Logistics

- Course outline
- HyFlex
- Zoom etiquette
- 3 problem sets & course project
  - grading comprises 50% psets + 50% project
- Office hours: Wed 4-5pm on Zoom
- TA: Chao Qin

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

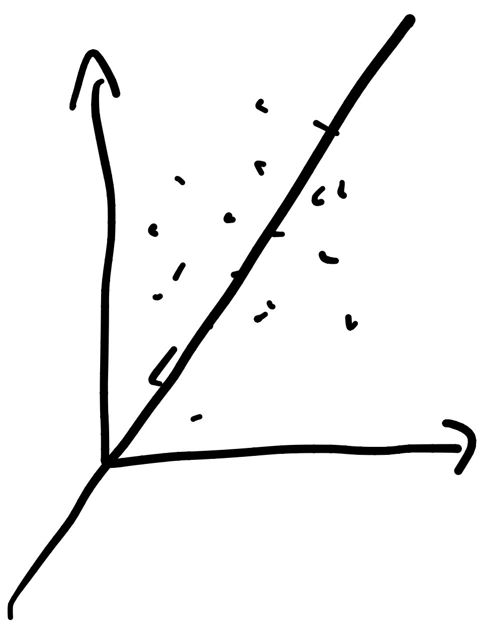
- Loss min. abstraction includes almost all canonical supervised learning problems

- Foundational framework in OR, statistics, and ML

Linear regression

$Z = (X, Y)$ $\quad$ Y: outcome $\quad$ X: covariate vector $\in \mathbb{R}^d$

$$\ell(\theta; X, Y) = (Y - \theta^T X)^2$$



If $\mathbb{E}XX^T \succ 0$, $\quad \theta^* = \arg\min_\theta \mathbb{E}\,\ell(\theta; X, Y)$

$$= (\mathbb{E}XX^T)^{-1}\mathbb{E}YX.$$

Robust regression : $\quad \ell(\theta; X, Y) = |Y - \theta^T X|.$

Maximum likelihood estimation

Likelihood model $\quad p_\theta(z)$

$$\min_{\theta \in \Theta} \; -\mathbb{E} \log p_\theta(z)$$

Conditional likelihood model $\quad p_\theta(y|x)$

$$\min_{\theta \in \Theta} \; -\mathbb{E} \log p_\theta(y|x)$$

Binary classification

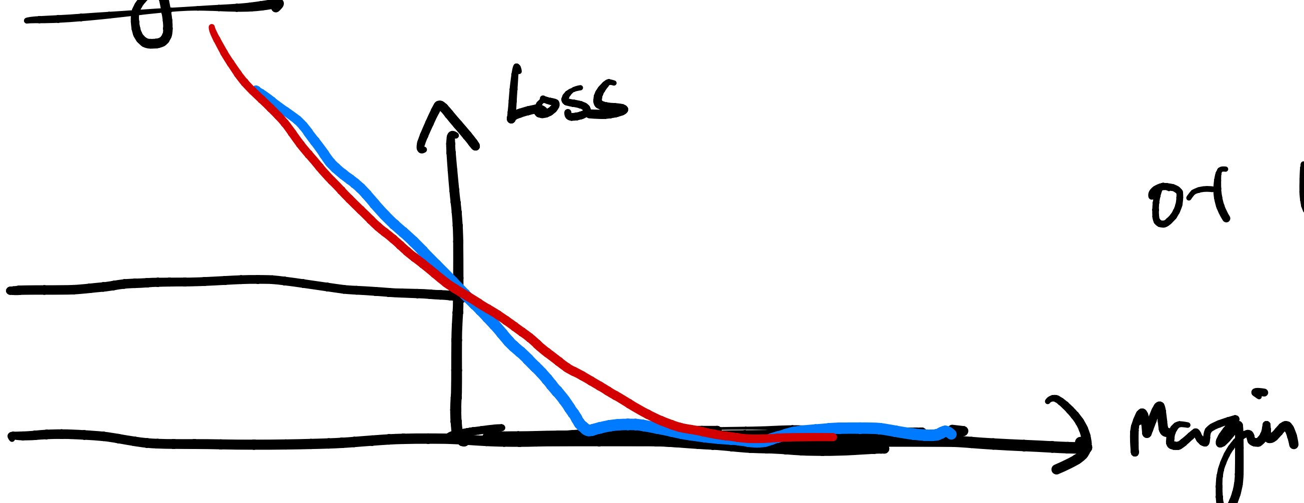
$Z = (X, Y)$ $\quad Y \in \{-1, 1\}$ $\quad X$: features $\in \mathbb{R}^d$

$\{ h_\theta(x) : \theta \in \Theta \}$: hypothesis class

Predict $\quad \mathrm{sgn}(h_\theta(x))$

0-1 loss : $\quad \mathbb{1}\{ \mathrm{sgn}(h_\theta(x)) \neq Y \} = \mathbb{1}\{ h_\theta(x) Y \le 0 \}$

Margin $\quad h_\theta(x) Y \quad$ "how right you are"



0-1 loss is non-smooth & non-convex

$\Rightarrow$ convex surrogates

Binary classification

Hinge loss : $\ell(\theta; X, Y) = (1 - Y h_\theta(X))_+$

Support Vector Machines $h_\theta(x) = \theta^T X$

$$\min_{\theta : \|\theta\|_2 \le r} \mathbb{E}(1 - Y\theta^T X)_+$$

Logistic Loss : $\ell(\theta; X, Y) = \log(1 + \exp(-Y h_\theta(x)))$

Logistic Regression $h_\theta(x) = \theta^T X$

$$\min_{\theta : \|\theta\|_p \le r} \mathbb{E} \log(1 + \exp(-Y h_\theta(x)))$$

$$\Theta = \{\theta : \|\theta\|_p \le r\}$$

Binary classification

Multi-class classification

$\theta = (\theta_1, \cdots, \theta_K) \in \mathbb{R}^{d \times K}$

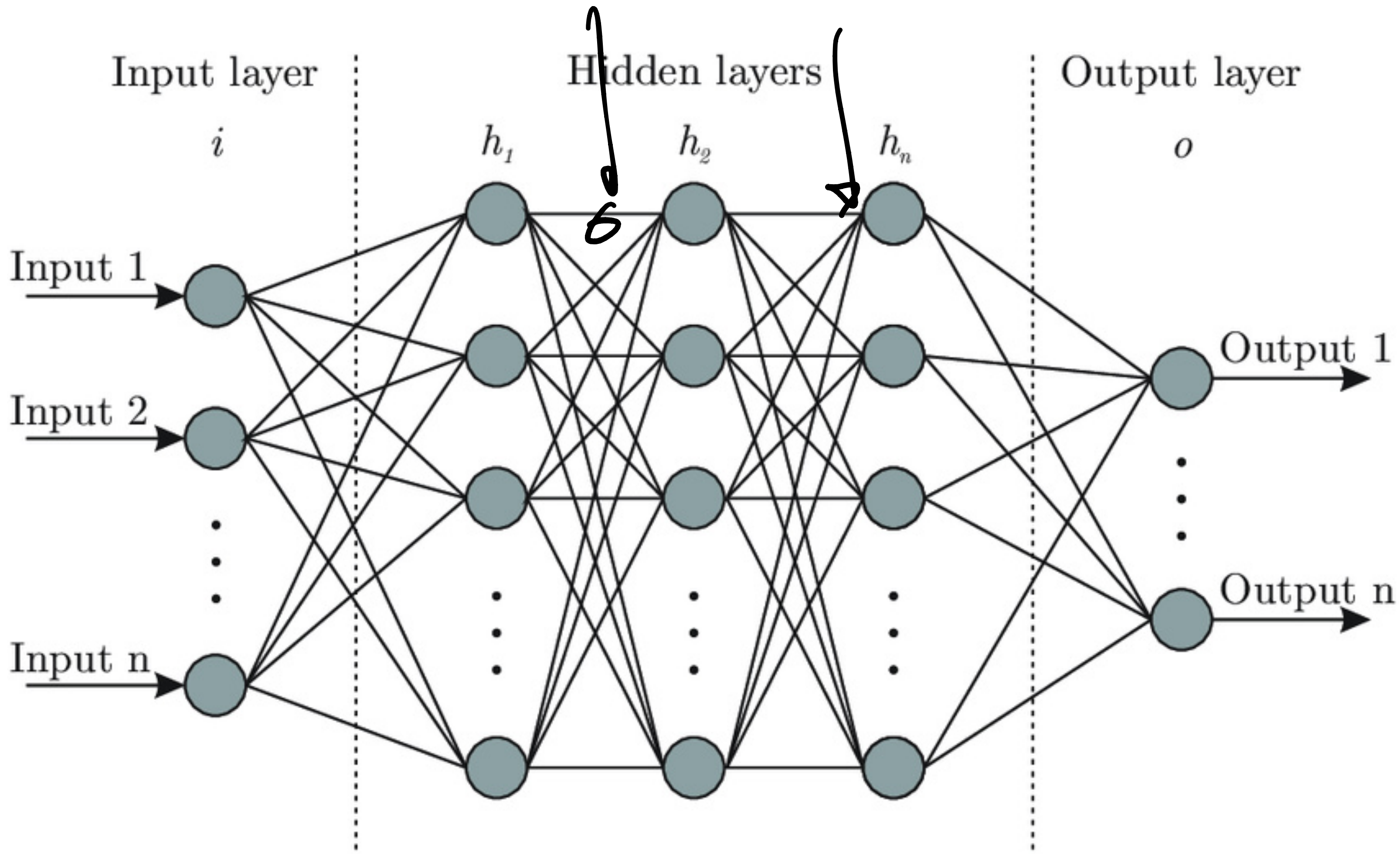
$Y \in \{1, \cdots, K\}$

Logit $\quad p_\theta(y|x) = \dfrac{\exp(\theta_y^T x)}{\sum_{k=1}^K \exp(\theta_k^T x)}$

Max log likelihood $\equiv \min_{\theta \in \Theta} -\mathbb{E} \log p_\theta(Y|X)$

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

Neural networks



instead of $\theta_k^T x$, $\underline{h_{\theta,k}(x)}$

Neural networks

$h_\theta(x) \in \mathbb{R}^K$

$\theta_1, \cdots, \theta_L$: weight matrices

$\sigma_1, \cdots, \sigma_L$: activations $\qquad \sigma_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ $\qquad d_L = K$

$d_0 = \dim(x)$

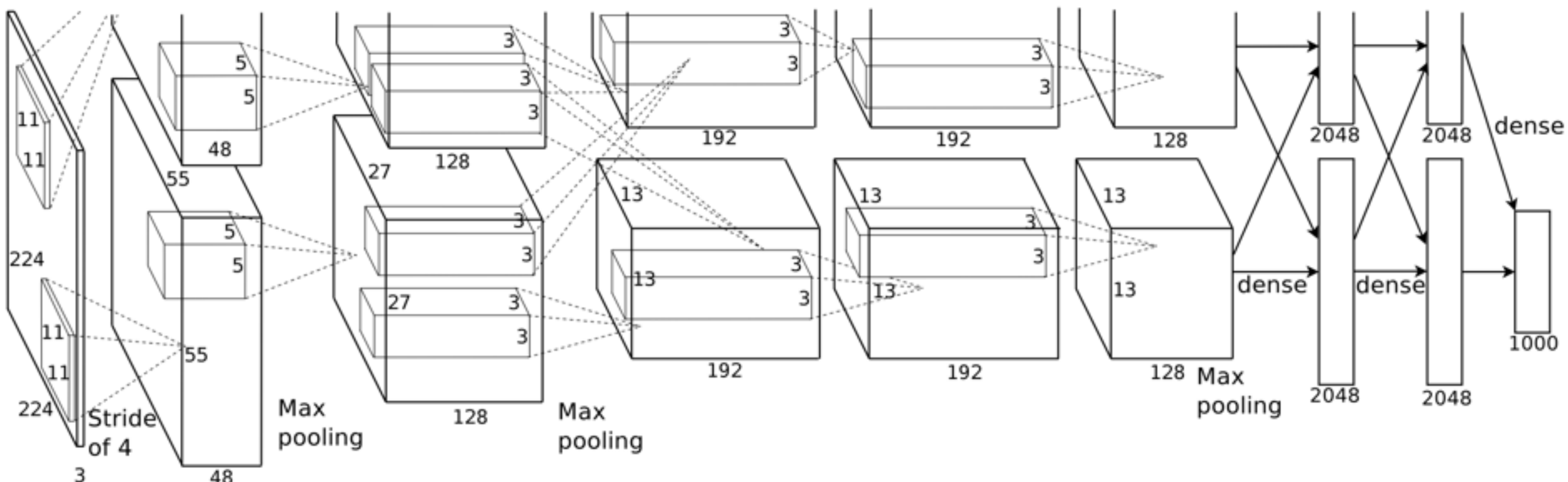
ReLU $\qquad \sigma(x)_j = \max(0, x_j)$

Max pooling : Take a bunch local coordinates, output its maximum

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

Final loss : $\qquad \ell(\theta; X, Y) = -\log \dfrac{\exp(h_{\theta,y}(X))}{\sum_{k=1}^K \exp(h_{\theta,k}(x))}$.

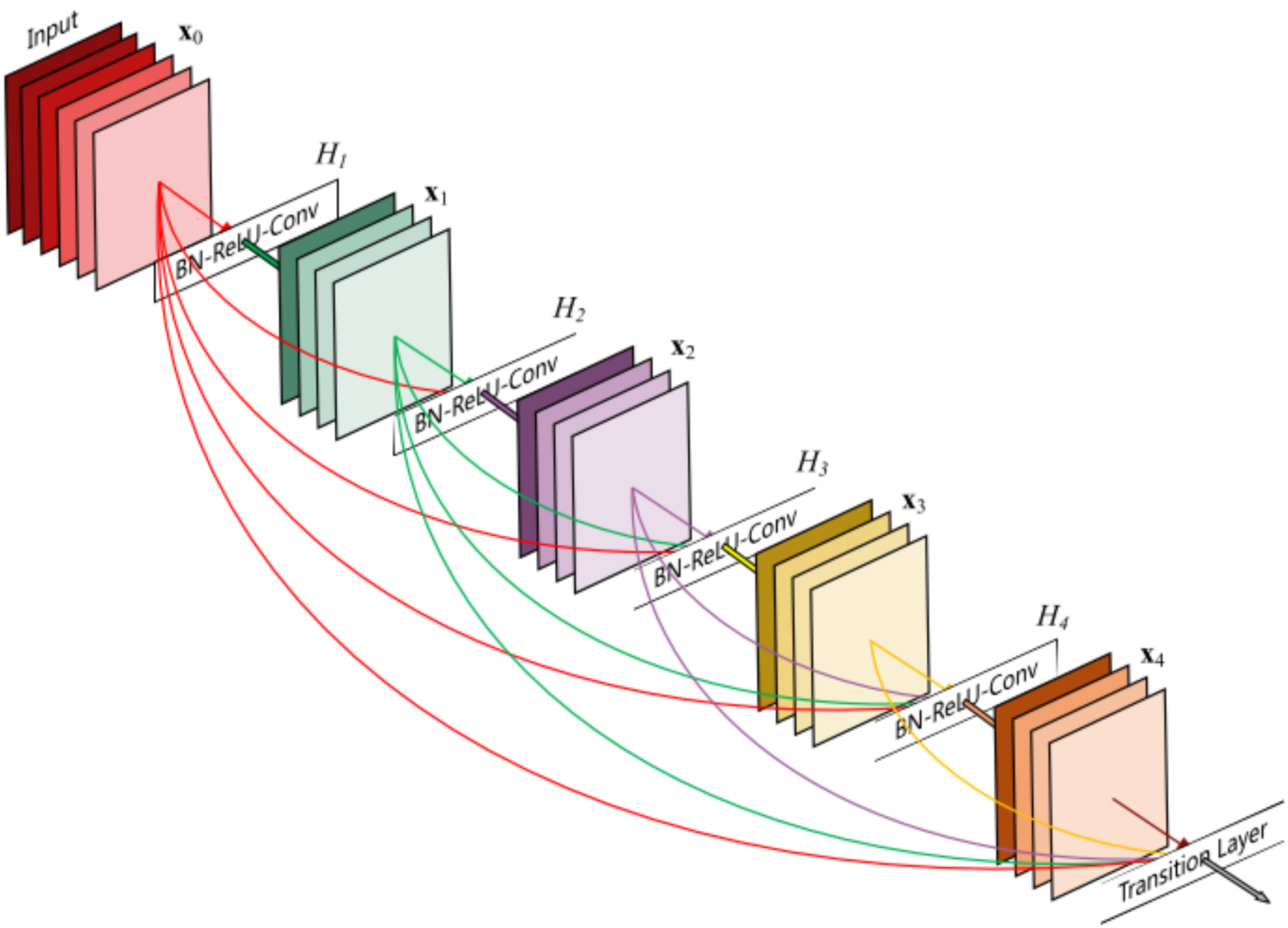
Convolutional nets



Krizhevsky et al. 2012

Residual nets

He et al. 2015



Newsvendor

Consider decision-maker deciding $\theta \in \mathbb{R}_+$: order quantity

Uncertainty Z : random demand

Order cost : c If $Z > \theta$, additional order cost $b \geq 0$

Holding cost : h

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

$\Theta = \mathbb{R}_{\geq 0}$ or $[0, M]$

Portfolio optimization

$\theta \in \mathbb{R}^d_+$: pfo weights

Z : random asset returns $\in \mathbb{R}^d$

$\ell(\theta; Z) = \theta^T Z$

$\Theta = \{\theta \in \mathbb{R}^d_+ : \theta^T \mathbb{1} = 1\}$

$$\begin{aligned} \min \quad & -\mathbb{E}\, \theta^T Z \\ \text{s.t.} \quad & \theta \in \Theta \end{aligned}$$

: risk-neutral

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 $Z_i \stackrel{\text{iid}}{\sim} P$

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

Optimization

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

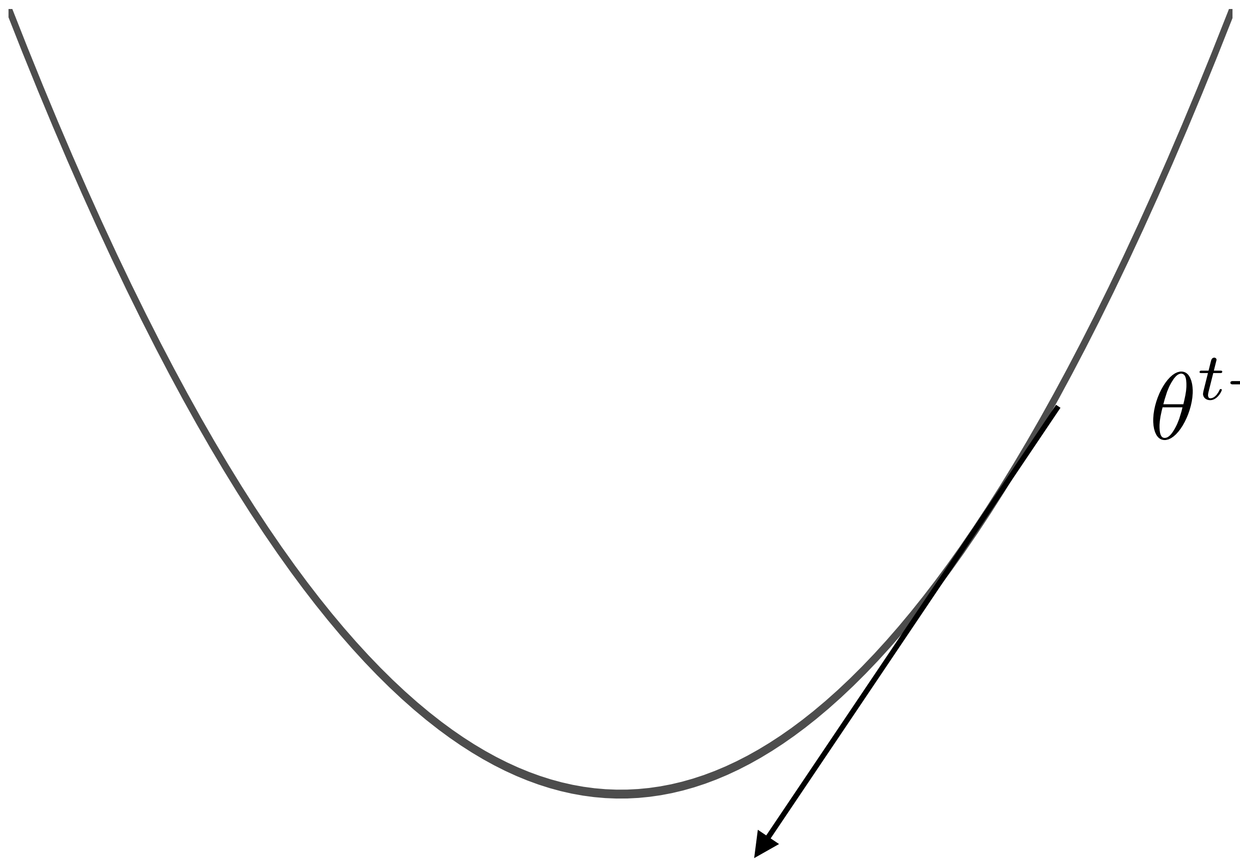
- How do we solve the ERM/SAA problem?
  - Let's say $\theta \mapsto \ell(\theta; Z)$ is convex
  - True for linear models [check for yourself!]

- 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

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

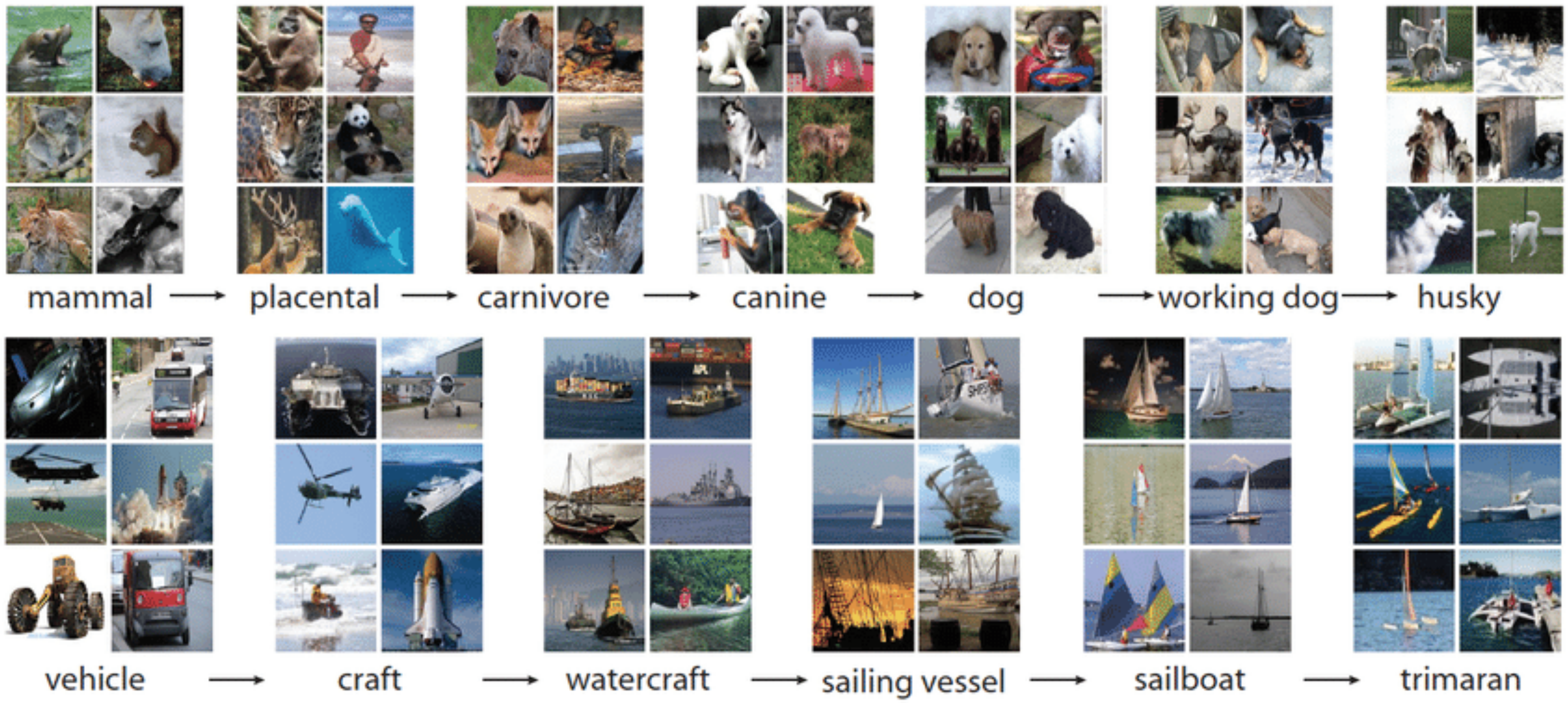


$$\theta^{t+1} \leftarrow \theta^t - \alpha_t \nabla_\theta \ell(\theta^t; Z_t)$$

Magic formula

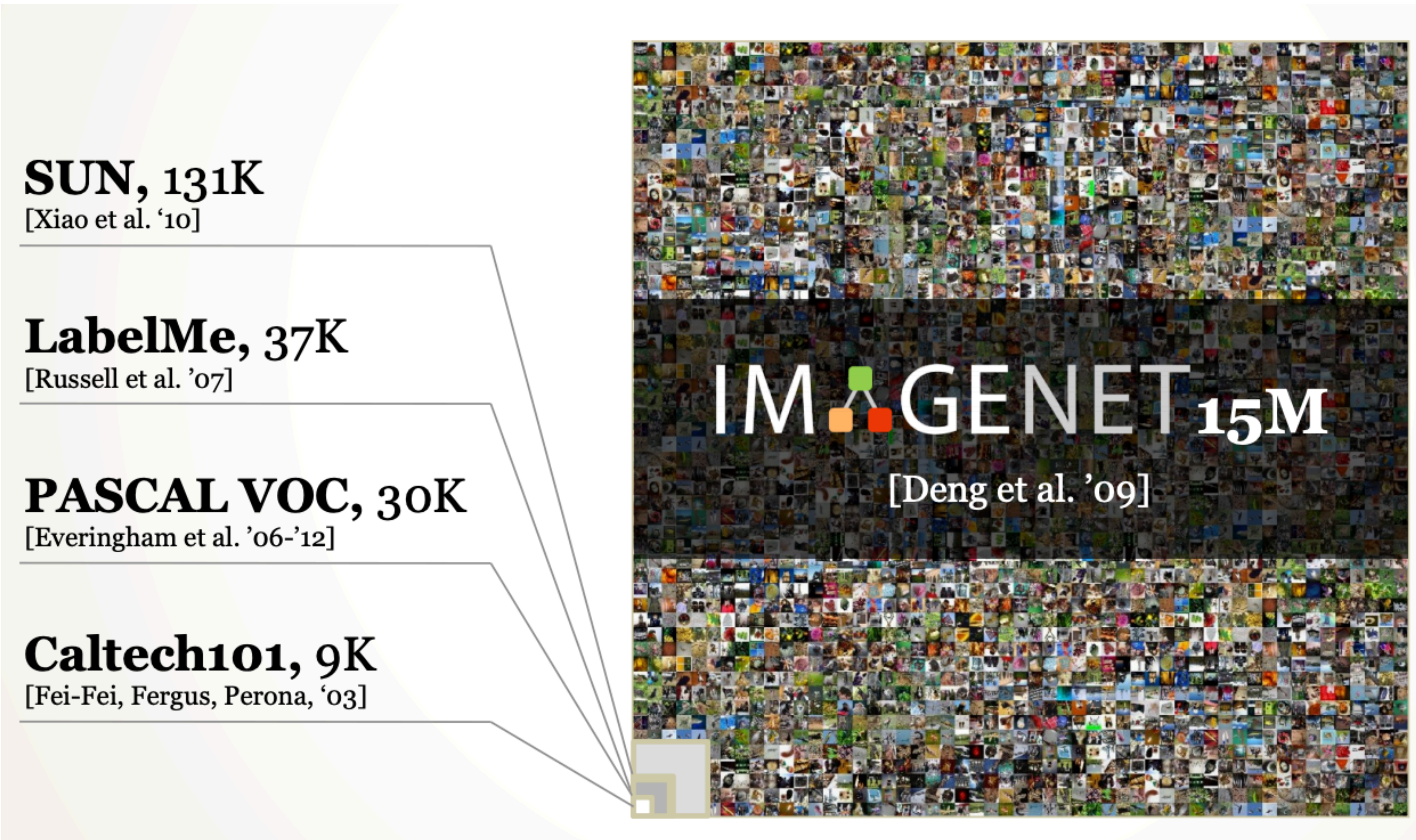
- Inductive bias: CNN, ResNet, RNN, LSTM, attention, transformers
- Big datasets
- Optimize some surrogate loss using SGD
- GPUs

Big datasets: ImageNet



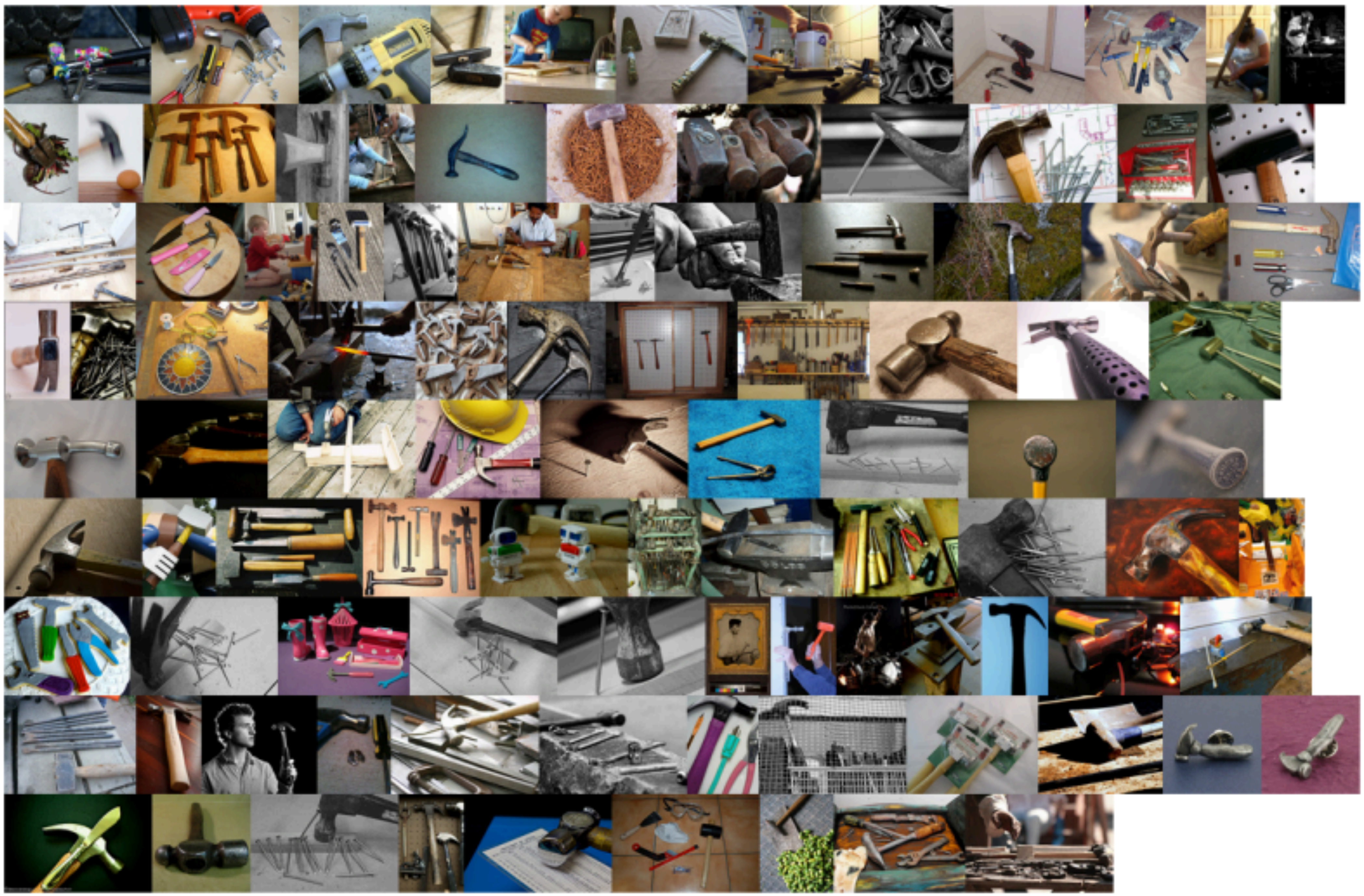
- 2012 classification challenge: 1.3M images, 1000 labels
- Collected through web search, verified via Mechanical Turk
- Hierarchy of labels

Big datasets: ImageNet



Slide from Fei-Fei Li

Hammers

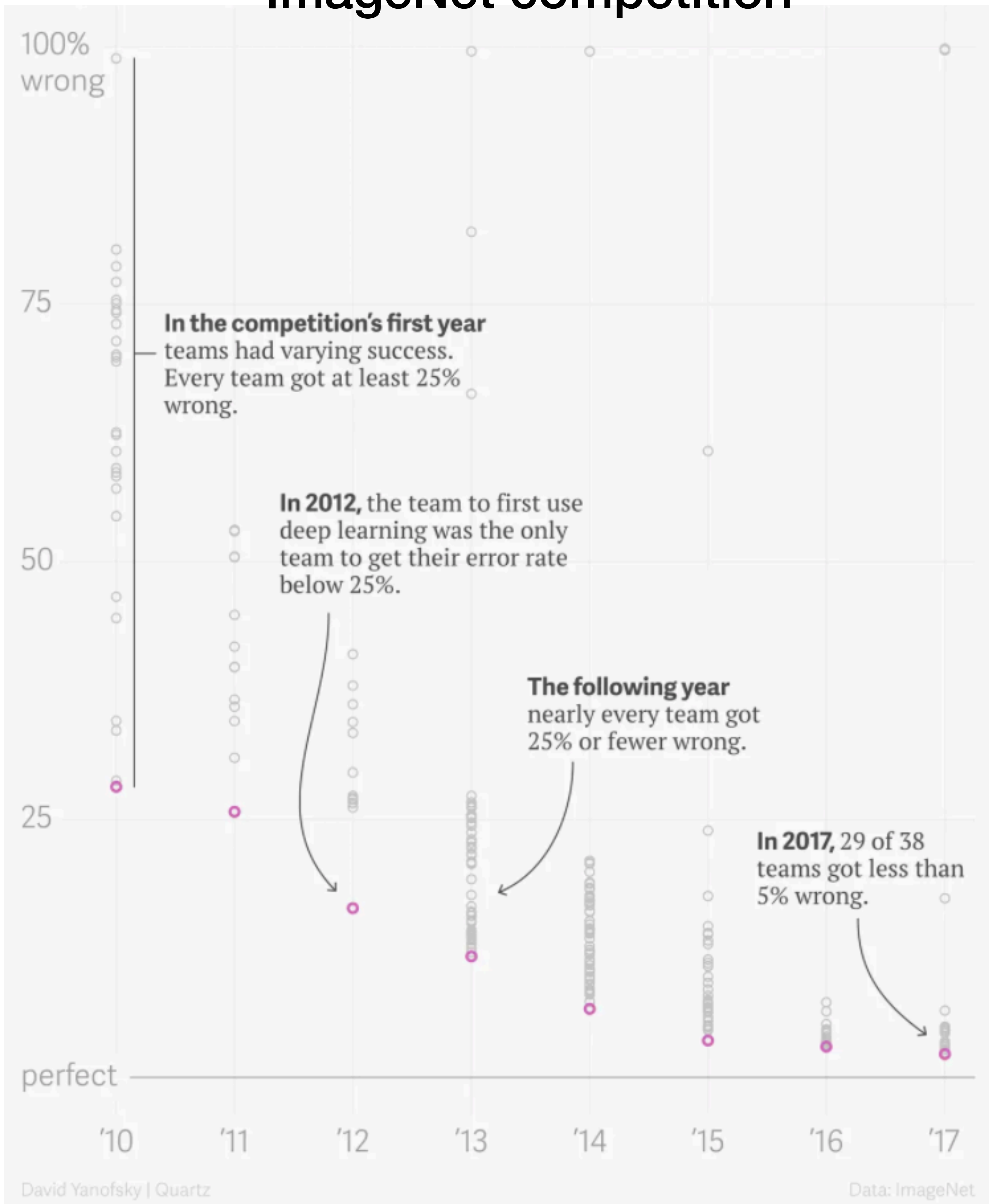


Slide from Jia Deng

Ladles

Slide from Jia Deng

ImageNet competition



Top-5 error

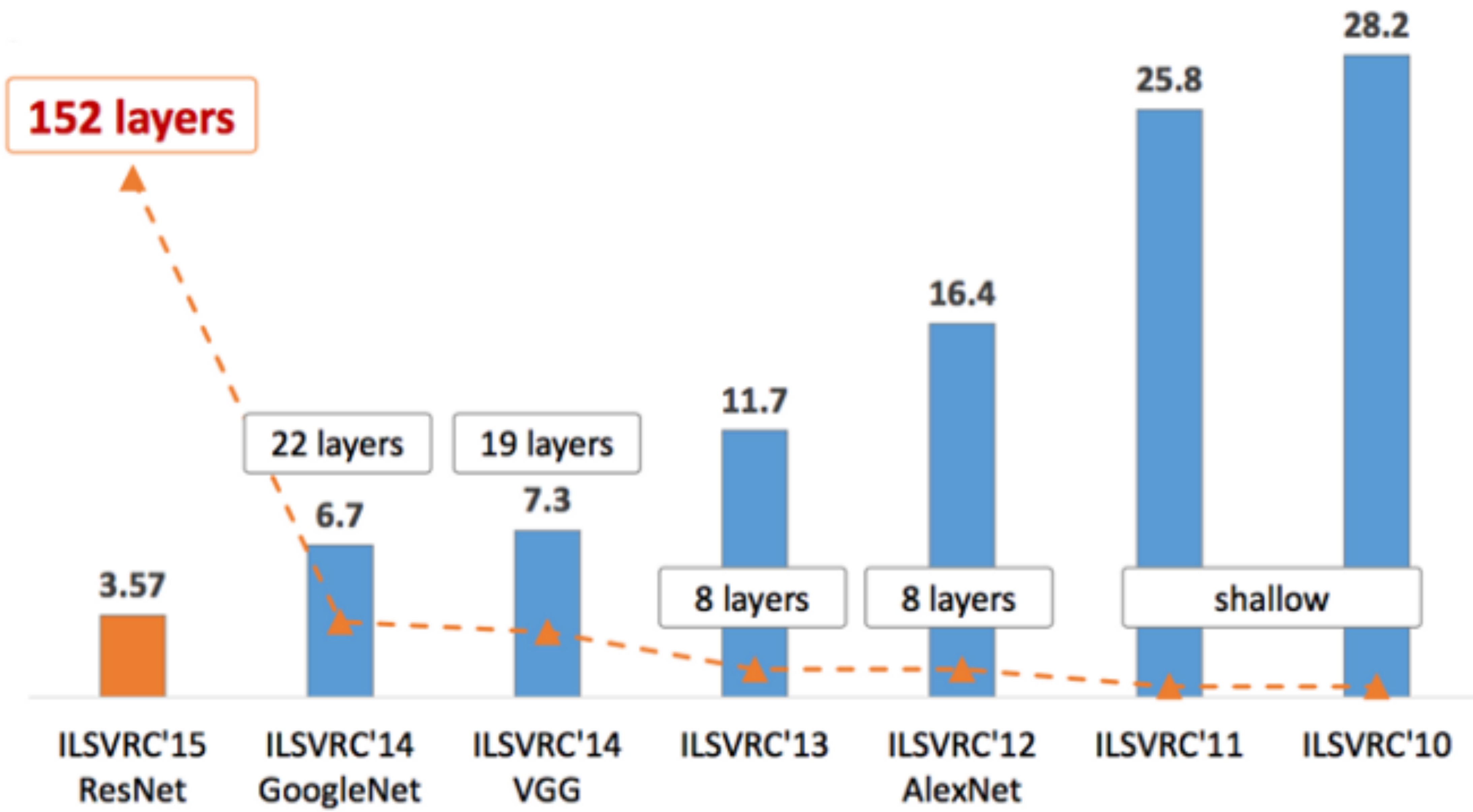


Figure from Siddharth Das

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

Slide from Chris Manning's NLP class CS224N

Now

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

fast.ai OpenAI Google AI OpenAI

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)

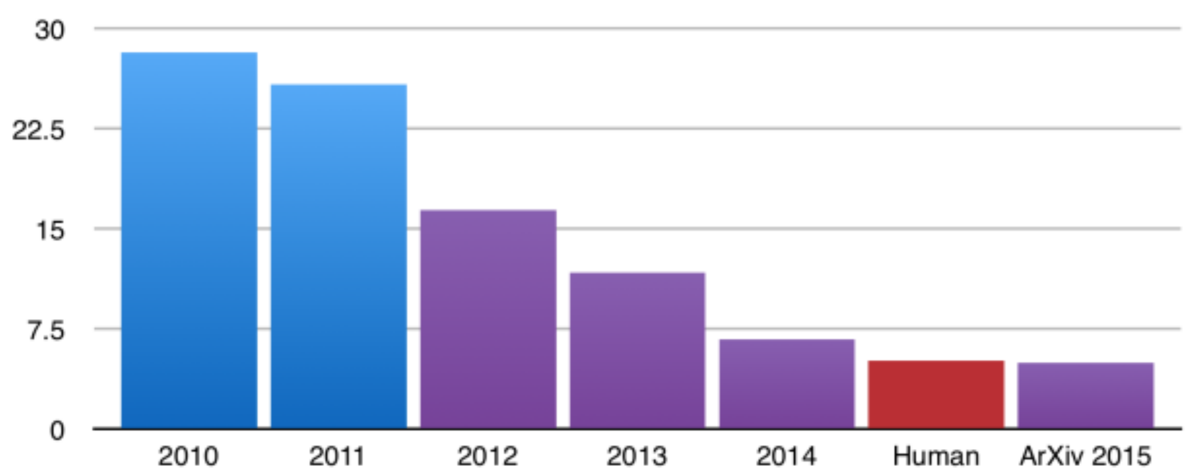
- Collecting enough labels and building the entire pipeline for deep learning is a HUGE effort

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

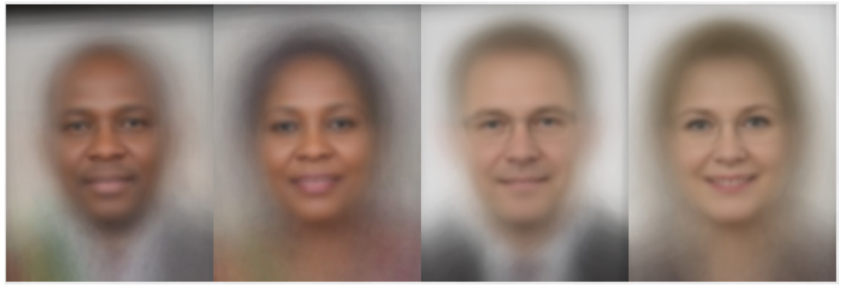
$$\text{minimize}_{\theta \in \Theta} \; \mathbb{E}_P[\ell(\theta; Z)]$$

- 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

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]

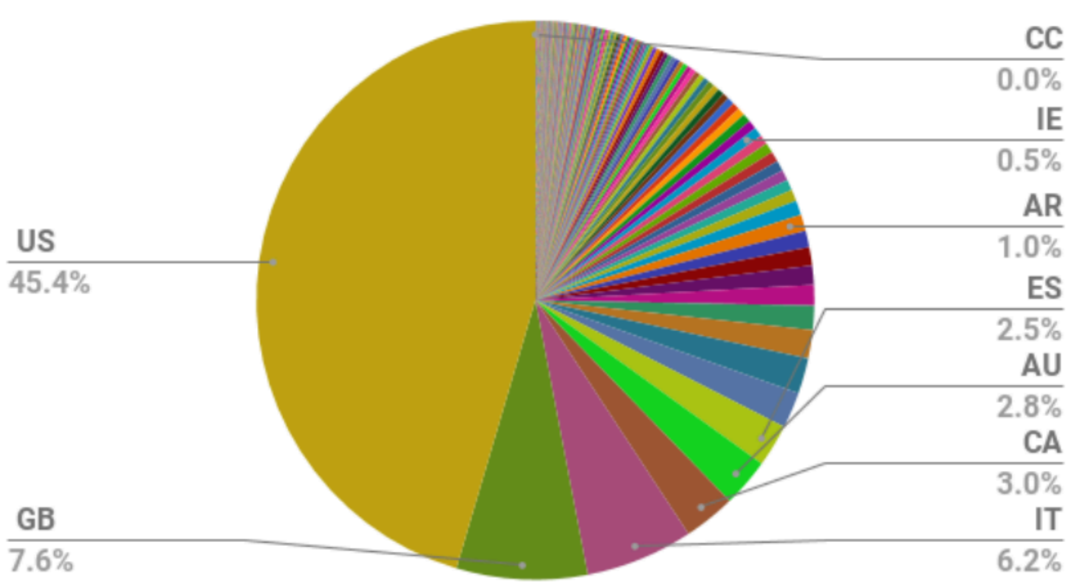
Lack of diversity in data

- "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]

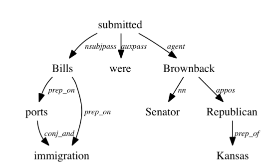
- Majority of image data from **US & Western Europe**

ImageNet: country of origin



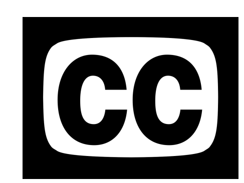
[Shankar et al. '17]

Other examples



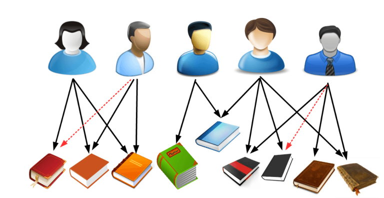
Dependency parsing

[Blodgett+ 16]



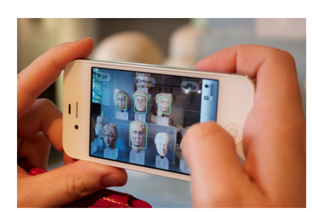
Captioning

[Tatman+ 17]



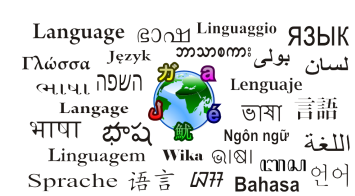
Recommender systems

[Ekstrand+ 17,18]



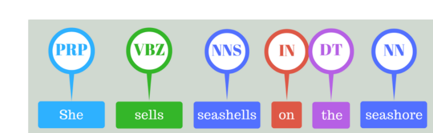
Face recognition

[Grother+ 11]



Language identification

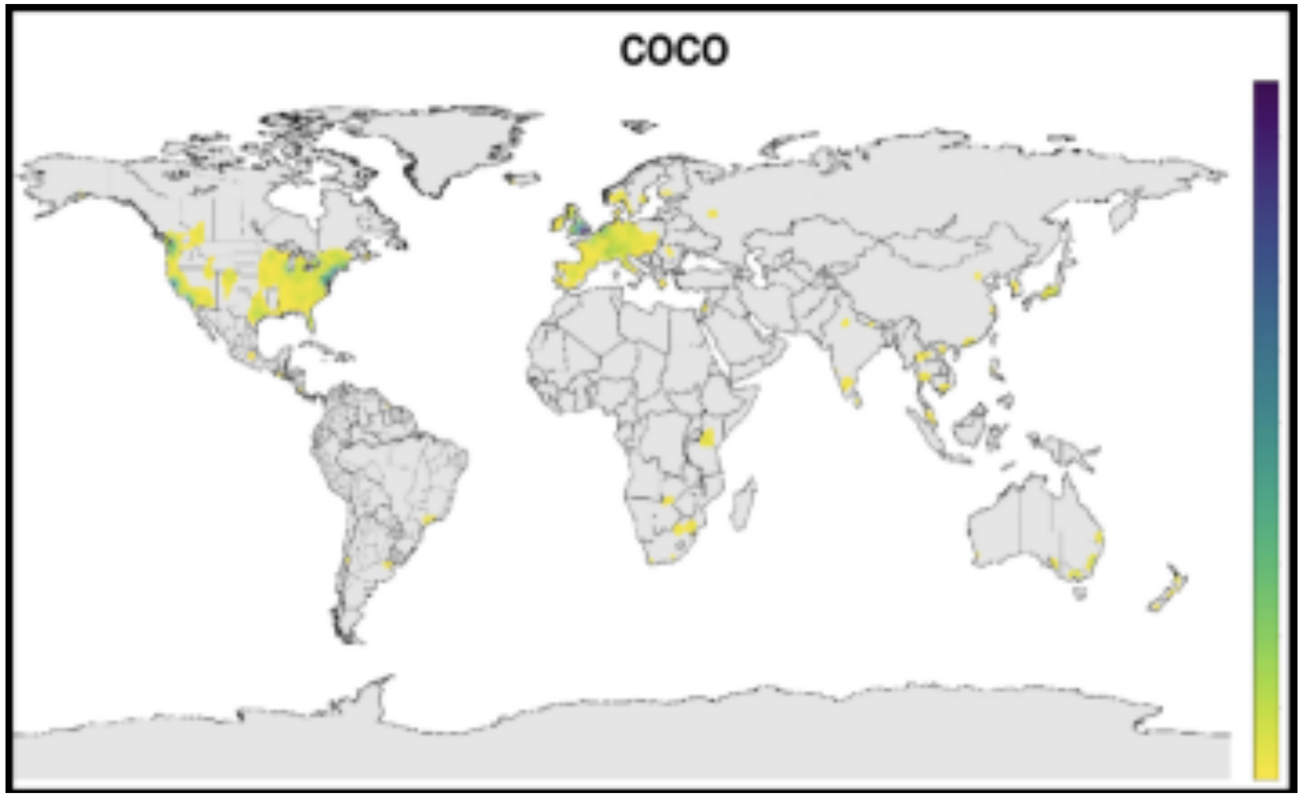
[Blodgett+ 16, Jurgens +17]

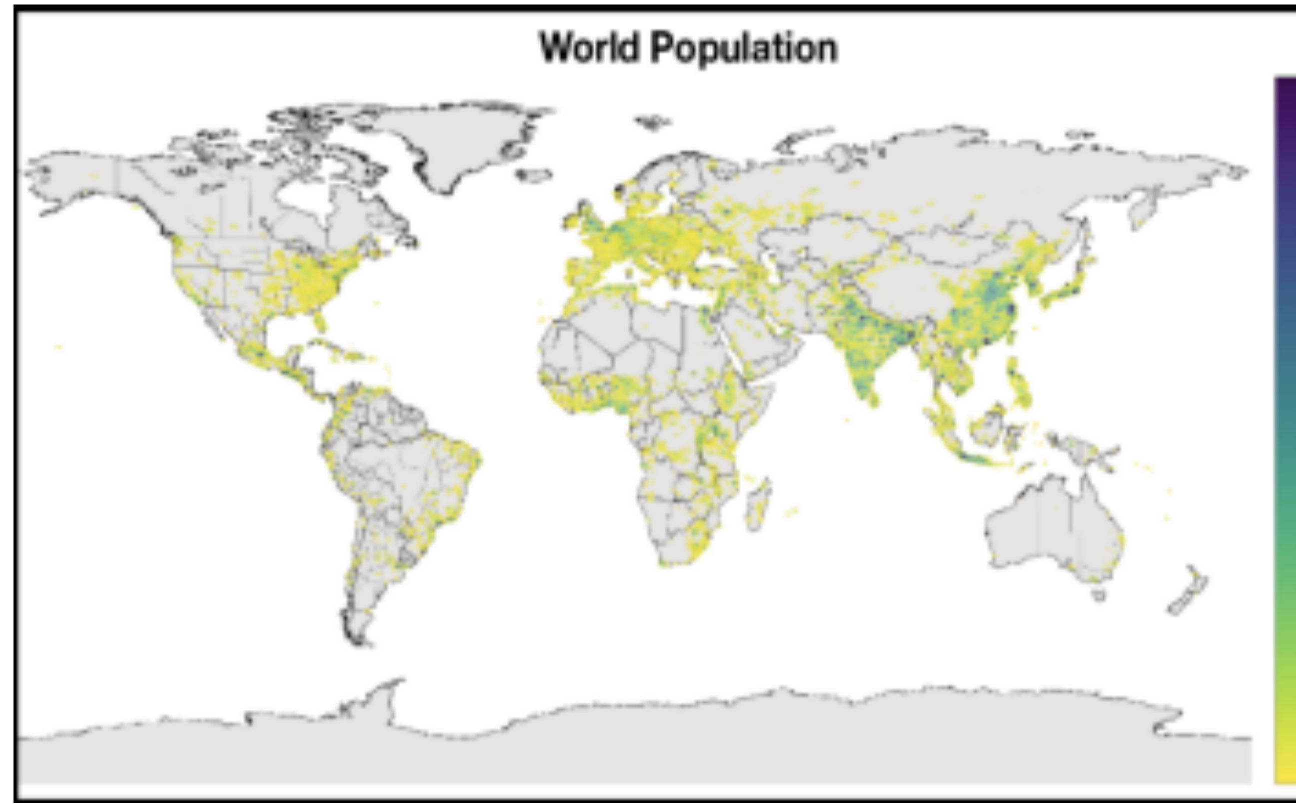


Part-of-speech tagging

[Hovy+ 15]

Lack of diversity in data





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

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

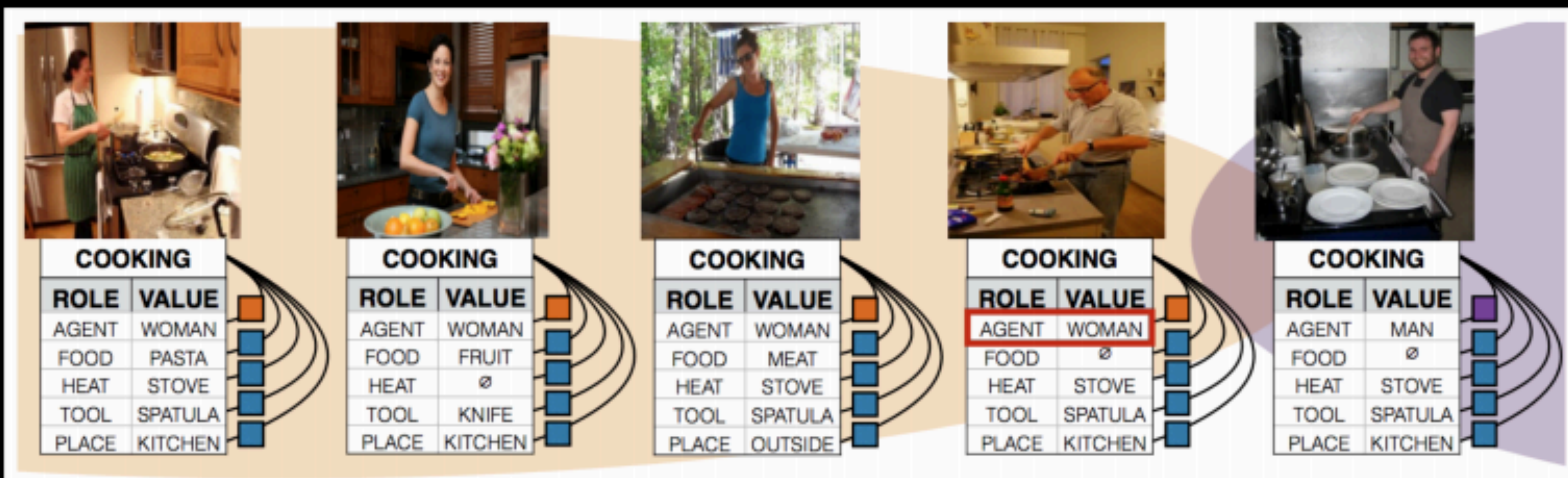
The disparity likely occurs because such technologies are based on machine learning systems that rely heavily on databases of English as spoken by white Americans.

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.

Who is seen? *How* are they seen?

Training data: 33% of cooking images have man in the agent role
Model predictions: 16% cooking images have man in the agent role



[Zhao et al. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints]
[Hendricks et al. Women also snowboard: Overcoming bias in captioning models.]

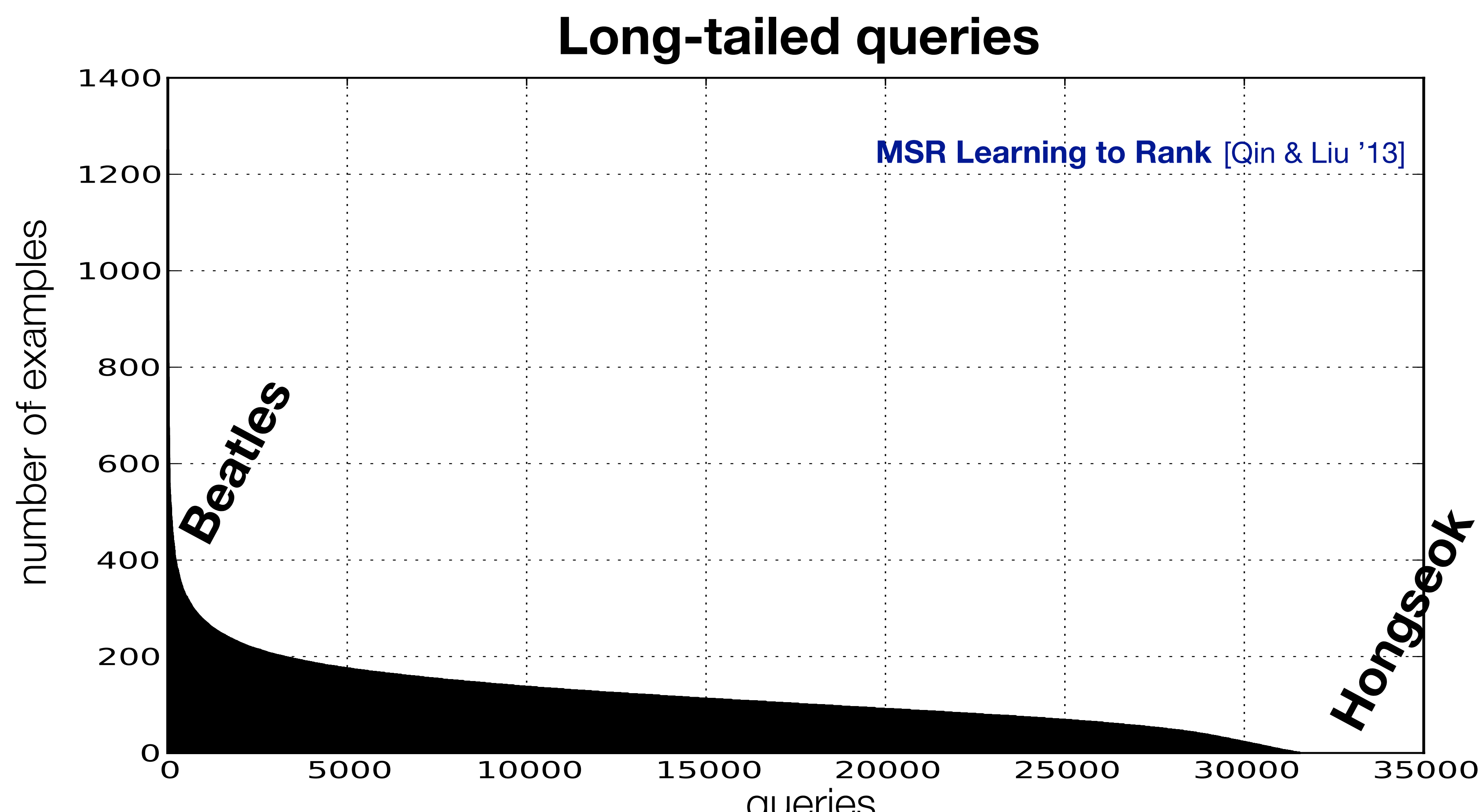
Slide from Timnit Gebru & Emily Denton's CVPR2020 tutorial

SMBC cartoon, Zach Wienersmith

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

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



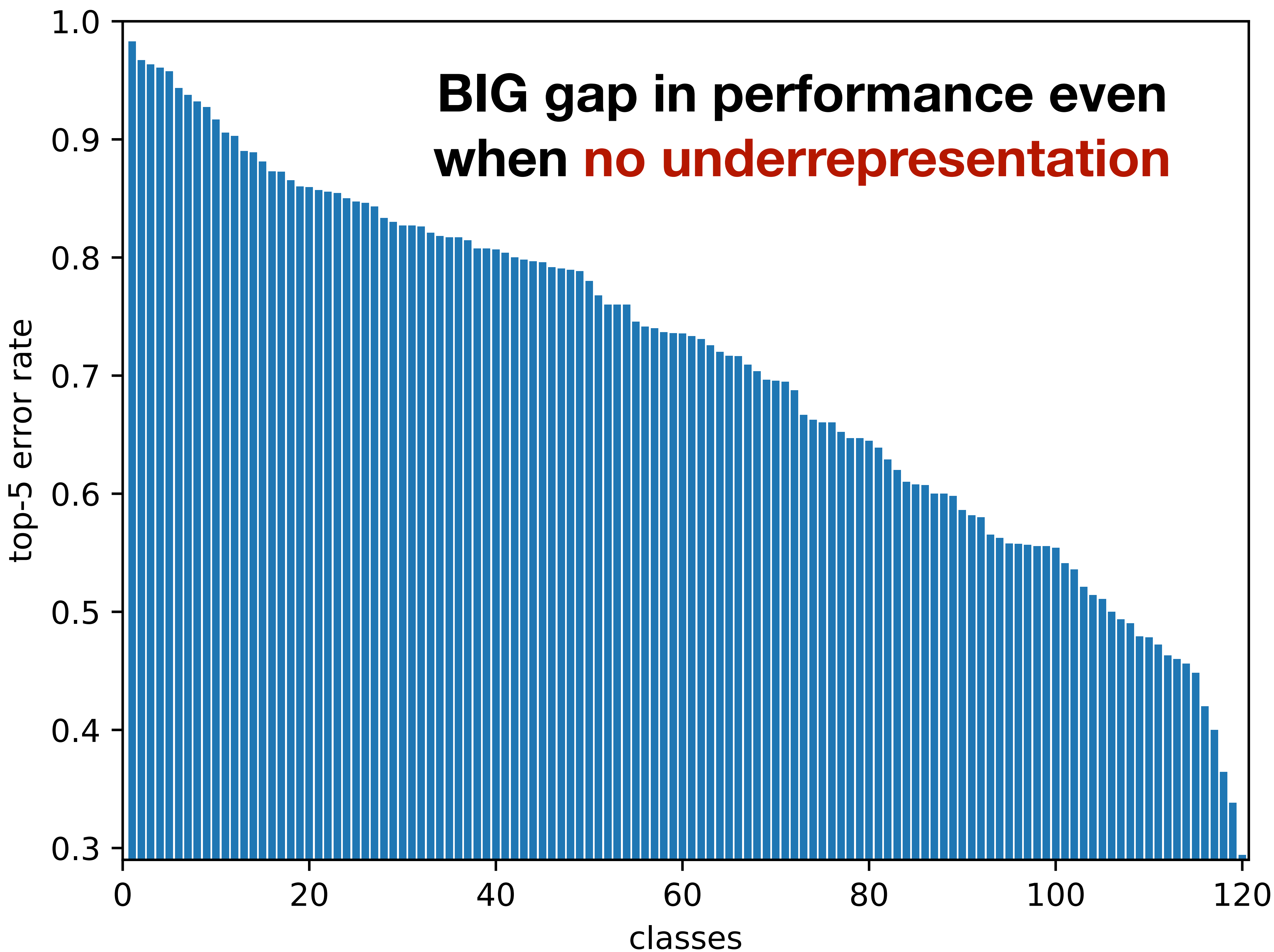
Fundamentally hard examples

- 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

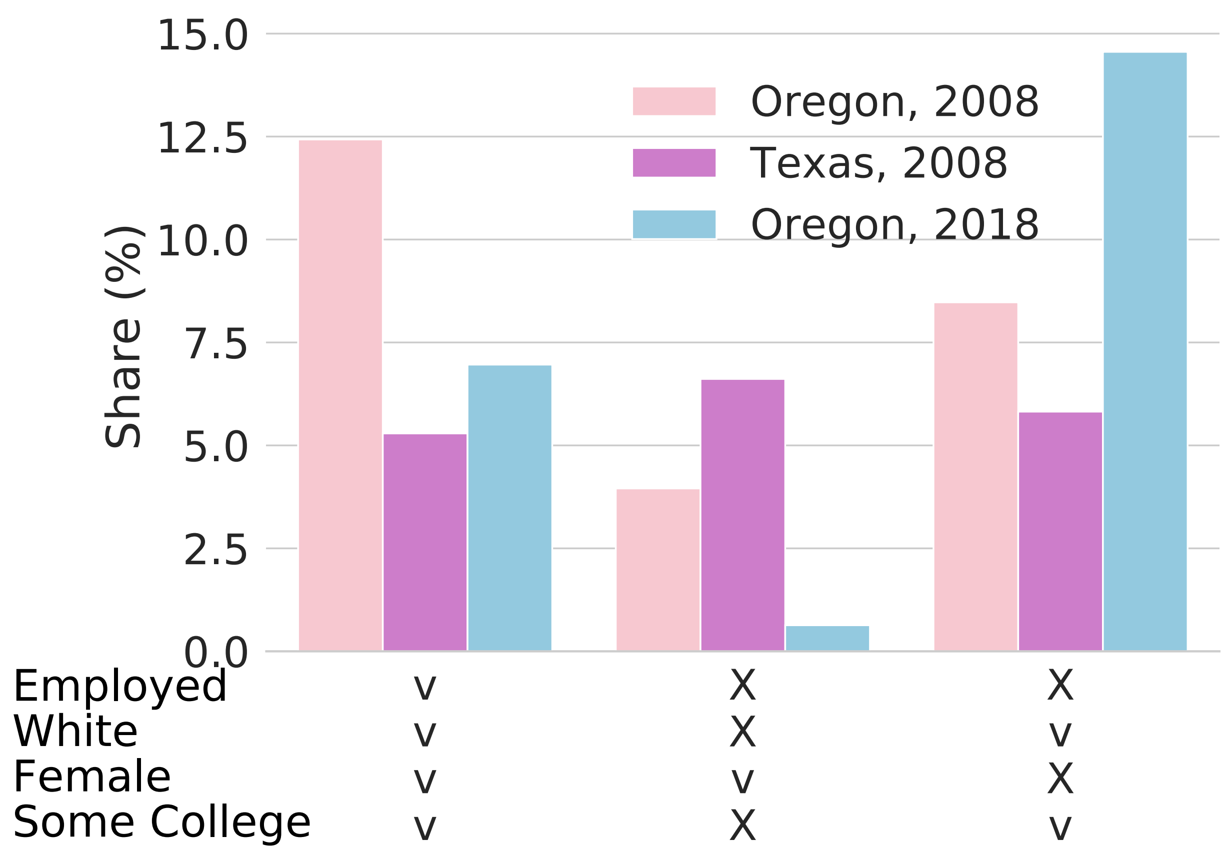


Hard **Easy**

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

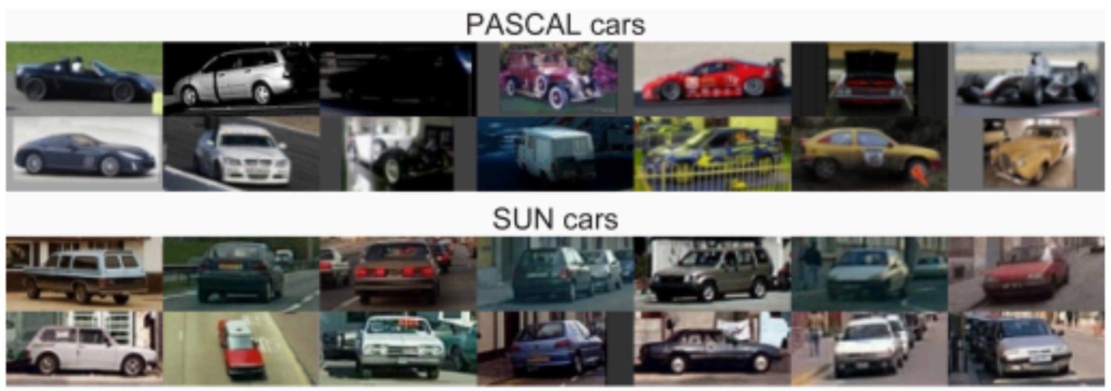
David J. Hand

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)]



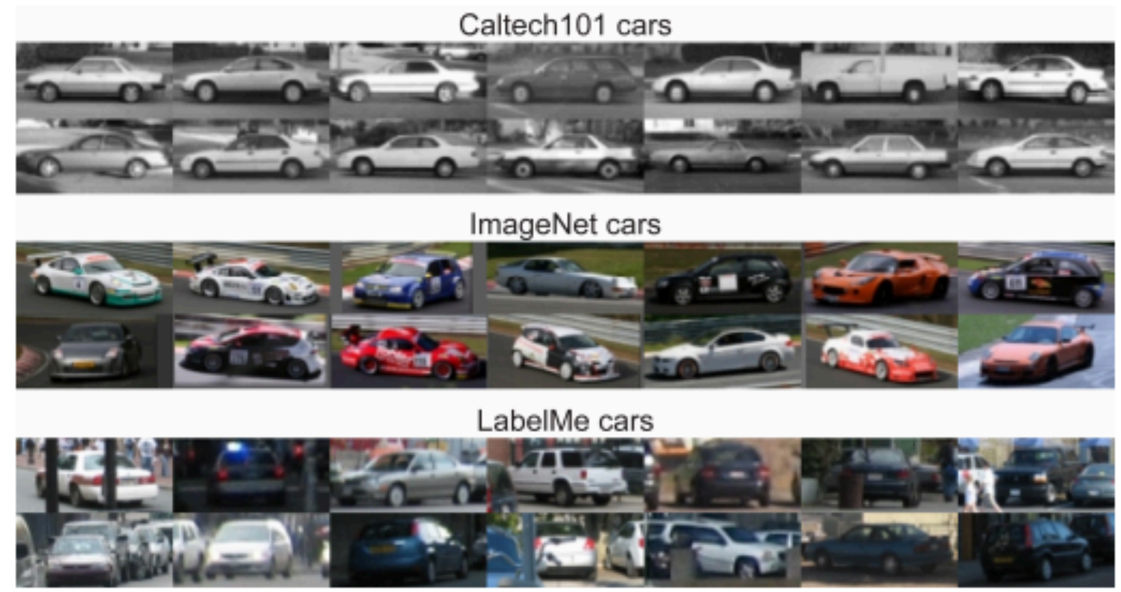
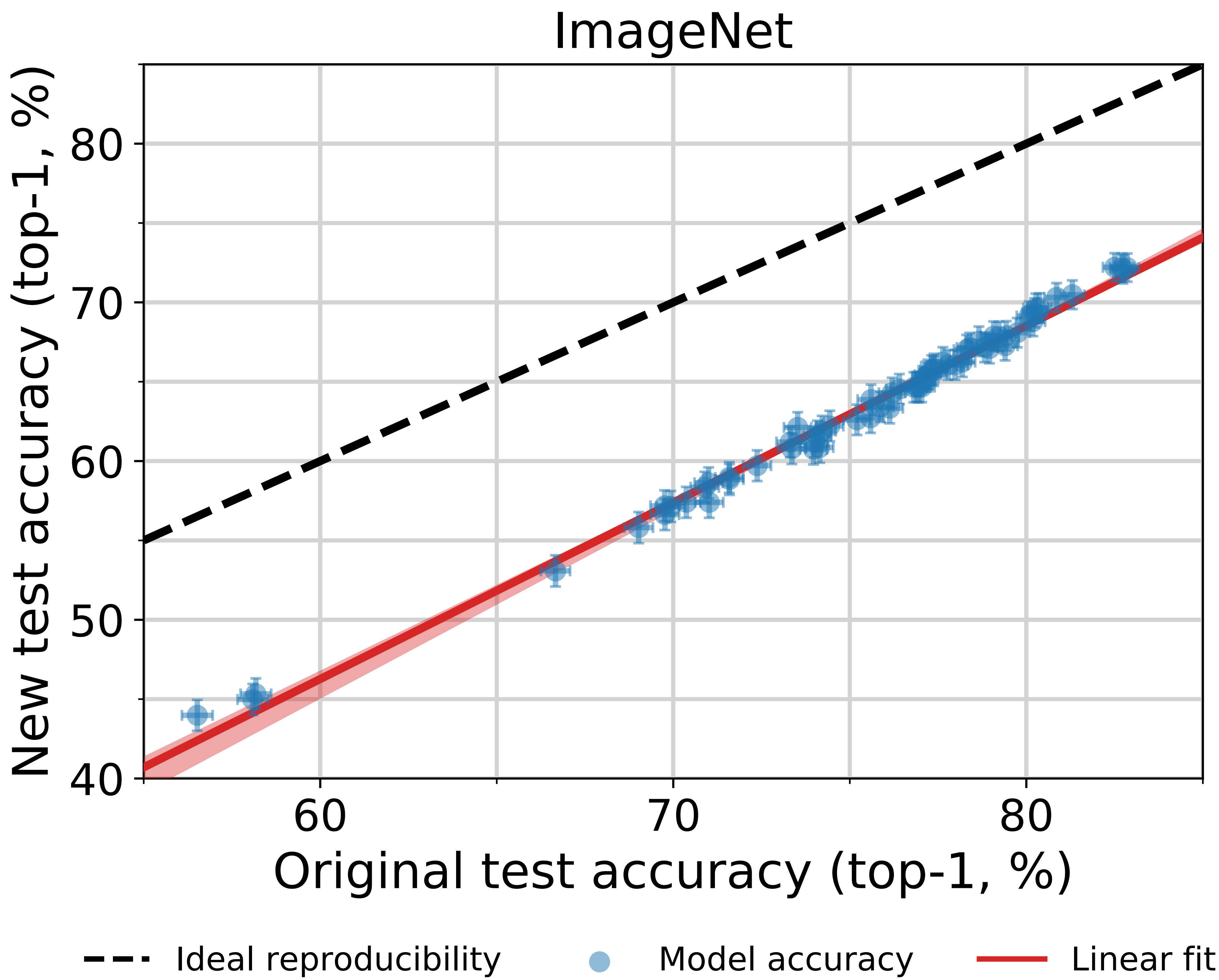


Table 1. Cross-dataset generalization. Object detection and classification performance (AP) for "car" and "person" when training on one dataset (rows) and testing on another (columns), i.e. each row is: training on one dataset and testing on all the others. "Self" refers to training and testing on the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

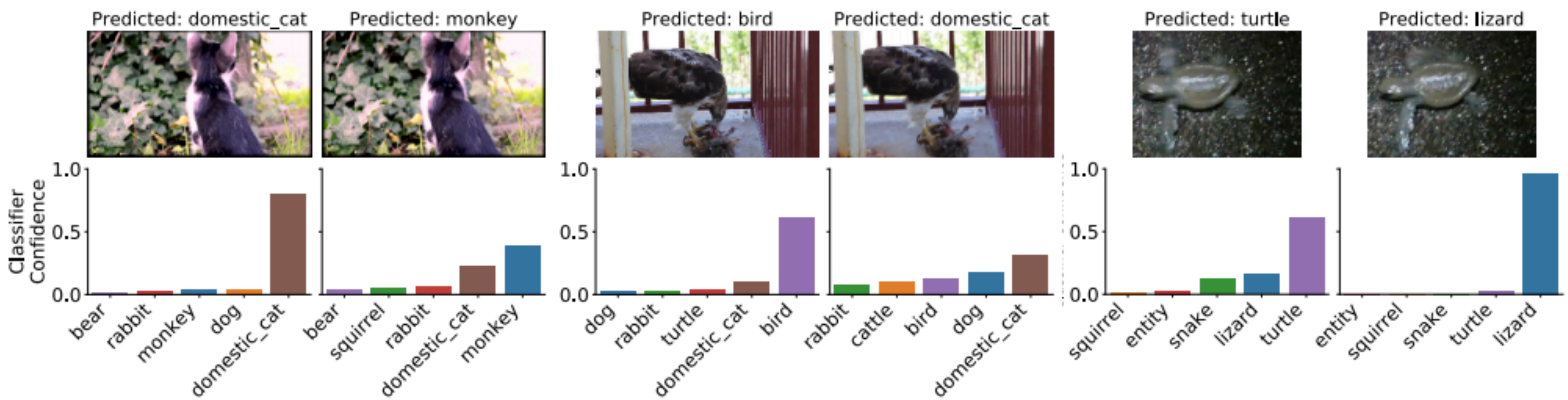


Figure 1: Three examples of natural perturbations from nearby video frames and resulting classifier predictions from a ResNet-152 model fine-tuned on ImageNet-Vid. While the images appear almost identical to the human eye, the classifier confidence changes substantially.

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

SOTA models are non-robust

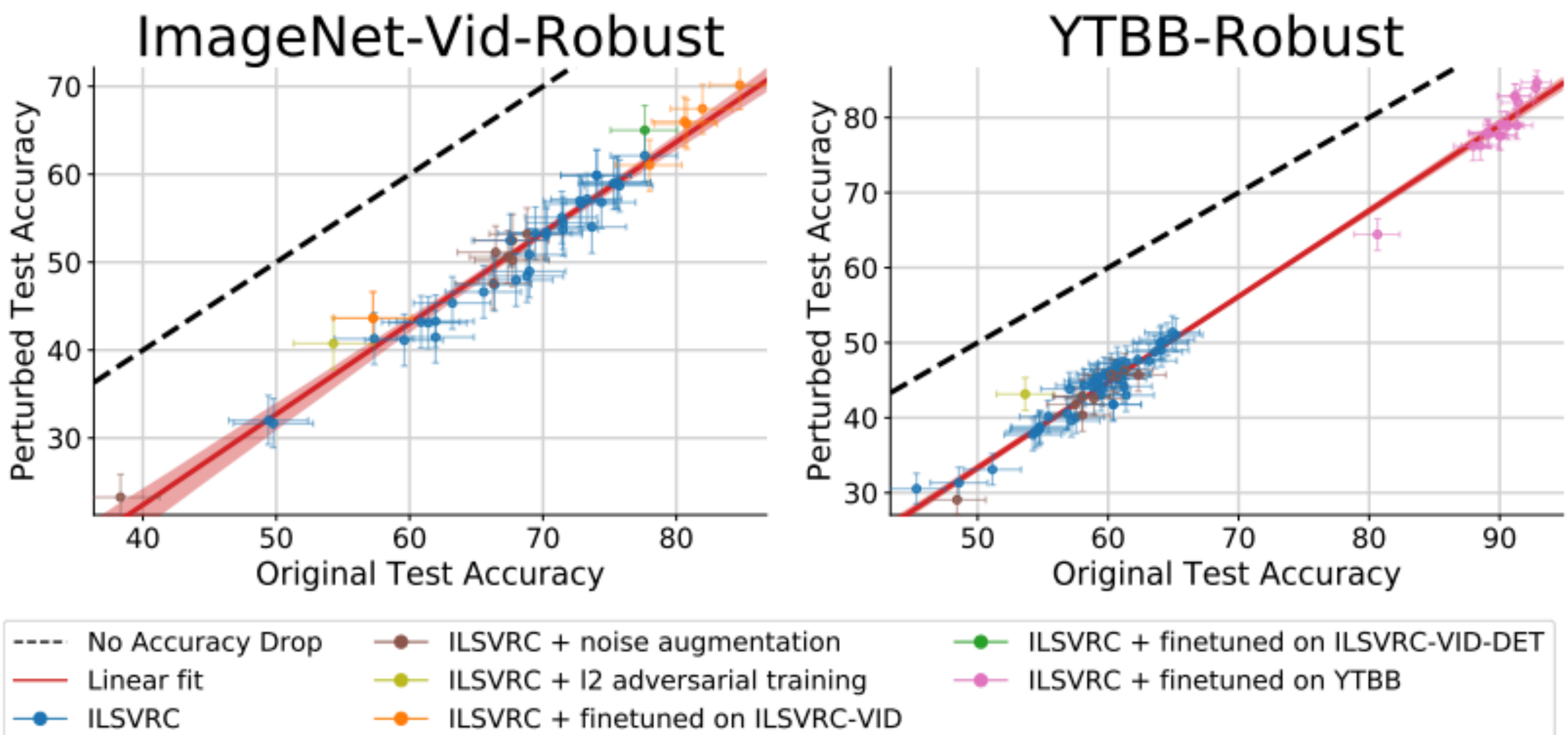


Figure 3: Model accuracy on original vs. perturbed images. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). Each perturbed frame was taken from a ten frame neighborhood of the original frame (approximately 0.3 seconds). All frames were reviewed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

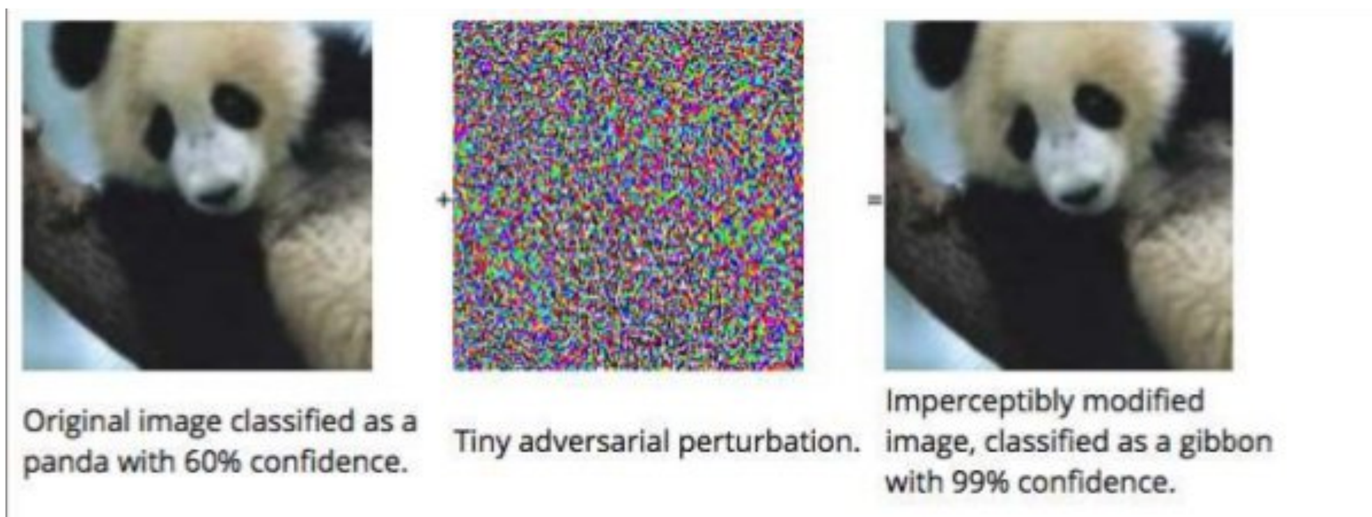
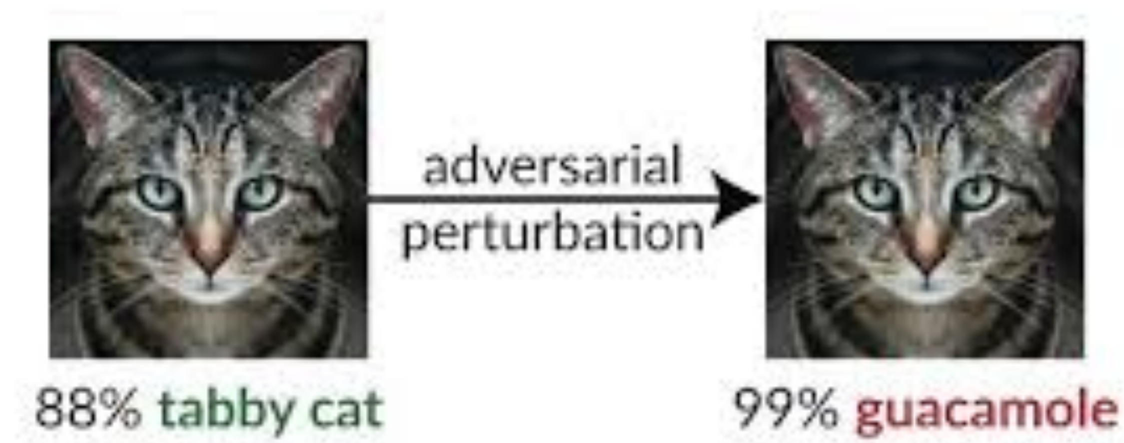


Fig 1. An *adversarial example* constructed by modifying this picture of a panda so that a machine learning model thinks it is a gibbon.

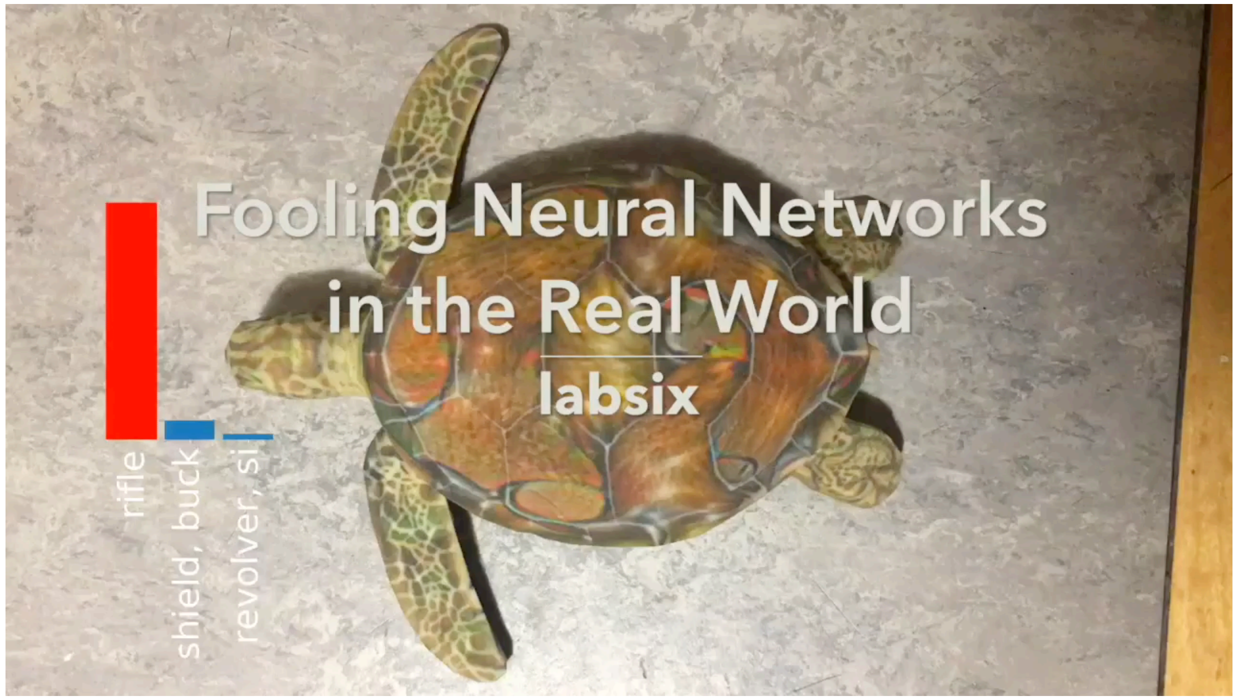
Goodfellow et al. (2015)



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

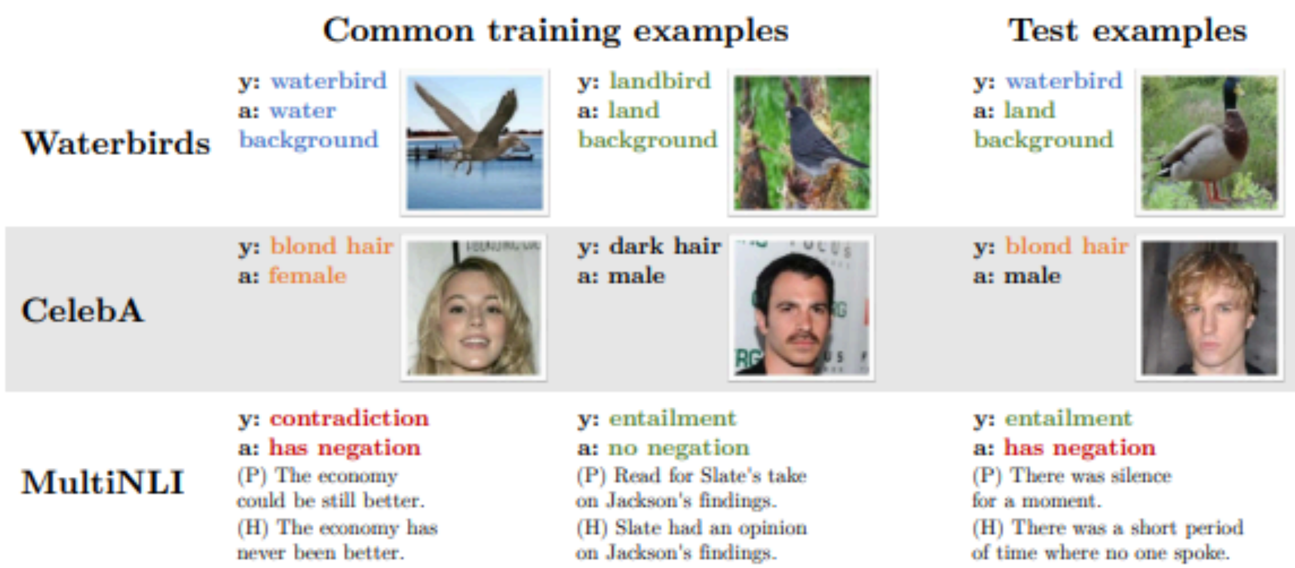


Figure 1: Representative training and test examples for the datasets we consider. The correlation between the label y and the spurious attribute a at training time does not hold at test time.

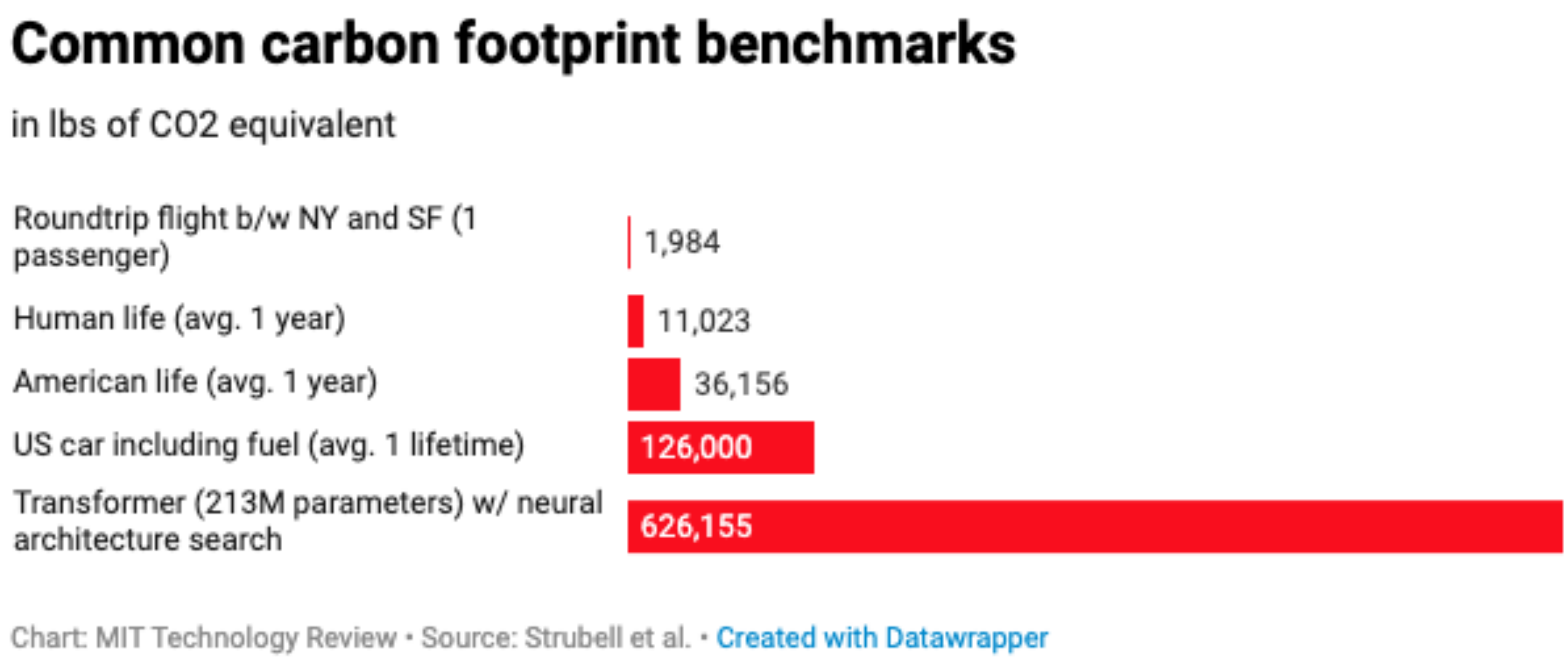
Accuracy:
100% train
=>
60% test

Sagawa et al. (2019)

- 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 REUTERS

Environmental concerns



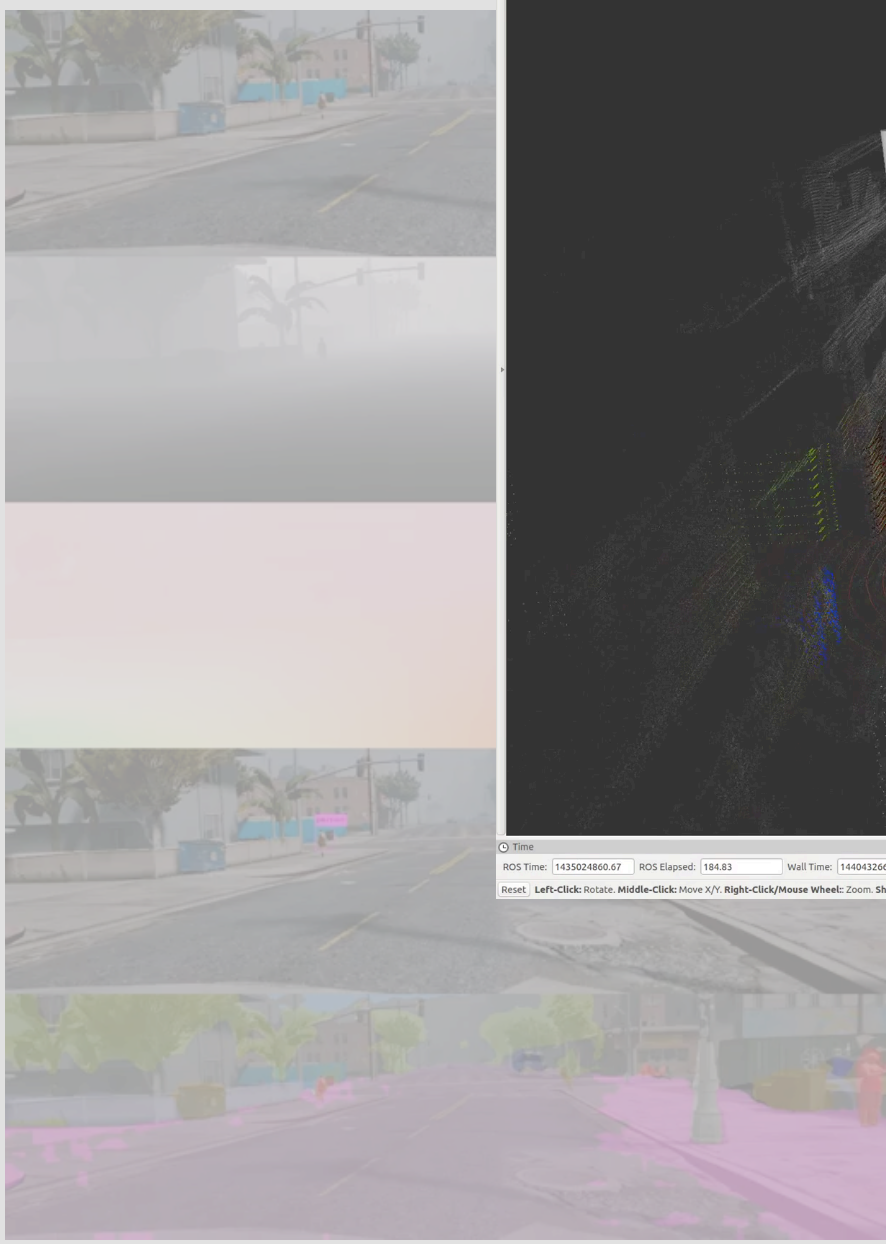
Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

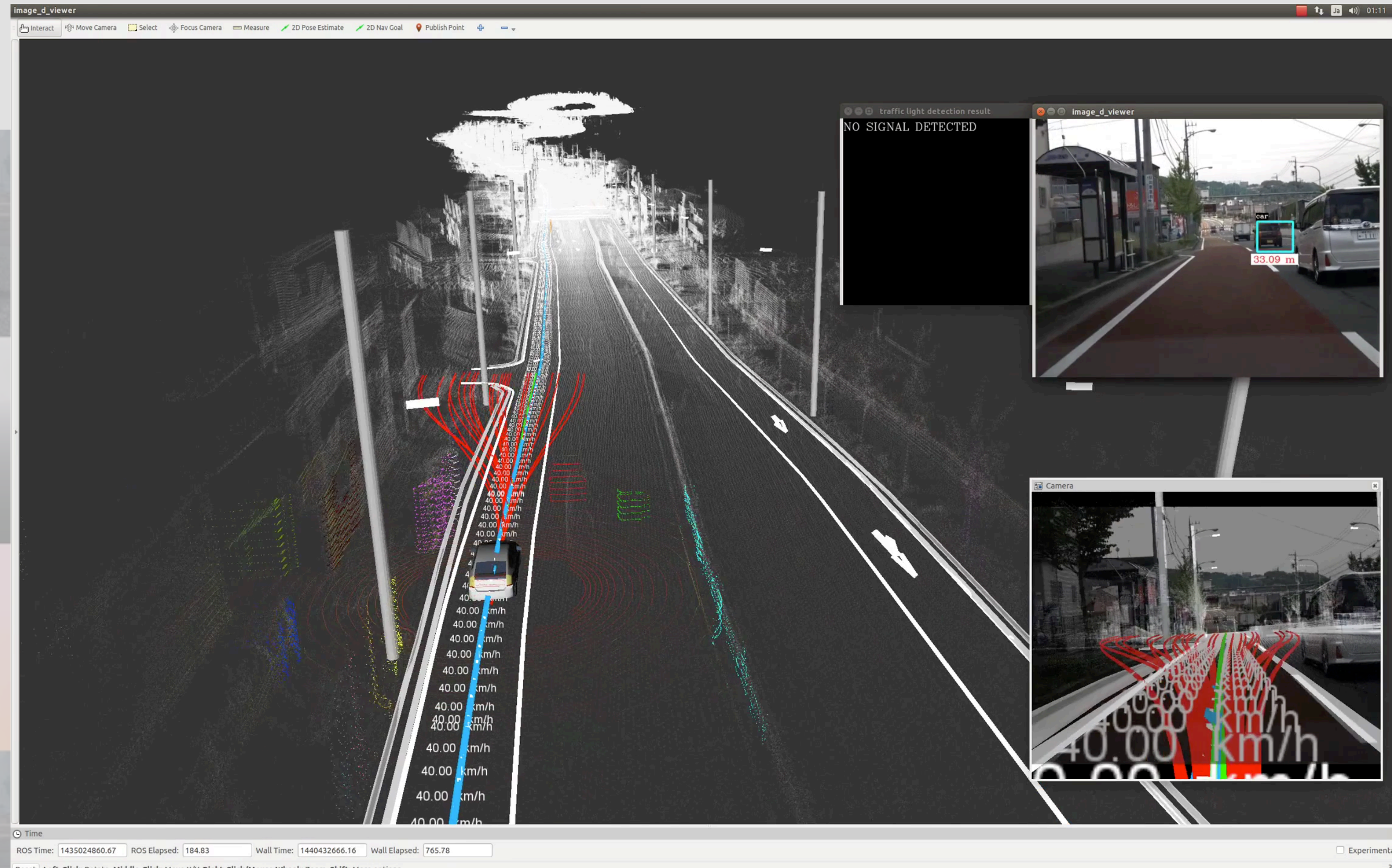
Lots of questions

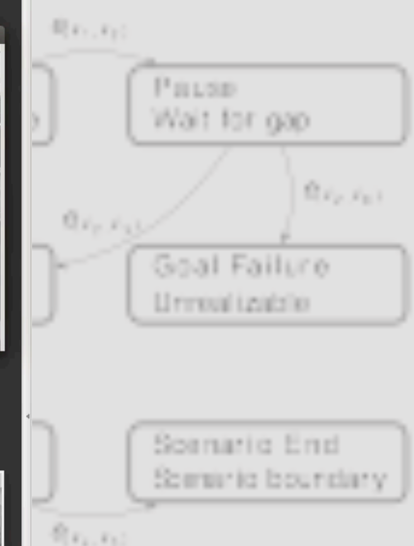
- How do we deal with unanticipated distributional shifts?
  - Modeling itself is nontrivial
- How do we learn causal structures?
- Ultimately, ML models work towards aiding downstream decisions
  - Prediction is not the ultimate goal
  - How to design models with this in mind?
- How do we evaluate the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

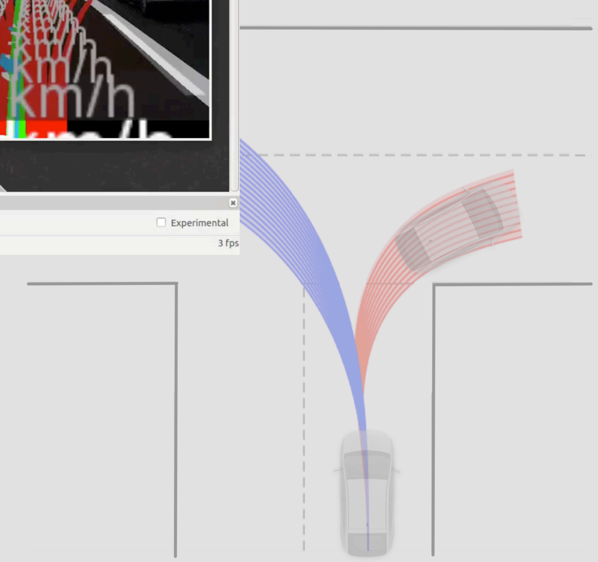
Complex system example: AVs

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

Lots of questions

- ML system interacts with (strategic) agents over time. How to model this interaction/dynamics?
- All modern platforms employ ML as a part of their pipeline
- 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!
  - One month on basic results in statistical learning, and how to prove them
- Then, survey recent works that aim to identify, model, improve upon aforementioned challenges
  - Focus is on *principled* methods, but we'll also discuss a range of practical issues
- Goal: Develop a critical view of topics surrounding reliability
  - Much remains to be done in ML
  - Discussions toward context-specific applications (e.g. healthcare, manufacturing, supply chains, finance, marketing…)
- Goal: Identify interfaces
  - mechanism design
  - sequential decision-making
  - simulation (e.g. rare-events)