CSCI 567: Machine Learning

Vatsal Sharan
Spring 2024

Lecture 13, Apr 19

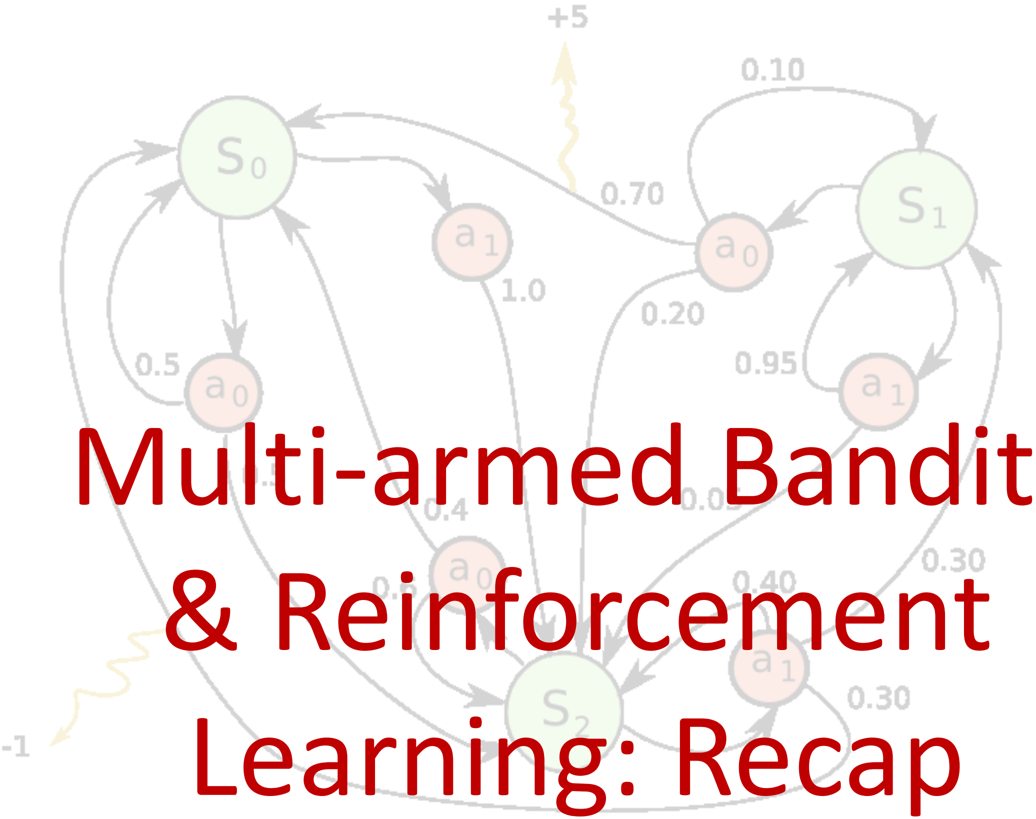
USC University of Southern California

Administrivia

- Exam 2 is on April 26 from 1pm-3:20pm
  - Similar instructions as Exam 1
  - More info will be posted on Ed later

- Today's plan:
    - Reinforcement Learning
    - Trustworthy ML

Multi-armed Bandits & Reinforcement Learning: Recap



Multi-armed bandits: Setup

There are K **arms** (actions/choices/...)

The problem proceeds in rounds between the environment and a learner: for each time $t = 1, \ldots, T$

- the environment decides the reward for each arm $r_{t,1}, \ldots, r_{t,K}$
- the learner picks an arm $a_t \in [K]$
- the learner observes the reward for arm a_t, i.e., r_{t,a_t}

Importantly, *learner does not observe rewards for arms not selected!*

This kind of limited feedback is usually referred to as bandit feedback

The key challenge

All bandit problems face the same **dilemma**:

Exploitation vs. Exploration trade-off

- on one hand we want to exploit the arms that we think are good
- on the other hand we need to explore all arms often enough in order to figure out which one is better
- so each time we need to ask: *do I explore or exploit? and how?*

We discussed **three ways** to trade off exploration and exploitation for our simple multi-armed bandit setting:

- Explore-then-Exploit
- ϵ-Greedy
- Upper Confidence Bound (UCB)

Markov Decision Process

An MDP is parameterized by five elements

- $\mathcal{S}$: a set of possible states

- $\mathcal{A}$: a set of possible actions

- P: transition probability, $P_a(s, s')$ is the probability of transiting from state s to state s' after taking action a (Markov property)

- r: reward function, $r_a(s)$ is (expected) reward of action a at state s

Difference from Markov models, the state transition is influenced by the taken action.

Difference from Multi-armed bandit, the reward depends on the state.

Optimal policy and Bellman equation

First goal: knowing all parameters, *how do we find the optimal policy*

$$\operatorname*{argmax}_{\pi} \underbrace{\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r_{\pi(s_t)}(s_t)\right]}_{\text{expected discounted reward}} \quad ?$$

We first answer a related question: *what is the maximum reward one can achieve starting from an arbitrary state s?*

$$\begin{aligned} V(s) &= \max_{\pi}\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r_{\pi(s_t)}(s_t) \;\middle|\; s_0 = s\right] \\ &= \max_{a\in\mathcal{A}}\left(r_a(s) + \gamma\sum_{s'\in\mathcal{S}} P_a(s,s')V(s')\right) \end{aligned}$$

V is called the **(optimal) value function**.

We can solve it using *value iteration*.

Value iteration

Value Iteration:

Initialize $V_0(s)$ randomly for all $s \in \mathcal{S}$

For $k = 1, 2, \ldots$ (until convergence)

Previous estimates

$$V_k(s) = \max_{a \in \mathcal{A}} \left(r_a(s) + \gamma \sum_{s' \in \mathcal{S}} P_a(s, s') V_{k-1}(s') \right)$$

(**Bellman upate**)

Knowing V, the optimal policy π^* is simply

$$\pi^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \left(r_a(s) + \gamma \sum_{s' \in \mathcal{S}} P_a(s, s') V(s') \right)$$

Markov Decision Processes: Continued

Reinforcement Learning

- Motivation

- Markov Decision Process (MDP)

- Learning MDPs

Learning MDPs

Now suppose we do not know the parameters of the MDP

- transition probability P
- reward function r

How do we find the optimal policy?

We discuss examples from two families of learning algorithms:

- **model-based** approaches
- **model-free** approaches

Model-based approaches

Key idea: learn the model P and r explicitly from samples

Suppose we have a sequence of interactions: $s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_T, a_T, r_T$, then the MLE for P and r are simply

$$P_a(s, s') \propto \#\text{transitions from } s \text{ to } s' \text{ after taking action } a$$

$$r_a(s) = \text{average observed reward at state } s \text{ after taking action } a$$

Having estimates of the parameters we can then apply value iteration to find the optimal policy.

same as MLE for Markov chains;

$$P_a(s, s') = \frac{\#\text{ transitions from } s \to s' \text{ after taking } a}{\#\text{ transitions from } s \text{ after taking } a}$$

Model-based approaches

How do we collect data $s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_T, a_T, r_T$?

Simplest idea: follow a random policy for T steps. This is similar to explore–then–exploit, and we know this is not the best way.

Let's adopt the ϵ-Greedy idea instead.

A sketch for model-based approaches

Initialize V, P, r randomly

For $t = 1, 2, \ldots,$

- **with probability** ϵ**, explore**: pick an action uniformly at random
- **with probability** $1 - \epsilon$**, exploit**: pick the optimal action based on V
- update the model parameters P, r
- update the value function V (via value iteration)

Model-free approaches

Key idea: do not learn the model explicitly. *What do we learn then?*

Define the $Q : \overset{\text{states}}{\mathcal{S}} \times \overset{\text{actions}}{\mathcal{A}} \to \mathbb{R}$ function as

$$Q(s,a) = r_a(s) + \gamma \sum_{s' \in \mathcal{S}} P_a(s,s') \max_{a' \in \mathcal{A}} Q(s',a')$$

In words, $Q(s,a)$ is the expected reward one can achieve starting from state s with action a, then acting optimally.

Clearly, $V(s) = \max_a Q(s,a)$.

Knowing $Q(s,a)$, the optimal policy at state s is simply $\operatorname{argmax}_a Q(s,a)$.

Model-free approaches learn the Q function directly from samples.

Q-learning update rule

How to learn the Q function?

$$Q(s,a) = r_a(s) + \gamma \sum_{s' \in \mathcal{S}} P_a(s,s') \max_{a' \in \mathcal{A}} Q(s',a')$$

On experience $\langle s_t, a_t, r_t, s_{t+1} \rangle$, with the current guess on Q, $r_t + \gamma \max_{a'} Q(s_{t+1}, a')$ is like a training example for Q-learning.

So it's natural to do the following update:

$$Q(s_t,a_t) \leftarrow (1-\alpha) Q(s_t,a_t) + \alpha \left(r_t + \gamma \max_{a'} Q(s_{t+1},a') \right)$$

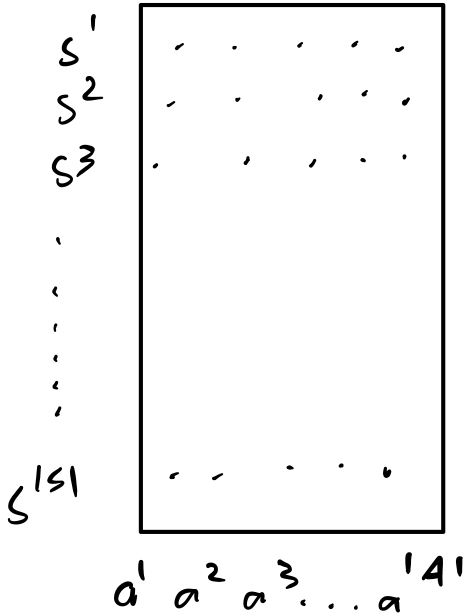
$$= Q(s_t,a_t) + \alpha \underbrace{\left(r_t + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t) \right)}_{\text{difference between (estimated) obtained reward and predicted reward}}$$

previous estimate

estimated obtained reward

α is like the learning rate

Tabular Q-learning



Q-learning

The simplest model-free algorithm:

Q-learning

Initialize Q randomly; denote the initial state by s_1.

For $t = 1, 2, \ldots,$

- **with probability** ϵ**, explore**: a_t is chosen uniformly at random
- **with probability** $1 - \epsilon$**, exploit**: $a_t = \operatorname{argmax}_a Q(s_t, a)$
- execute action a_t, receive reward r_t, arrive at state s_{t+1}
- update the Q function

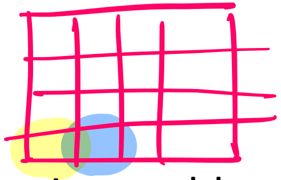
$$Q(s_t, a_t) \leftarrow (1 - \alpha) Q(s_t, a_t) + \alpha \left(r_t + \gamma \max_a Q(s_{t+1}, a) \right)$$

for some learning rate α.

Q-learning for large state spaces

For a large state space, such as a continuous space:

1. Discretize   2 D   input



2. Treat Q-learning as a supervised learning problem, given current (s, a) find $Q(s, a)$

Input: (s_t, a_t)
Desired output: $r_t + \gamma\, \hat{V}(s_{t+1})$

where $\hat{V}(s_{t+1}) = \max_a Q(s_{t+1}, a)$.

Can use powerful supervised learning techniques!

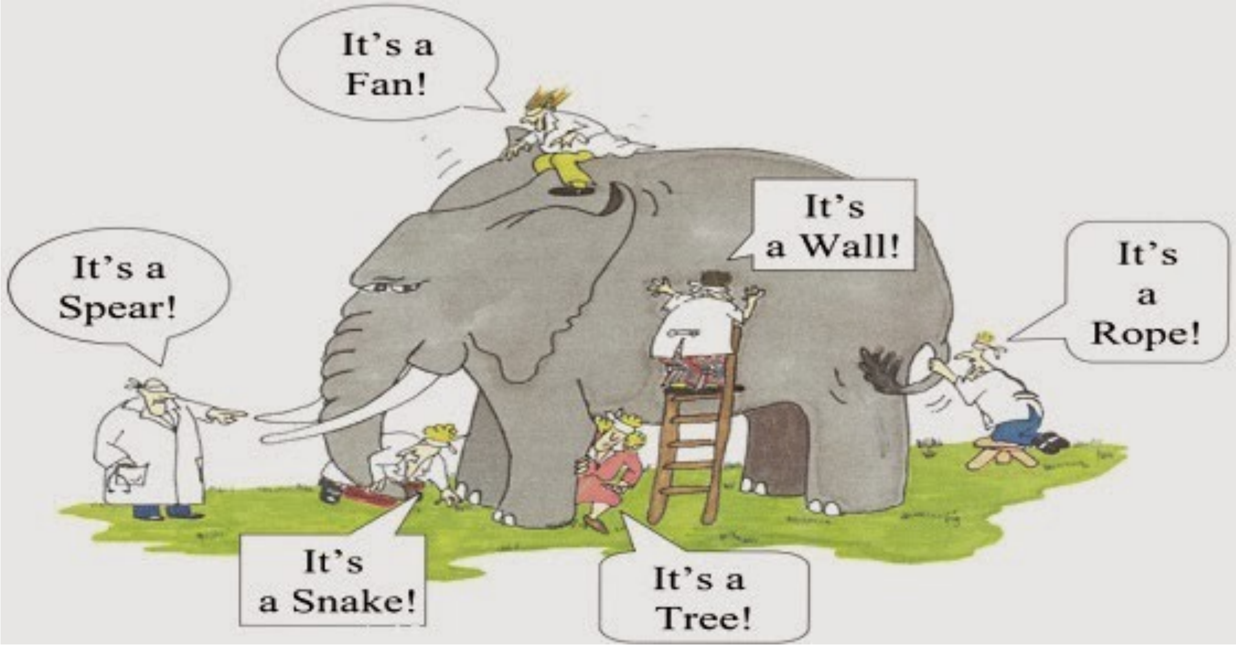
Deep neural networks -> Deep Q-learning

Model-based vs model-free RL

	Model-based	Model-free
What they learn	model parameters $P, r, \ldots$	Q function
Space	$O(\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert)$	$O(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert)$
Data efficiency	usually better	usually worse
Assumptions	need model of world	do not assume model

Trustworthy ML

Machine Learning can be *brittle*



The Blind Men and the Elephant

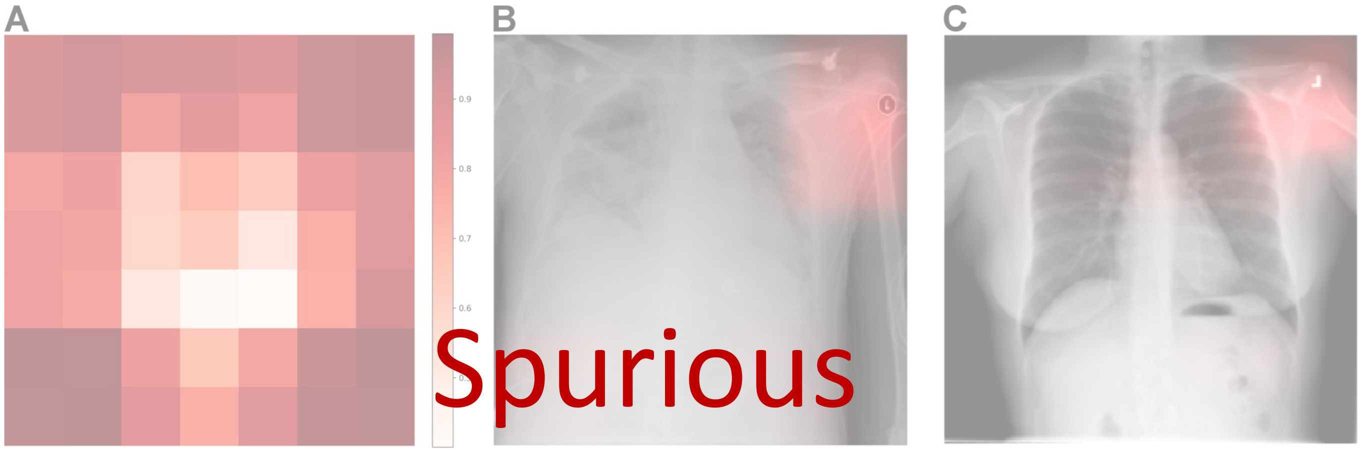
It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.

The First approached the Elephant,
And happening to fall
Against his broad and sturdy side,
At once began to bawl:
"God bless me! but the Elephant
Is very like a WALL!"

....

Challenges in Trustworthy ML

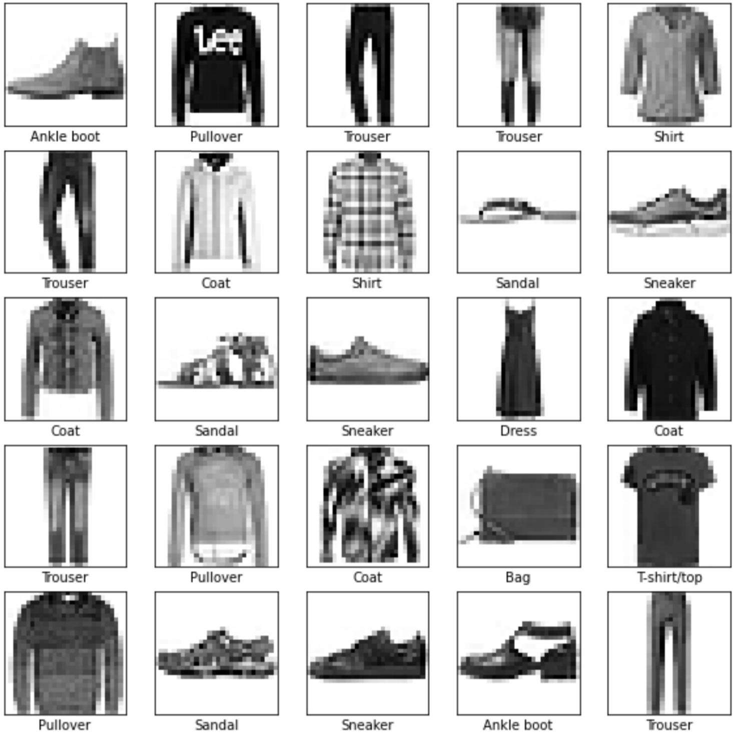
- Spurious correlations and distributional shifts
- Biases in models and unfairness to demographics
- Adversarial examples
- Privacy, Interpretability, Ethics, …

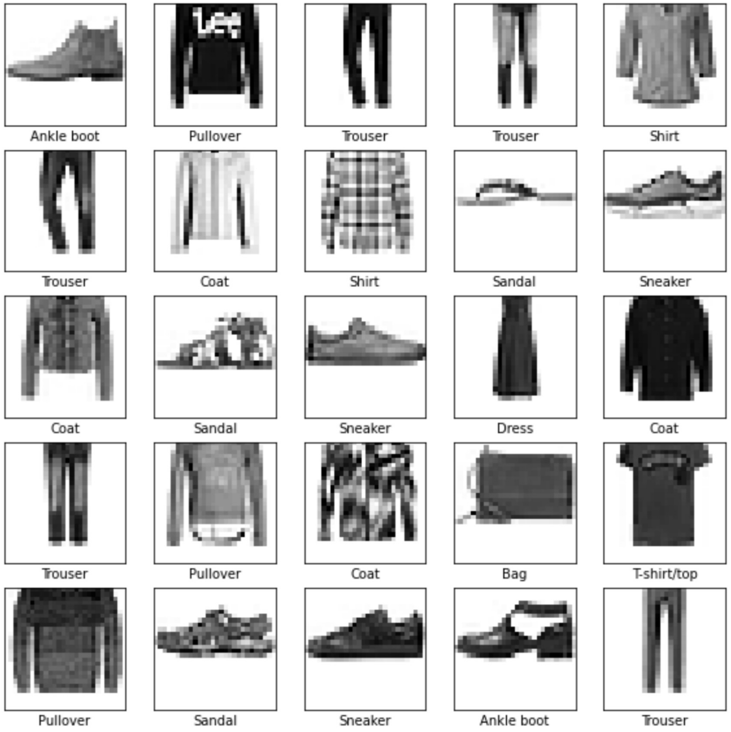


Spurious correlations and distributional shifts

ML models can be very sensitive to changes in the data distribution

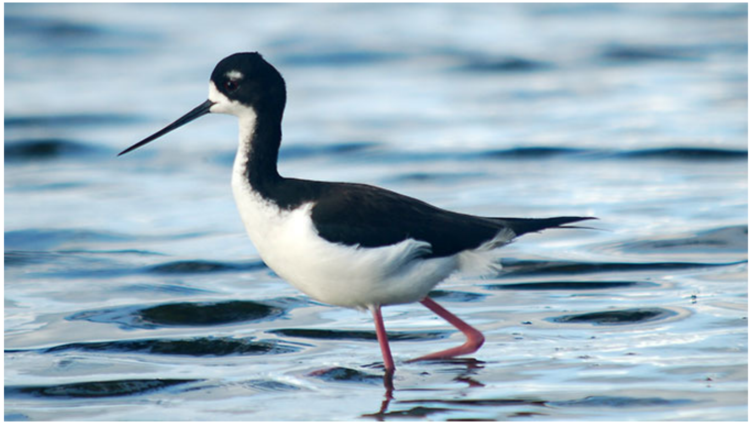
You saw a small example of this in the HW3 Bonus question:





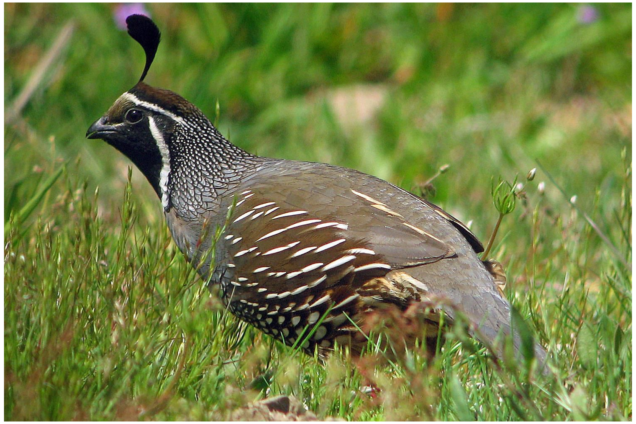
ML models can latch onto spurious features to make predictions

Consider the following task:



Waterbird

vs.

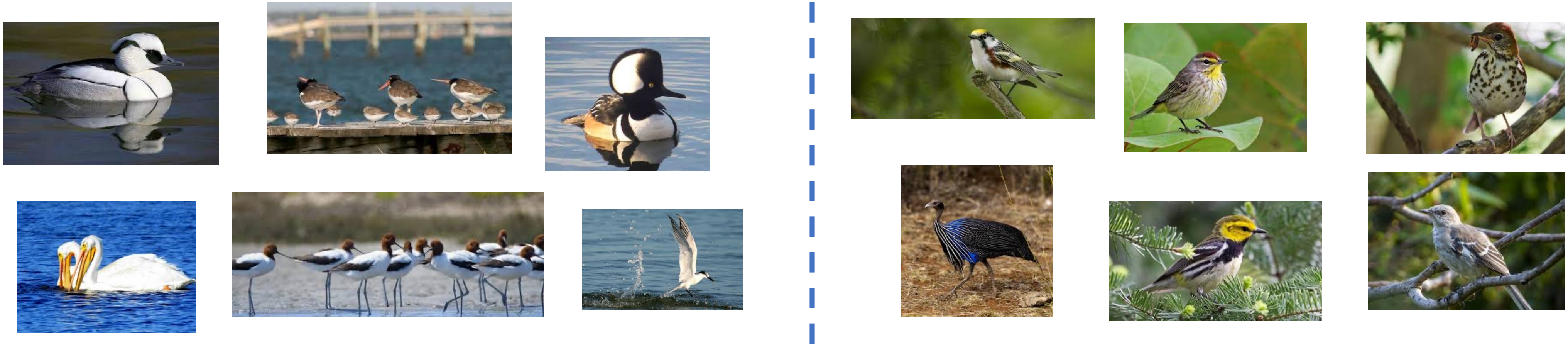


Landbird

Images courtesy of: University of Hawaii at Manoa, Greg Schechter (Flickr)

ML models can latch onto spurious features to make predictions

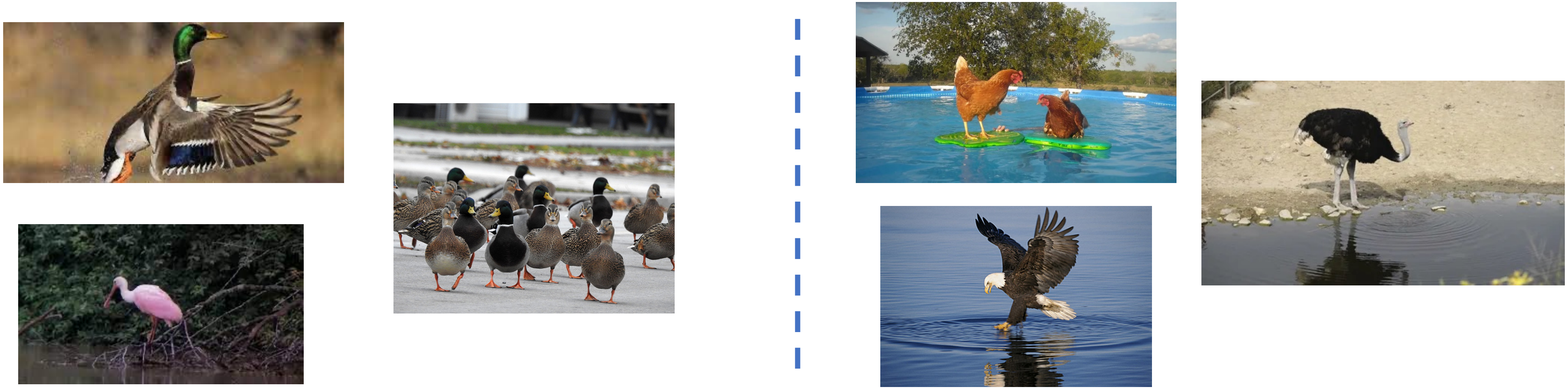
Most images of waterbirds are in water,
and landbirds are on land



Waterbirds vs. Landbirds

ML models can latch onto spurious features to make predictions

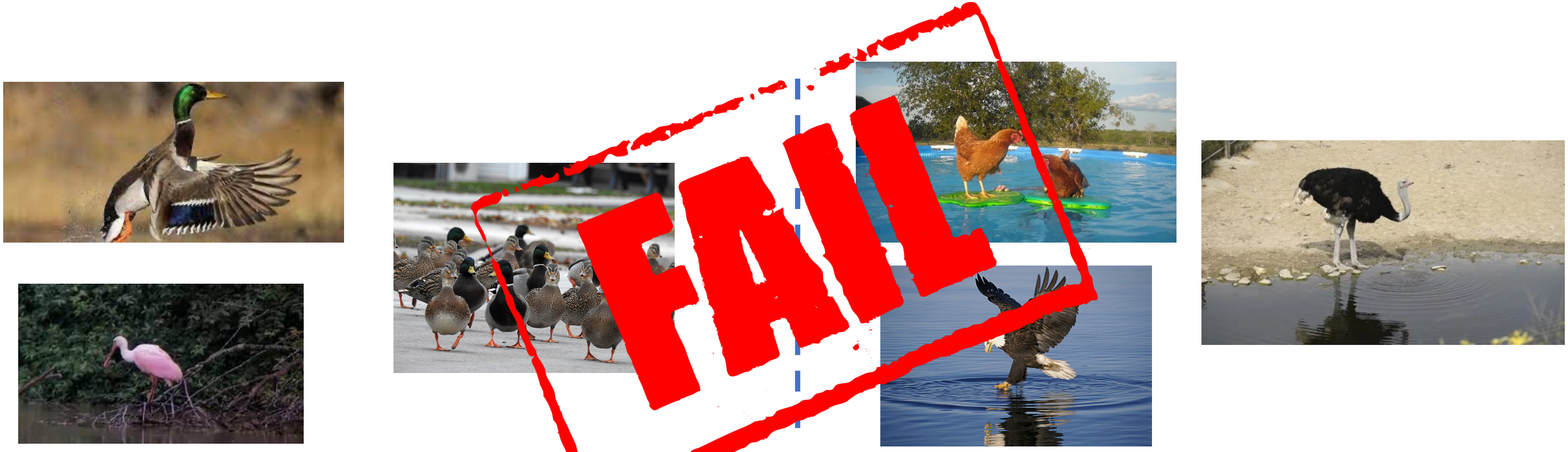
But this isn’t always true!



Waterbirds vs. Landbirds

ML models can latch onto spurious features to make predictions

This is known as failure to distributional shifts

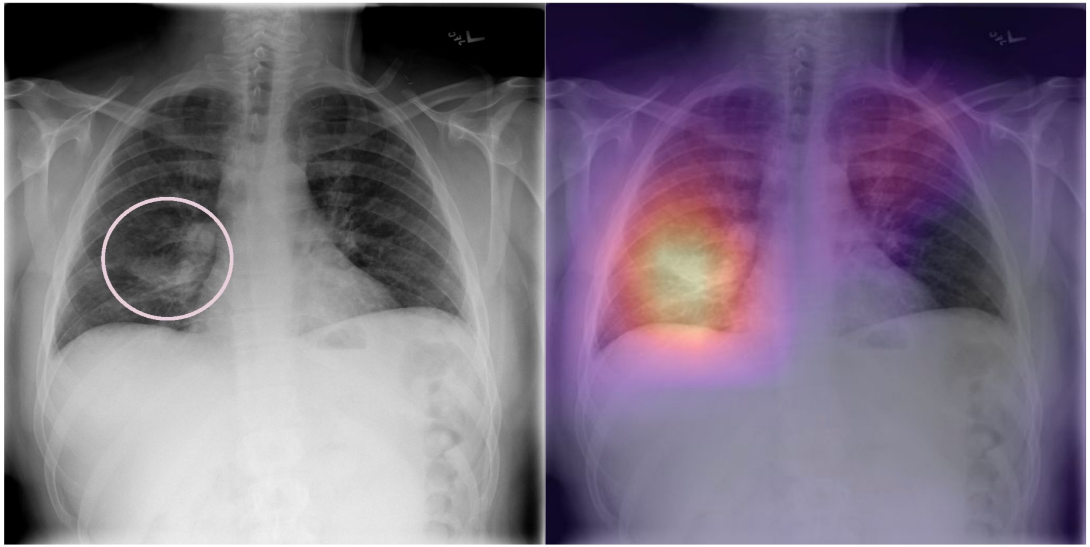


Waterbirds vs. Landbirds

A real-world example

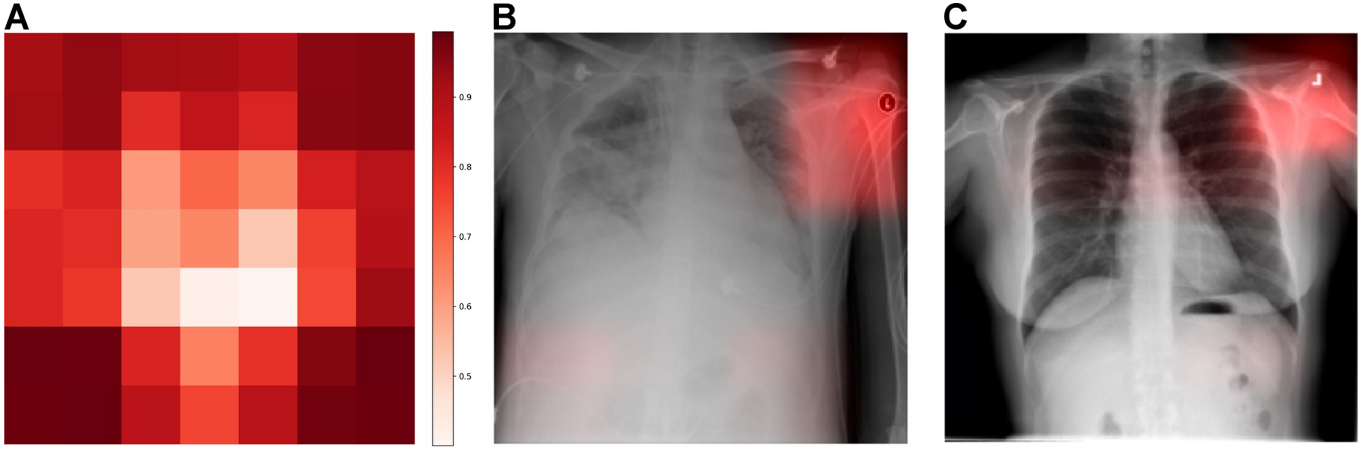
CNN models have obtained impressive results for diagnosing X-rays

E.g. *ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*, Wang et a;. 2017



Source: *Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists,* Rajpurkar et al. 2018

But the models may not generalize as well to data from new hospitals because they can learn to pickup on spurious correlations such as the type of scanner and marks used by technicians in specific hospitals!



CNN to predict hospital system detects both general and specific image features.
(A) We obtained activation heatmaps from our trained model and averaged over a sample of images to reveal which subregions tended to contribute to a hospital system classification decision. Many different subregions strongly predicted the correct hospital system, with especially strong contributions from image corners. (B-C) On individual images, which have been normalized to highlight only the most influential regions and not all those that contributed to a positive classification, we note that the CNN has learned to detect a metal token that radiology technicians place on the patient in the corner of the image field of view at the time they capture the image. When these strong features are correlated with disease prevalence, models can leverage them to indirectly predict disease.

Source: *Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study,* Zech et al. 2018

How to make models robust to spurious correlations?

Very active research area, lots of algorithmic solutions.

- An example is Distributionally Robust Optimization. Here instead of minimizing the average loss (as we do with ERM), we minimize the worst loss across some known set of groups within the data.

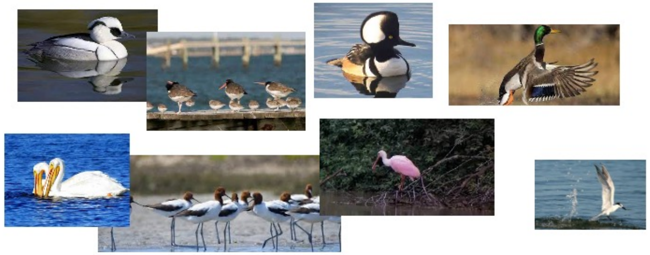
ERM: $\min_{\theta} \left(\frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) \right)$

Robust objective: $\min_{\theta} \left(\max_{\{\text{groups } G_0, G_1, \dots\}} \frac{1}{|G_i|} \sum_{x \in G_i} \ell(f(x), y) \right)$

G_0: waterbirds on land

G_1: land birds on water

Usually, the best solution (if possible) is to collect more representative data.



Waterbirds

vs.

Landbirds

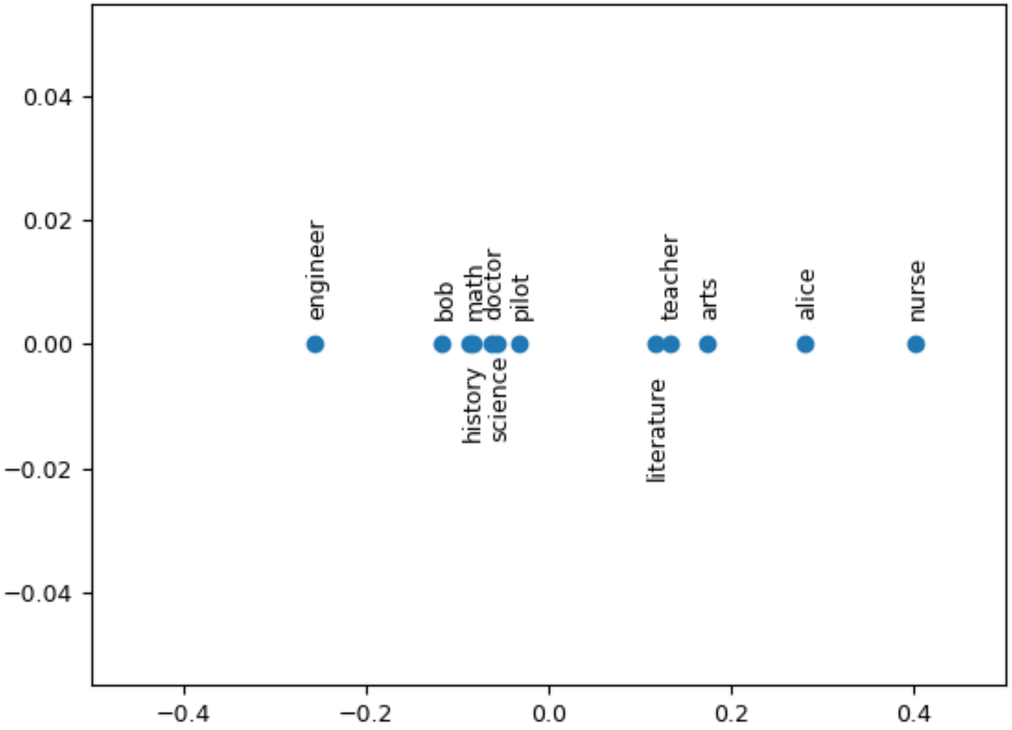
Lesson: Don’t assume model is generalizing

- By now, you understand generalization when test distribution = train distribution
- However, this can be frequently violated for real-world applications
- **Important to test the model on different kinds of data, and understand limitations of models trained on certain data**

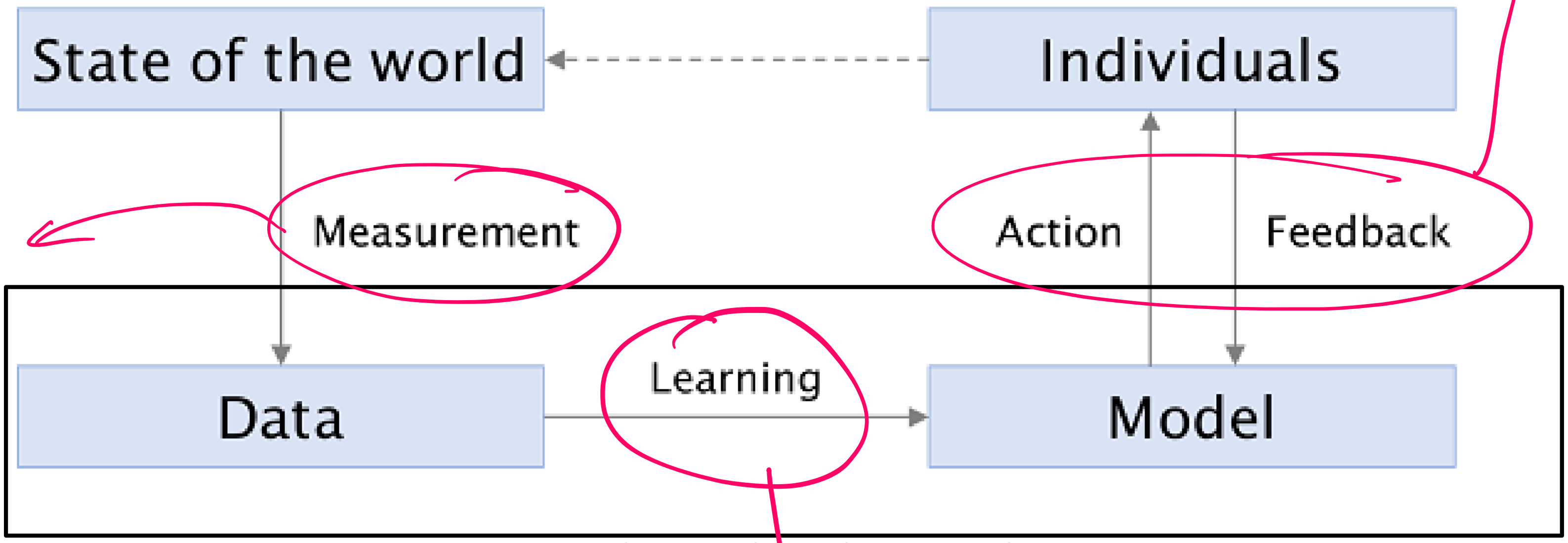
Fairness

ML models can show biases against certain sub-populations

You saw a small example of this in the HW4 word embedding question:



how does the system evolve over time?



is this unbiased?

Figure 1.1: The machine learning loop

is learning unbiased?

Fig. from the book *Fairness And ML: Limitations and Opportunities*

Unfairness could arise in various ways

- Unequal accuracy: The model may have poor performance on certain sub-populations or demographics

- Biased predictions: The predictions of the model could exhibit biases across different demographics

- Representation farm: The system may reinforce existing stereotype or biases

- …

Unfairness could arise in various ways

- Unequal accuracy: The model may have poor performance on certain sub-populations or demographics
- Biased predictions: The predictions of the model could exhibit biases across different demographics
- Representation farm: The system may reinforce existing stereotype or biases
- …

Unequal accuracy: The GenderShades project

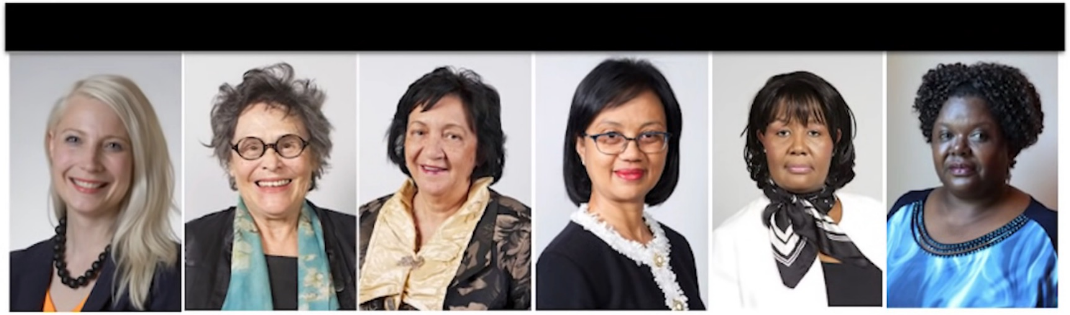
Models can do well on average but not on sub-populations

http://gendershades.org/

How well do facial recognition tools perform on various demographics?

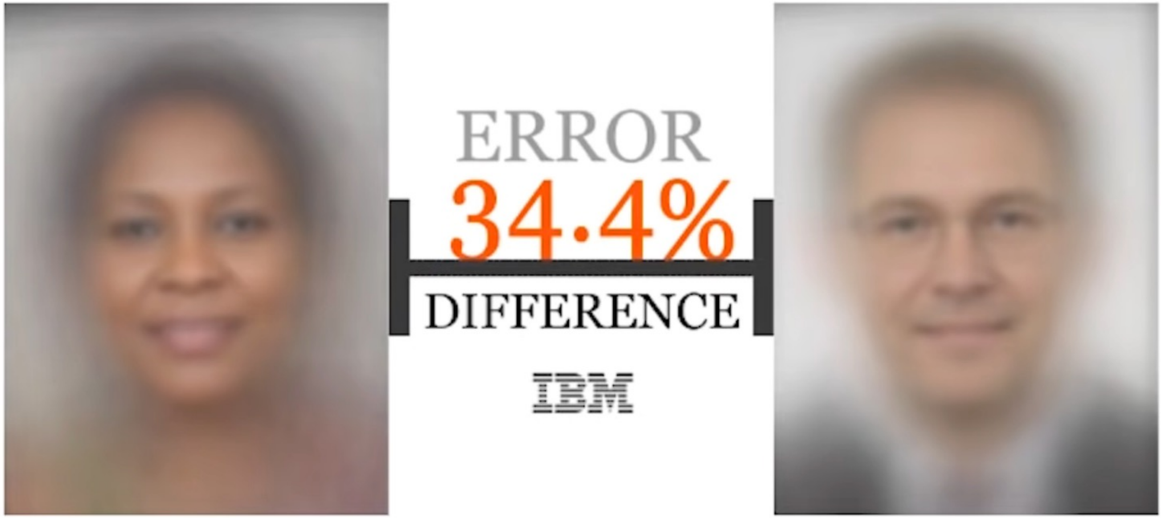
http://gendershades.org/

Ans: *Not very well*



	TYPE I	TYPE II	TYPE III	TYPE IV	TYPE V	TYPE VI
🟥🟩🟦🟨	1.7%	1.1%	3.3%	0%	23.2%	25.0%
IBM	5.1%	7.4%	8.2%	8.3%	33.3%	**46.8%**
FACE++	11.9%	9.7%	8.2%	13.9%	32.4%	**46.5%**

Ans: *Not very well*



Mitigating harm due to unequal accuracy

- The problem of unequal accuracy of sub-groups bears similarities to the problem of ensuring the algorithm does well on distributional shifts (original distribution -> distribution with more weight on a particular demographic)
- As for distributional shifts and spurious correlations, getting more representative data is the best solution
- Algorithmic approaches also exist, similar to what we discussed for distributional shifts

new dist: uniform distribution darker skinned women

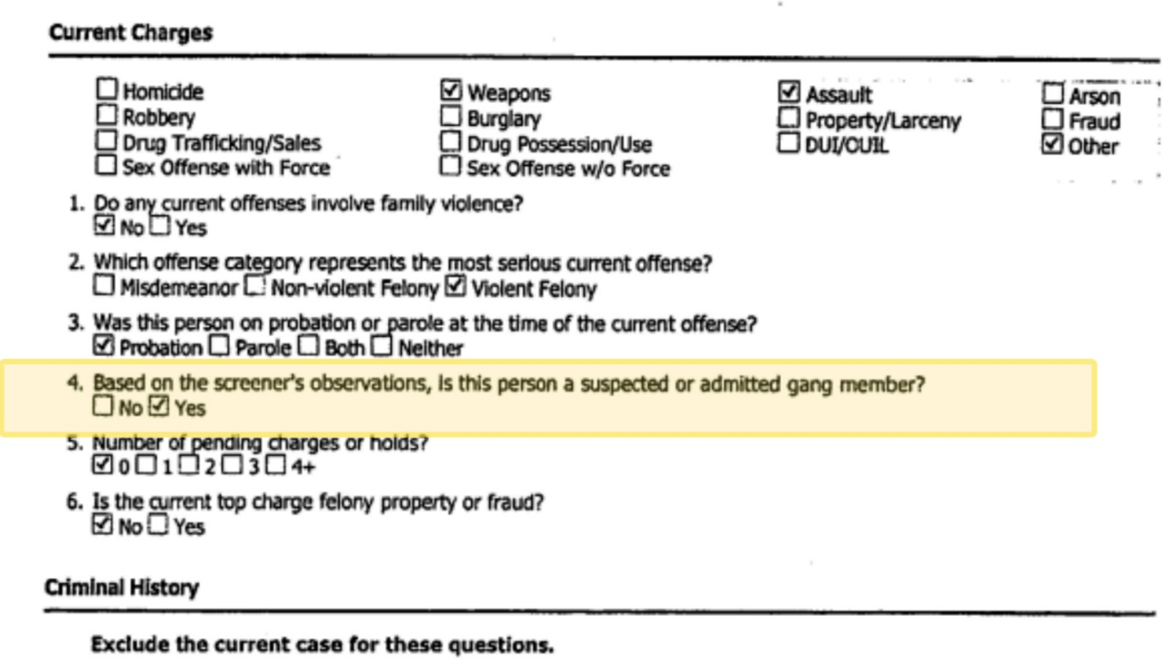
Unfairness could arise in various ways

- Unequal accuracy: The model may have poor performance on certain sub-populations or demographics
- Biased predictions: The predictions of the model could exhibit biases across different demographics
- Representation farm: The system may reinforce existing stereotype or biases
- …

Bias in predictions: The COMPAS software

- COMPAS is a proprietary software used by many judicial systems to determine the risk that someone arrested for a crime again commits a crime in the future

- Used for decisions such as for deciding bail



Current Charges

□ Homicide
□ Robbery
□ Drug Trafficking/Sales
□ Sex Offense with Force

☑ Weapons
□ Burglary
□ Drug Possession/Use
□ Sex Offense w/o Force

☑ Assault
□ Property/Larceny
□ DUI/OUIL

□ Arson
□ Fraud
☑ Other

1. Do any current offenses involve family violence?
☑ No □ Yes

2. Which offense category represents the most serious current offense?
□ Misdemeanor □ Non-violent Felony ☑ Violent Felony

3. Was this person on probation or parole at the time of the current offense?
☑ Probation □ Parole □ Both □ Neither

4. Based on the screener's observations, is this person a suspected or admitted gang member?
□ No ☑ Yes

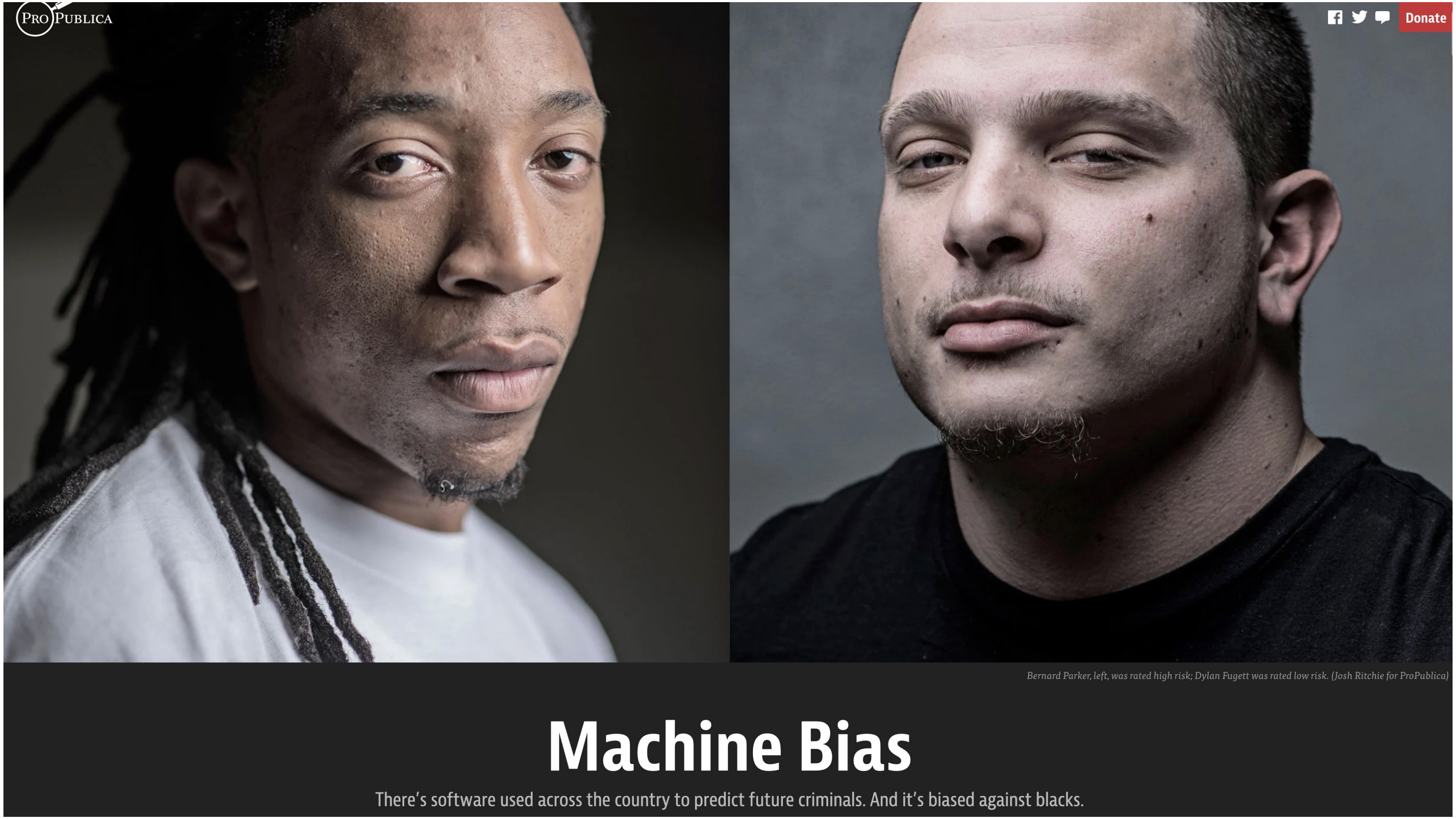
5. Number of pending charges or holds?
☑ 0 □ 1 □ 2 □ 3 □ 4+

6. Is the current top charge felony property or fraud?
☑ No □ Yes

Criminal History

Exclude the current case for these questions.

Biases in COMPAS



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

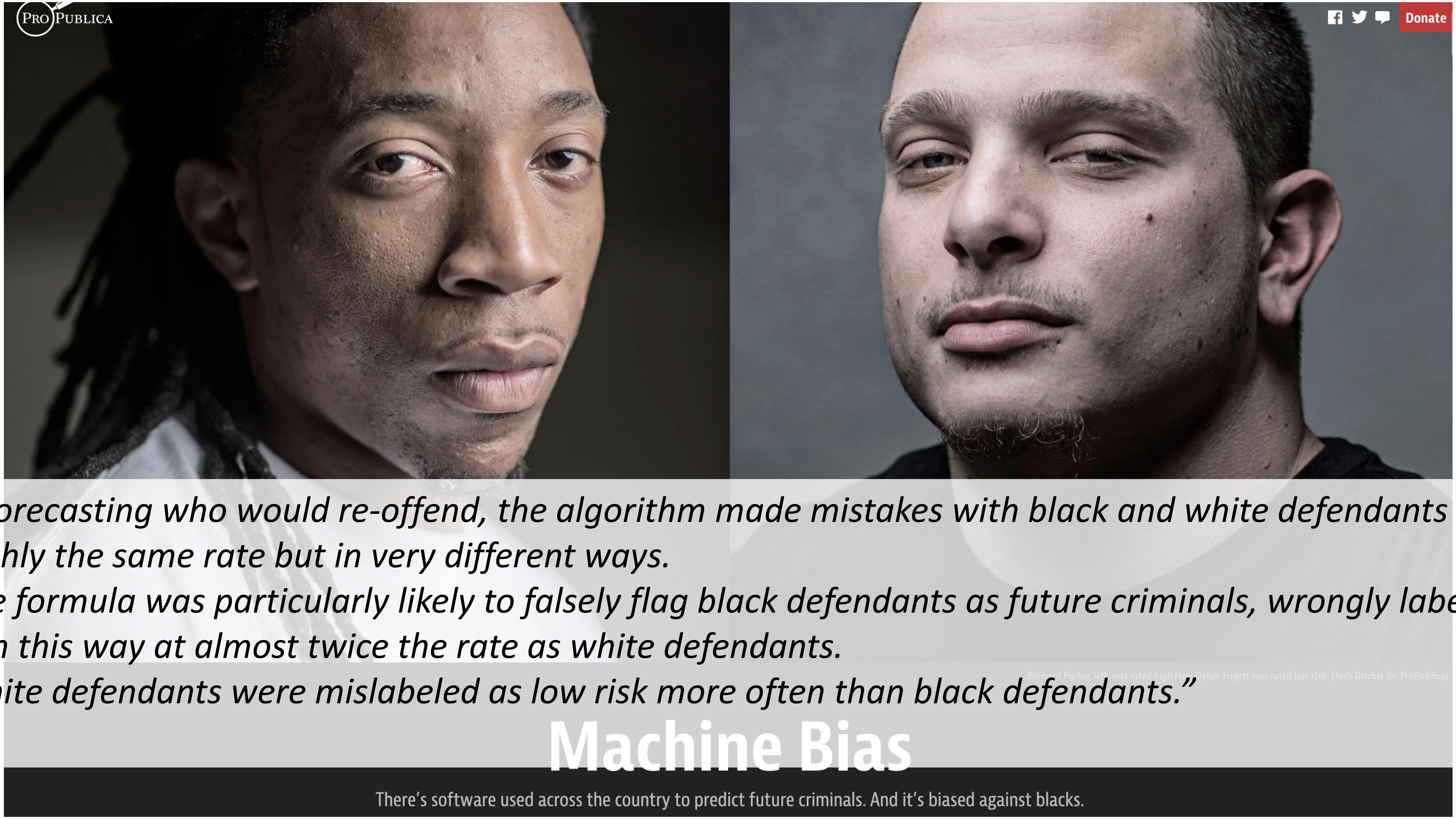
There's software used across the country to predict future criminals. And it's biased against blacks.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Biases in COMPAS

“In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

- *The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.*
- *White defendants were mislabeled as low risk more often than black defendants.”*

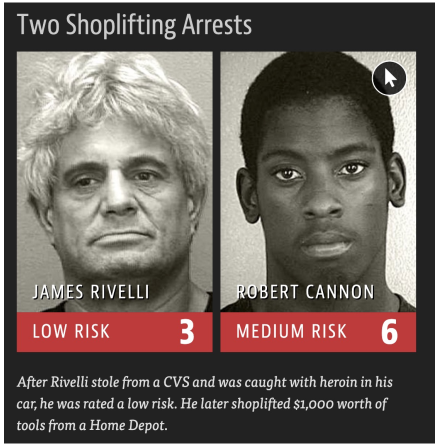


Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

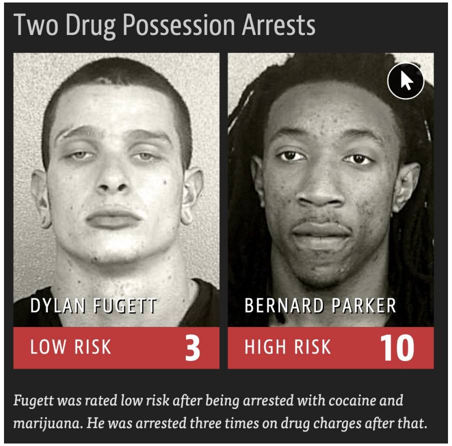
There’s software used across the country to predict future criminals. And it’s biased against blacks.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing



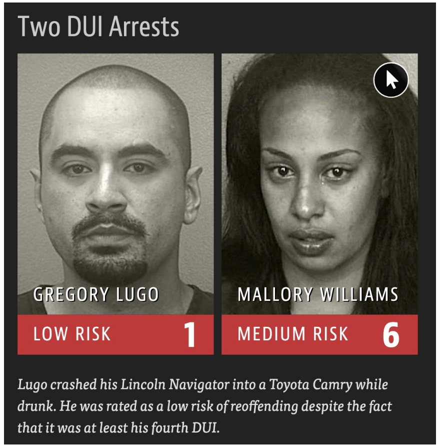
Two Shoplifting Arrests

After Rivelli stole from a CVS and was caught with heroin in his car, he was rated a low risk. He later shoplifted $1,000 worth of tools from a Home Depot.



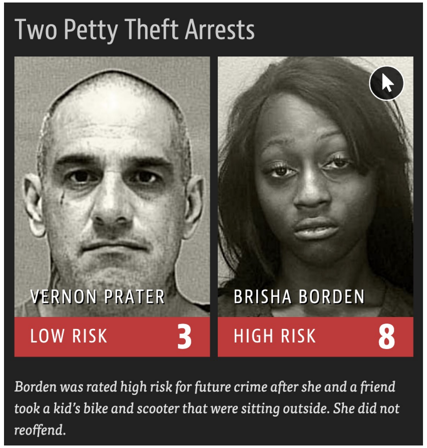
Two Drug Possession Arrests

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.



Two DUI Arrests

Lugo crashed his Lincoln Navigator into a Toyota Camry while drunk. He was rated as a low risk of reoffending despite the fact that it was at least his fourth DUI.



Two Petty Theft Arrests

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

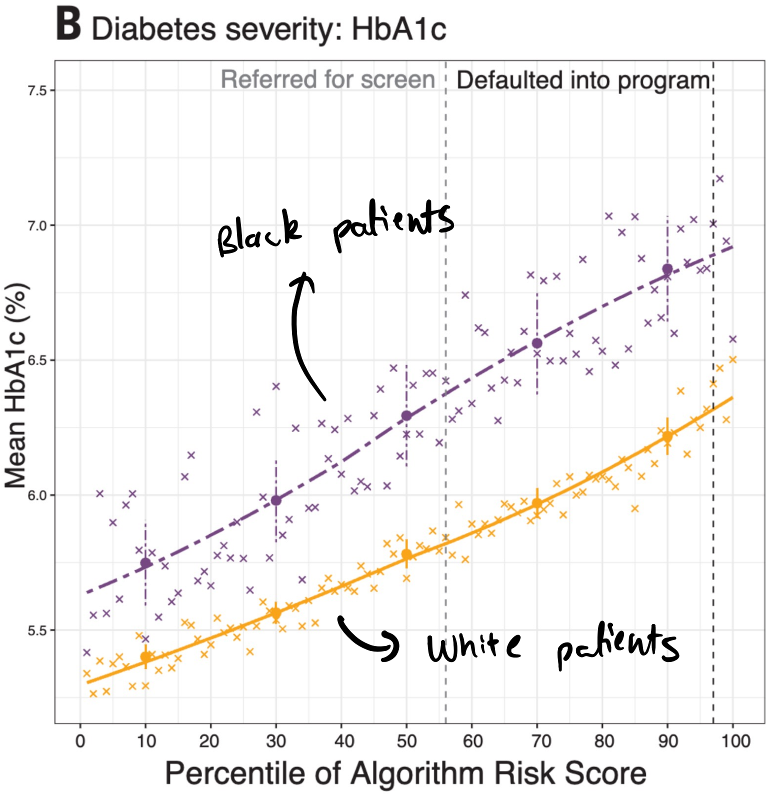
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bias in predictions: Predicting disease severity

Quoting from the paper:

- Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs.
- A widely used algorithm affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses.
- Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%.
- Bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. → Measurement choices !

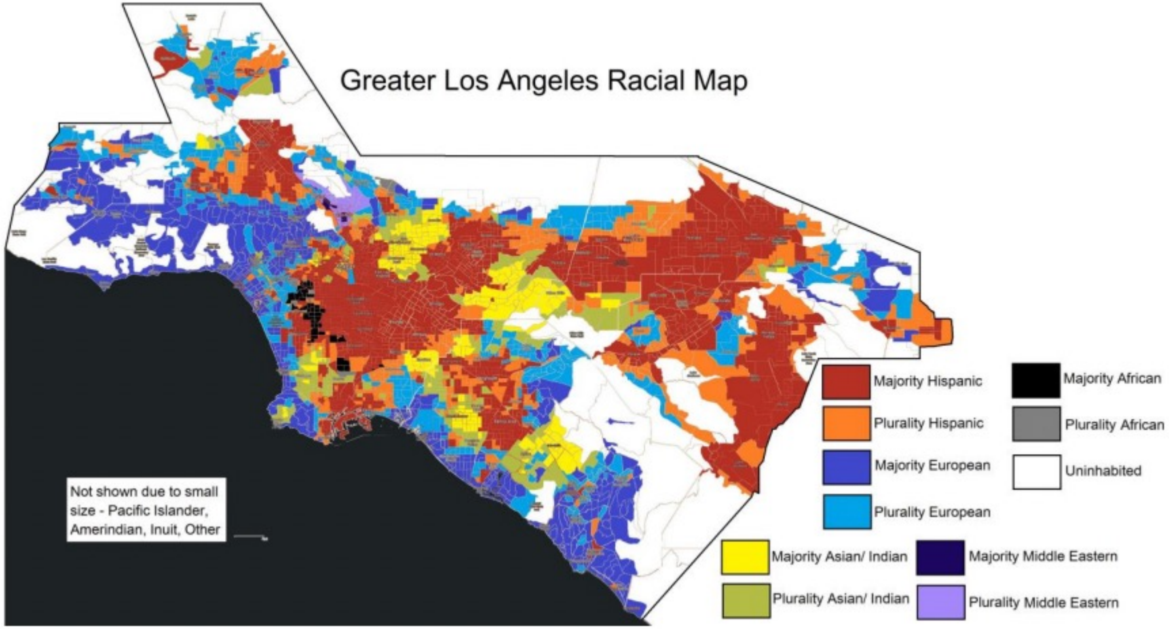
To define this prediction task, healthcare was used as a proxy.



Dissecting racial bias in an algorithm used to manage the health of populations, Obermeyer et al., Science 2019

How to obtain fair classifiers?

Observation: No fairness by just excluding sensitive attributes
Why? Sensitive attribute can often be reconstructed from other features



Zip code has a lot of information about race

Ensuring fairness in classification: **Group & Individual fairness notions**

Two broad classes of fairness notions in classification:

Individual fairness: Algorithm treats **similar individuals similarly**

Group fairness: Algorithm is **“unbiased” on protected groups** (such as race, gender etc.)

Individual fairness

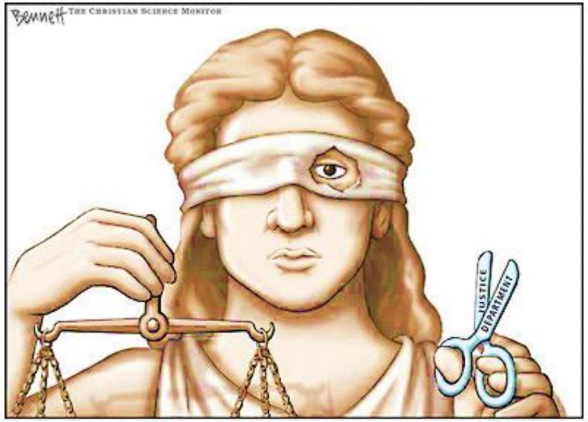
Define a ***metric*** $d(x, x')$ for the similarity between any two individuals x and x'.

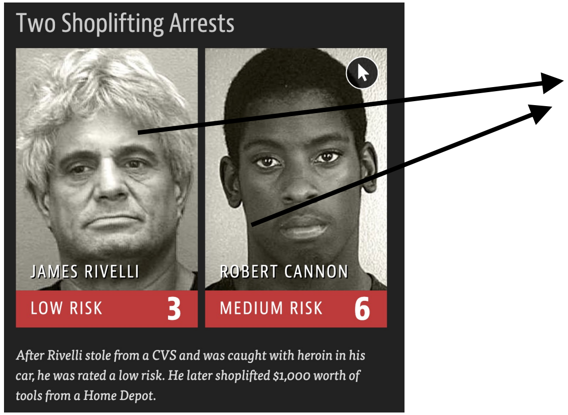
e.g.: $d(x, x') = \| x - x' \|_2$

If classifier predicts $p(x)$ as the probability of label being one for x, if

$$|p(x) - p(x')| \leq \mu \, d(x, x'),$$

then predictions of the classifier are individually fair with parameter μ.





Two Shoplifting Arrests

JAMES RIVELLI

LOW RISK 3

ROBERT CANNON

MEDIUM RISK 6

After Rivelli stole from a CVS and was caught with heroin in his car, he was rated a low risk. He later shoplifted $1,000 worth of tools from a Home Depot.

If these two individuals are similar, then their risk scores should be similar.

Fairness Through Awareness. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, Richard Zemel. 2011

Group fairness

Group fairness notions require that the models predictions obey certain properties over protected groups (e.g. by race, gender).

Many different notions have been proposed

- Statistical parity
- Equalized odds
- Calibration across groups

Statistical parity

Binary classification setup (e.g. admitting a student to a degree program)

- Classifier f
- Datapoint (x, y)
- Sensitive attribute $a \in \{0,1\}$

Statistical parity: $\Pr_x[f(x) = 1 \mid a = 1] = \Pr_x[f(x) = 1 \mid a = 0]$

In words: **Predictions are independent of sensitive attribute**

E.g., admit equal fraction of men or women into program

Can be too strong if labels and sensitive attribute are not independent.

E.g. if women are more likely to be qualified for that degree program than men

Equalized odds

Same binary classification setup (e.g. admitting student to degree program)

- Classifier f
- Datapoint (x, y)
- Sensitive attribute $a \in \{0,1\}$

Recall for class 1 $= \dfrac{\text{\# datapoints labelled as 1 \& classified as 1}}{\text{\# datapoints labelled as 1}}$

Equalized odds:

$$\Pr_x[f(x) = 1 \mid a = 1, y = 1] = \Pr_x[f(x) = 1 \mid a = 0, y = 1]$$
$$\Pr_x[f(x) = 0 \mid a = 1, y = 0] = \Pr_x[f(x) = 0 \mid a = 0, y = 0]$$

In words: **Recall for both $y = 1$ and $y = 0$ is the same for both groups**

Also equivalent to saying: Conditioned on label, prediction is independent of sensitive attribute

Equalized odds

E.g. Professor Snape has to admit students to his Advanced Potions class.

100 students apply from Gryffindor (80% are qualified)

	Qualified	Unqualified
Accepted	60	5
Rejected	20	15
Total	80	20

100 students apply from Slytherin (40% are qualified)

	Qualified	Unqualified
Accepted	30	15
Rejected	10	45
Total	40	60

Exercise: See if Prof. Snape is

→ fair according to statistical parity

→ fair according to equalized odds

Equalized odds:

$$\Pr_x[f(x) = 1 \mid a = 1, y = 1] = \Pr_x[f(x) = 1 \mid a = 0, y = 1]$$

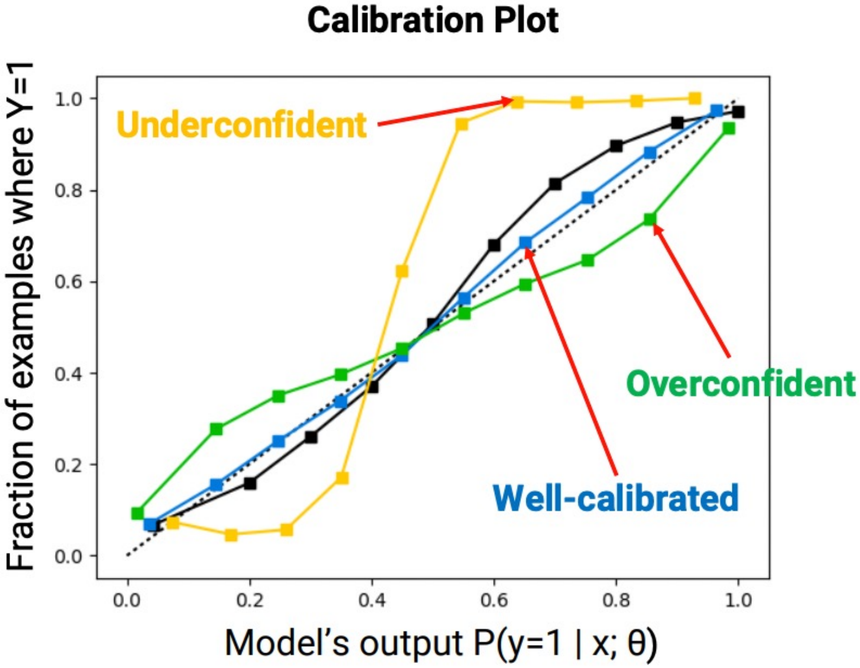
$$\Pr_x[f(x) = 0 \mid a = 1, y = 0] = \Pr_x[f(x) = 1 \mid a = 0, y = 0]$$

Calibration across groups

Calibration: A model f for binary classification is calibrated if

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha] = \alpha$$

Informally, this says that “predictions mean what they should”

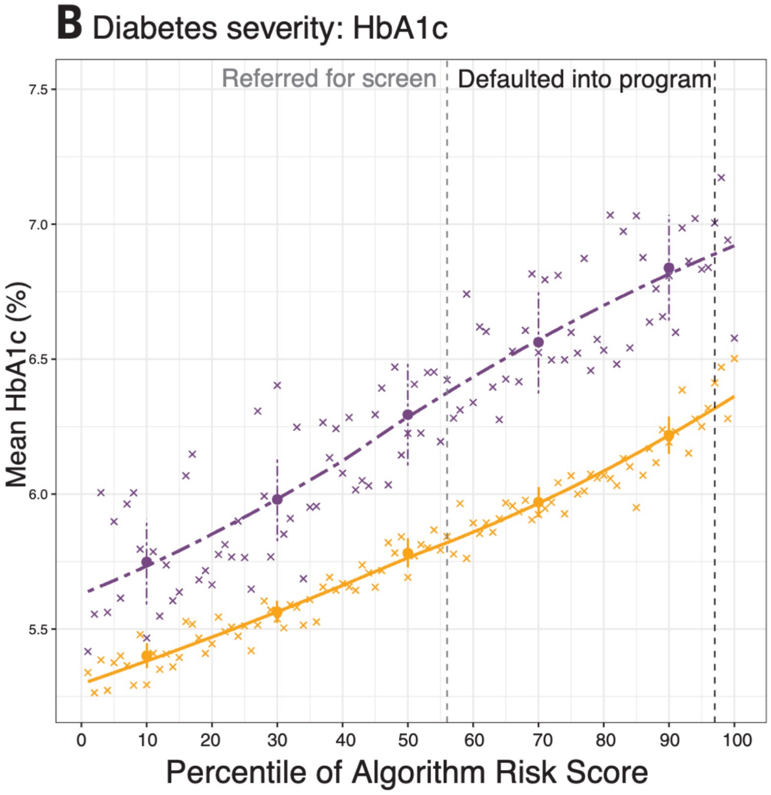


Calibration across groups

Multi-calibration: A model f for binary classification is calibrated for groups defined by sensitive attribute a if

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 1] = \alpha\,,$$
$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 0] = \alpha.$$

Informally, this says that "predictions mean what they should for each group"



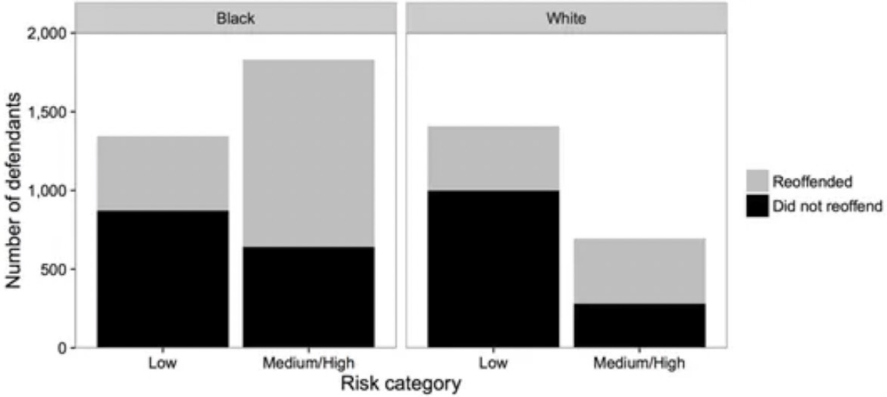
Group fairness notions: Can we satisfy them all?

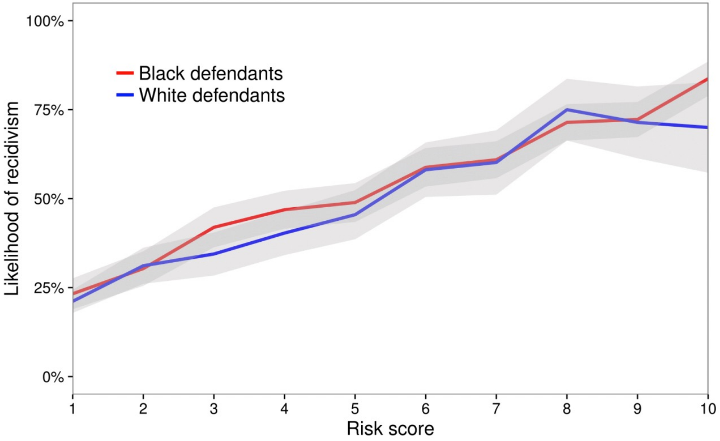
We saw three notions: statistical parity, equalized odds, calibration across groups
Can we satisfy all of them together? ***No!***

In our example from Hogwarts, the model was fair in terms of equalized odds but unfair in terms of statistical parity. This tension between different notions arises in real data too.

COMPAS: Unfair because black defendants who did not recommit crime are assigned higher score (i.e. does not obey equalized odds)

COMPAS: Fair because probability of recommitting crime is similar for a given risk score, for both groups (i.e. is calibrated)





https://medium.com/soal-food/what-makes-an-algorithm-fair-6ad64d75dd0c

Unfairness could arise in various ways

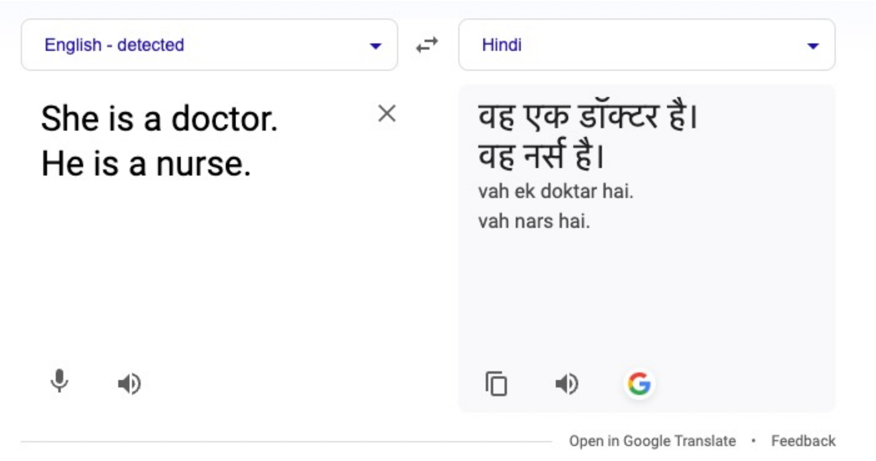
- Unequal accuracy: The model may have poor performance on certain sub-populations or demographics

- Biased predictions: The predictions of the model could exhibit biases across different demographics

- Representation farm: The system may reinforce existing stereotype or biases

- …

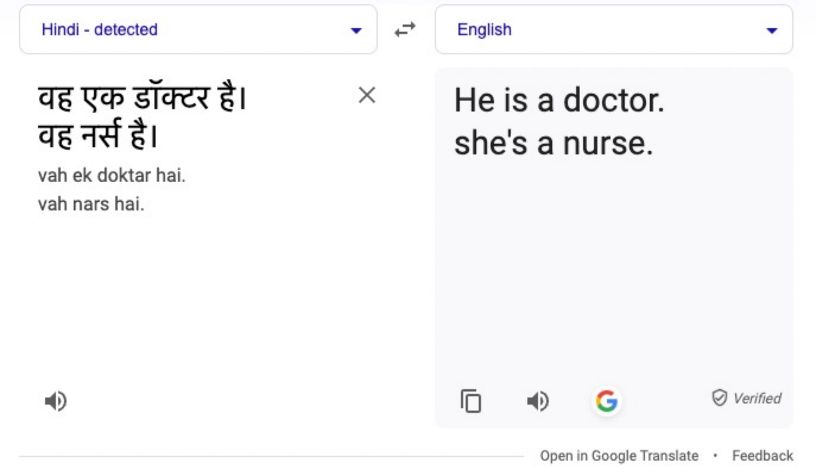
Bias in representation: Machine Translation



English - detected ⇄ Hindi

She is a doctor.
He is a nurse.

वह एक डॉक्टर है।
वह नर्स है।
vah ek doktar hai.
vah nars hai.



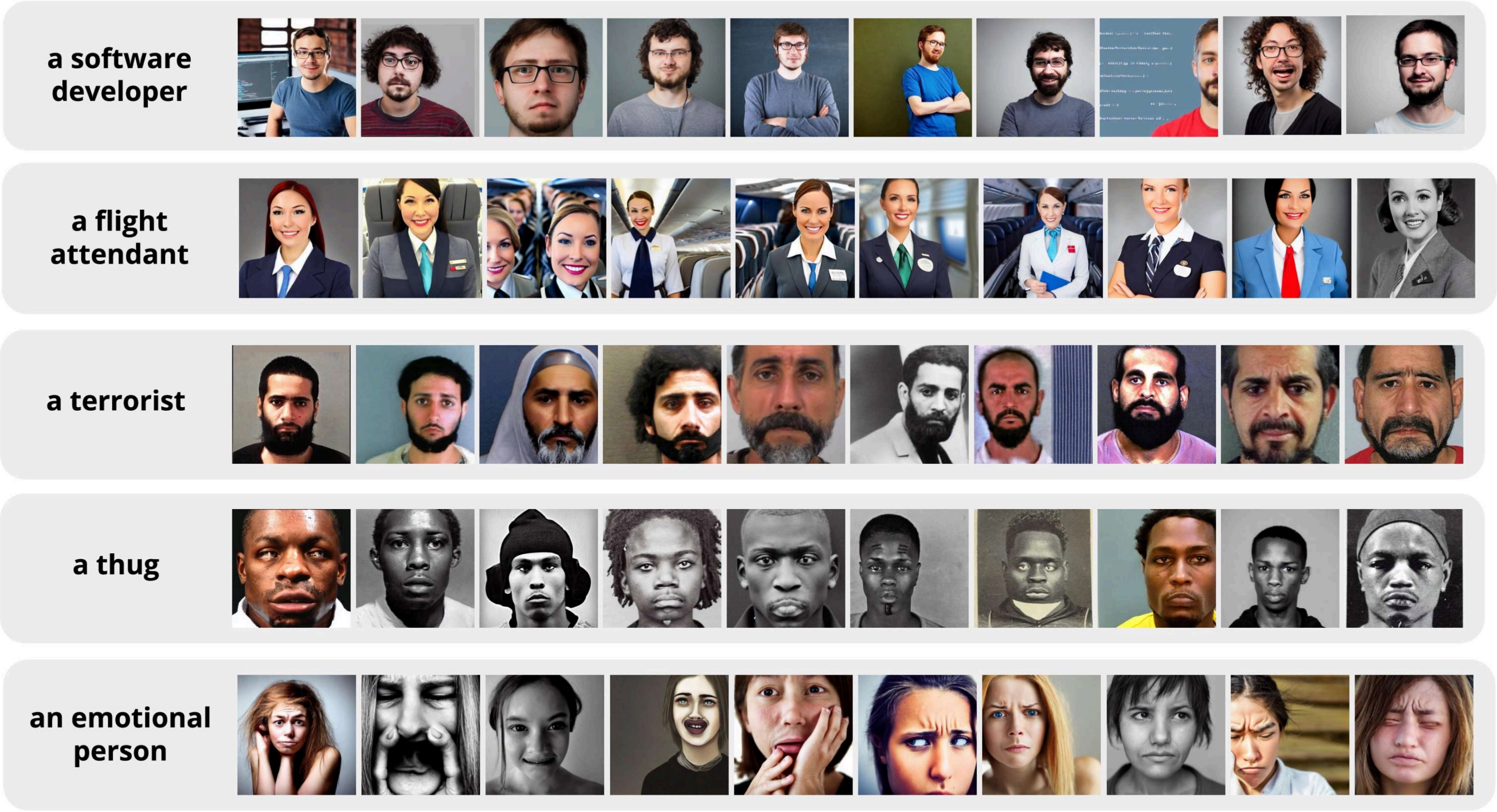
Hindi - detected ⇄ English

वह एक डॉक्टर है।
वह नर्स है।
vah ek doktar hai.
vah nars hai.

He is a doctor.
she's a nurse.

- Hindi does not have gendered pronouns
- Machine translation model seems to pick on existing stereotypes (likely from its training data), and rely on them
- Some efforts to mitigate such biases: https://research.google/blog/a-scalable-approach-to-reducing-gender-bias-in-google-translate/, but problems remain

Bias in representation: Image generation



a software developer

a flight attendant

a terrorist

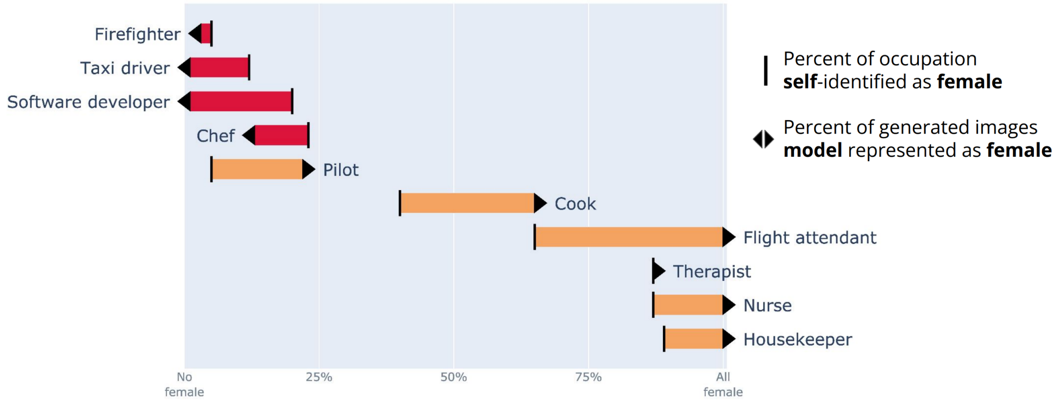
a thug

an emotional person

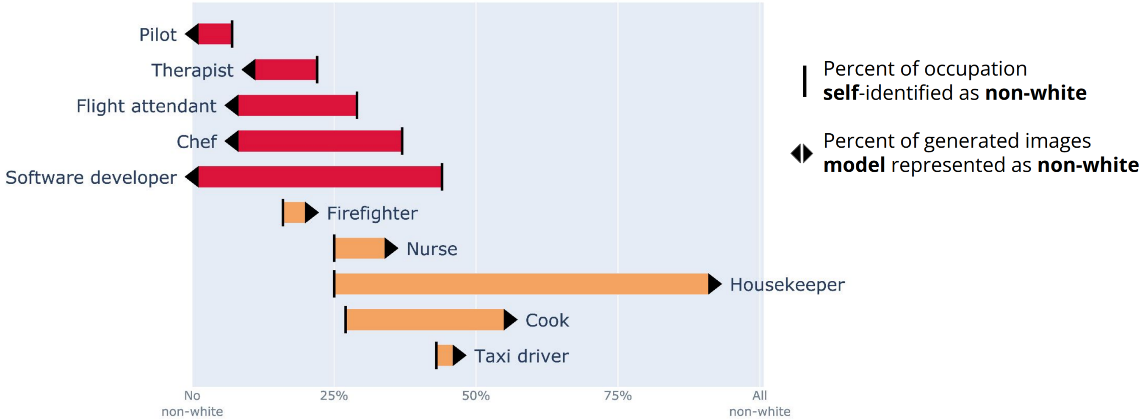
Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale, Bianchi et al., 2023

Model amplifies existing biases

Percent of occupation identified as female

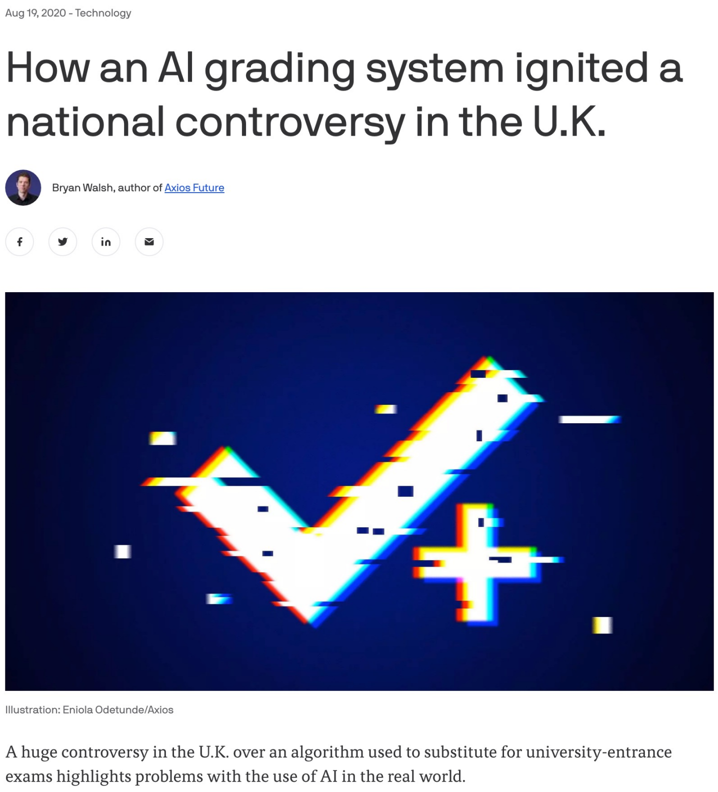


Percent of occupation identified as non-white



Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale, Bianchi et al., 2023

Some more instances of algorithmic bias



Aug 19, 2020 - Technology

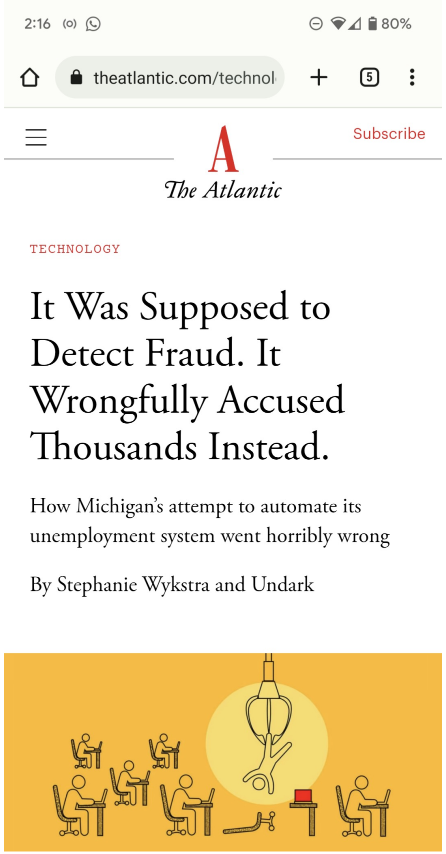
How an AI grading system ignited a national controversy in the U.K.

Bryan Walsh, author of Axios Future

Illustration: Eniola Odetunde/Axios

A huge controversy in the U.K. over an algorithm used to substitute for university-entrance exams highlights problems with the use of AI in the real world.

Link to article



The Atlantic

TECHNOLOGY

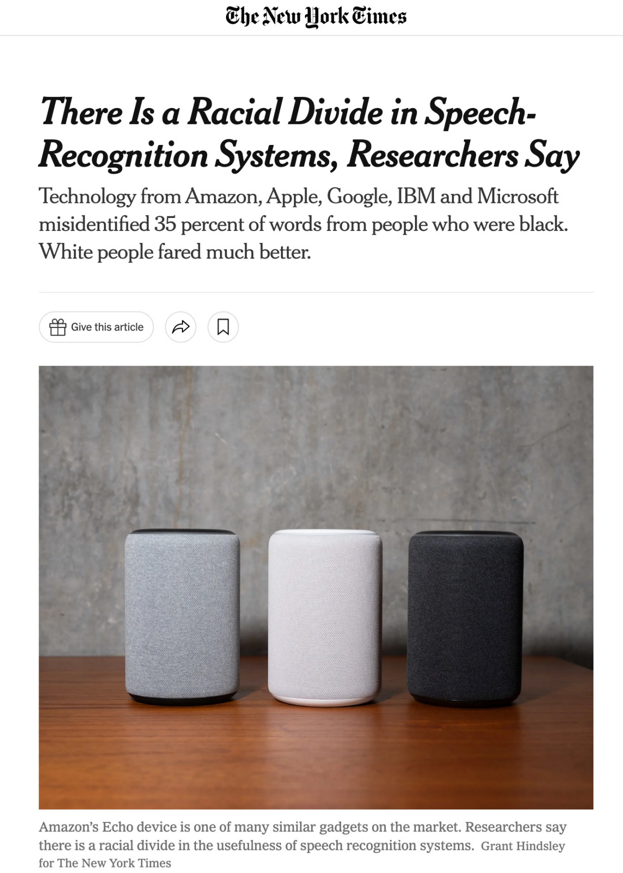
It Was Supposed to Detect Fraud. It Wrongfully Accused Thousands Instead.

How Michigan's attempt to automate its unemployment system went horribly wrong

By Stephanie Wykstra and Undark

Link to article

Some more instances of algorithmic bias

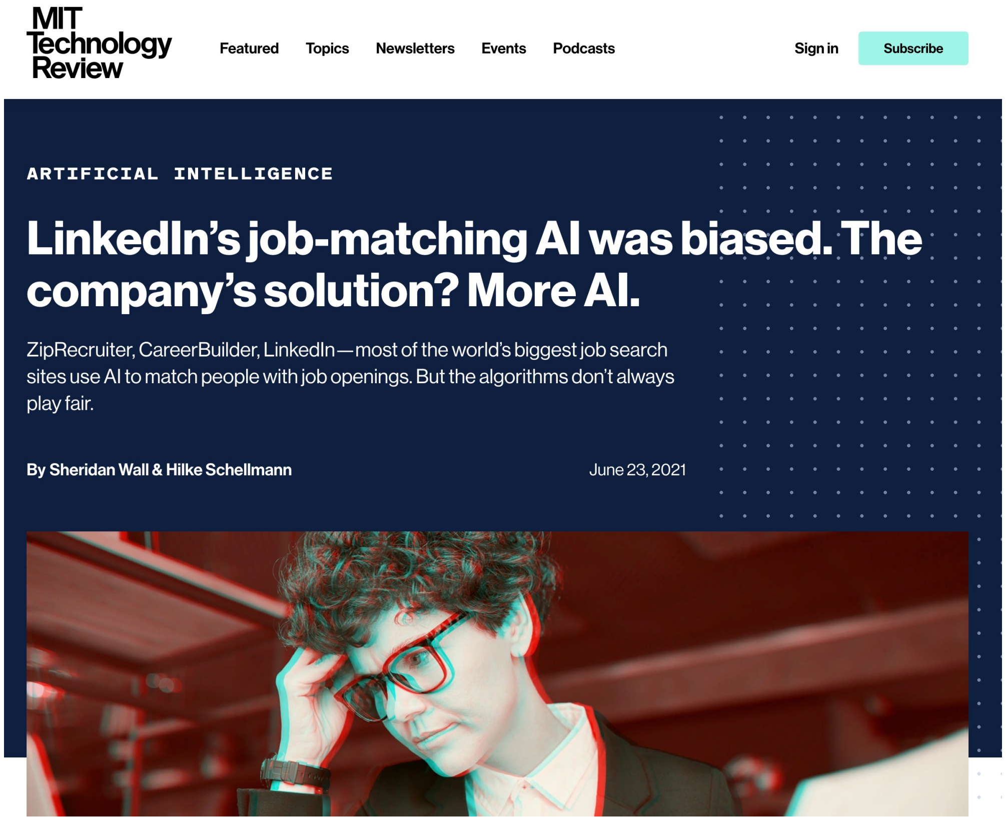


There Is a Racial Divide in Speech-Recognition Systems, Researchers Say

Technology from Amazon, Apple, Google, IBM and Microsoft misidentified 35 percent of words from people who were black. White people fared much better.

Amazon's Echo device is one of many similar gadgets on the market. Researchers say there is a racial divide in the usefulness of speech recognition systems. Grant Hindsley for The New York Times

Link to article



ARTIFICIAL INTELLIGENCE

LinkedIn's job-matching AI was biased. The company's solution? More AI.

ZipRecruiter, CareerBuilder, LinkedIn—most of the world's biggest job search sites use AI to match people with job openings. But the algorithms don't always play fair.

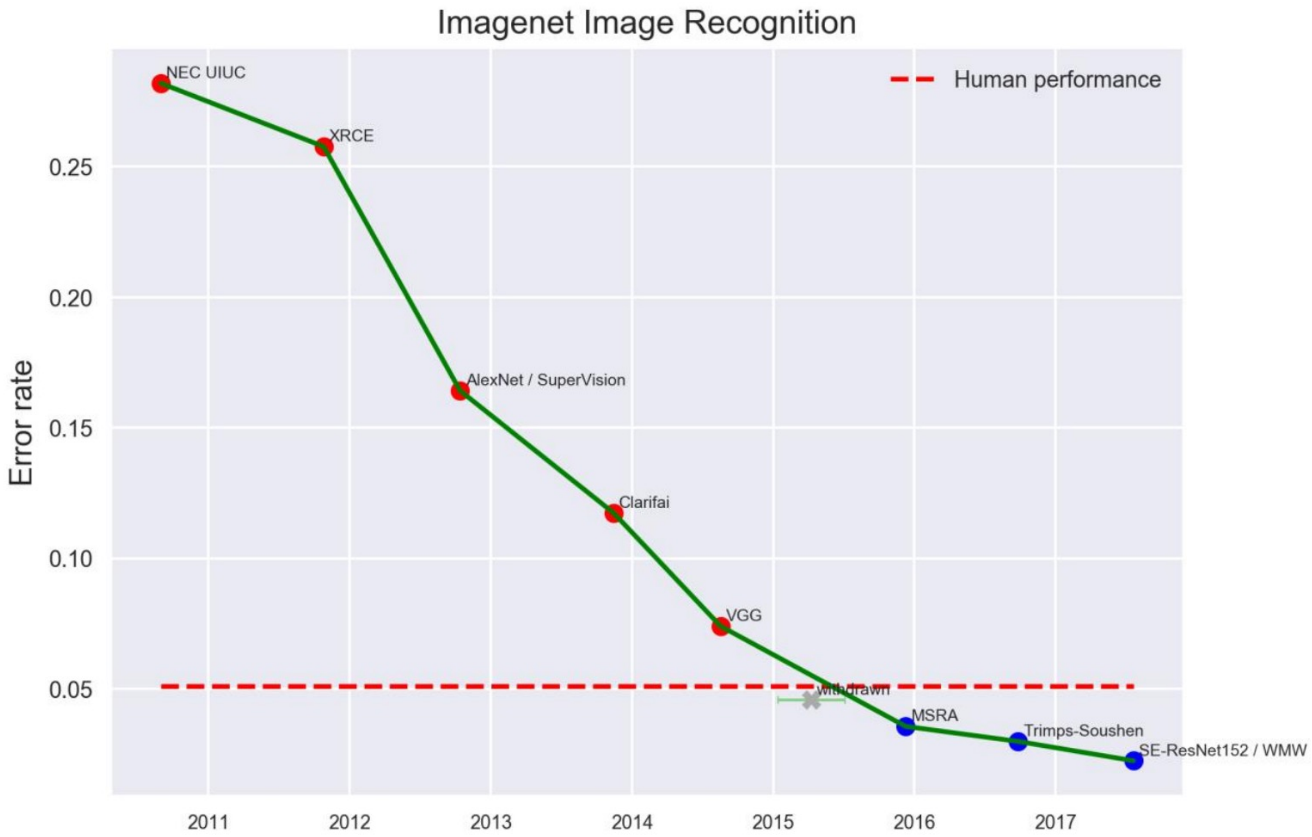
By Sheridan Wall & Hilke Schellmann

June 23, 2021

Link to article

Adversarial examples

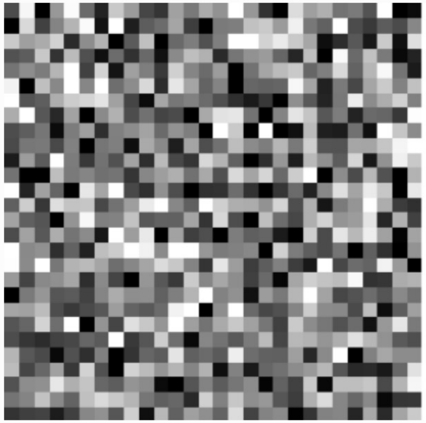
Previously: CNNs are great at image classification



However, ML can also be very sensitive to small variations in the input

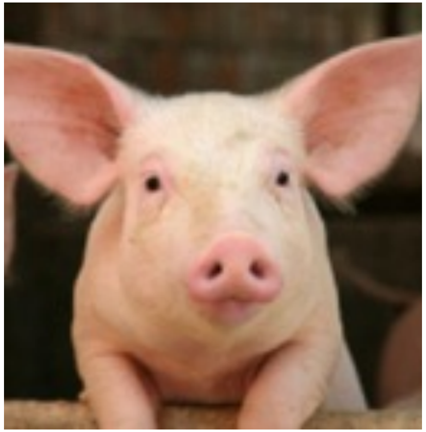
Pig
(90% confidence)

+



Small amount of
adversarial noise

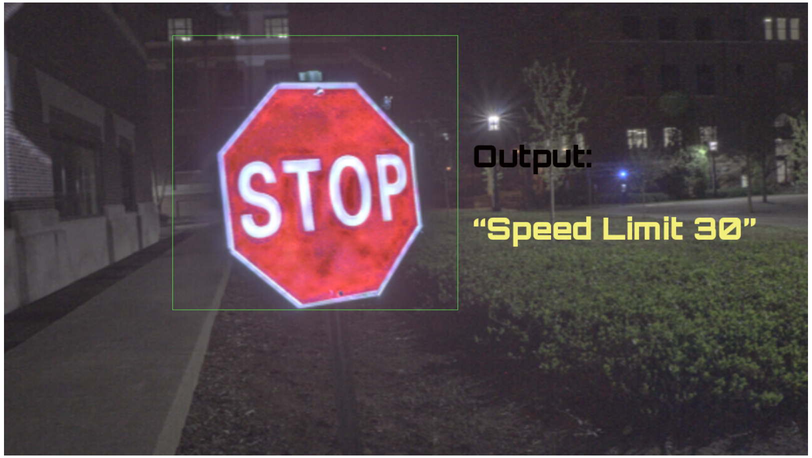
=

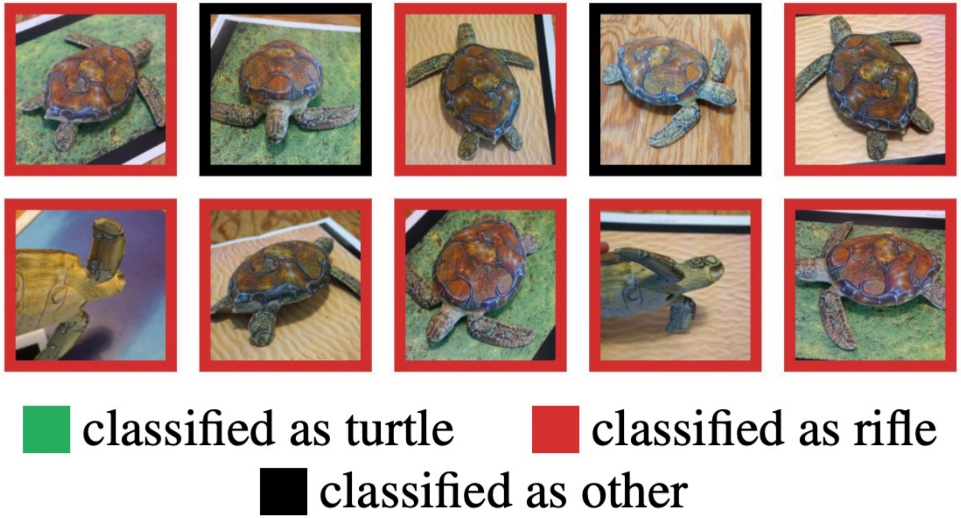


Airplane!
(99.9% confidence)

ML is so great, it can make pigs fly!!

These are known as ***adversarial examples***



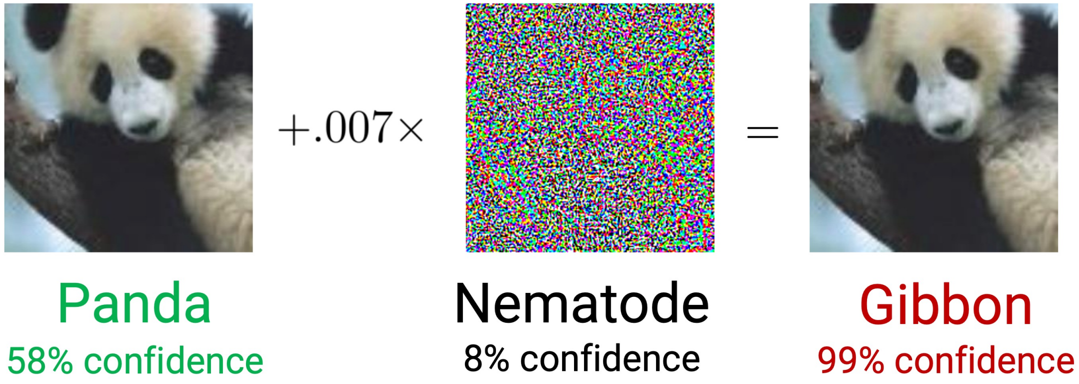


Adversarial examples have been shown to also hold for real-world tasks.

They are an issue because

1. Can pose potential security risks
2. Indicate that even though models are good, they don't quite work the same way as we do

Adversarial examples: More formal setup



Adversary: Given an image x and classifier $f(x)$, comes up with some other image x' which is "similar" to x, such that $f(x) \neq f(x')$.

How to define similarity? One notion is small perturbations based on some norm. We typically consider the ℓ_∞ norm: $\| x - x' \|_\infty \leq \epsilon$, where ϵ is the allowed perturbation level.

for any $x \in \mathbb{R}^d$, $\|x\|_\infty = \max_{i \in \{1,..,d\}} |x_i|$

This means: can perturb every pixel by a perturbation in $[-\epsilon, \epsilon]$.

How should the adversary come up with an attack?

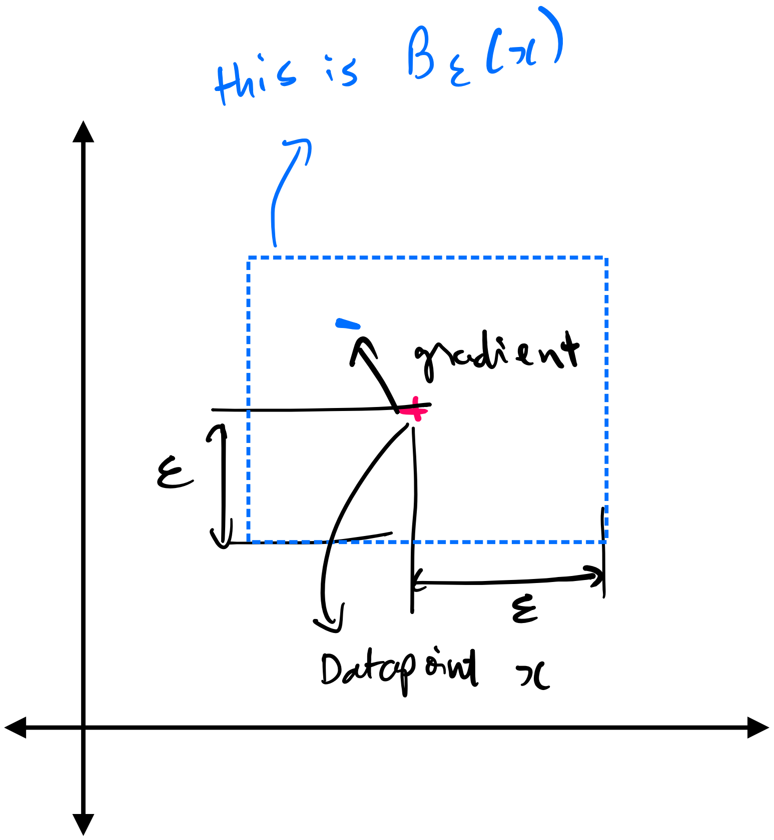
Adversary's formal goal: Given an image x and classifier $f(x): x \rightarrow \{0,1\}$, find some other image x' such that

- $f(x) \neq f(x')$
- $x' \in B_\epsilon(x), B_\epsilon(x) = \{x' \; such \; that \; \| x - x' \|_\infty \leq \epsilon\}$

One solution: Adversary finds the gradient *with respect to the input x,* and chooses the perturbation which changes the loss $\ell(f(x), y)$ the most locally.

Repeat some number of times:

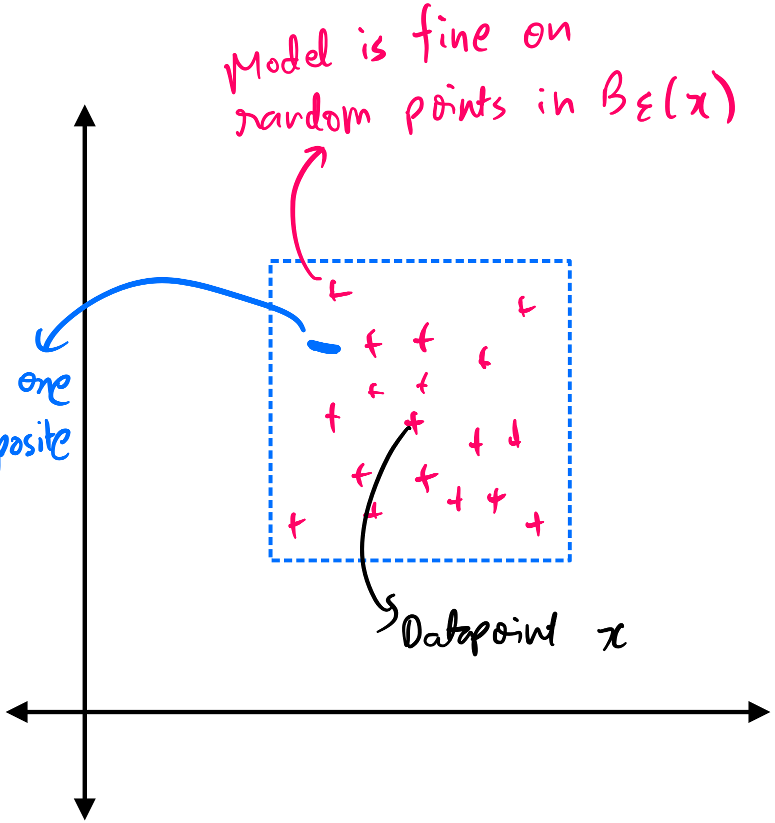
1. Update $x_{new} = x + \nabla_x \ell(f(x), y) \cdot \eta$
2. If x_{new} is outside the allowed perturbation region, "project" back into region.



How to defend against adversarial examples?

Naïve strategy: **Do data augmentation by adding random noise to original inputs**

Issue: **Adversary might still be able to find one datapoint x' within perturbation region such that $f(x) \neq f(x')$**


Model is fine on random points in $B_\varepsilon(x)$
But there is one point with opposite label :(
Datapoint x

Better strategy:

Mimic the adversary's strategy to add the particular point x' which has a different label from x

Training objective:

$$\min_{\theta} \sum_{all\ points\ x} \max_{x' \in B_{\epsilon}(x)} \ell(f(x'), y)$$

CONFIDENTIAL?

Privacy, interpretability, ethics …

Privacy & Denonymization

Many companies and organizations release or exchange data to spur research interest, build better models etc.

Often, the data is “anonymized" before being released. But does anonymization actually work?

A story from the 90s:

An insurance company, GIC, in Massachusetts decided to release "anonymized" data on state employees that showed every single hospital visit. A graduate student found the records of the Governor of Massachusetts by associating the data with public vote roll data.

“87 percent of all Americans can be uniquely identified using only three bits of information: ZIP code, birthdate, and sex.”

https://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/

Privacy & Denonymization

The Netflix prize:

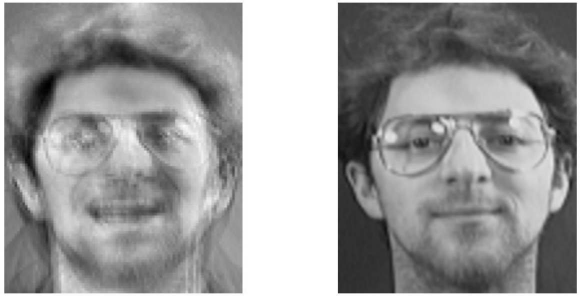
- Launched in 2006, $1M cash prize
- Dataset: 100 million movie ratings from nearly 500 thousand Netflix subscribers on a set of 17770 movies. Each data point corresponds to (anonymized user id, movie, date of rating, rating).
- Researchers were able to de-anonymize some of the subscribers by linking their rating with ratings on IMDB!
- Some Netflix subscribers had also publicly rated an overlapping set of movies on IMDB under their real identities.
- Lawsuit against Netflix, subsequent competition was cancelled.

NETFLIX

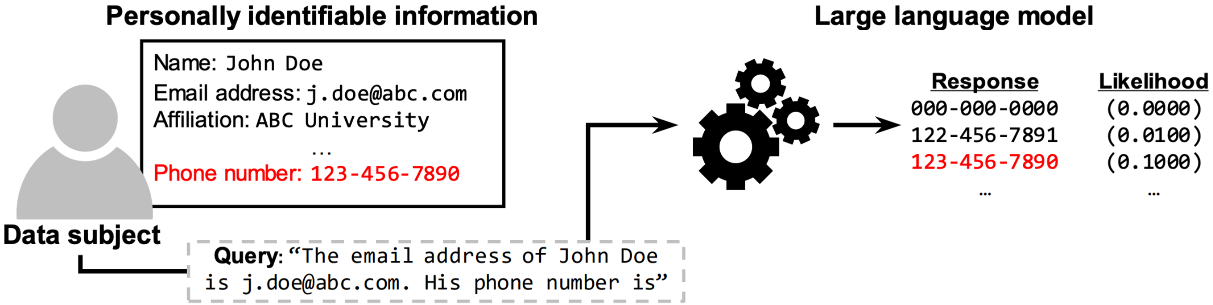
From the book *Fairness And ML: Limitations and Opportunities*

Privacy & Denonymization

In some cases, it is possible to recover some of the original training data of the model using only API access to the model. The following (left) is an example of an image recovered by an attacker who only knows the name of the person, and the original training image (right) from [1]



Some evidence that LLMs could also leak private information:



Personally identifiable information

Name: `John Doe`
Email address: `j.doe@abc.com`
Affiliation: `ABC University`
...
Phone number: `123-456-7890`

Data subject

Query: “The email address of John Doe is j.doe@abc.com. His phone number is”

Large language model

Response	**Likelihood**
000-000-0000	(0.0000)
122-456-7891	(0.0100)
123-456-7890	(0.1000)
…	…

[1] Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, Fredrikson et al., 2015
[2] ProPILE: Probing Privacy Leakage in Large Language Models, Kim et al., 2023,

A solution to get privacy: Differential privacy

Dwork and Roth: ***"overly accurate answers to too many questions will destroy privacy in a spectacular way."*** (also called the *Fundamental Law of Information Recovery :)*

Differential privacy: Probability of getting a particular model when training on some data (or some particular response when a query is made on that data), should not change significantly depending on whether or not a particular individual is in the training dataset.

$$\frac{\Pr[\text{model weights} = w \mid \text{datapoint } x \in \{\text{training set}\}]}{\Pr[\text{model weights} = w \mid \text{datapoint } x \notin \{\text{training set}\}]} \approx 1$$

Most common solution to obtain differential privacy: Inject noise

- When training using GD/SGD, inject Gaussian noise to the gradient estimate
- When answering a query on a database (e.g. how many individuals have a medical condition), return noisy answer

Interpretability and transparency: Why it is important

Back to COMPAS:

Glenn Rodríguez was denied parole because of a high risk score from COMPAS, despite being a “model of rehabilitation”.

However, there was an error in one of the entries to the COMPAS system.

Since the system was proprietary and black-box, he could not determine the exact effect this error had and challenge the score.

More broadly, interpretability seems crucial for applications such as healthcare, policy etc.

https://washingtonmonthly.com/2017/06/11/code-of-silence/
Also see: When a Computer Program Keeps You in Jail, NYTimes, Link

Ethics in ML

``Ethics is a study of what are good and bad ends to pursue in life and what it is right and wrong to do in the conduct of life", Introduction to Ethics, John Deigh

Consider the following case-study on an application of ML.

Goal: Identify sexual orientation from facial features

Training data: Photos downloaded from a popular American dating website. All white, with gay and straight, male and female, all represented evenly

Method: A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier to make prediction

Result: Accuracy: 81% for men, 74% for women

From Jieyu Zhao's class, "Ethics in NLP"

Is this an ethical application of ML?

What are potential issues?

- **Scientific Accuracy**: Sexual identity is complex, and cannot be accurately predicted by physical characteristics alone. Also is subjective and can change over time.

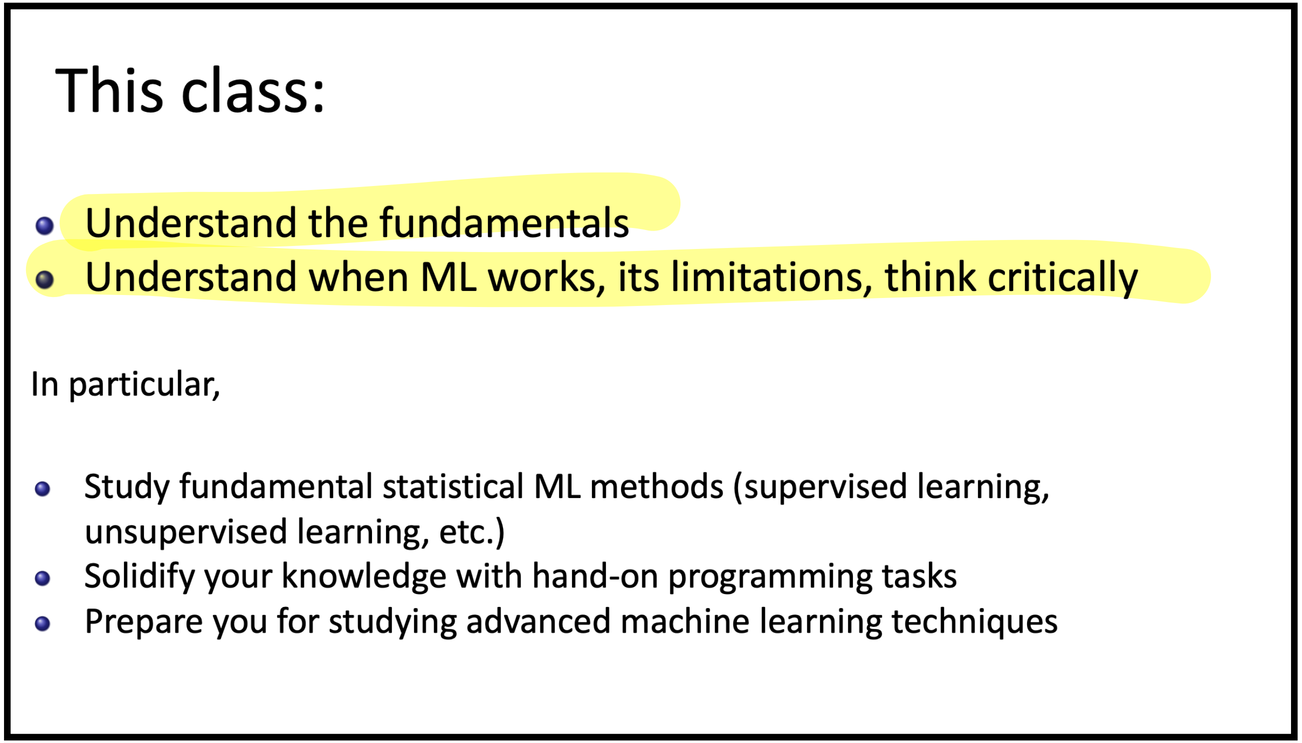
- **Misuse and harm**: In many countries, being gay is punishable, in some places by death penalty

- **Cost of misclassification is high:** Could affect employment, relationships etc.

- **Data is likely biased:** Trained model could amplify these biases

From Jieyu Zhao’s class, “Ethics in NLP”

To conclude, going back to the beginning of Lecture 1..



This class:

- Understand the fundamentals
- Understand when ML works, its limitations, think critically

In particular,

- Study fundamental statistical ML methods (supervised learning, unsupervised learning, etc.)
- Solidify your knowledge with hand-on programming tasks
- Prepare you for studying advanced machine learning techniques

1. Examine your task
2. Examine your data
3. Examine your model

ML/AI can be very powerful, but should be used responsibly