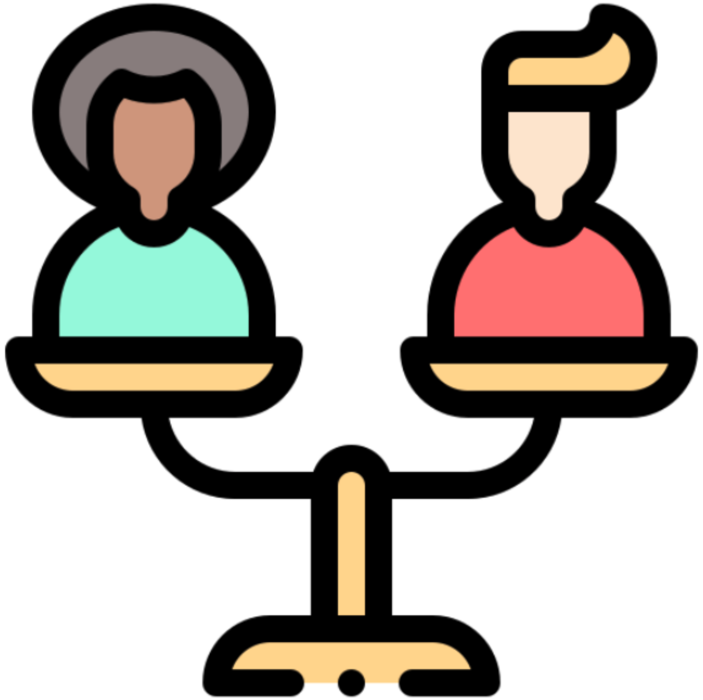
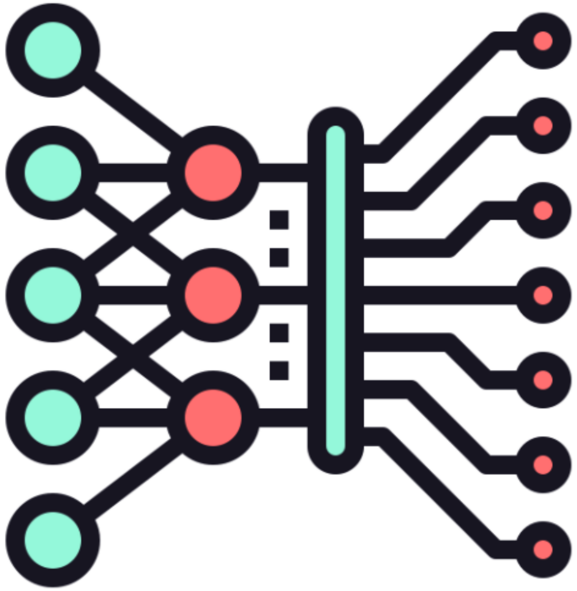
Fairness in Machine Learning





ML algorithms have failed to uphold basic notions of fairness throughout the history of machine learning

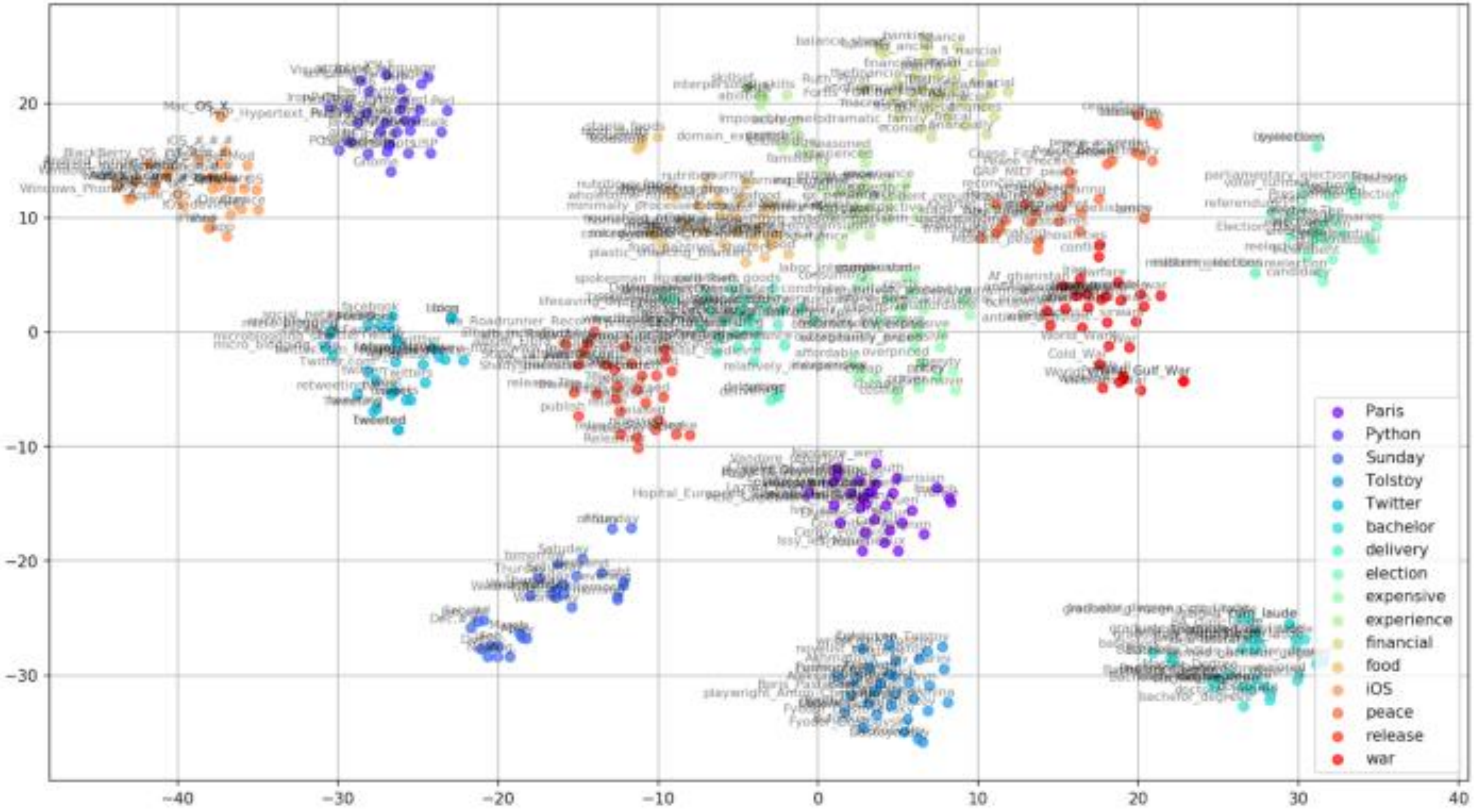
Language models

Goal:

- Learning word associations
- Applications:
  - Complete sentences
  - Find synonyms
  - Word analogies: “a is to b as x is to y”

Ex: word2vec. i) Computes distance between any pair of words, then ii) embeds in it a 2 or 3D space

Word2vec embedding

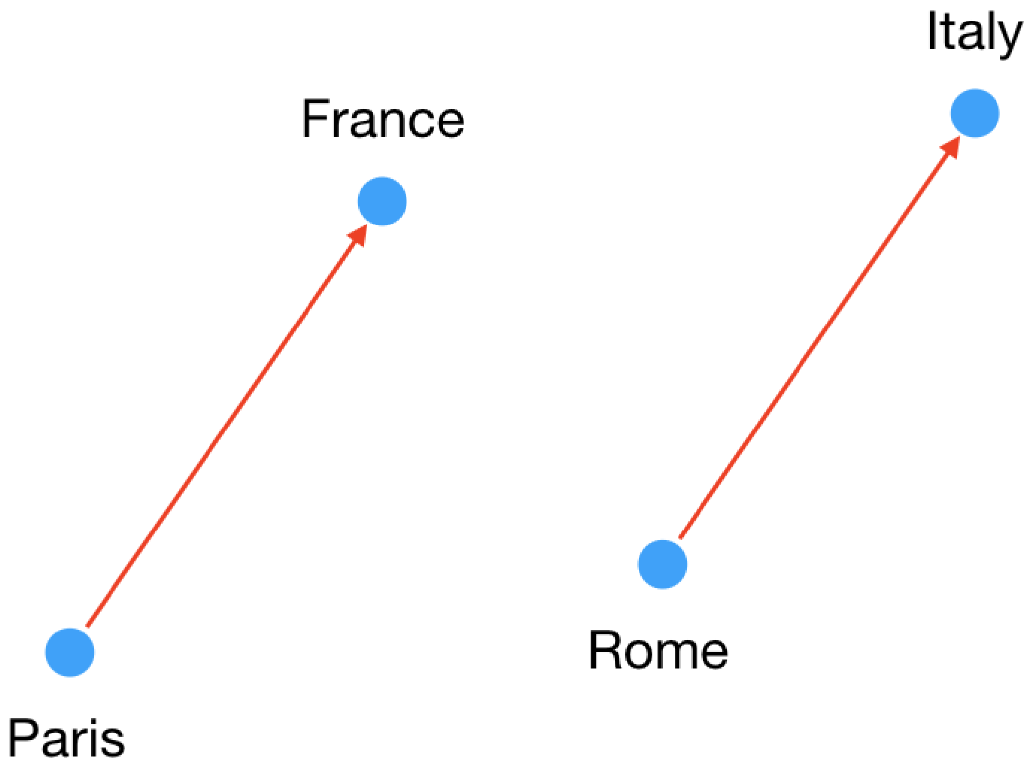


Word2vec – word analogies

Ex: “Paris is to France like 🟦 is to Italy”

Word2vec – word analogies

Ex: “Paris is to France like Rome is to Italy”



Word2vec – word analogies

2016: T. Bolukbasi, K-W Chang, Zou, Saligrama, Kalai

Started investigating failures of the word2vec model

Used word2vec to make analogies across genders:

Word2vec – word analogies

2016: T. Bolukbasi, K-W Chang, Zou, Saligrama, Kalai

Started investigating failures of the word2vec model

Used word2vec to make analogies across genders:

- “Man is to X as woman is to Y”

Word2vec – word analogies

2016: T. Bolukbasi, K-W Chang, Zou, Saligrama, Kalai

Started investigating failures of the word2vec model

Used word2vec to make analogies across genders:

- “Man is to X as woman is to Y”
- Set X = “Computer Programmer”

Word2vec – word analogies

2016: T. Bolukbasi, K-W Chang, Zou, Saligrama, Kalai

Started investigating failures of the word2vec model

Used word2vec to make analogies across genders:

- “Man is to X as woman is to Y”
- Set X = “Computer Programmer”
- Y = “Homemaker”

Word2vec – word analogies

2016: T. Bolukbasi, K-W Chang, Zou, Saligrama, Kalai

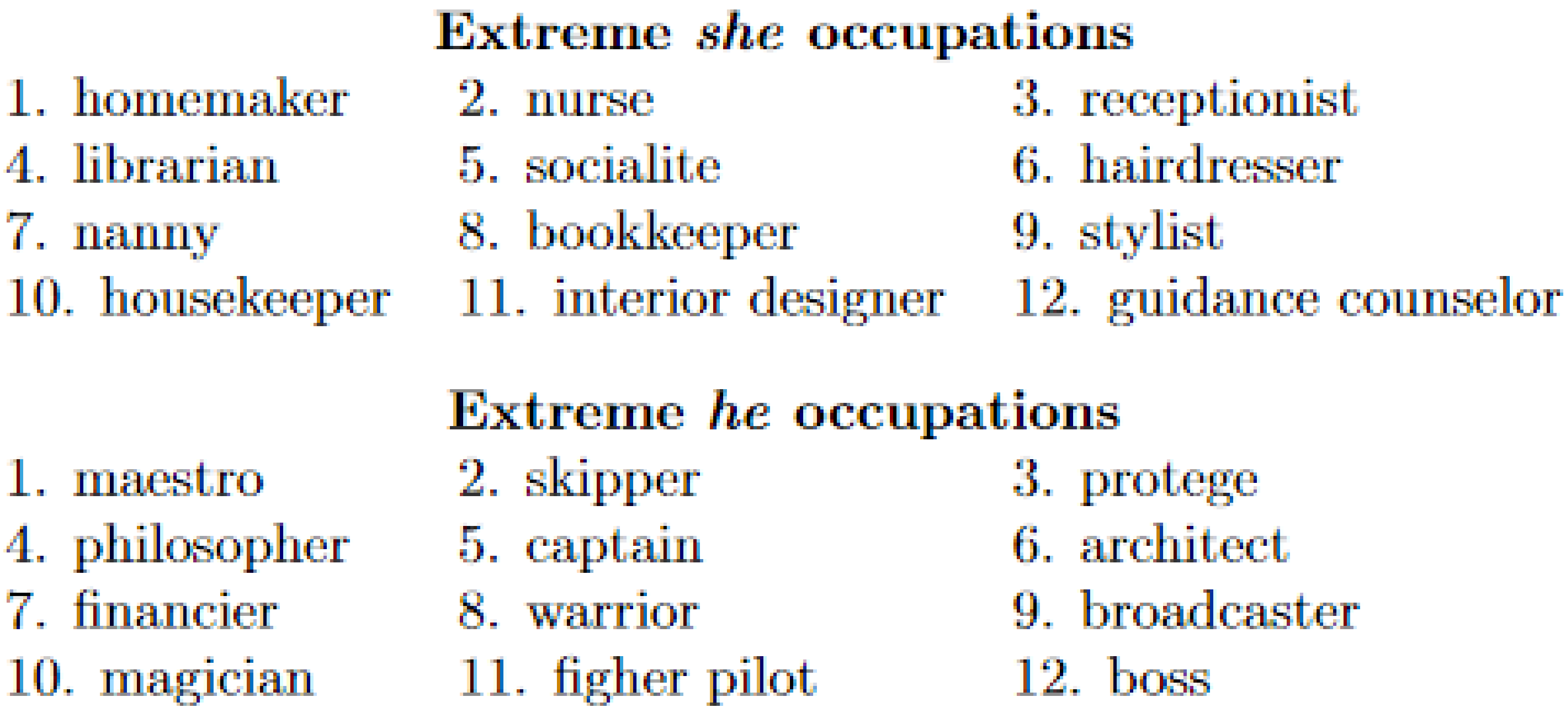
Started investigating failures of the word2vec model

Used word2vec to make analogies across genders:

- “Man is to X as woman is to Y”
- Set X = “Computer Programmer”
- Y = “Homemaker”

“Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”

Man is to Computer Programmer as Woman is to Homemaker?



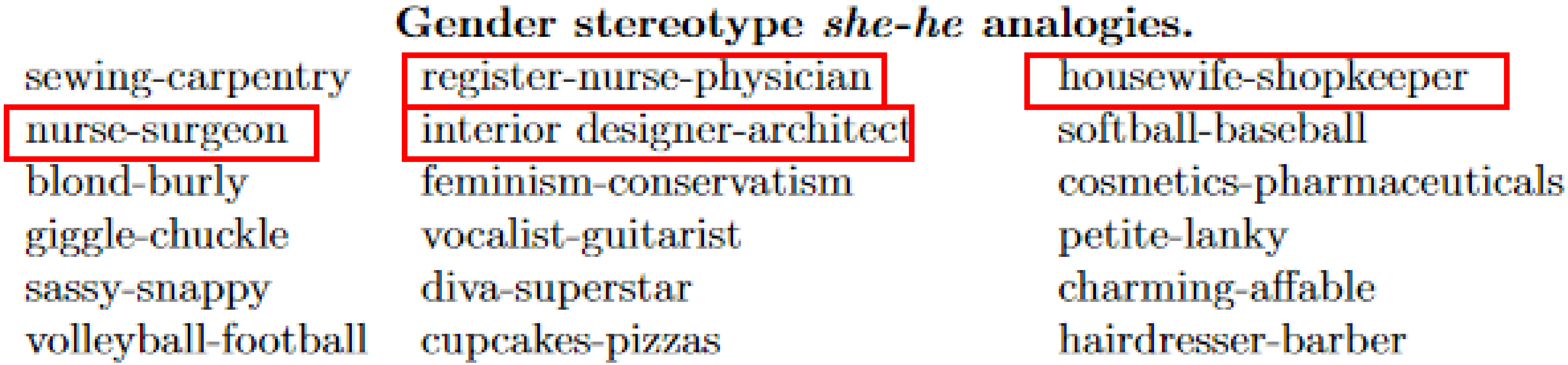
Extreme *she* occupations

1. homemaker
2. nurse
3. receptionist
4. librarian
5. socialite
6. hairdresser
7. nanny
8. bookkeeper
9. stylist
10. housekeeper
11. interior designer
12. guidance counselor

Extreme *he* occupations

1. maestro
2. skipper
3. protege
4. philosopher
5. captain
6. architect
7. financier
8. warrior
9. broadcaster
10. magician
11. figher pilot
12. boss

Man is to Computer Programmer as Woman is to Homemaker?



Gender stereotype *she-he* analogies.

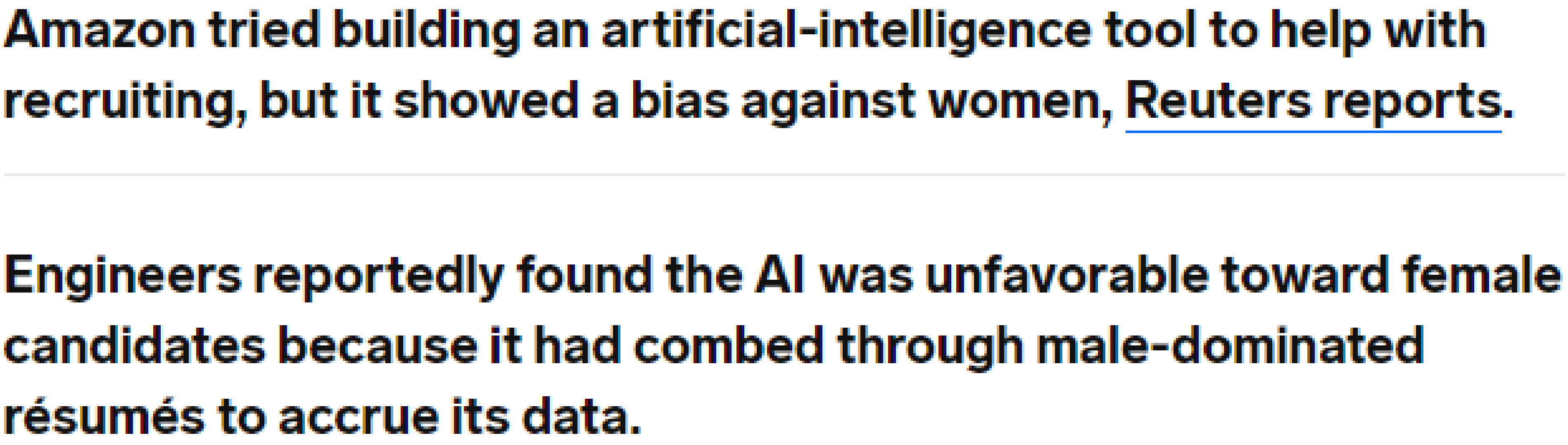
sewing-carpentry
nurse-surgeon
blond-burly
giggle-chuckle
sassy-snappy
volleyball-football

register-nurse-physician
interior designer-architect
feminism-conservatism
vocalist-guitarist
diva-superstar
cupcakes-pizzas

housewife-shopkeeper
softball-baseball
cosmetics-pharmaceuticals
petite-lanky
charming-affable
hairdresser-barber

What’s the harm?

Language models can be used to make decisions about individuals, and in turn can be discriminatory

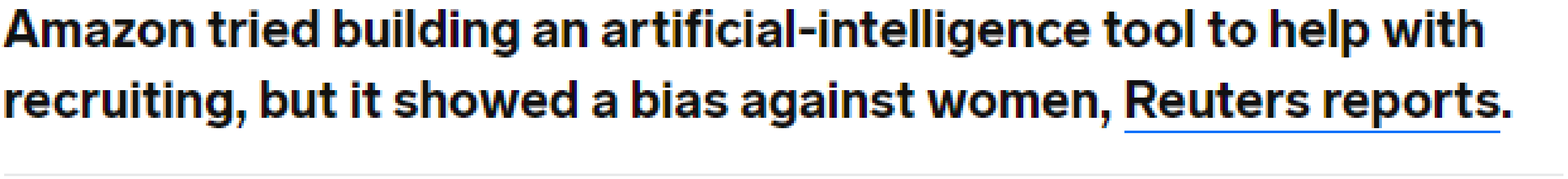


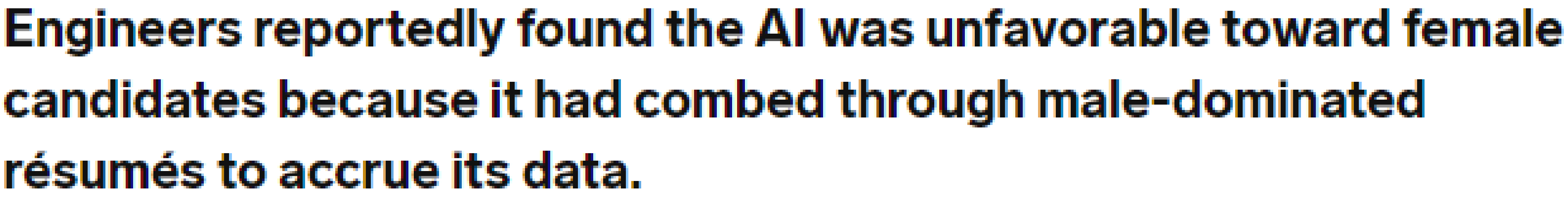
- Amazon tried building an artificial-intelligence tool to help with recruiting, but it showed a bias against women, Reuters reports.

- Engineers reportedly found the AI was unfavorable toward female candidates because it had combed through male-dominated résumés to accrue its data.

What’s the harm?

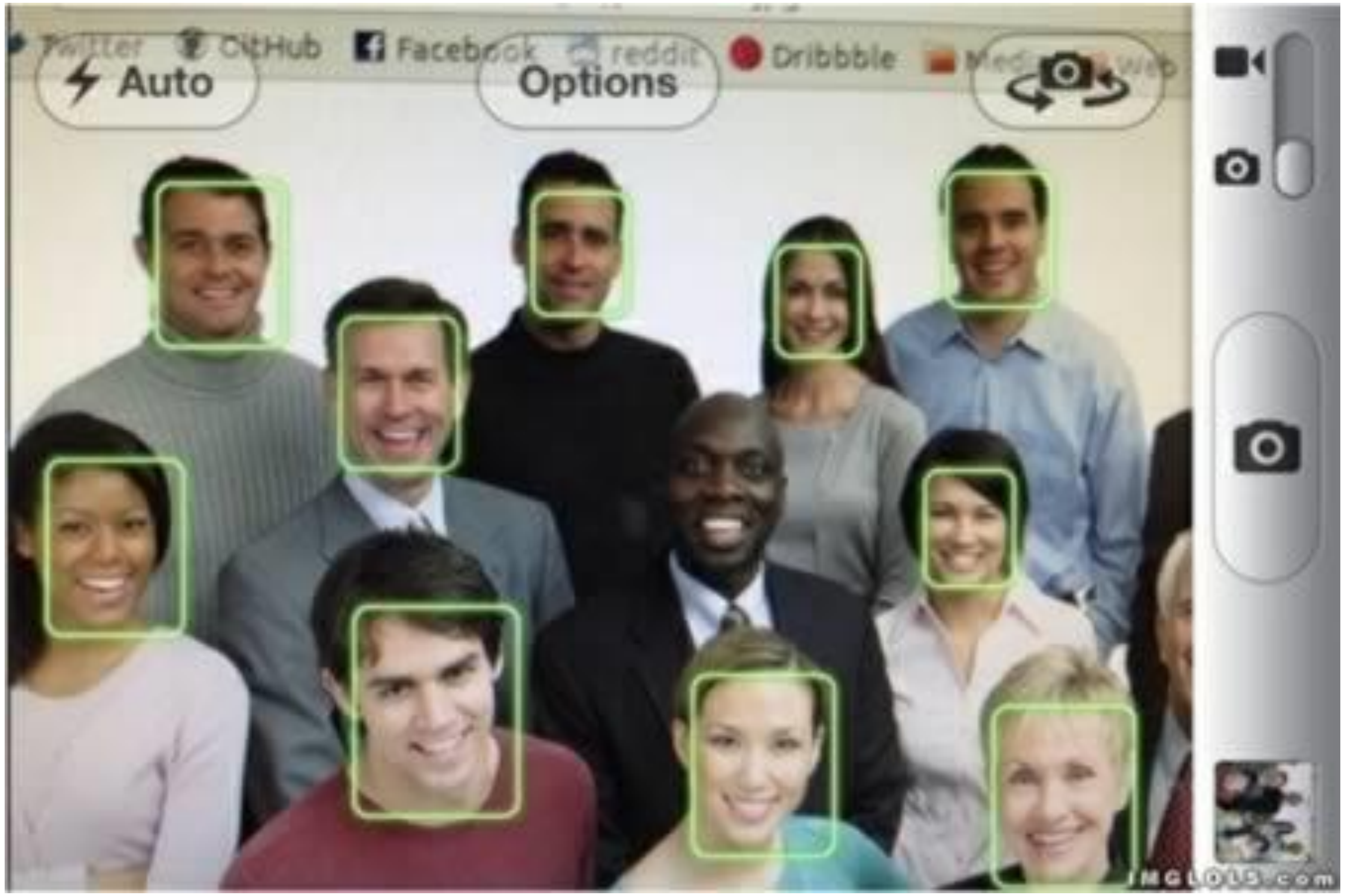
Language models can be used to make decisions about individuals, and in turn can be discriminatory

- 
  Amazon tried building an artificial-intelligence tool to help with recruiting, but it showed a bias against women, Reuters reports.

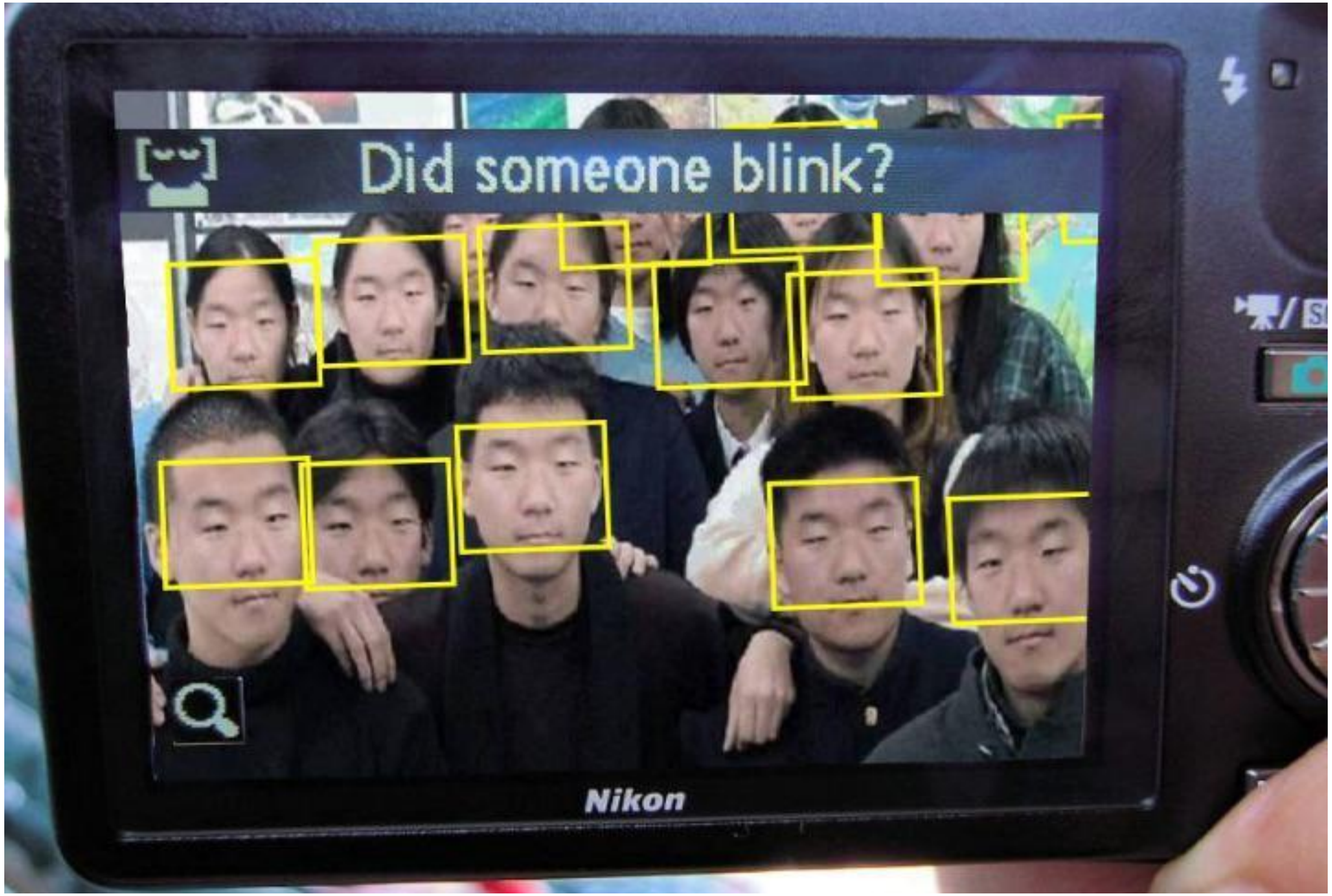
- 
  Engineers reportedly found the AI was unfavorable toward female candidates because it had combed through male-dominated résumés to accrue its data.

Facial recognition

Facial recognition



Facial recognition



Facial recognition

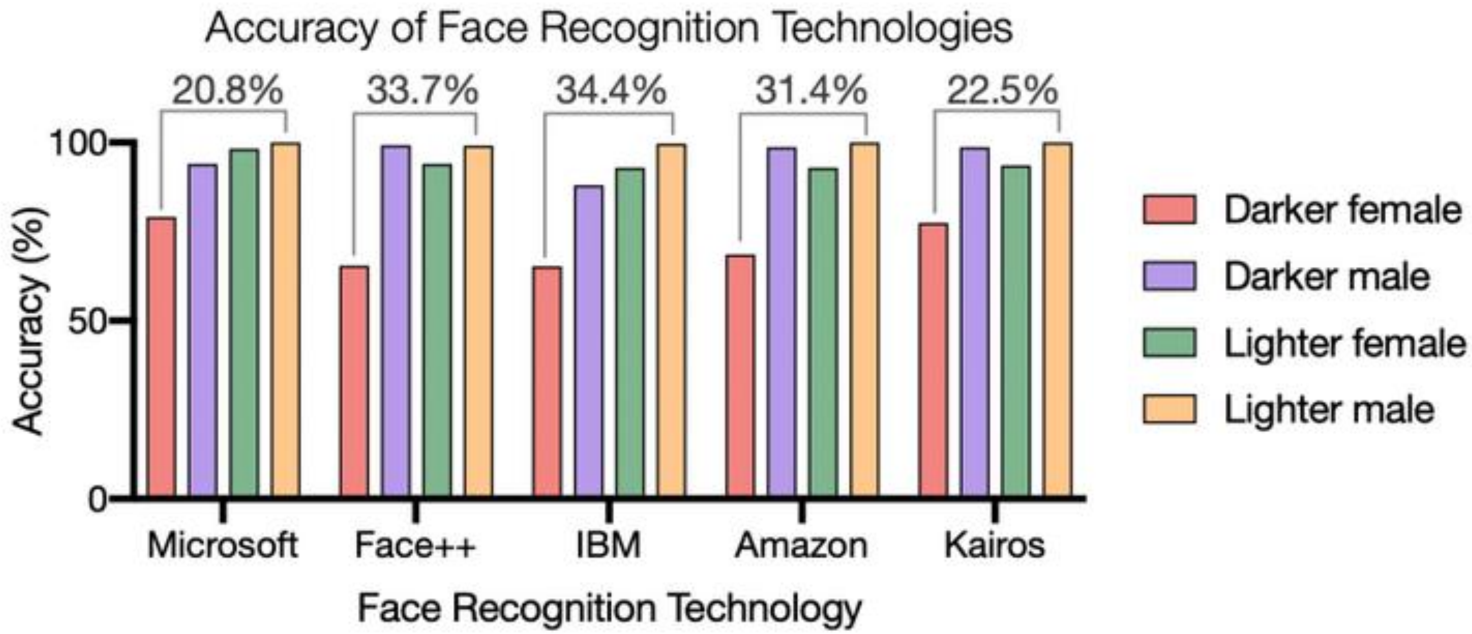
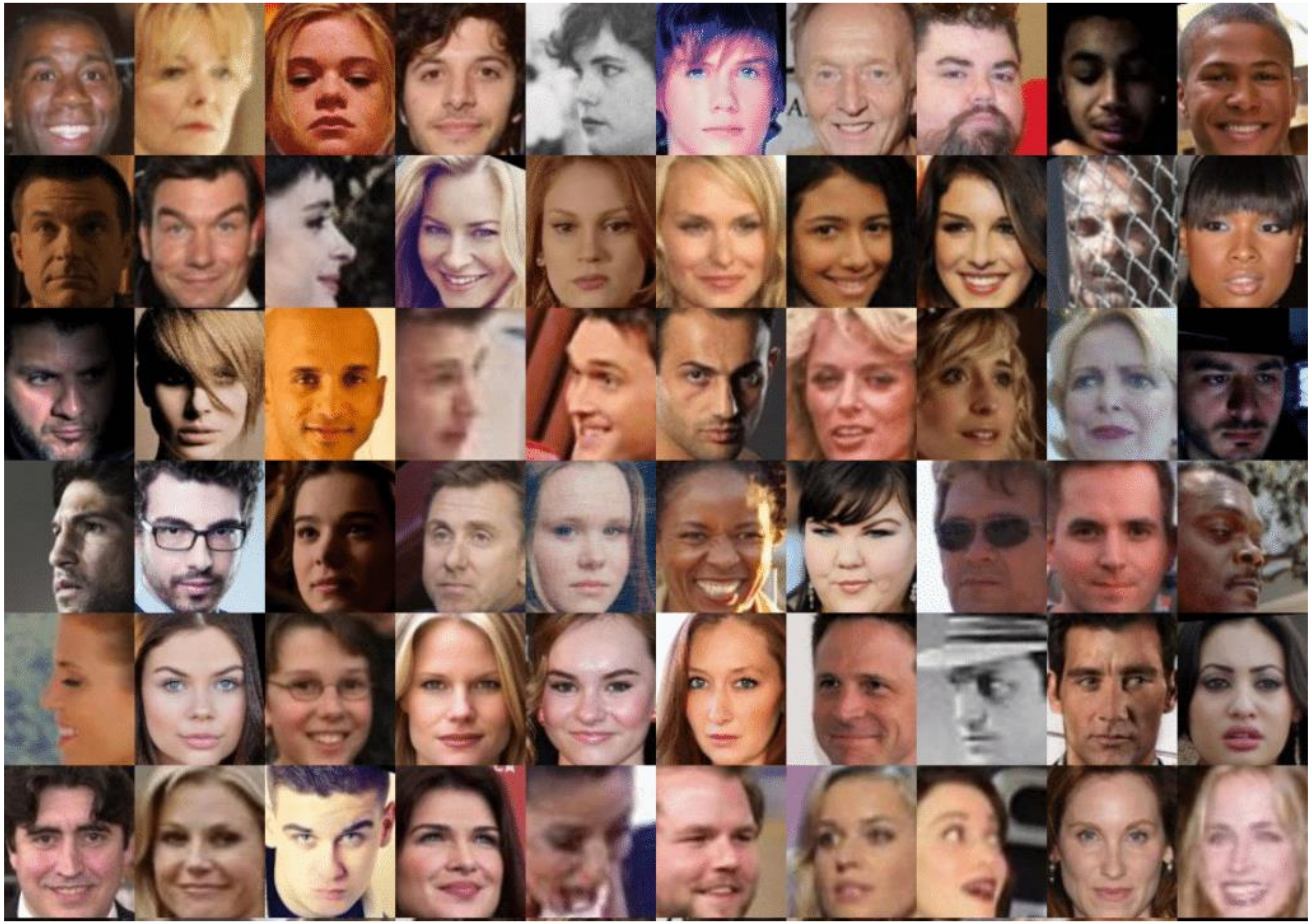
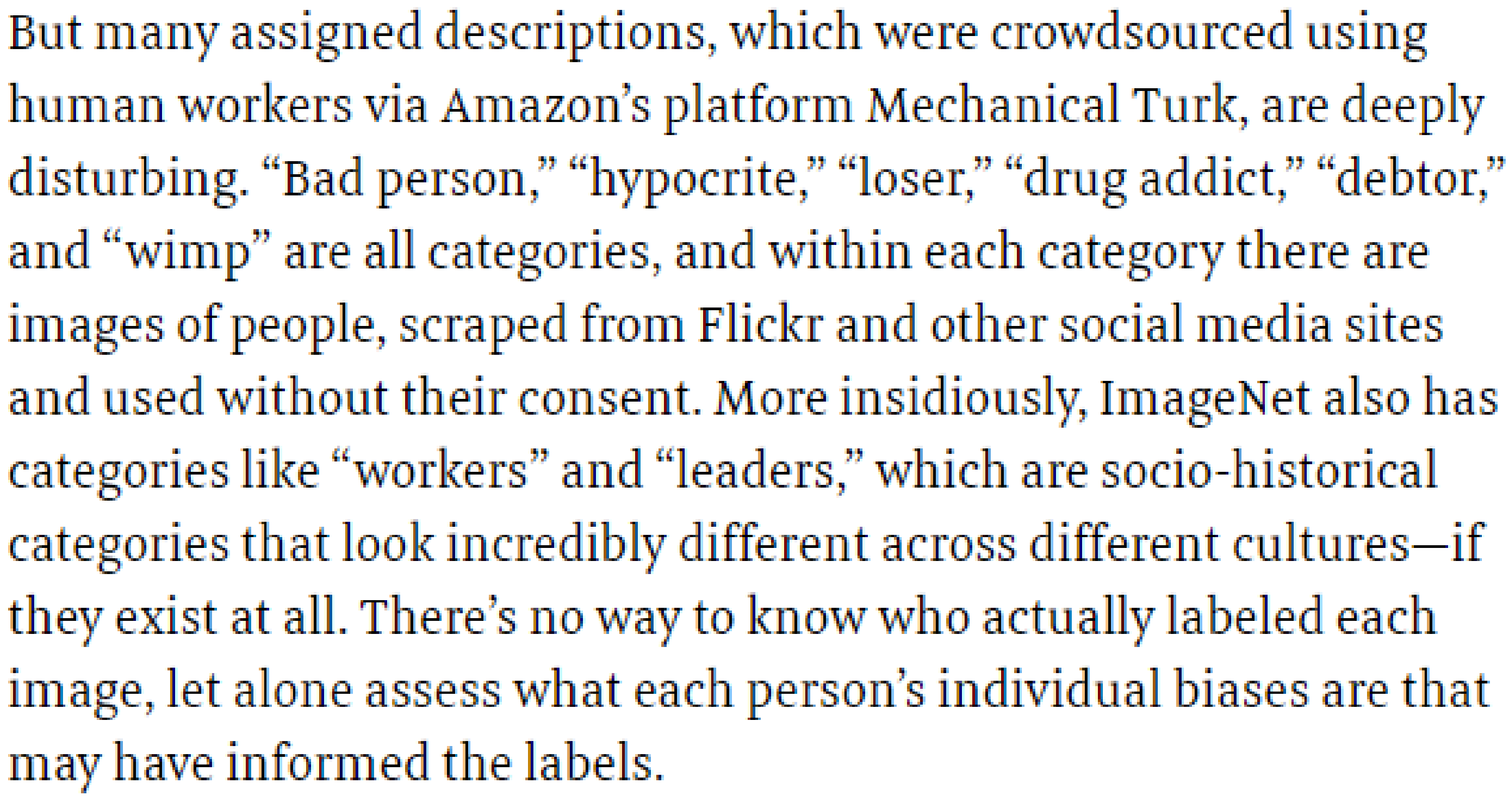


Figure 1: Auditing five face recognition technologies. *The* ***Gender Shades project*** *revealed* ***discrepancies*** *in the classification accuracy of face recognition technologies for different skin tones and sexes. These algorithms consistently demonstrated the poorest accuracy for darker-skinned females and the highest for lighter-skinned males.*

Facial recognition – why does this happen?



Facial recognition – why does this happen?



But many assigned descriptions, which were crowdsourced using human workers via Amazon's platform Mechanical Turk, are deeply disturbing. "Bad person," "hypocrite," "loser," "drug addict," "debtor," and "wimp" are all categories, and within each category there are images of people, scraped from Flickr and other social media sites and used without their consent. More insidiously, ImageNet also has categories like "workers" and "leaders," which are socio-historical categories that look incredibly different across different cultures—if they exist at all. There's no way to know who actually labeled each image, let alone assess what each person's individual biases are that may have informed the labels.

Facial recognition bias and consequences

Facial recognition is used today by:

- Airport and airline screening
- Public housing
- Employers
- Law enforcement
- Military drones

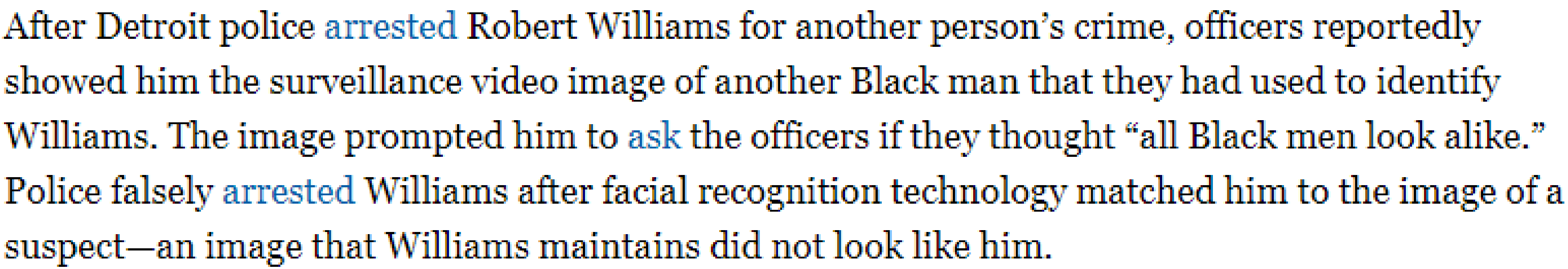
Since facial recognition algorithms are unfair, the harm that results from using them are disparate

Ex: arresting criminals based on facial recognition

Facial recognition bias can have big consequences

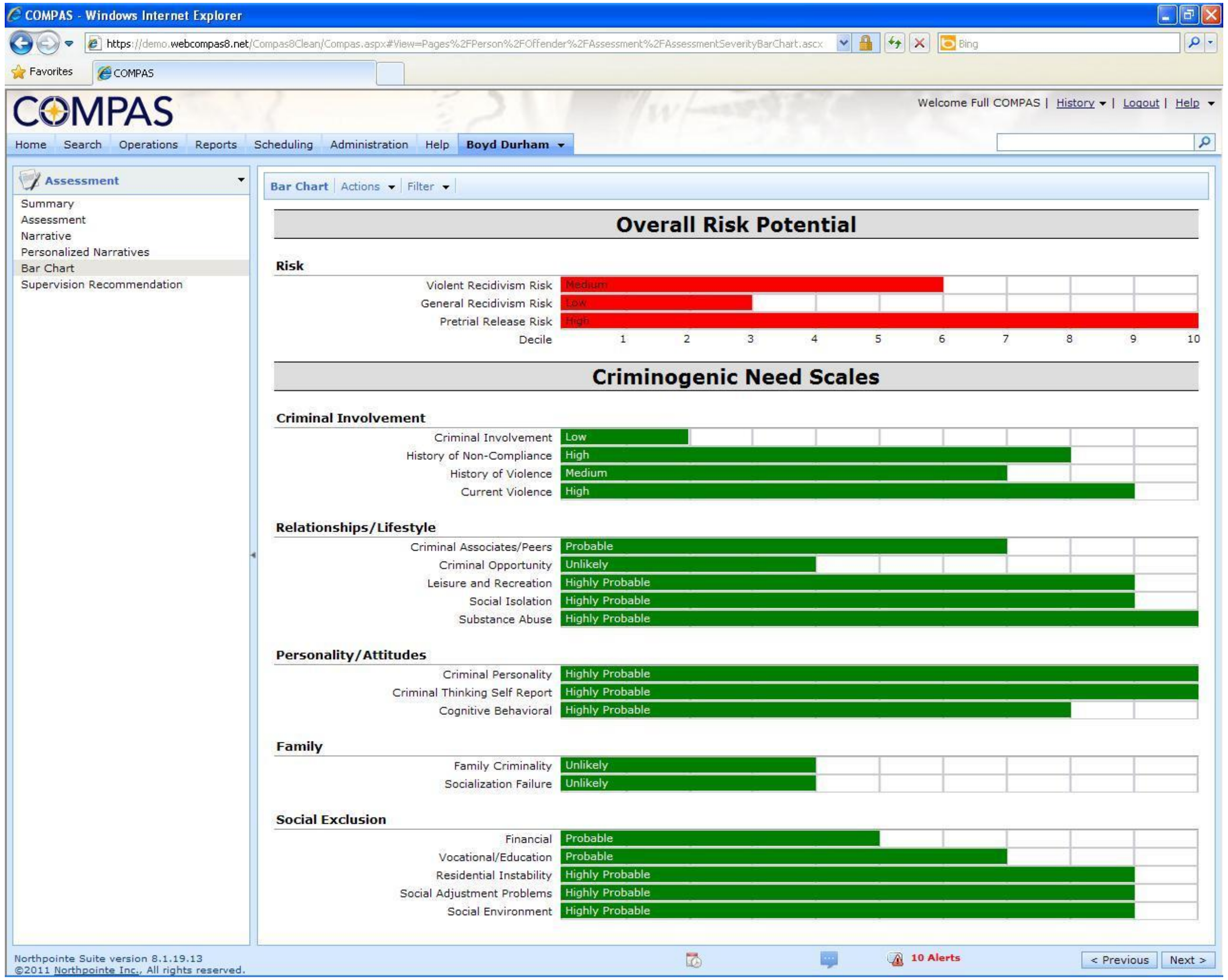
These algorithmic biases have major real-life implications. Several levels of law enforcement and U.S. Customs and Border Protection use facial recognition technology to support policing and airport screenings, respectively. This technology sometimes determines who receives housing or employment offers. One analyst at the American Civil Liberties Union reportedly warned that false matches "can lead to missed flights, lengthy interrogations, watch list placements, tense police encounters, false arrests, or worse." Even if developers can make the algorithms equitable, some advocates fear that law enforcement will employ the technology in a discriminatory manner, disproportionately harming marginalized populations.

Facial recognition bias can have big consequences



After Detroit police arrested Robert Williams for another person's crime, officers reportedly showed him the surveillance video image of another Black man that they had used to identify Williams. The image prompted him to ask the officers if they thought "all Black men look alike." Police falsely arrested Williams after facial recognition technology matched him to the image of a suspect—an image that Williams maintains did not look like him.

Algorithms in criminal justice: COMPAS



COMPAS - Windows Internet Explorer
https://demo.webcompas8.net/Compas8Clean/Compas.aspx#View=Pages%2FPerson%2FOffender%2FAssessment%2FAssessmentSeverityBarChart.ascx
Bing
Favorites
COMPAS
COMPAS
Welcome Full COMPAS | History | Logout | Help
Home
Search
Operations
Reports
Scheduling
Administration
Help
Boyd Durham
Assessment
Summary
Assessment
Narrative
Personalized Narratives
Bar Chart
Supervision Recommendation
Bar Chart
Actions
Filter

Overall Risk Potential

Risk

Violent Recidivism Risk	Medium
General Recidivism Risk	Low
Pretrial Release Risk	High

Decile
1
2
3
4
5
6
7
8
9
10

Criminogenic Need Scales

Criminal Involvement

Criminal Involvement	Low
History of Non-Compliance	High
History of Violence	Medium
Current Violence	High

Relationships/Lifestyle

Criminal Associates/Peers	Probable
Criminal Opportunity	Unlikely
Leisure and Recreation	Highly Probable
Social Isolation	Highly Probable
Substance Abuse	Highly Probable

Personality/Attitudes

Criminal Personality	Highly Probable
Criminal Thinking Self Report	Highly Probable
Cognitive Behavioral	Highly Probable

Family

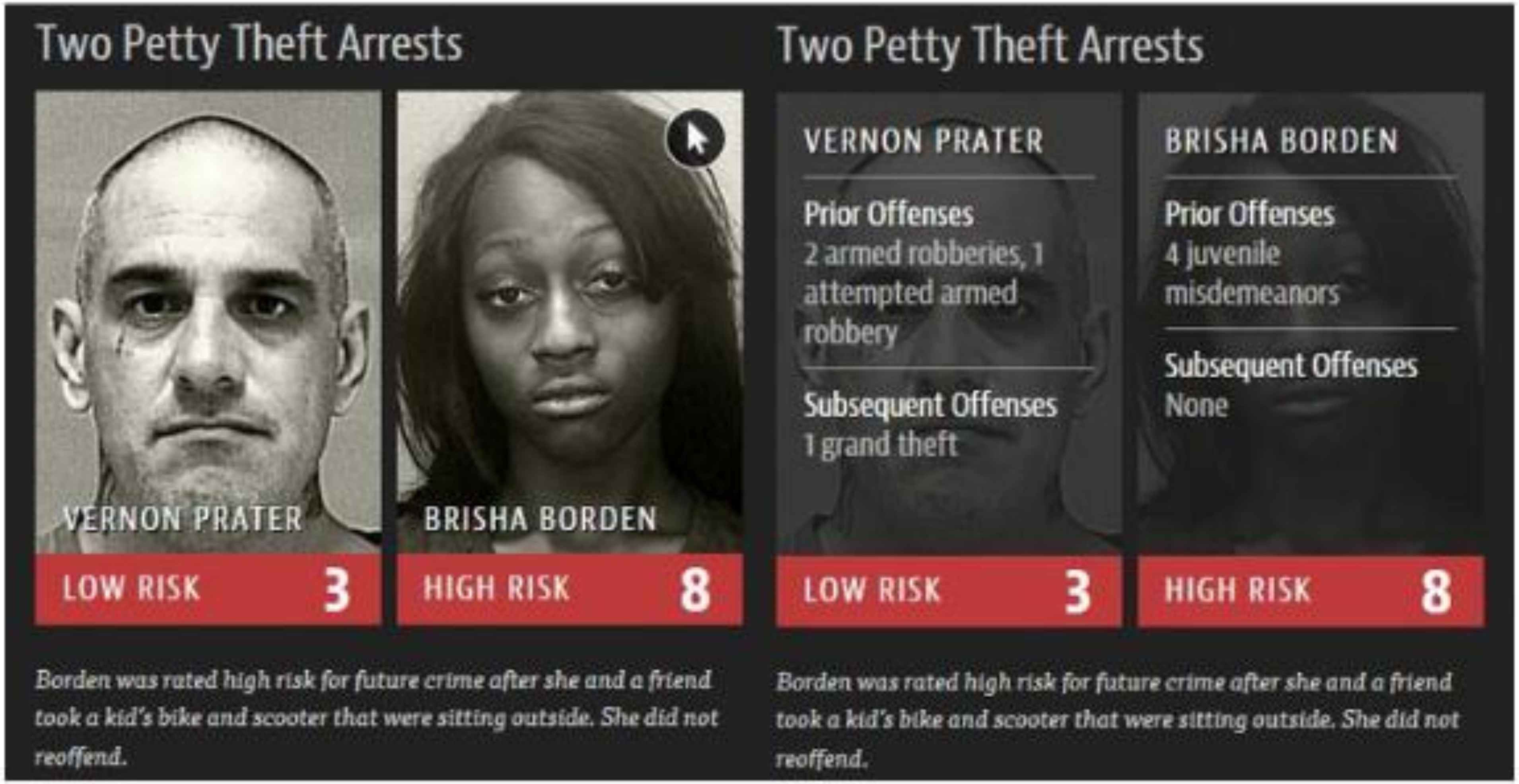
Family Criminality	Unlikely
Socialization Failure	Unlikely

Social Exclusion

Financial	Probable
Vocational/Education	Probable
Residential Instability	Highly Probable
Social Adjustment Problems	Highly Probable
Social Environment	Highly Probable

Northpointe Suite version 8.1.19.13
©2011 Northpointe Inc., All rights reserved.

Algorithms in criminal justice: COMPAS



Two Petty Theft Arrests

VERNON PRATER

LOW RISK 3

BRISHA BORDEN

HIGH RISK 8

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.

Two Petty Theft Arrests

VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

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.

Algorithms in loan/mortgage decisions

Algorithms are a big part of loan/mortgage/credit decisions today

Loan decisions are biased:

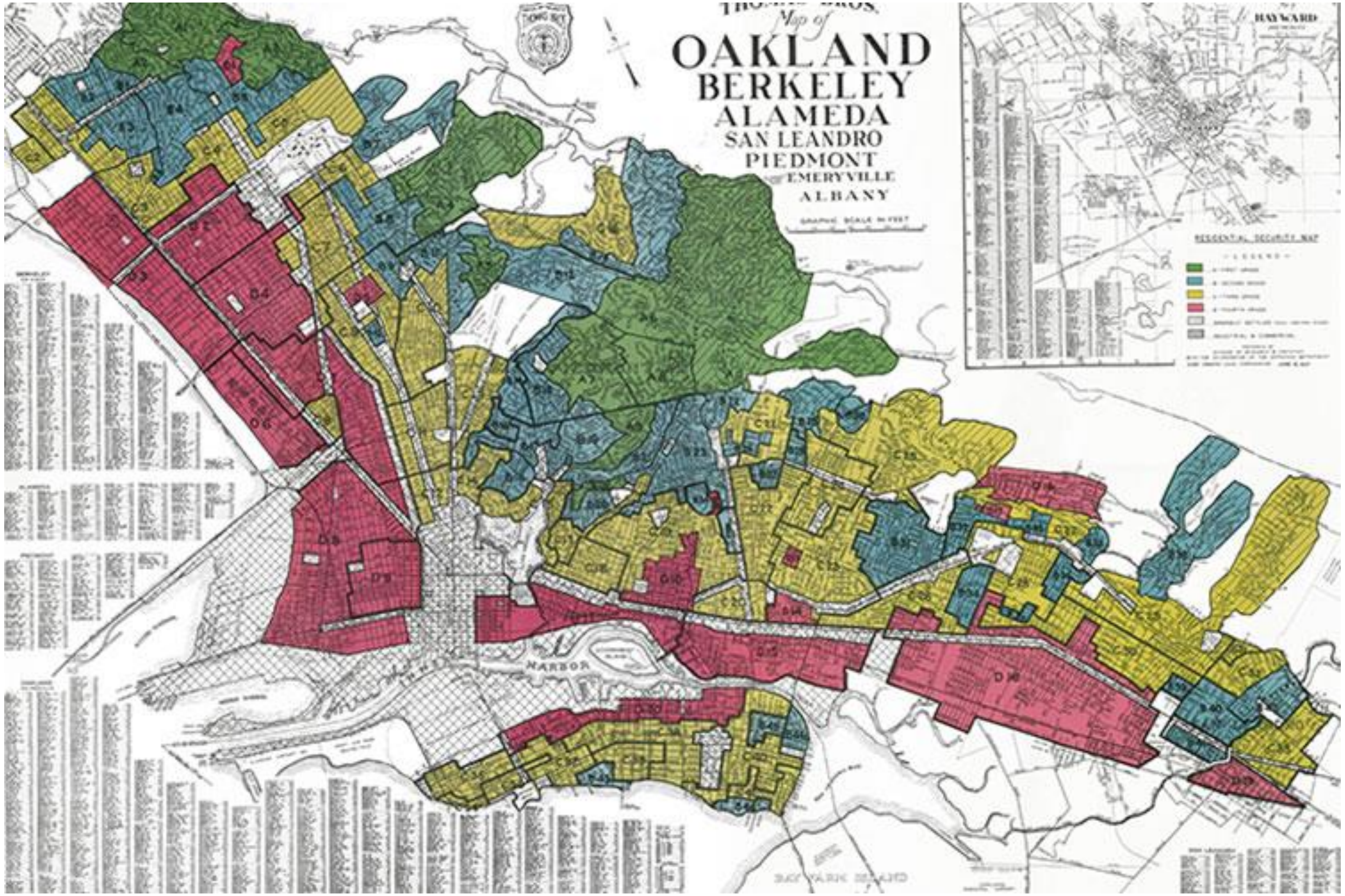
At equal financial situation, people of color are less likely to get the same loan/mortgage.

Why?

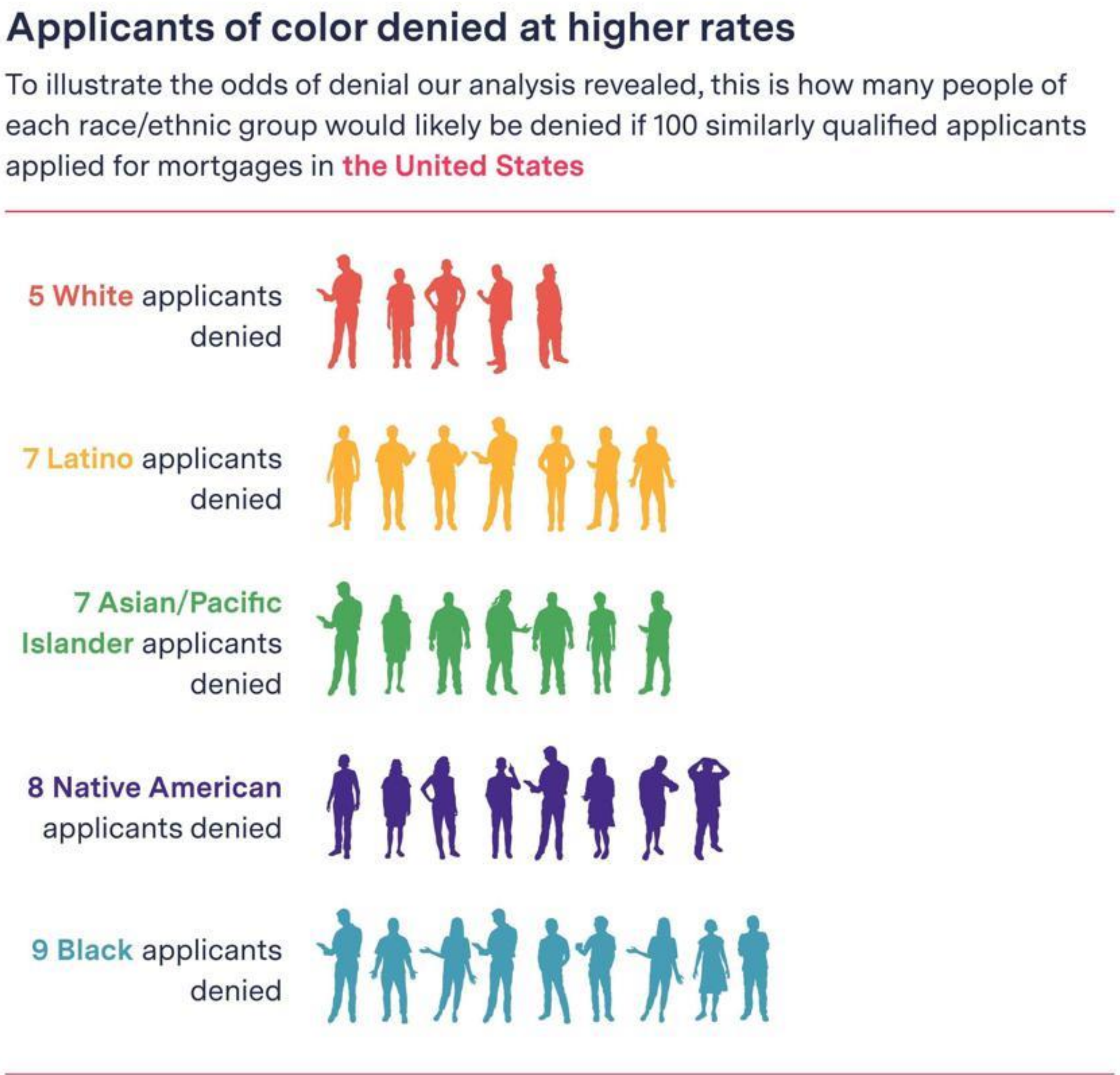
- Historically, loan decisions have been racist ➔ biased training data
- Algorithms can learn to mimic that human bias/racism

Algorithms in loan/mortgage decisions

Redlining



Algorithms in loan/mortgage decisions



Applicants of color denied at higher rates

To illustrate the odds of denial our analysis revealed, this is how many people of each race/ethnic group would likely be denied if 100 similarly qualified applicants applied for mortgages in **the United States**

5 White applicants denied

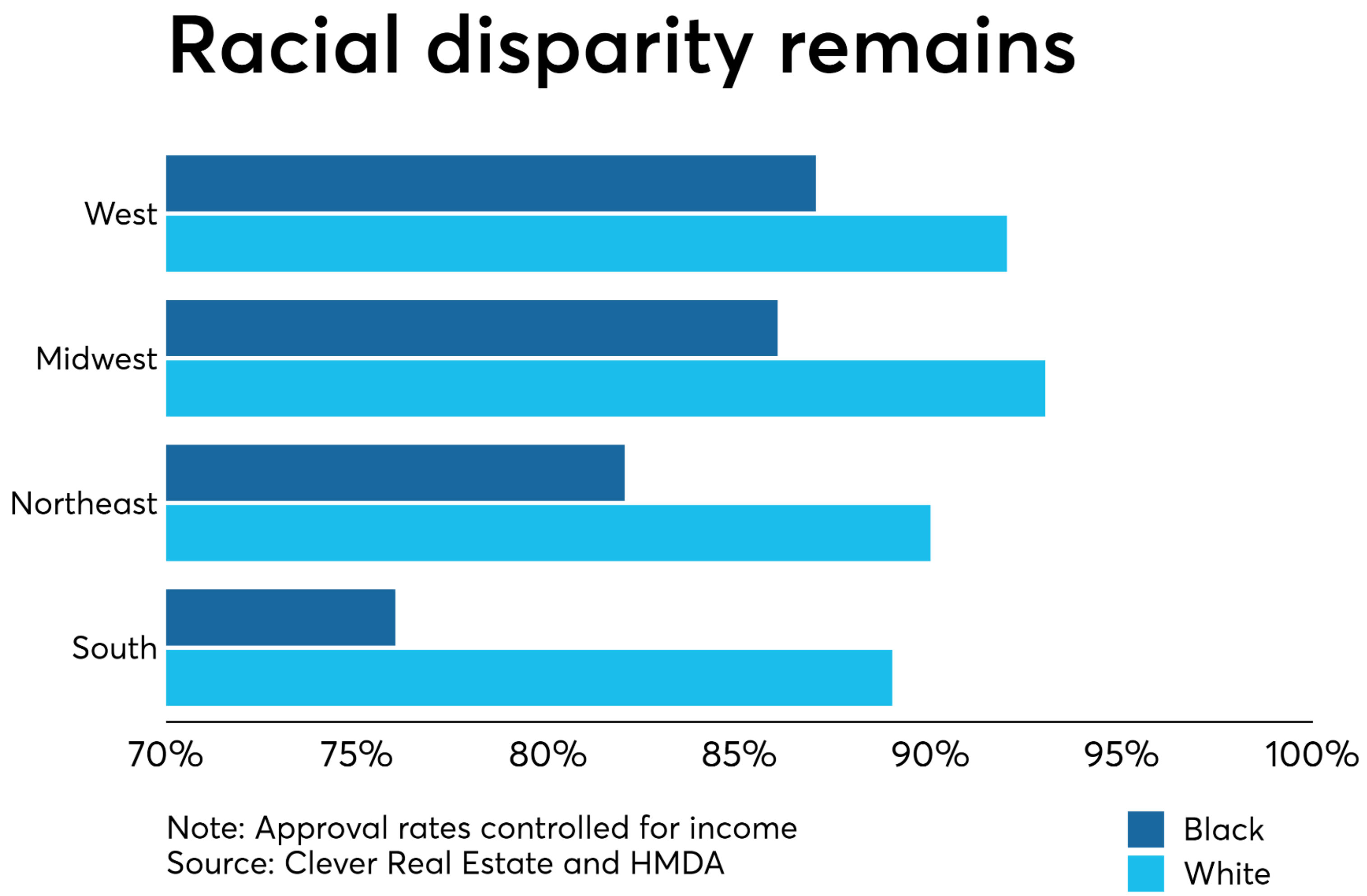
7 Latino applicants denied

7 Asian/Pacific Islander applicants denied

8 Native American applicants denied

9 Black applicants denied

Algorithms in loan/mortgage decisions



Racial disparity remains

Note: Approval rates controlled for income
Source: Clever Real Estate and HMDA

Algorithms in loan/mortgage decisions

Loan algorithms not only perpetuate bias…

Algorithms in loan/mortgage decisions

Loan algorithms not only perpetuate bias…

… But they can also amplify it

Algorithms in loan/mortgage decisions

Loan algorithms not only perpetuate bias…

… But they can also amplify it

Less likely to approve loans for people of color

➔ get little data about likeliness to repay loan

➔ makes approving loans for POCs risky and “undesirable”

missing label problem

Algorithms in loan/mortgage decisions

Loan algorithms not only perpetuate bias…

… But they can also amplify it

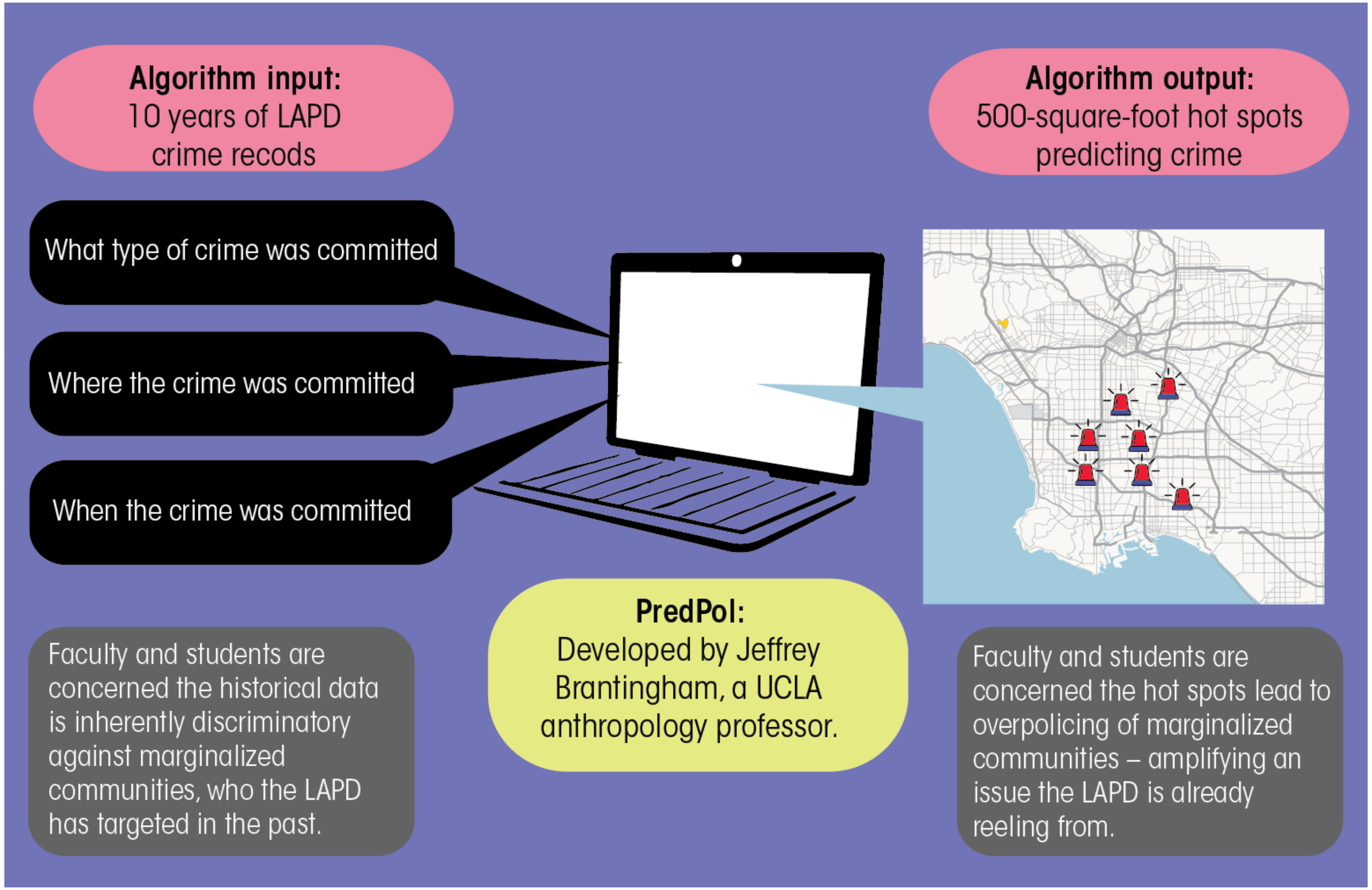
Less likely to approve loans for people of color

➔ get little data about likeliness to repay loan

➔ makes approving loans for POCs risky and “undesirable”

Vicious cycle/feedback loop ➔ less and less fairness over time

Predictive policing: PredPol



Algorithm input:
10 years of LAPD
crime recods

What type of crime was committed

Where the crime was committed

When the crime was committed

Algorithm output:
500-square-foot hot spots
predicting crime

PredPol:
Developed by Jeffrey
Brantingham, a UCLA
anthropology professor.

Faculty and students are concerned the historical data is inherently discriminatory against marginalized communities, who the LAPD has targeted in the past.

Faculty and students are concerned the hot spots lead to overpolicing of marginalized communities – amplifying an issue the LAPD is already reeling from.

Predictive policing & bad feedback loops

Racist data showing more people of colors arrested in the past

➔ Spend more police resources in areas with more POCs

➔ Arrest more POCs (not necessarily because of higher crime rate, but higher detection rate)

Predictive policing & bad feedback loops

Racist data showing more people of colors arrested in the past

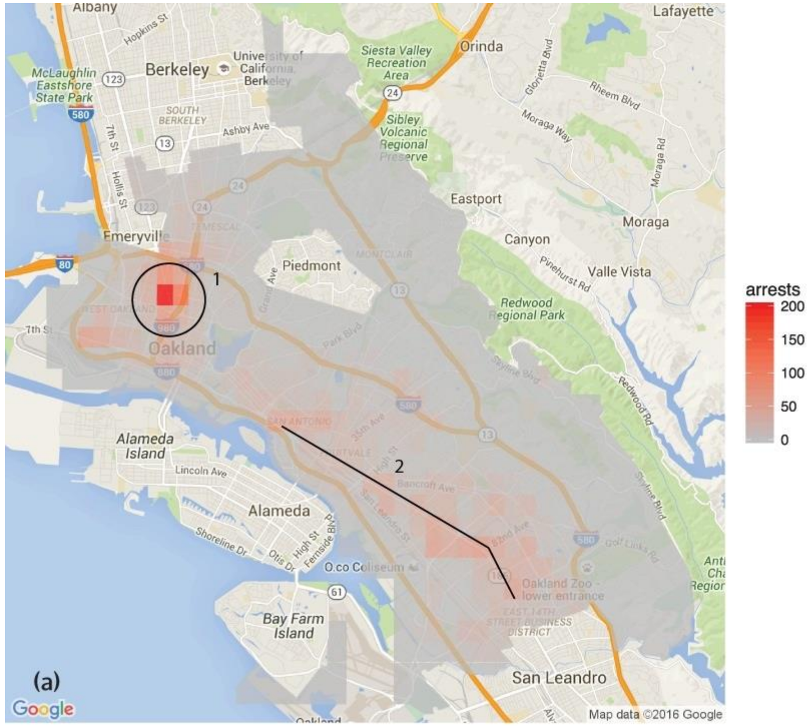
➔ Spend more police resources in areas with more POCs

➔ Arrest more POCs (not necessarily because of higher crime rate, but higher detection rate)

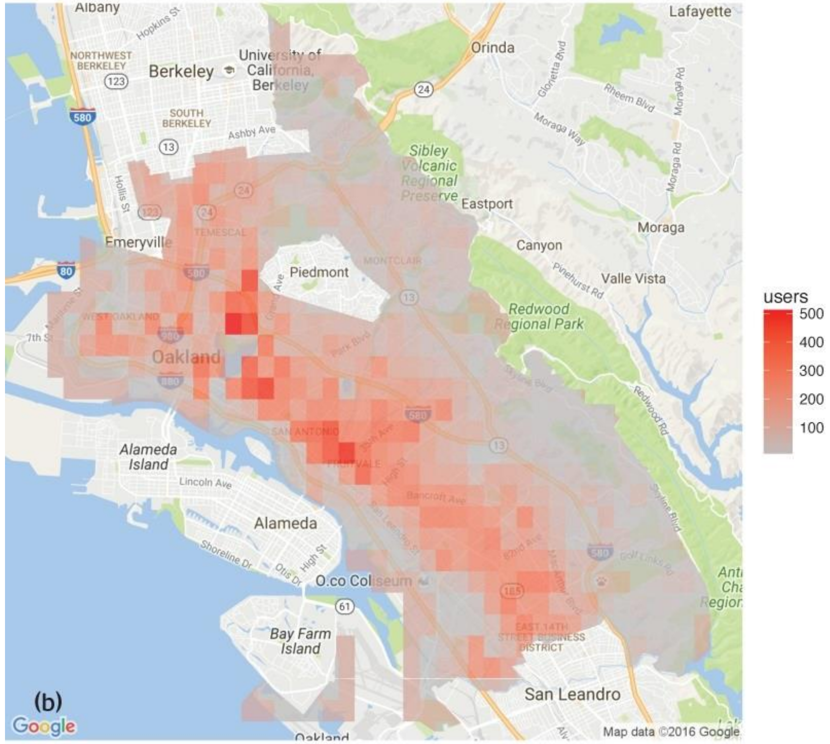
Predictive policing algorithms can get stuck in negative feedback loops that allocate more to certain areas independently of crime rates

missing label problem again

Predictive policing & bad feedback loops



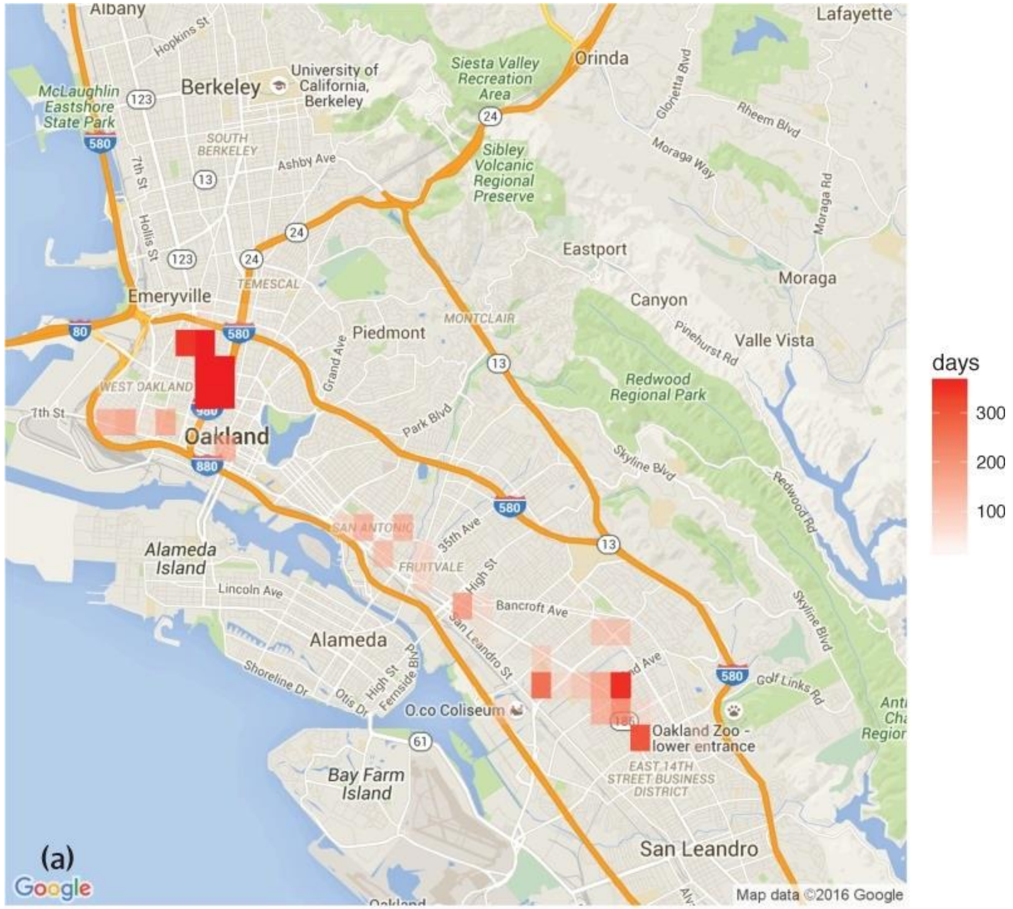
Drug arrests made by the Oakland Police Department, 2010

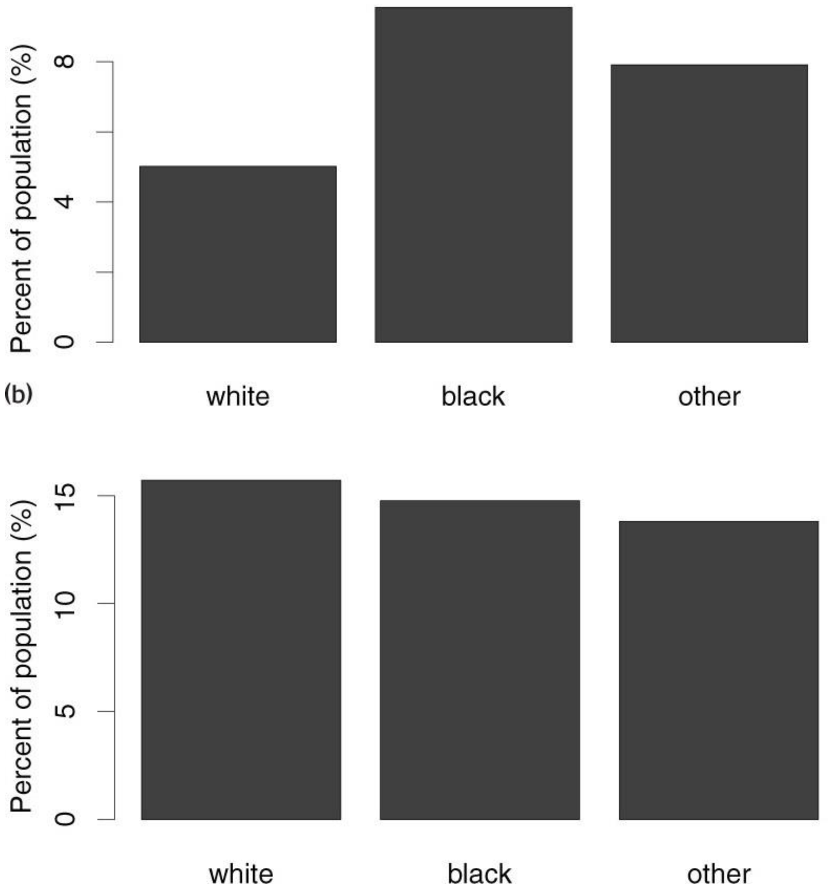


Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

“To predict and serve?” Kristian Lum, William Isaac

Predictive policing & bad feedback loops



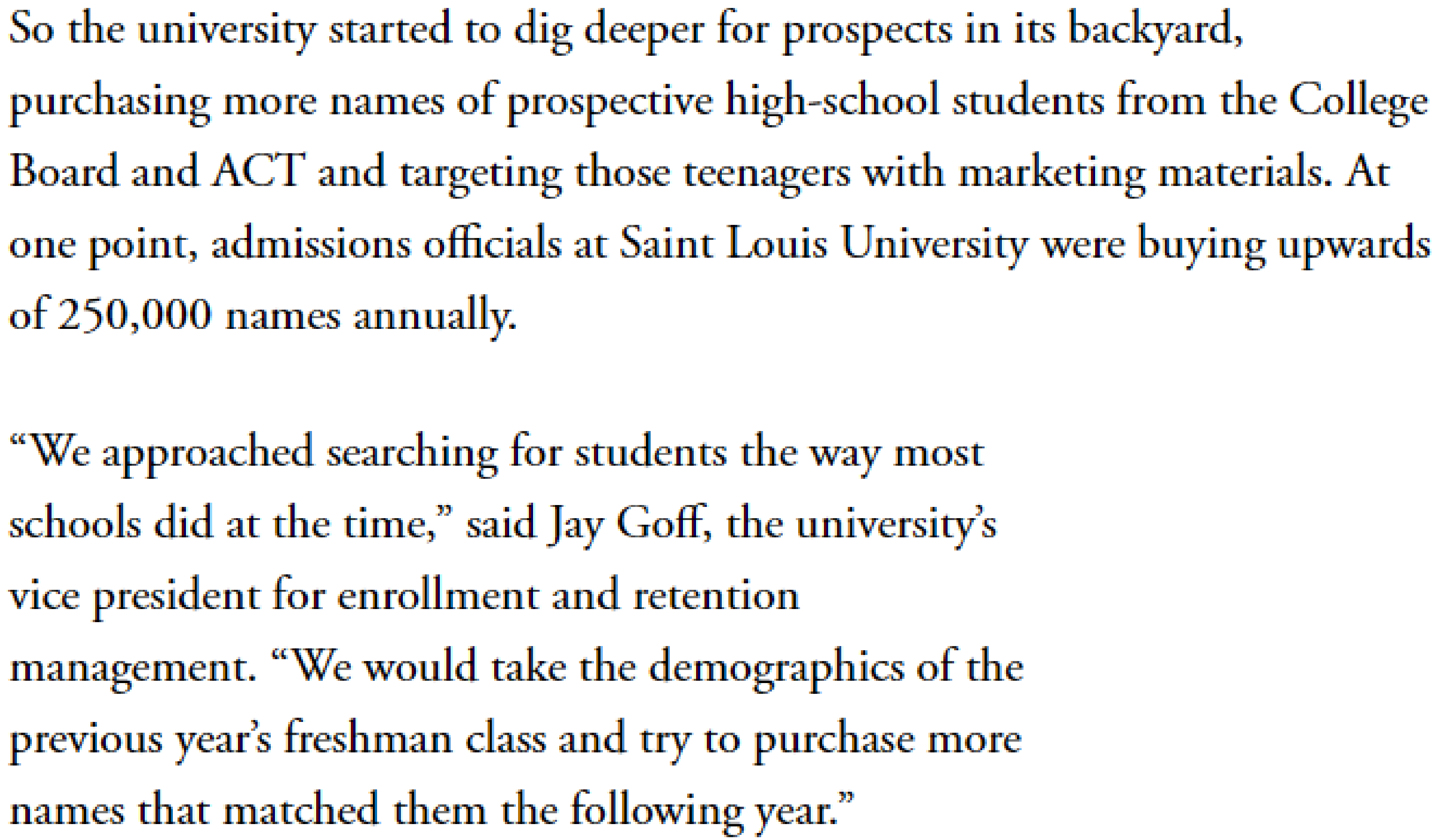


Populations targeted by PredPol

Estimated drug use

“To predict and serve?” Kristian Lum, William Isaac

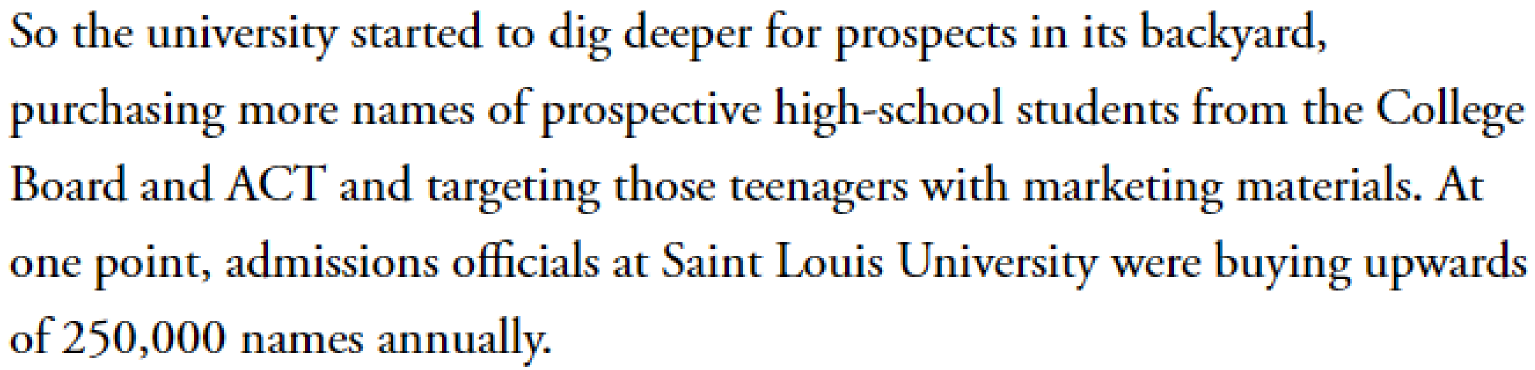
Big Data in College Admission Marketing



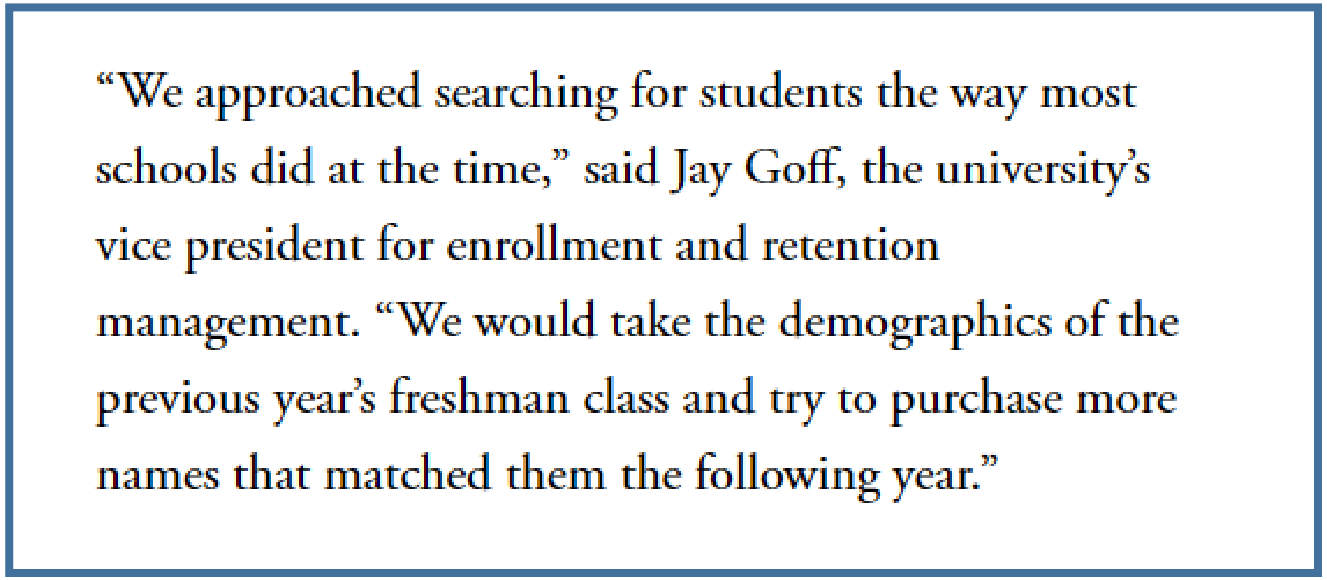
So the university started to dig deeper for prospects in its backyard, purchasing more names of prospective high-school students from the College Board and ACT and targeting those teenagers with marketing materials. At one point, admissions officials at Saint Louis University were buying upwards of 250,000 names annually.

"We approached searching for students the way most schools did at the time," said Jay Goff, the university's vice president for enrollment and retention management. "We would take the demographics of the previous year's freshman class and try to purchase more names that matched them the following year."

Big Data in College Admission Marketing



So the university started to dig deeper for prospects in its backyard, purchasing more names of prospective high-school students from the College Board and ACT and targeting those teenagers with marketing materials. At one point, admissions officials at Saint Louis University were buying upwards of 250,000 names annually.



"We approached searching for students the way most schools did at the time," said Jay Goff, the university's vice president for enrollment and retention management. "We would take the demographics of the previous year's freshman class and try to purchase more names that matched them the following year."

Targeting majority population

→ More applicants from majority population

→ More admitted students from majority population

→ Target the majority population more

Why is fairness not “trivial”?

What is even the right notion of fairness?

Two legal doctrines in the U.S.:

Disparate Treatment vs **Disparate Impact**

What is even the right notion of fairness?

Two legal doctrines in the U.S.:

Disparate Treatment vs **Disparate Impact**

- Intentional discrimination
- Taking race into account in decisions

What is even the right notion of fairness?

Two legal doctrines in the U.S.:

Disparate Treatment

- Intentional discrimination
- Taking race into account in decisions

vs

Disparate Impact

- Avoidable harm or discrimination
- Possibly indirect

What is even the right notion of fairness?

Two legal doctrines in the U.S.:

Disparate Treatment vs **Disparate Impact**

But disparate treatment and impact are often incompatible!

(Un)fairness through unawareness

Common critique:

“Fairness should be easy. How can it be unfair if we do not take sensitive attributes like race into account?”

(Un)fairness through unawareness

Common critique:

“Fairness should be easy. How can it be unfair if we do not take sensitive attributes like race into account?”

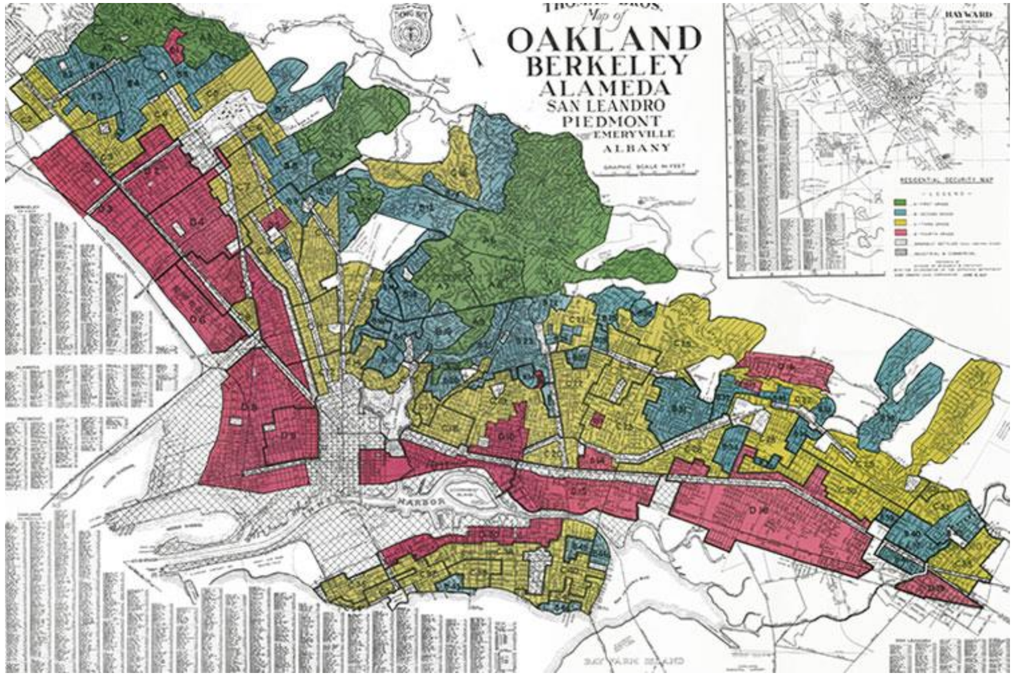
Issue #1:

- Algorithms are never truly “race-blind” or “unaware”
- Correlation across features can reveal information about sensitive attribute, even if sensitive attribute is not explicitly used

(Un)fairness through unawareness

Common critique:

“Fairness should be easy. How can it be unfair if we do not take sensitive attributes like race into account?”



Back to redlining example:

- Can use zip code/etc. as a proxy for race
- Even if race is hidden, can use zip code to discriminate

(Un)fairness through unawareness

Common critique:

“Fairness should be easy. How can it be unfair if we do not take sensitive attributes like race into account?”

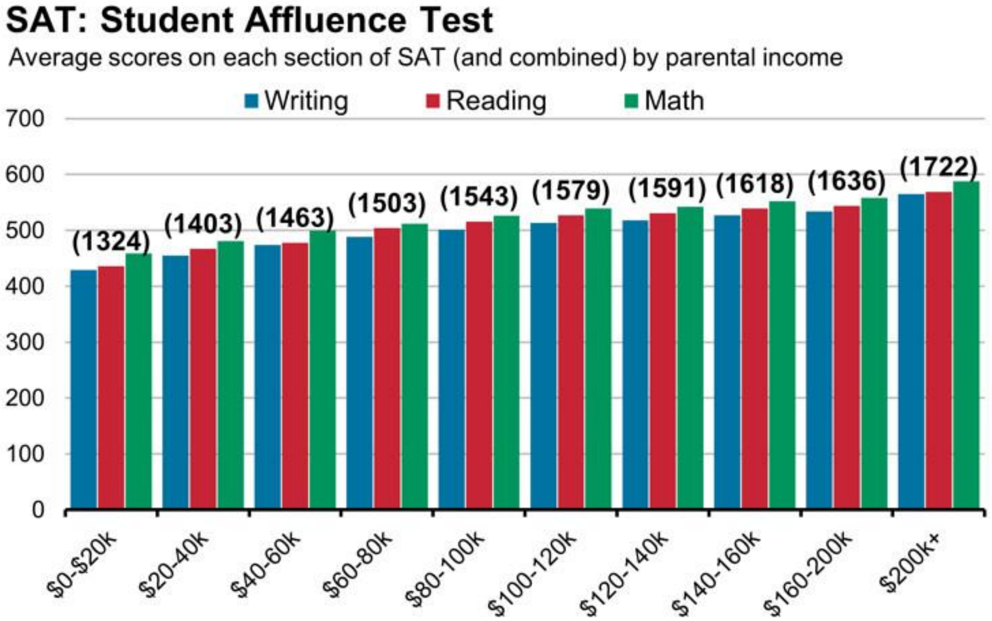
Issue #2:

- Observed features are only proxies for someone’s “true” attributes
- Same value for a given observed feature may have different meanings across different populations, and cannot be treated the same

(Un)fairness through unawareness

Common critique:

“Fairness should be easy. How can it be unfair if we do not take sensitive attributes like race into account?”



SAT: Student Affluence Test

Average scores on each section of SAT (and combined) by parental income

Source: FairTest, College Board | WSJ.com

(Un)fairness through unawareness

Common critique:

“Fairness should be easy. How can it be unfair if we do not take sensitive attributes like race into account?”

Why would wealthier populations have advantages on SAT?

- Access to better preparation for SAT ➔ higher scores at everything else equal
- Can take the SAT several times, until get desired score

So to design “fair” algorithms, we often need to take fairness explicitly into account…

Many possible sources of unfairness

Previous examples showed different reasons for unfairness:

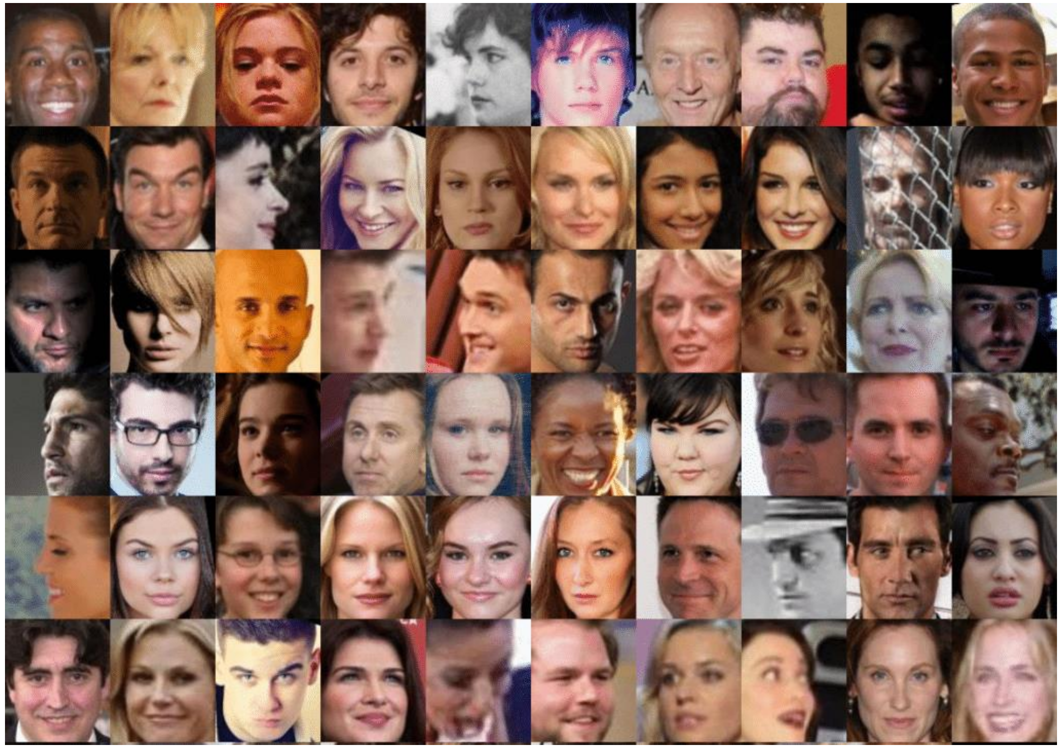
- Data bias (probably the most obvious one)

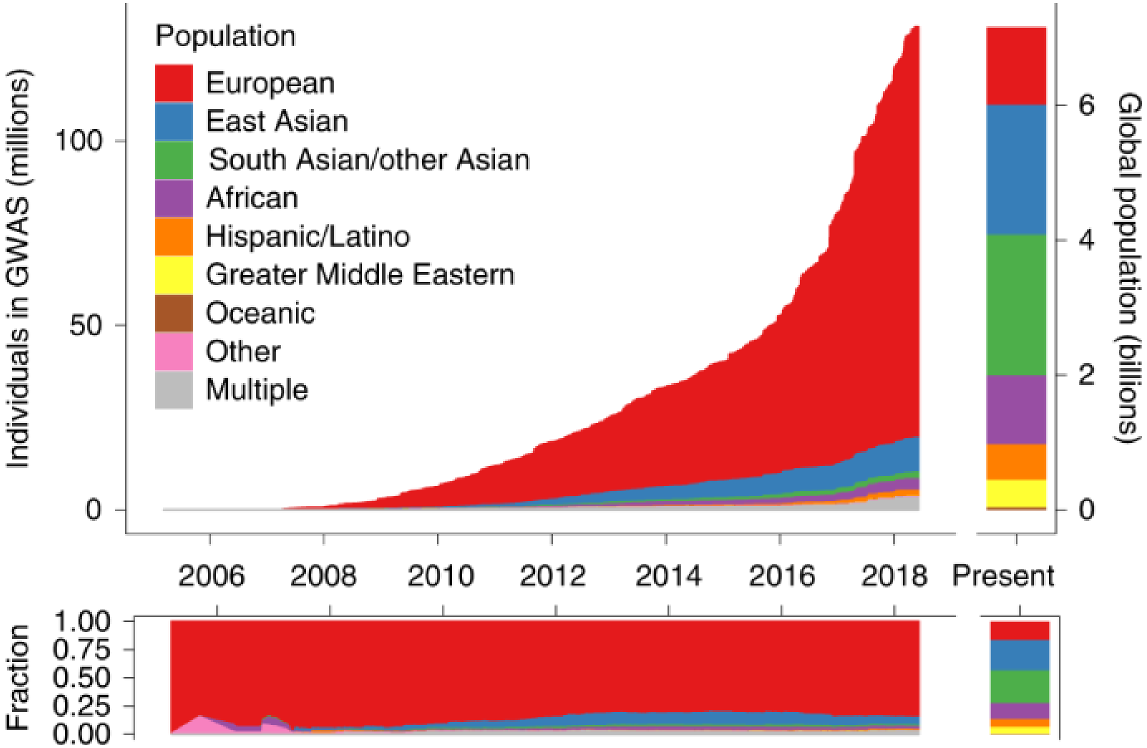
Sources of unfairness:

Previous examples showed different reasons for unfairness:

- Data bias
- Imbalance in the quantity of data across groups

Sources of unfairness: data imbalance



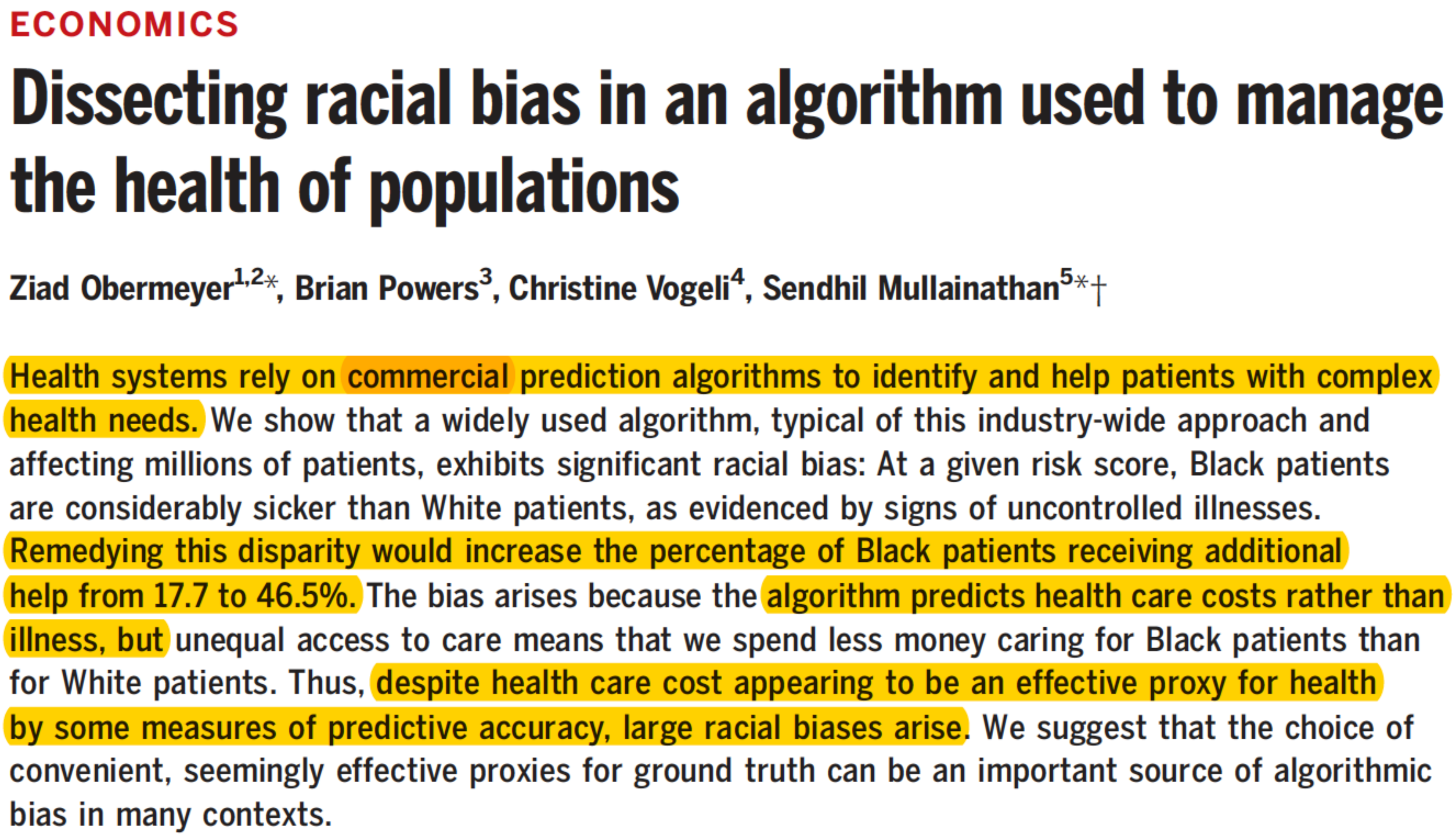


Sources of unfairness

Previous examples showed different reasons for unfairness:

- Data bias
- Imbalance in the quantity of data across groups
- Only access to proxies rather than true, intrisic attributes

Sources of unfairness: proxies



ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and 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%. The 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. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

Sources of unfairness

Previous examples showed different reasons for unfairness:

- Data bias
- Imbalance in the quantity of data across groups
- Only access to proxies rather than true, intrisic attributes
- Learning objective causes unanticipated bias

Sources of unfairness: learning objective

Example 1: ad auction

- Two type of agents: A and B.
- Advertiser 1 has value (1,1), advertiser 2 (0,1.1). Single ad slot.

Sources of unfairness: learning objective

Example 1: ad auction

- Two type of agents: A and B.
- Advertiser 1 has value (1,1), advertiser 2 (0,1.1). Single ad slot.

Objective: max social welfare

- If type = A, allocate to advertiser 1
- If type = B, allocate to advertiser 2

Sources of unfairness: learning objective

Example 1: ad auction

- Two type of agents: A and B.
- Advertiser 1 has value (1,1), advertiser 2 (0,1.1). Single ad slot.

Objective: max social welfare

- If type = A, allocate to advertiser 1
- If type = B, allocate to advertiser 2

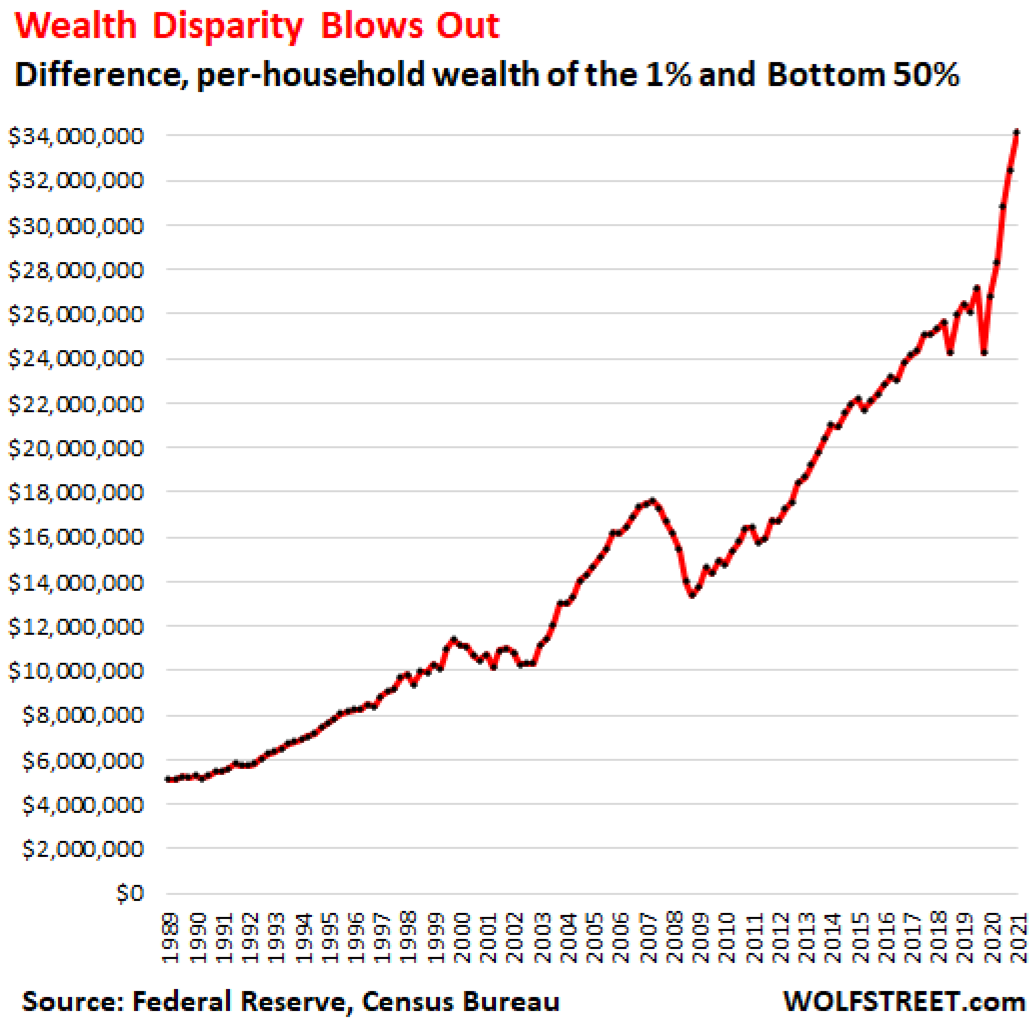
A = male, B = female; advertiser 1 = higher ed, ad 2 = maternity clothes

Sources of unfairness

Previous examples showed different reasons for unfairness:

- Data bias
- Imbalance in the quantity of data across groups
- Only access to proxies rather than true, intrinsic attributes
- Learning objective causes unanticipated bias
- Feedback loops/long-term and composed effects

Sources of unfairness: feedback loops



Wealth Disparity Blows Out

Difference, per-household wealth of the 1% and Bottom 50%

Source: Federal Reserve, Census Bureau WOLFSTREET.com

“The rich get richer”

More wealth

➔ better access to better loans or mortgages

➔ more wealth

Sources of unfairness

Previous examples showed different reasons for unfairness:

- Data bias
- Imbalance in the quantity of data across groups
- Only access to proxies rather than true, intrisic attributes
- Learning objective causes unanticipated bias
- Feedback loops/long-term and composed effects
- Possibly many other reasons…

Fairness issues do not simply arise at the algorithmic level, and algorithms probably cannot solve all fairness issues…

…But algorithms are widely used in important life decisions, and we should strive to make them as fair as possible

Algorithmic Fairness

How can we prevent our algorithms for making unfair decisions, and mimicking or perpetuating bias?