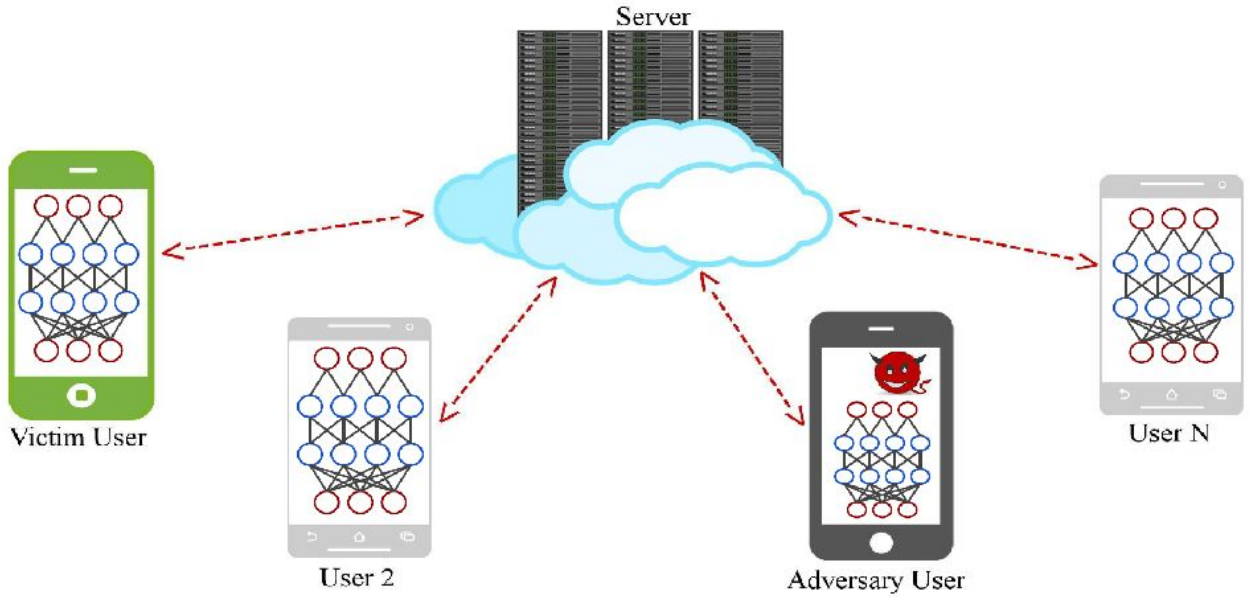
Lecture 12: Bias/Fairness

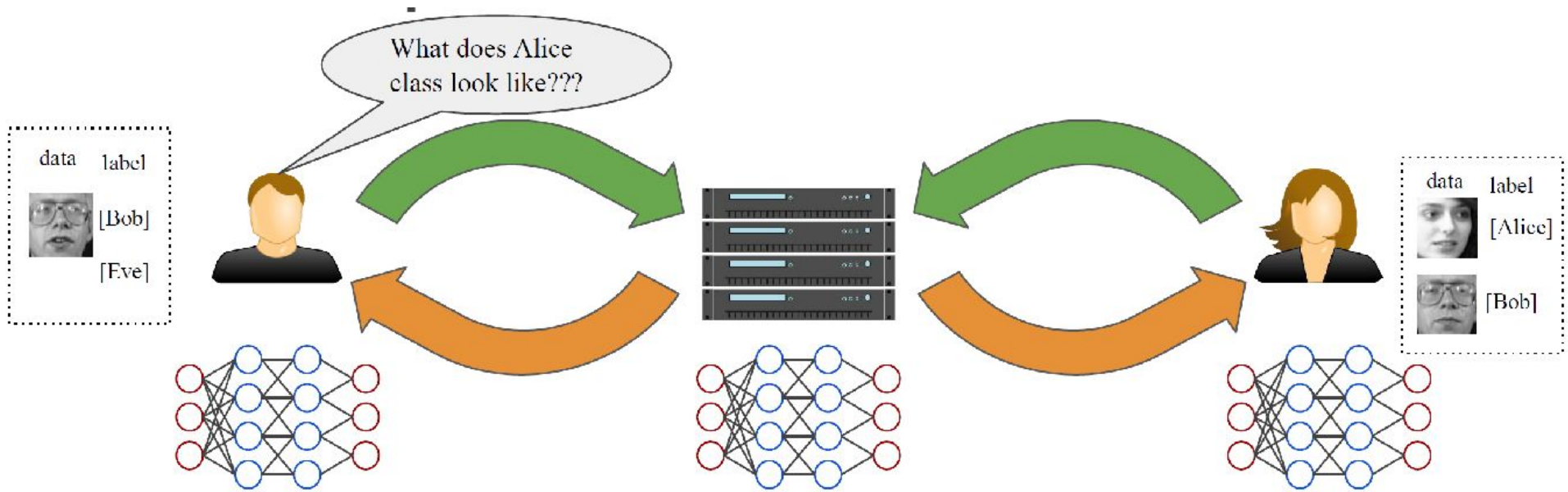
Hitaj, Briland, Giuseppe Ateniese, and Fernando Pérez-Cruz. "Deep models under the GAN: information leakage from collaborative deep learning." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017.

Decentralized Learning Scheme (collaborative/federated)



- Indirectly influencing the learning of other participants, allows potentially anyone to be an adversary

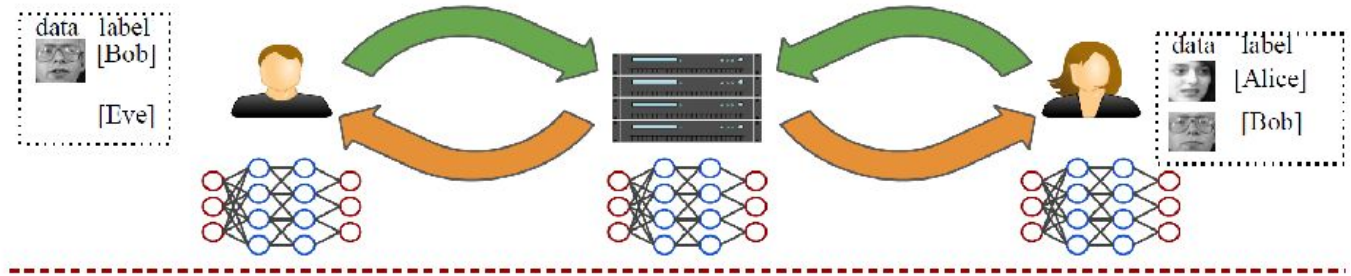
https://documentcloud.adobe.com/link/track?uri=urn:aaid:scds:US:7e7367a0-d8c1-46d9-a30c-cba881d1af4c

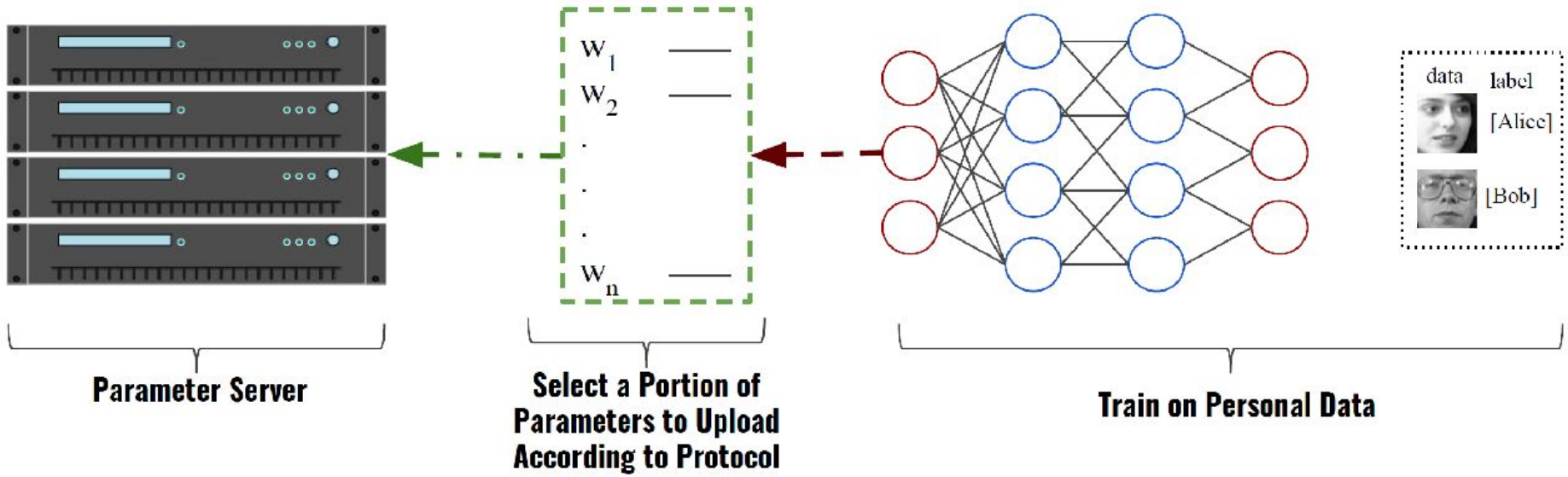


Agreement

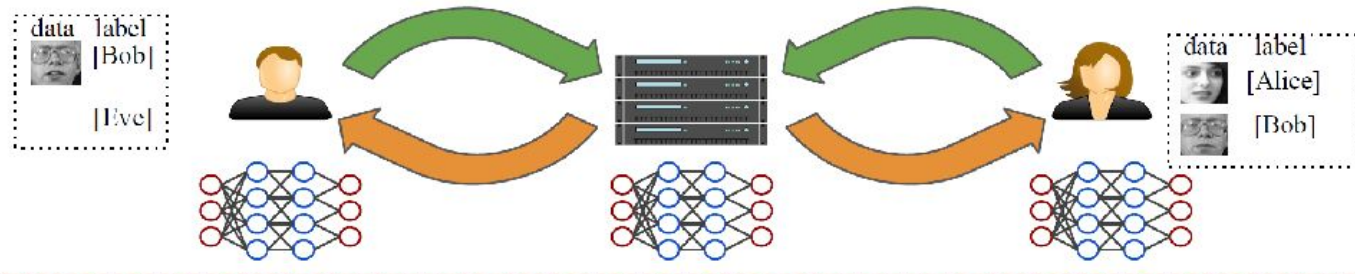
- Common learning objective
- Architecture of the model
- Labels/Classes present

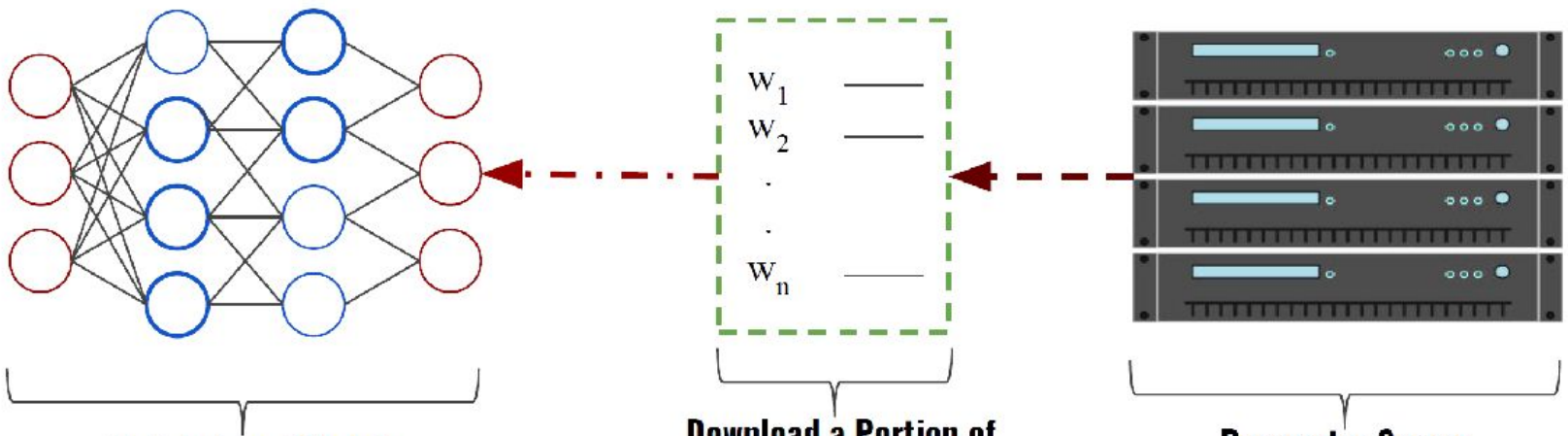
Victim's Turn



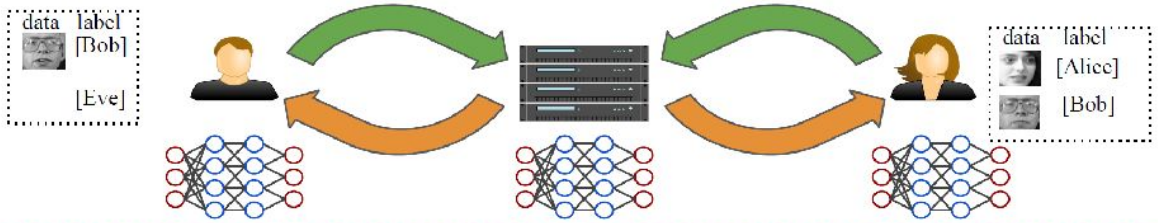


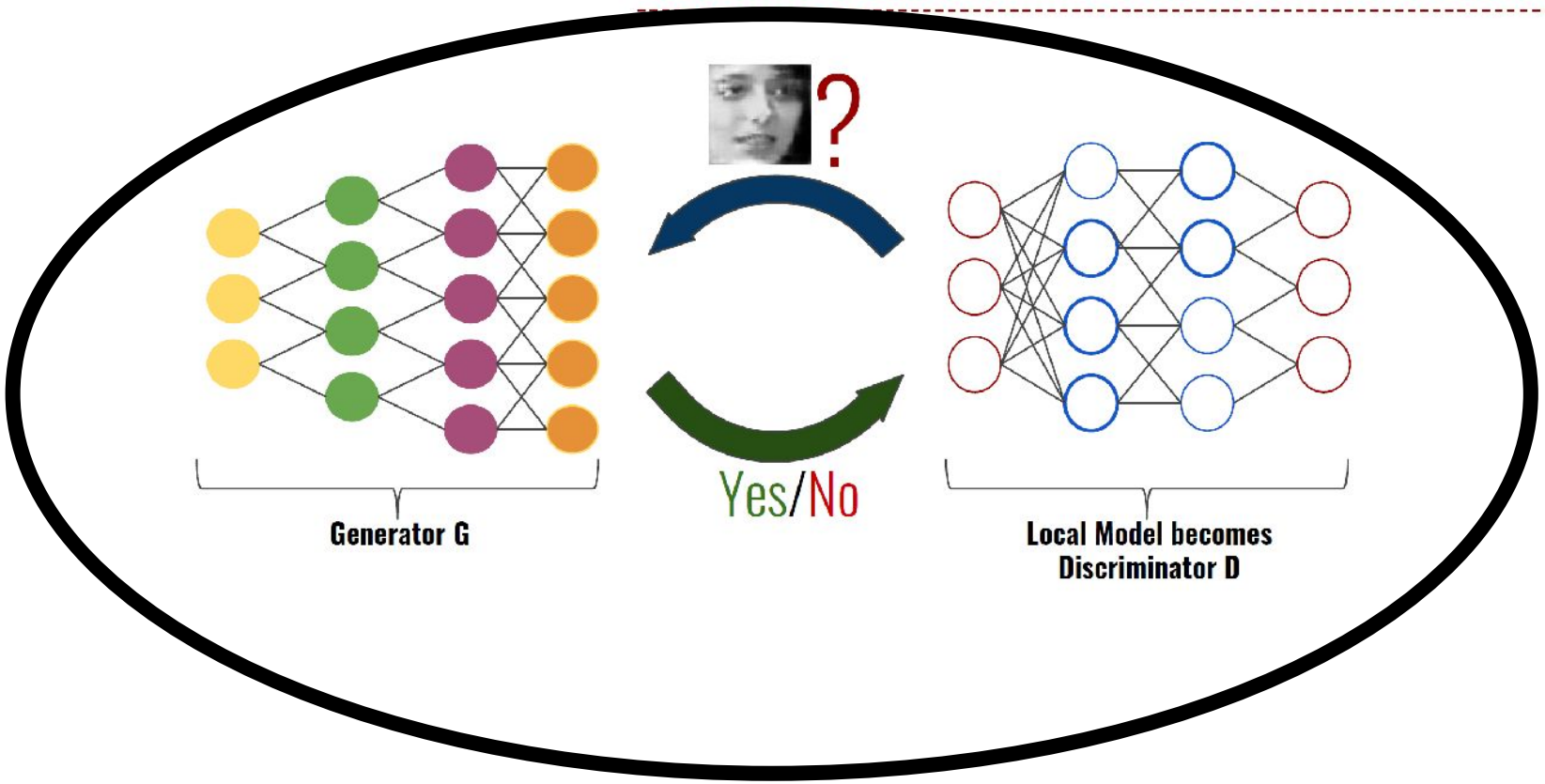
Adversary's Turn





Adversary's Turn





"GAN"

Generative Models

Given training data, generate new samples from same distribution

Training data ~ $p_{data}(x)$

Generated samples ~ $p_{model}(x)$

Want to learn $p_{model}(x)$ similar to $p_{data}(x)$

Generative Models

Given training data, generate new samples from same distribution

Training data ~ $p_{data}(x)$

Generated samples ~ $p_{model}(x)$

Want to learn $p_{model}(x)$ similar to $p_{data}(x)$

Addresses density estimation, a core problem in unsupervised learning

Several flavors:

- Explicit density estimation: explicitly define and solve for $p_{model}(x)$
- Implicit density estimation: learn model that can sample from $p_{model}(x)$ w/o explicitly defining it

Slides from F-F Li's class

Generative Adversarial Networks

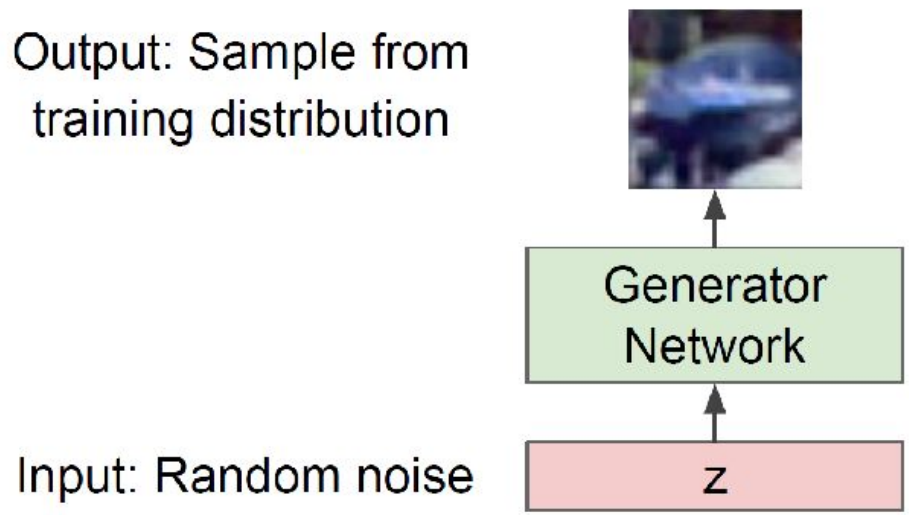
Ian Goodfellow et al., "Generative Adversarial Nets", NIPS 2014

Problem: Want to sample from complex, high-dimensional training distribution. No direct way to do this!

Solution: Sample from a simple distribution, e.g. random noise. Learn transformation to training distribution.

Q: What can we use to represent this complex transformation?

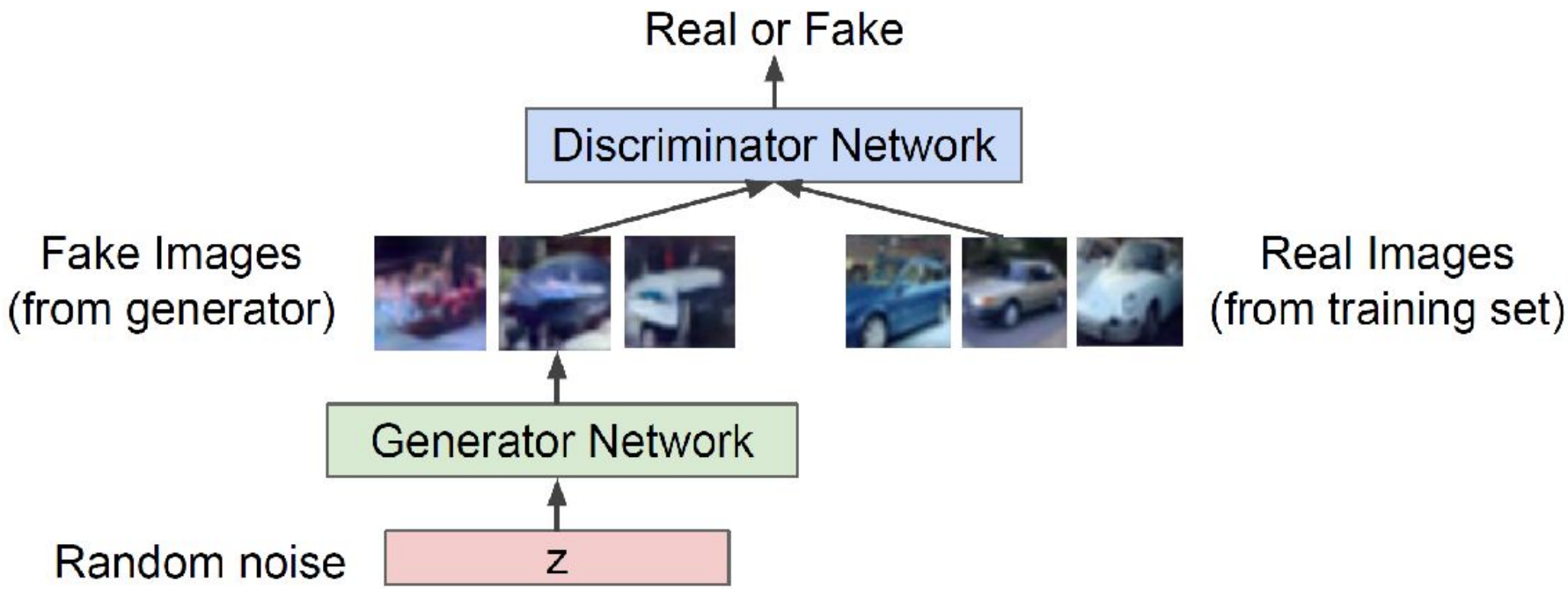
A: A neural network!



Training GANs: Two-player game

Ian Goodfellow et al., "Generative Adversarial Nets", NIPS 2014

Generator network: try to fool the discriminator by generating real-looking images
Discriminator network: try to distinguish between real and fake images



Training GANs: Two-player game

Ian Goodfellow et al., "Generative Adversarial Nets", NIPS 2014

Generator network: try to fool the discriminator by generating real-looking images
Discriminator network: try to distinguish between real and fake images

Train jointly in **minimax game**

Discriminator outputs likelihood in (0,1) of real image

Minimax objective function:

$$\min_{\theta_g} \max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

Discriminator output for real data x

Discriminator output for generated fake data G(z)

- Discriminator (θ_d) wants to **maximize objective** such that D(x) is close to 1 (real) and D(G(z)) is close to 0 (fake)
- Generator (θ_g) wants to **minimize objective** such that D(G(z)) is close to 1 (discriminator is fooled into thinking generated G(z) is real)

Training GANs: Two-player game

Ian Goodfellow et al., "Generative Adversarial Nets", NIPS 2014

Minimax objective function:

$$\min_{\theta_g} \max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

Alternate between:

1. **Gradient ascent** on discriminator

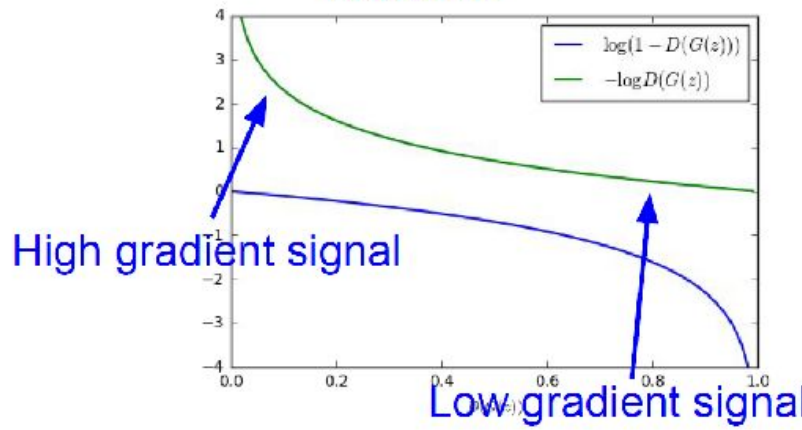
$$\max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

2. **Instead: Gradient ascent** on generator, different objective

$$\max_{\theta_g} \mathbb{E}_{z \sim p(z)} \log(D_{\theta_d}(G_{\theta_g}(z)))$$

Instead of minimizing likelihood of discriminator being correct, now maximize likelihood of discriminator being wrong.
Same objective of fooling discriminator, but now higher gradient signal for bad samples => works much better! Standard in practice.

Aside: Jointly training two networks is challenging, can be unstable. Choosing objectives with better loss landscapes helps training, is an active area of research.



Training GANs: Two-player game

Ian Goodfellow et al., "Generative Adversarial Nets", NIPS 2014

Putting it together: GAN training algorithm

for number of training iterations **do**

  for k steps **do**

  - Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  - Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
  - Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D_{\theta_d}(x^{(i)}) + \log(1 - D_{\theta_d}(G_{\theta_g}(z^{(i)}))) \right]$$

  end for

  - Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  - Update the generator by ascending its stochastic gradient (improved objective):

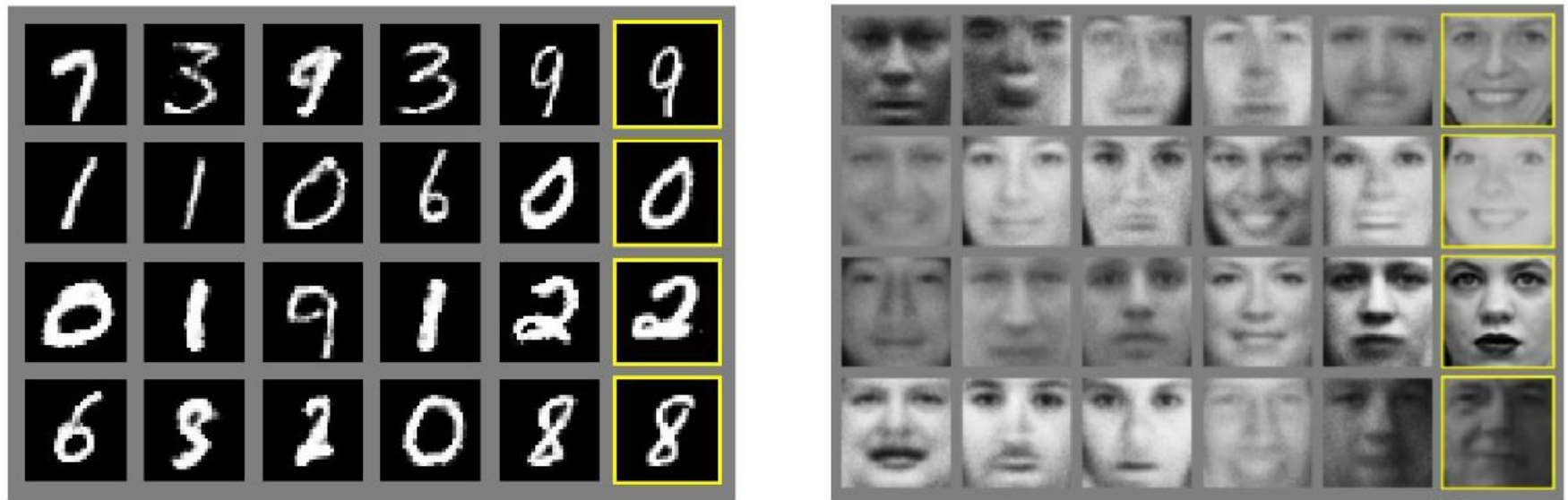
$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log(D_{\theta_d}(G_{\theta_g}(z^{(i)})))$$

end for

Ian Goodfellow et al., "Generative Adversarial Nets", NIPS 2014

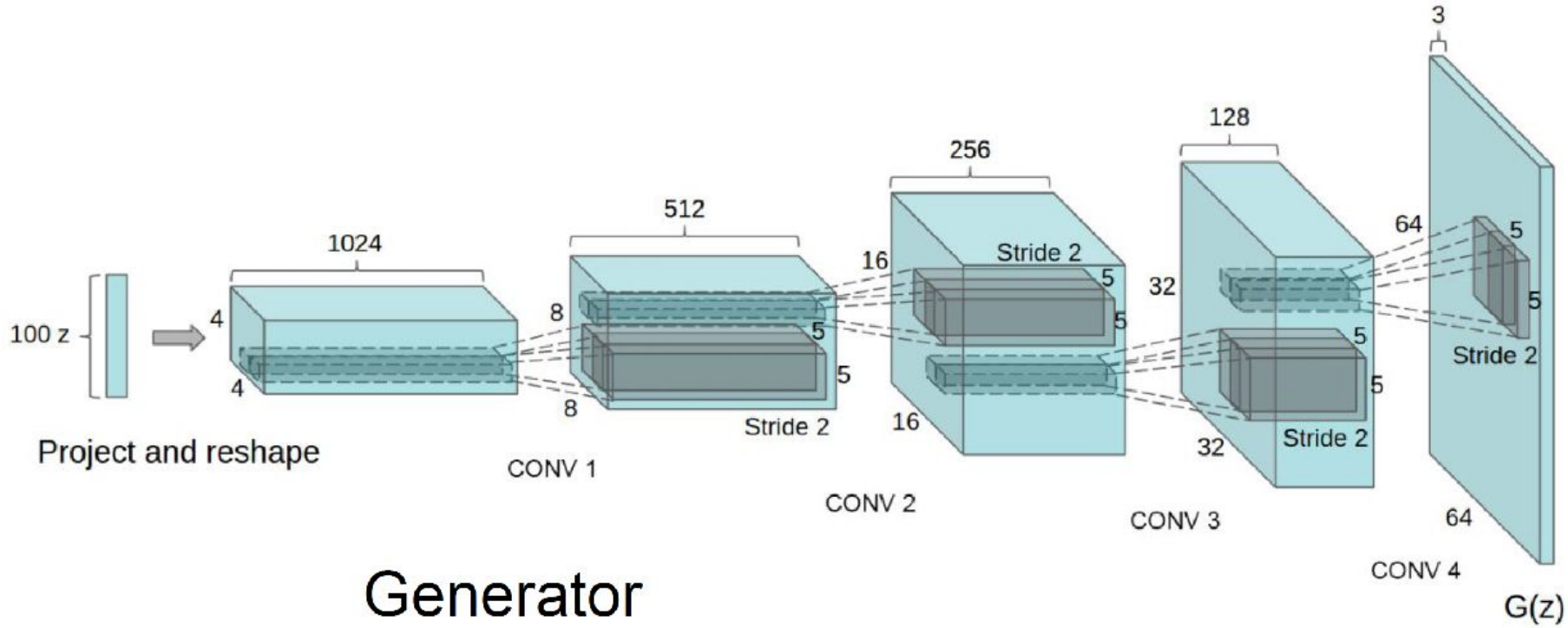
Generative Adversarial Nets

Generated samples



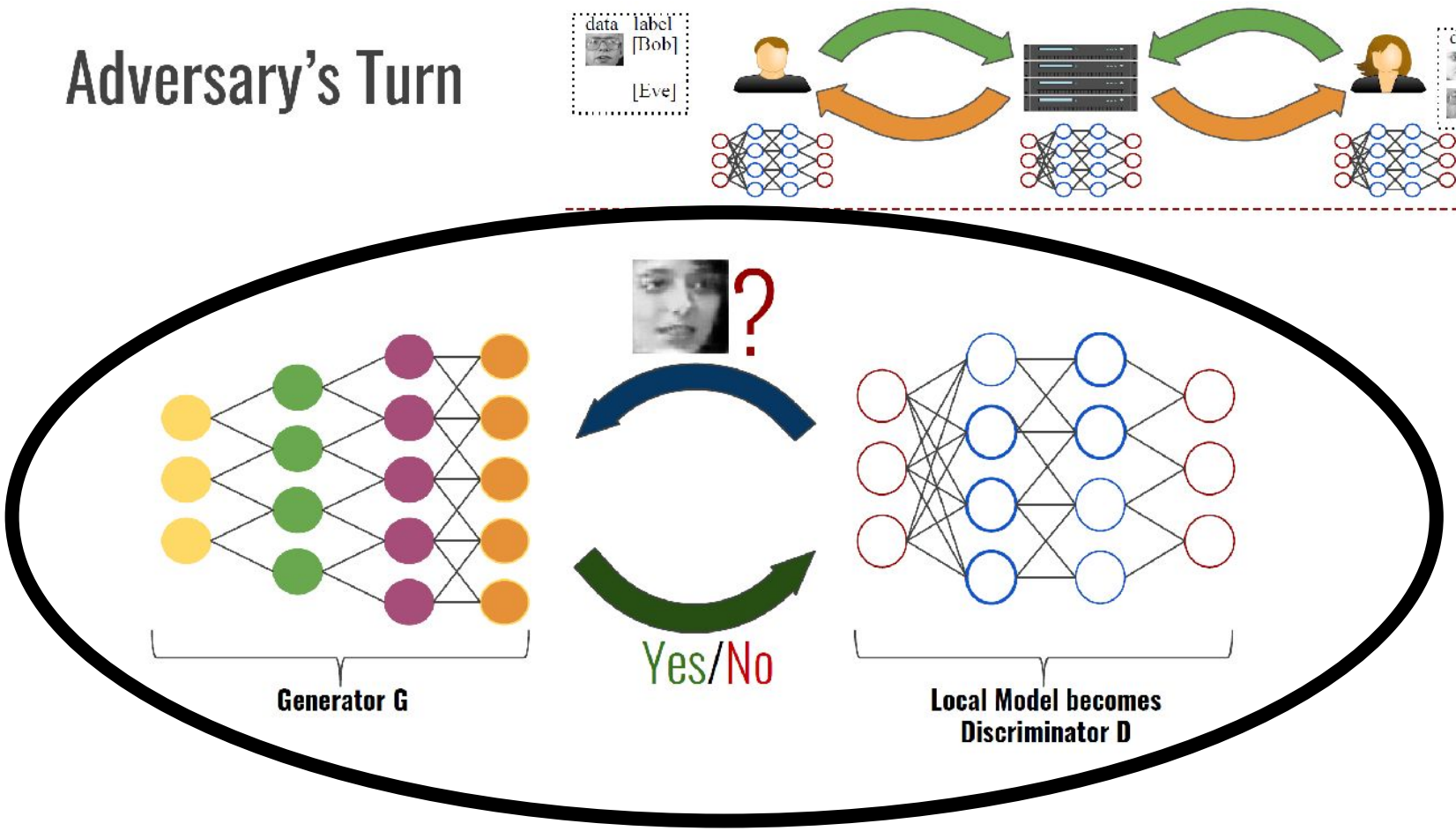
Nearest neighbor from training set

Generative Adversarial Nets: Convolutional Architectures



Generator

Radford et al, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”, ICLR 2016



"GAN"

Advanced Attack: Use GAN to forces Alice to reveal increasingly discriminative features about herself!

Experiments without Differential Privacy

Actual Images

Generated Data

Original vs Generated

Experiments with Differential Privacy

Actual Images

Generated Data

Original vs Generated

Is Differential Privacy Broken?

Algorithms for….

The Upshot

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

Claire Cain Miller @clairecm JUNE 25, 2015

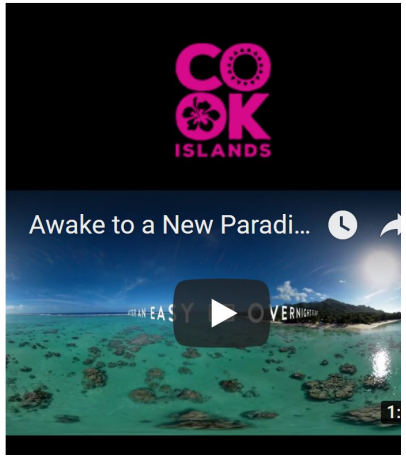
Hiring and recruiting might seem like some of the least likely jobs to be automated. The whole process seems to need human skills that computers lack, like making conversation and reading social cues.

But people have biases and predilections. They make hiring decisions, often unconsciously, based on similarities that have nothing to do with the job requirements — like whether an applicant has a friend in common, went to the same school or likes the same sports.

That is one reason researchers say traditional job searches are broken. The question is how to make them better.

AI is helping job candidates bypass resume bias and black holes

MARK NEWMAN, HIREVUE NOVEMBER 9, 2016 12:10 PM



hiring

Algorithms for….

Predictive Policing

10.27.2015

SARAH BRAYNE, ALEX ROSENBLAT, and DANAH BOYD

Introduction[1]

Predictive policing refers to the use of analytical techniques by law enforcement to make statistical predictions about potential criminal activity.[2] Predictive policing can involve either predicting events (i.e., forecasting when and where crimes are likely to occur) or people (i.e., individuals likely to be victims or perpetrators of crimes). Instead of relying on an officer's 'hunch' about an area, "predictive policing uses the power of 'big data' to isolate patterns."[3] A 2013 RAND report offers a taxonomy of predictive methods, identifying four categories of predictive policing:

- Methods for predicting crimes
- Methods for predicting offenders
- Methods for predicting perpetrators' identities
- Methods for predicting victims[4]

This primer offers an overview of what is currently known about predictive policing and highlights unanswered questions about the implications of predictive policing.

Predictive policing

Algorithms for….

Predictive Policing

10.27.2015

SARAH BRAYNE, ALEX ROSENBLAT, and DANAH BOYD

Introduction[1]

Predictive policing refers to the use of analytical techniques by law enforcement to make statistical predictions about potential criminal activity.[2] Predictive policing can involve either predicting events (i.e., forecasting when and where crimes are likely to occur) or people (i.e., individuals likely to be victims or perpetrators of crimes). Instead of relying on an officer's 'hunch' about an area, "predictive policing uses the power of 'big data' to isolate patterns."[3] A 2013 RAND report offers a taxonomy of predictive methods, identifying four categories of predictive policing:

- Methods for predicting crimes
- Methods for predicting offenders
- Methods for predicting perpetrators' identities
- Methods for predicting victims[4]

This primer offers an overview of what is currently known about predictive policing and highlights unanswered questions about the implications of predictive policing.

Predictive policing

Algorithms for….

Long Reads

Big data meets Big Brother as China moves to rate its citizens

The Chinese government plans to launch its Social Credit System in 2020. The aim? To judge the trustworthiness – or otherwise – of its 1.3 billion residents

Imagine a world where many of your daily activities were constantly monitored and evaluated: what you buy at the shops and online; where you are at any given time; who your friends are and how you interact with them; how many hours you spend watching content or playing video games; and what bills and taxes you pay (or not). It's not hard to picture, because most of that already happens, thanks to all those data-collecting behemoths like Google, Facebook and Instagram or health-tracking apps such as Fitbit. But now imagine a system where all these behaviours are rated as either positive or negative and distilled into a single number, according to rules set by the government. That would create your Citizen Score and it would tell everyone whether or not you were trustworthy. Plus, your rating would be publicly ranked against that of the entire population and used to determine your eligibility for a mortgage or a job, where your children can go to school - or even just your chances of getting a date.

Loan decisions

But….

The Upshot

HIDDEN BIAS

When Algorithms Discriminate

Claire Cain Miller @clairecm JULY 9, 2015

The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

But can computer programs be discriminatory?

There is a widespread belief that software and algorithms that rely on data are objective. But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result, say researchers in computer science, ethics and law, algorithms can reinforce human prejudices.

Google's online advertising system, for instance, showed an ad for high-income jobs to men much more often than it showed the ad to women, a new study by Carnegie Mellon University researchers found.

Research from Harvard University found that ads for arrest records were significantly more likely to show up on searches for distinctively black names or a historically black fraternity. The Federal Trade Commission said advertisers are able to target people who live in low-income neighborhoods

Are algorithms fair and balanced?

Bias in predictive policing

Machine Bias

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

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

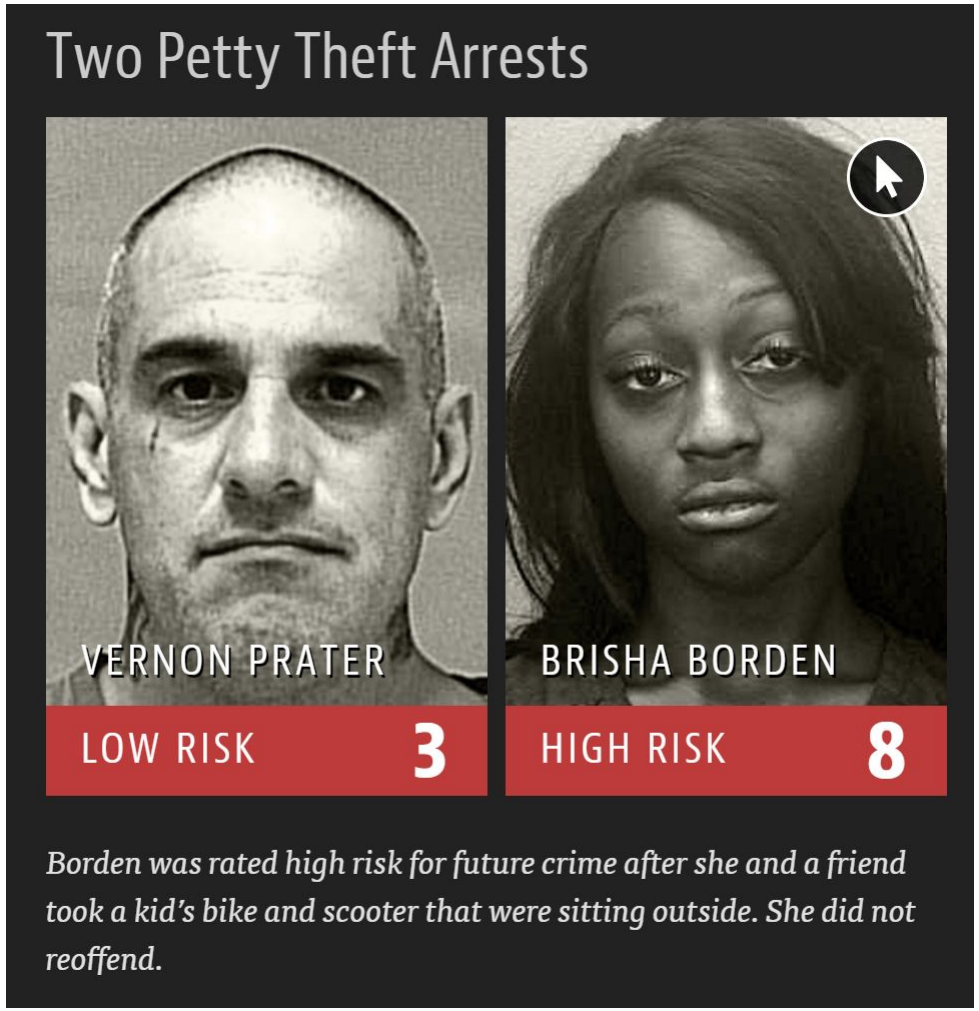
ON A SPRING AFTERNOON IN 2014, **Brisha Borden** was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

The previous summer, 41-year-old **Vernon Prater** was picked up for shoplifting $86.35 worth of tools from a nearby Home Depot store. Prater was the more seasoned criminal. He had already been convicted of armed robbery and attempted armed robbery, for which he served five years in prison, in addition to another armed robbery charge. Borden had a record, too, but it was for misdemeanors committed when she was a juvenile.

Predicted Risk Scores



"Two years later, we know the computer algorithm got it exactly backward. Borden has not been charged with any new crimes. Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars' worth of electronics."

Aren't ML models meant to "discriminate"

- Practical irrelevance
  - Race, gender or sexual orientation might be practically irrelevant, for instance, for job employment

- Moral irrelevance
  - Even if there are statistical differences in job performance, we might stipulate for moral reasons that we do not want to discriminate on certain grounds, e.g., individuals with impairmants

Legal Basis for Fairness

Domains in which fairness is mandated by law:

- **Credit** (Equal Credit Opportunity Act)
- **Education** (Civil Rights Act of 1964; Education Amendments of 1972)
- **Employment** (Civil Rights Act of 1964)
- **Housing** (Fair Housing Act)
- **'Public Accommodation'** (Civil Rights Act of 1964)

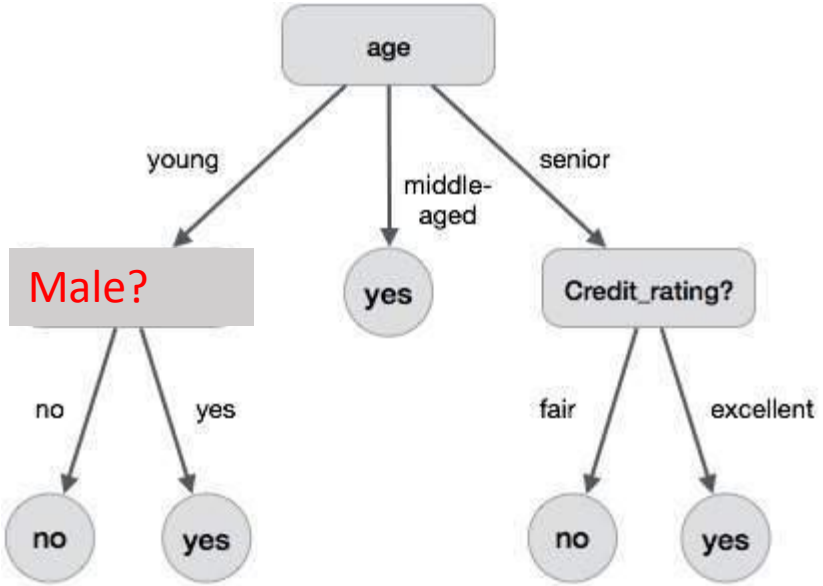
Characteristics on which discrimination is prohibited:

Race (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964); **National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967); **Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

Discrimination Law Doctrines

Disparate Treatment

Considering a protected characteristic during decision making, *even if it is ignored by the decision making algorithm*, constitutes disparate treatment



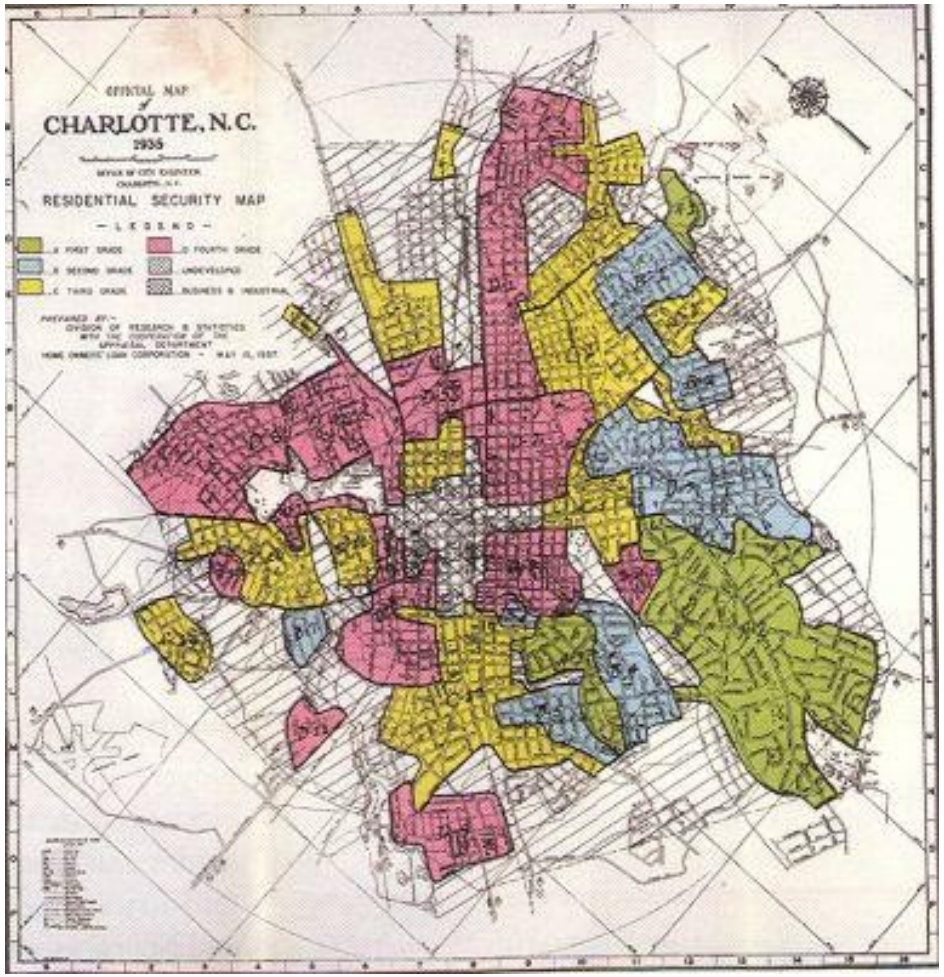
Discrimination Law Doctrines

Disparate Treatment

Considering a protected characteristic during decision making, *even if it is ignored by the decision making algorithm*, constitutes disparate treatment

Proxies for protected characteristics are not allowed either

Red-Lining



Zip codes used as proxies for predominantly black neighborhoods

Discrimination Law Doctrines

Disparate **Impact**

Practices that adversely affect one group of people of a protected characteristic more than another,
even if rules applied are formally neutral (or neutral on their face)

Prior to 1974, New Bedford had two distinct police categories: males were "police officers" and females were "police women." At that time, male applicants were ineligible for positions as police officers if they failed to meet a minimum height requirement of five feet six inches. In February 1974 the city abandoned these separate job categories, and thereafter both men and women competed for positions as "police officers." Women applicants were then also required to meet the five feet six-inch height minimum.

Since the minimum height requirement excludes far more women than men from competing for positions as police officers, the requirement has a disparate impact on women.

Proving Disparate Impact

4/5th rule: "A selection rate for any race, sex, ethnic group which is less than four-fifths (**4/5**) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact…"

Justification: In response, defendant can argue that the rule being used is "job related" or a "business necessity"

Alternate: petitioner must then describe a different procedure that would reduce disparity

Treatment Vs. Impact

Twenty white city firefighters at the New Haven Fire Department claimed discrimination under Title VII of the Civil Rights Act of 1964 after they had passed the test for promotions to management positions and the city declined to promote them. New Haven officials invalidated the test results because none of the black firefighters who took it scored high enough to be considered for the positions.

City officials said that they feared a lawsuit over the test's disproportionate exclusion of certain racial groups from promotion under "disparate impact" head of liability.

Supreme Court Decision

Supreme court decided 5-4 in favor of the petitioners

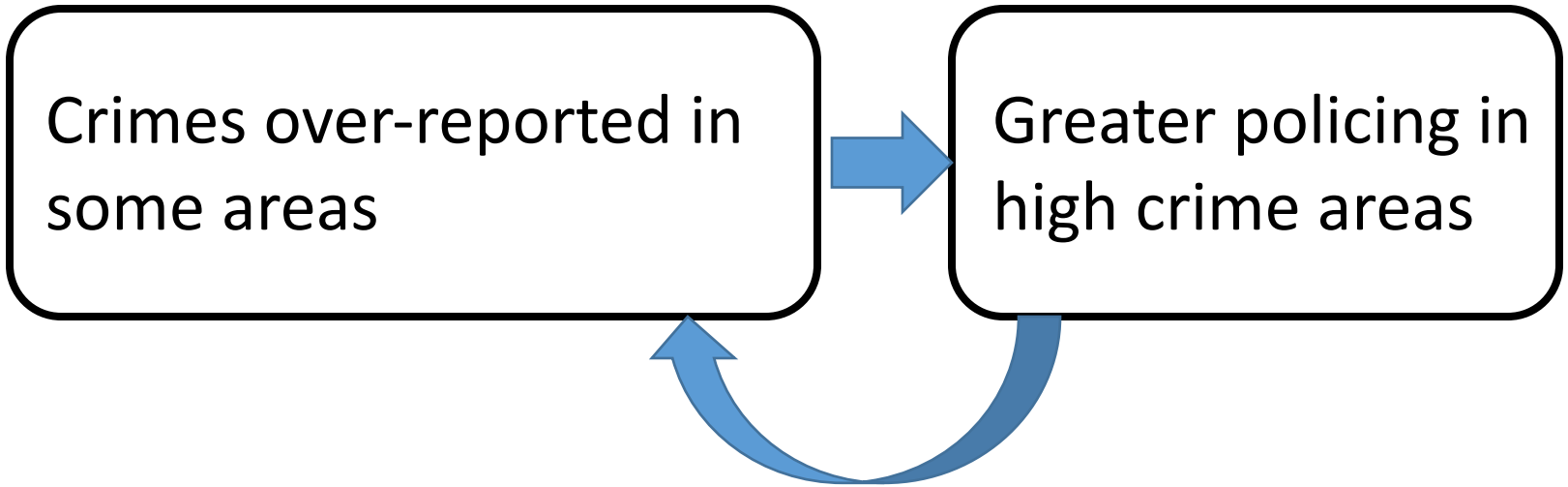
Justice Kennedy writing for the majority: “reached the statutory construction that, in instances of conflict between the disparate-treatment and disparate-impact provisions, permissible **justifications for disparate treatment must be grounded in the strong-basis-in-evidence** standard.”

Justice Ginsberg writing for the minority: “ignores substantial evidence of multiple flaws in the tests New Haven used. The Court similarly **fails to acknowledge the better tests used in other cities**, which have yielded less racially skewed outcomes.”

How machine learn to discriminate?

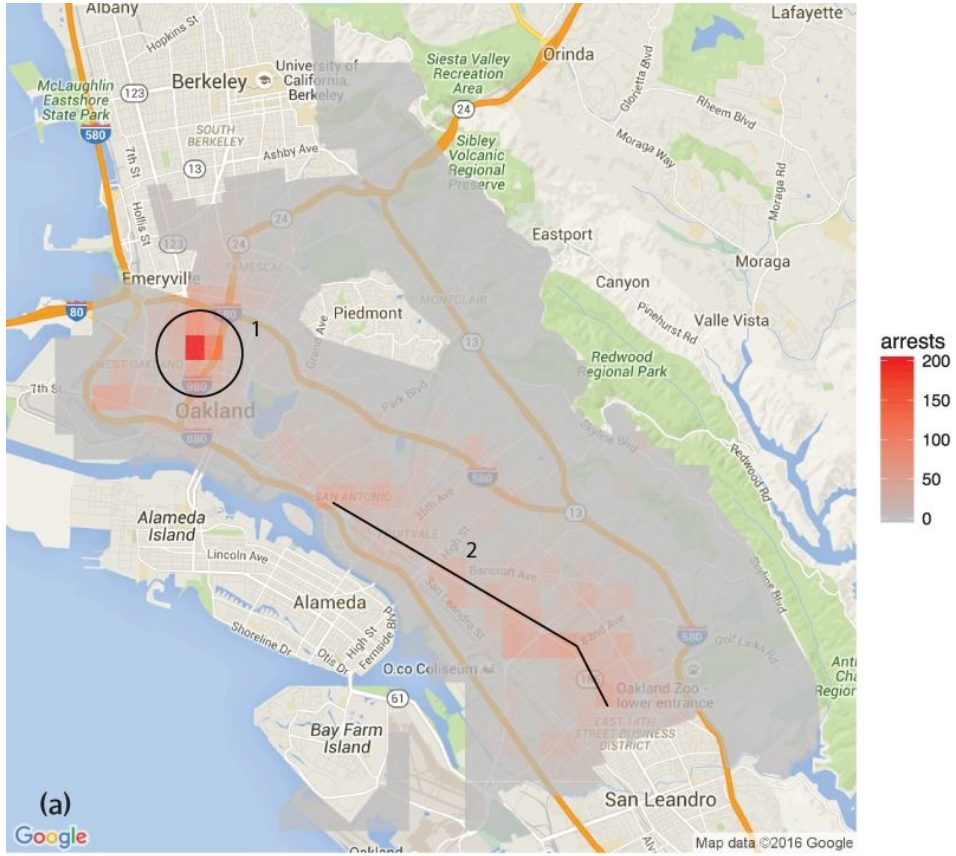
Skewed Samples

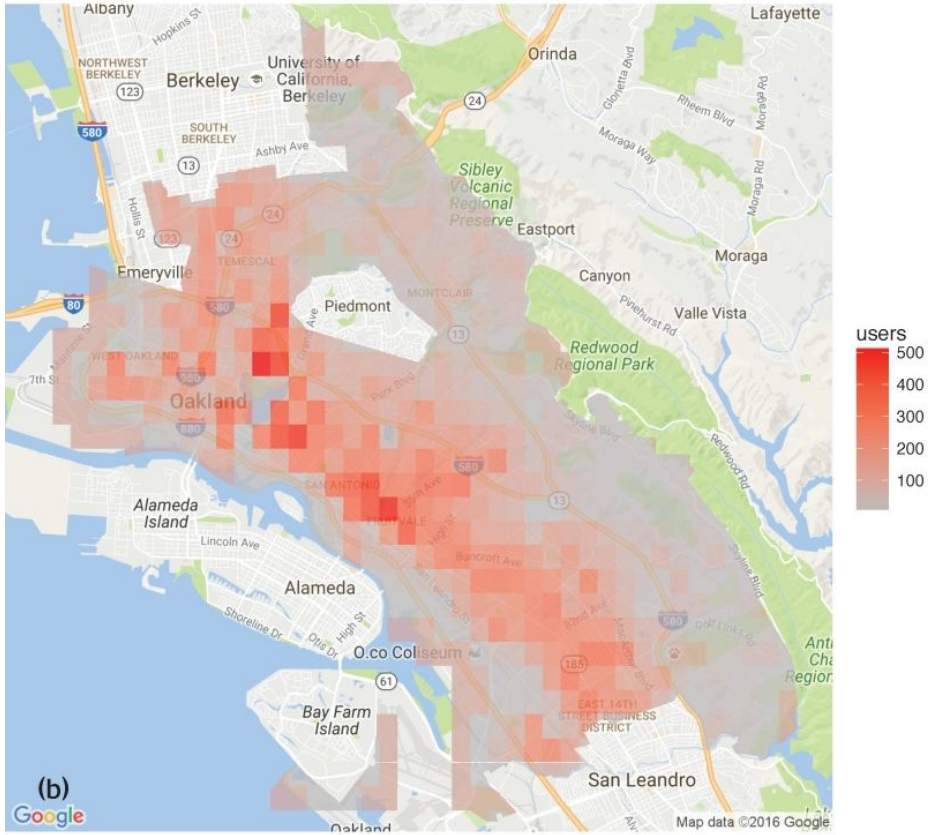
Predictive policing algorithms based on reported incidents of crime, but there can be human bias in the way that humans observe and report crimes



Confirmation bias

Evidence of Skewed Samples





https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x

How machine learn to discriminate?

Tainted Samples

Incorrect labels for supervised learning problems

Consider job hiring decisions based on previous job review scores assigned by managers. Hiring algorithm inherits biases, for example, gender or racial biases, in review scores generated by human managers

How machine learn to discriminate?

Limited Features

Features highly informative for majority group, but less informative for minority groups

Even if model has high accuracy overall, might have high errors for minorities

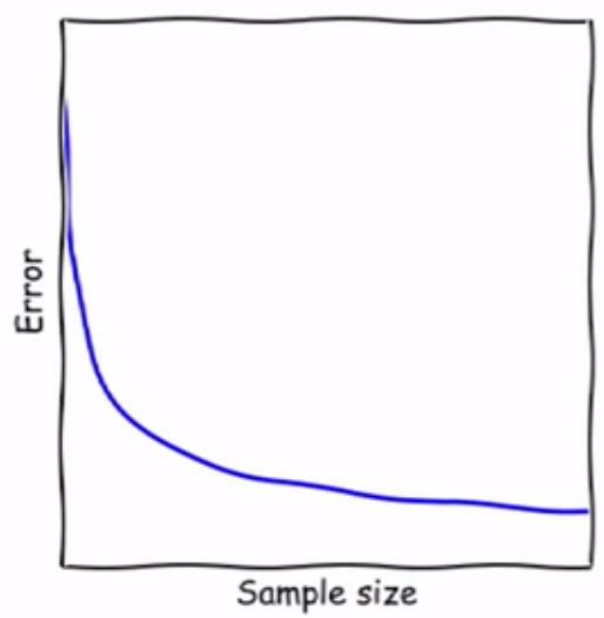
Distribution of errors is biased

How machine learn to discriminate?

Sample Size Disparity

Majority groups over-represented in training data

Distribution of errors is biased



Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on Fairness, Accountability and Transparency*. 2018.

Discrimination in Face Recognition

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Commercial face recognition software consistently perform poorly on darker skinned women

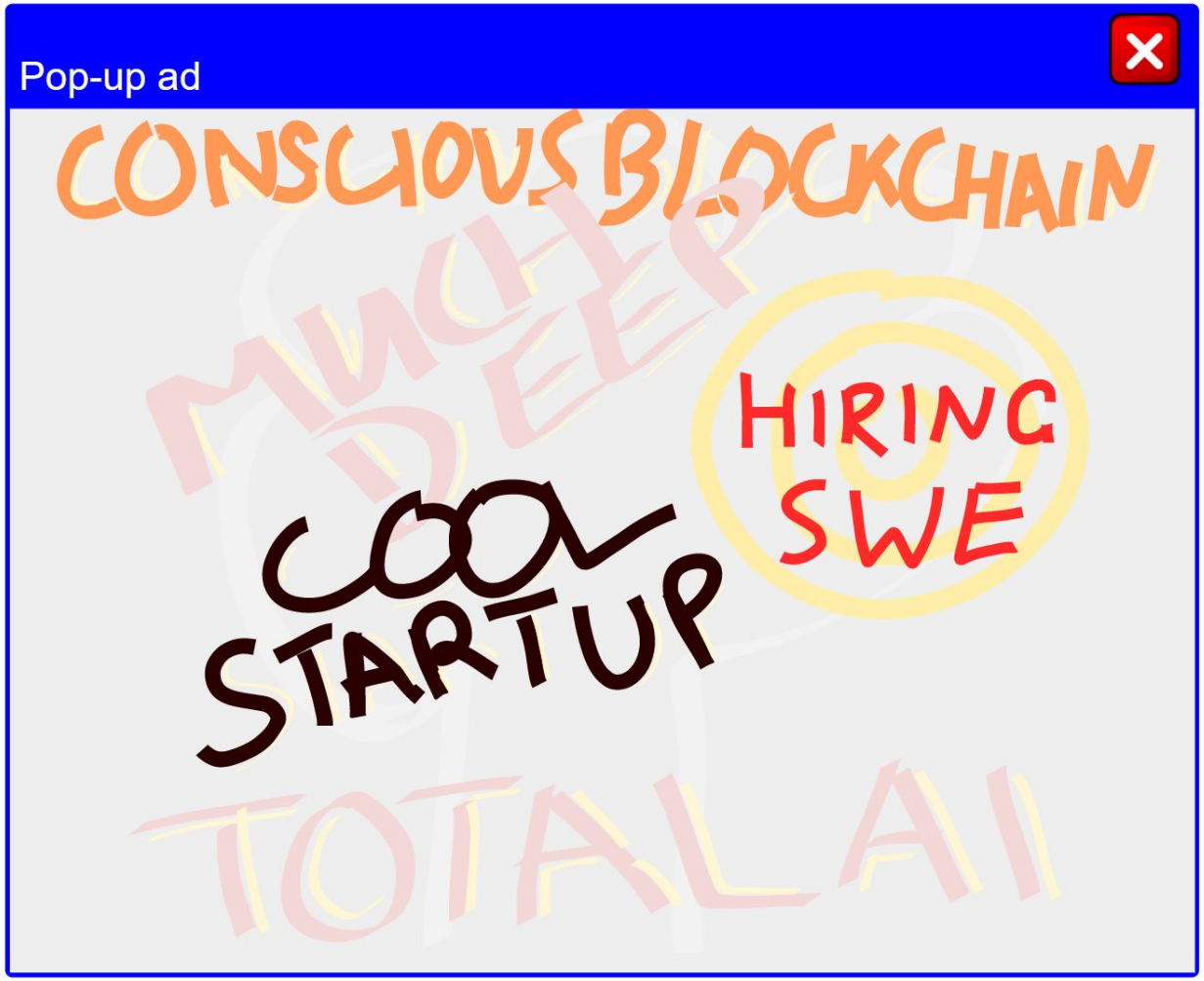
How machine learn to discriminate?

Proxies

Features might be correlated with protected attributes; even if protected attributes are not directly used, ML model might inherently infer and use them from correlated features

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Running Example



Formal setup

- X features of an individual (browsing history etc.)
- A sensitive attribute (here, gender)
- $C = c(X, A)$ predictor (here, show ad or not)
- Y target variable (here, SWE)

Note: random variables in the same probability space

Notation: $\mathbb{P}_a\{E\} = \mathbb{P}\{E \mid A = a\}.$

Formal setup

Score function is any random variable $R = r(X, A) \in [0, 1]$.

Can be turned into (binary) predictor by thresholding

Example: *Bayes optimal* score given by $r(x, a) = \mathbb{E}[Y \mid X = x, A = a]$

Three fundamental criteria

Independence: C independent of A

Separation: C independent of A conditional on Y

Sufficiency: Y independent of A conditional on C

Lots of other criteria are related to these

First criterion: Independence

Require C and A to be independent, denoted $C \perp A$

That is, for all groups a, b and all values c:

$$\mathbb{P}_a\{C = c\} = \mathbb{P}_b\{C = c\}$$

Variants of independence

Sometimes called *demographic parity*, *statistical parity*

When C is binary $0/1$-variables, this means $\mathbb{P}_a\{C = 1\} = \mathbb{P}_b\{C = 1\}$ for all groups a, b.

Approximate versions:

$$\frac{\mathbb{P}_a\{C = 1\}}{\mathbb{P}_b\{C = 1\}} \geq 1 - \epsilon$$

$$|\mathbb{P}_a\{C = 1\} - \mathbb{P}_b\{C = 1\}| \leq \epsilon$$

Achieving independence

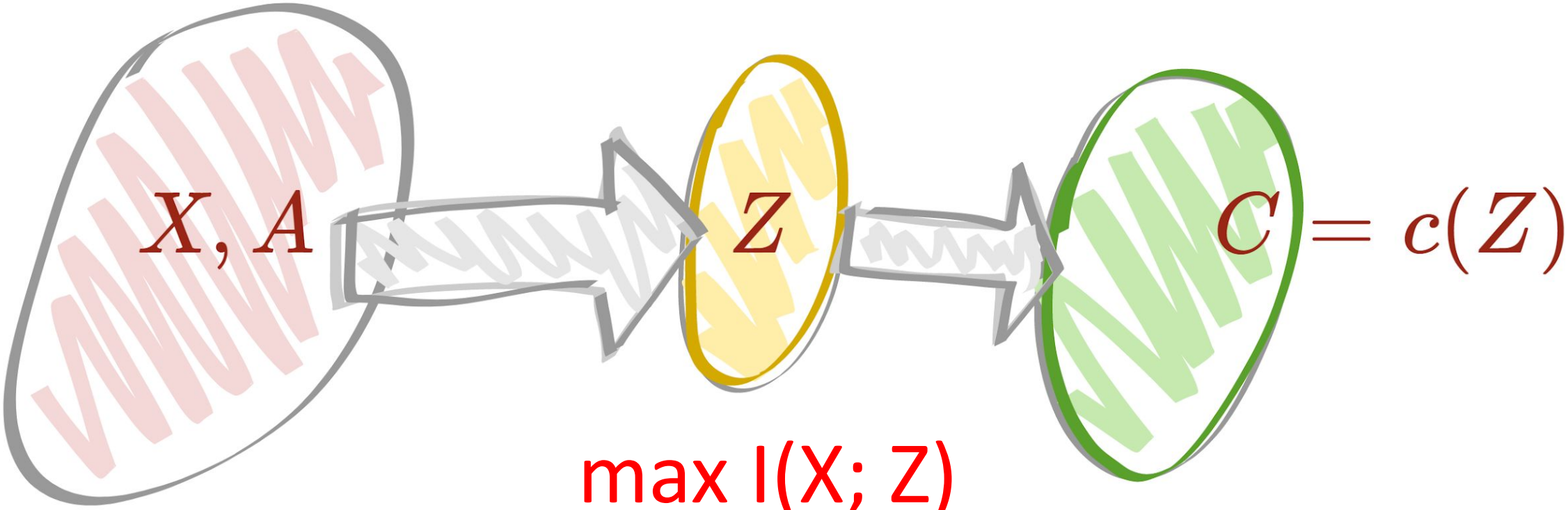
Post-processing: Feldman, Friedler, Moeller, Scheidegger, Venkatasubramanian (2014)

Training time constraint: Calders, Kamiran, Pechenizkiy (2009)

Pre-processing: Via representation learning — Zemel, Yu, Swersky, Pitassi, Dwork (2013) and Louizos, Swersky, Li, Welling, Zemel (2016); Via feature adjustment — Lum-Johndrow (2016)

Many more...

Representation learning approach



max I(X; Z)

min I(A; Z)

Shortcomings of independence

Ignores possible correlation between in Y and A.
In particular, rules out perfect predictor $C = Y$.

Premits **laziness**:
Accept the qualified in one group, random people in other

Rich data for majority group, limited samples for minority group. Effectively, model makes random decisions for minority groups

Allows to trade false negatives for false positives.

Conflates desirable long-term goal with algorithmic constraint

In the long run we might desire gender parity in SWE, but do we achieve it by imposing algorithmic constraints now?

Second criterion: Separation

Require R and A to be independent *conditional on target variable* Y, denoted $R \perp A \mid Y$

Definition. Random variable R *separated* from A if $R \perp A \mid Y$.

Proposed in H, Price, Srebro (2016);
Zafar, Valera, Rodriguez, Gummadi (2016)

Desirable properties of separation

Optimality compatibility

$R = Y$ is allowed

Penalizes lazyness

Incentive to reduce errors uniformly in all groups

Recall, neither of these is achieved by independence.

Equalize Acceptance Rates

We say the random variables (R, A, Y) satisfy **separation** if the sensitive characteristics A are statistically independent to the prediction R given the target value Y, and we write $R \perp A|Y$.

We can also express this notion with the following formula:

$$P(R = r|Y = q, A = a) = P(R = r|Y = q, A = b) \quad \forall r \in R \quad q \in Y \quad \forall a, b \in A$$

This means that the probability of being classified by the algorithm in each of the groups is equal for two individuals with different sensitive characteristics given that they actually belong in the s

Another equivalent expression, in the case of a binary target rate, is that the true positive rate and the false positive rate are equal (and therefore the false negative rate and the true negative ra characteristics:

$$P(R = 1|Y = 1, A = a) = P(R = 1|Y = 1, A = b) \quad \forall a, b \in A$$

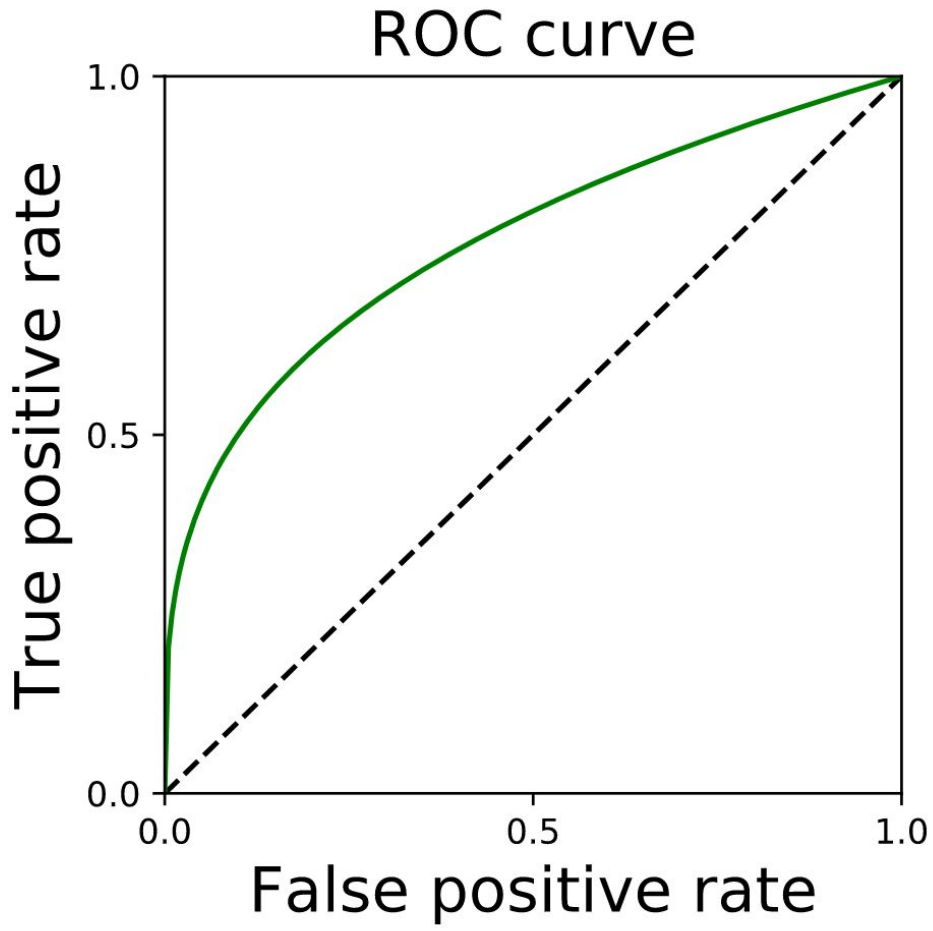
$$P(R = 1|Y = 0, A = a) = P(R = 1|Y = 0, A = b) \quad \forall a, b \in A$$

Achieving separation

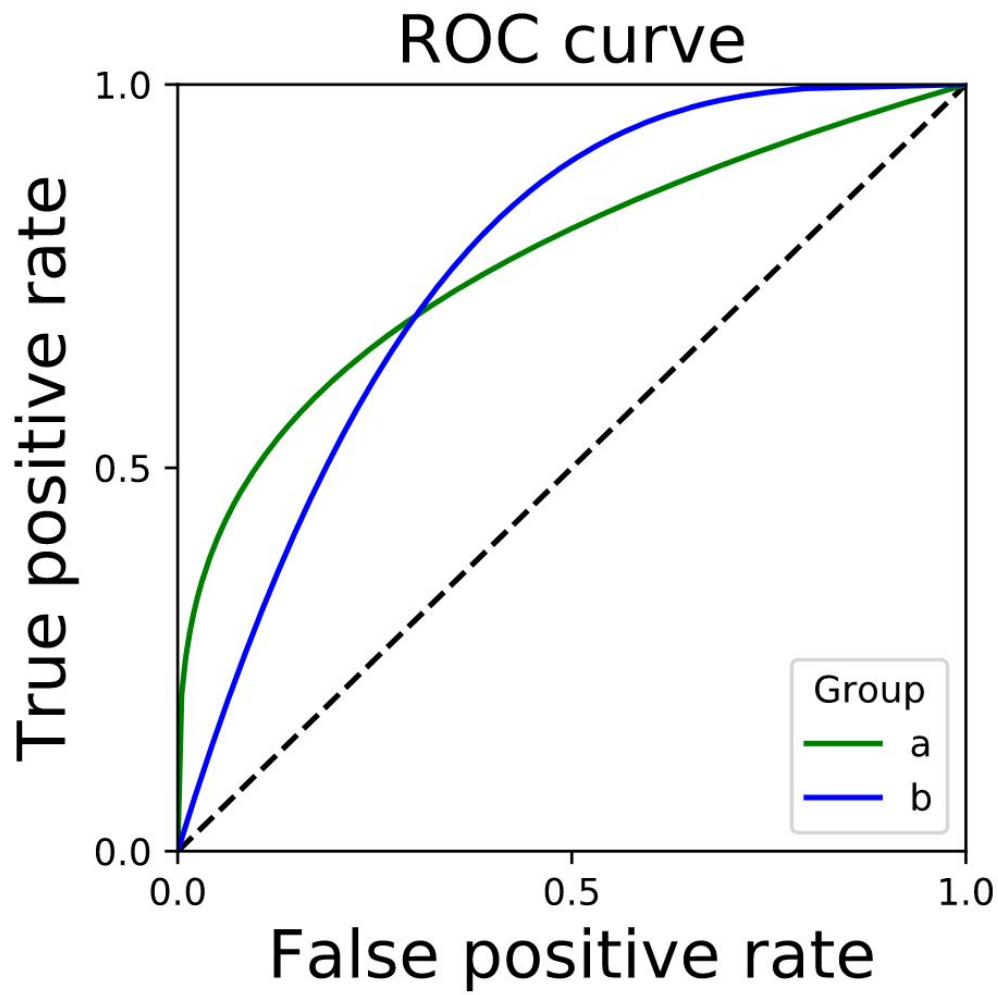
Method from H, Price, Srebro (2016):
Post-processing correct of score function

Post-processing: Any thresholding of R (possibly depending on A)
No retraining/changes to R

Given score R, plot (TPR, FPR) for all possible thresholds



Look at ROC curve for each group



Feasible region: Trade-offs realizable in all groups