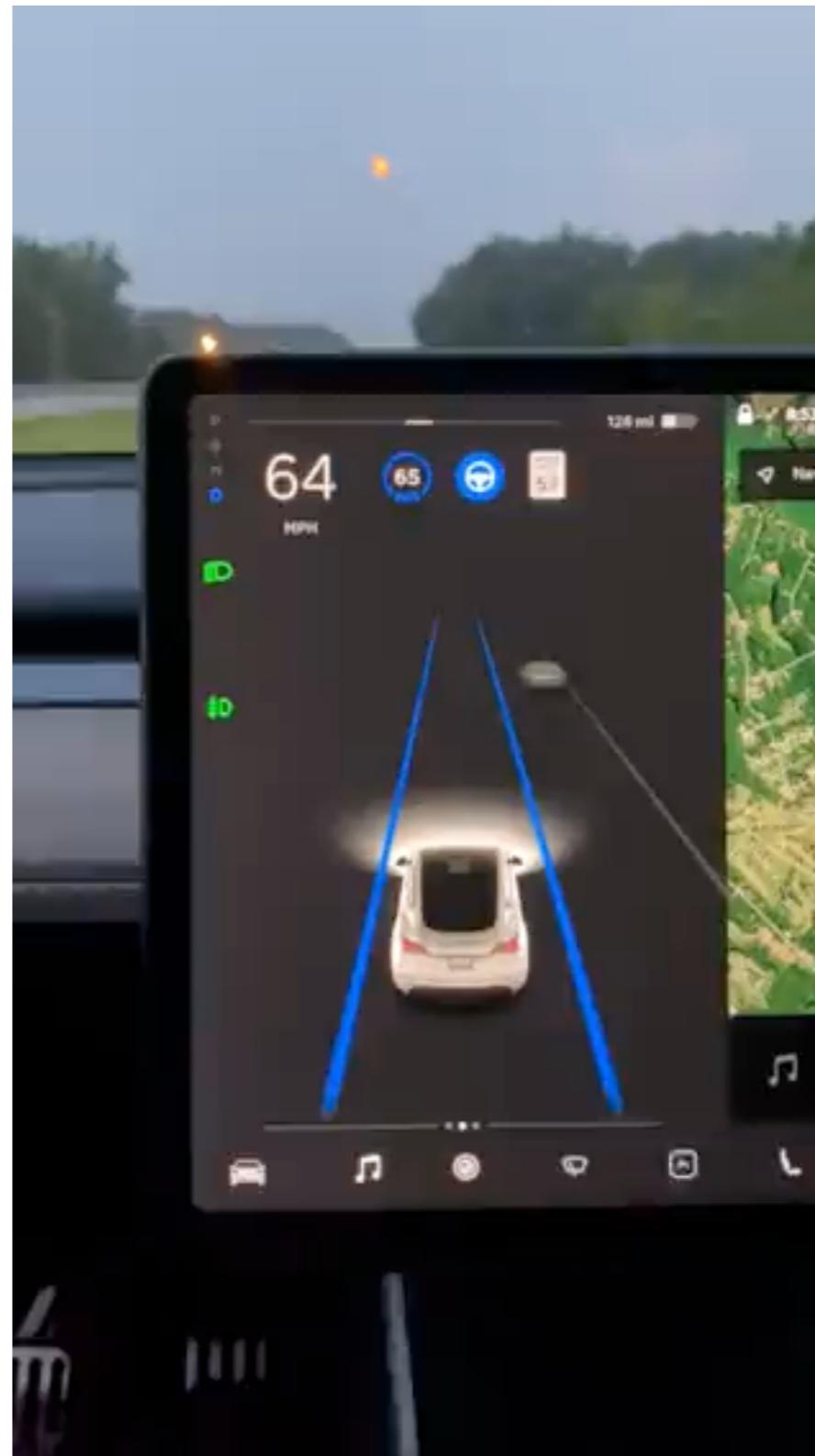
Tesla

- 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!

