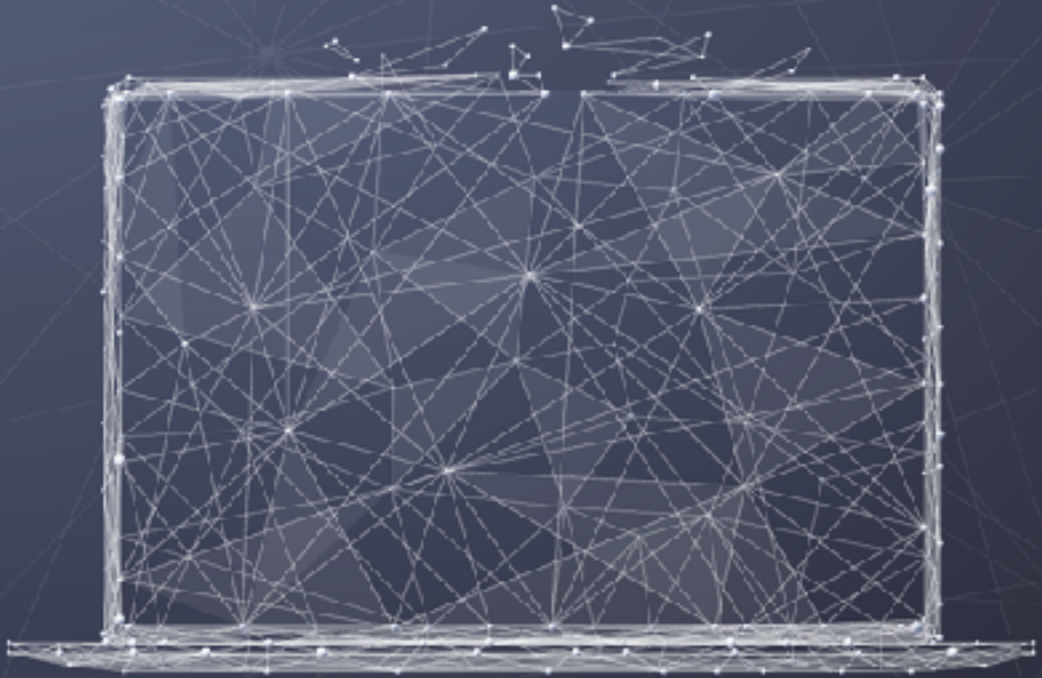
Data Science
Foundations of Decision Making

Fairness and bias



PURDUE UNIVERSITY | College of Science

Why data science methods are helpful

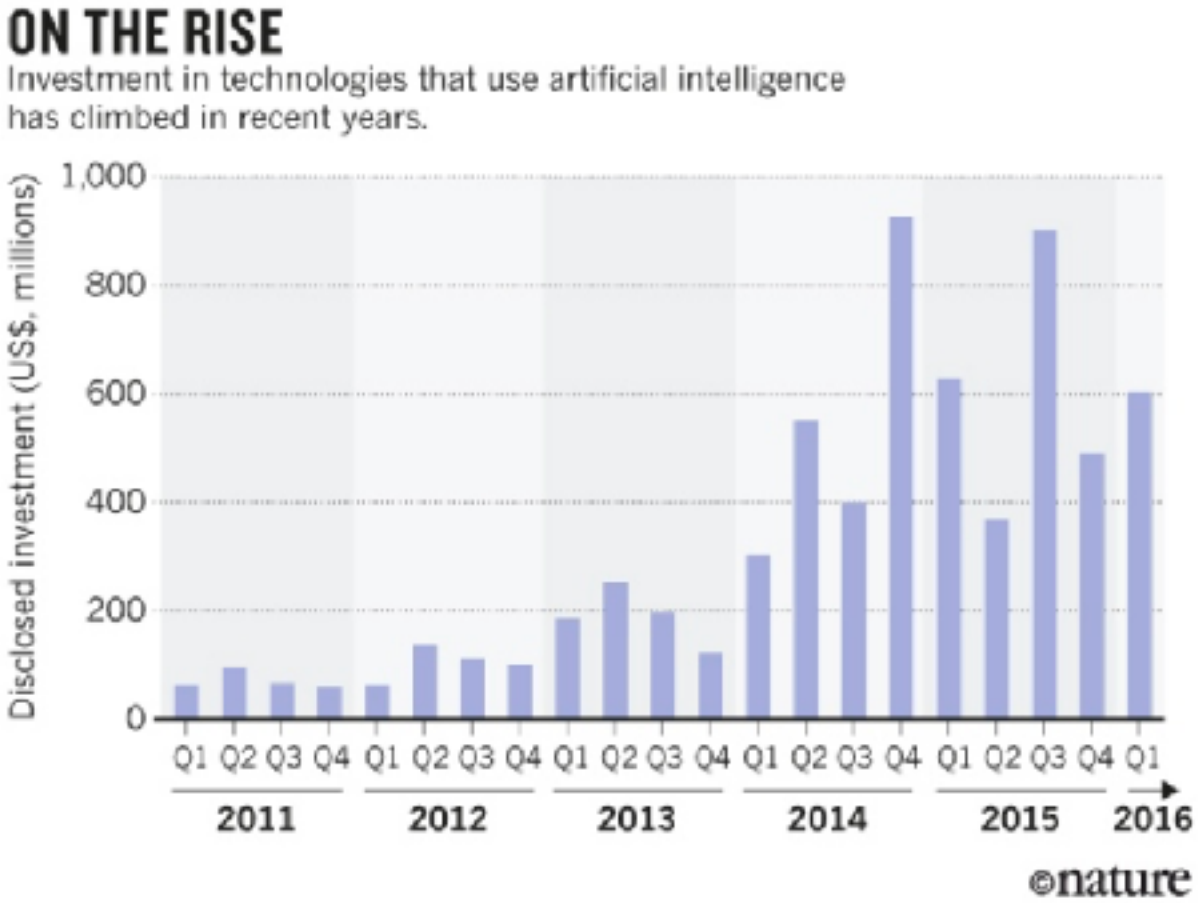
- Provide algorithms for tasks that were not previously amenable to automation (e.g. image analysis)
- Provide advantages over humans doing similar tasks:
    - Inexpensive/scalable/fast
    - Consistent and verifiable
    - More accurate*
    - More fair*, less subject to social biases

*Maybe, sometimes

PURDUE UNIVERSITY | College of Science

Source: Lawrence Hunter, Colorado School of Mines

Automatic prediction systems are on the rise



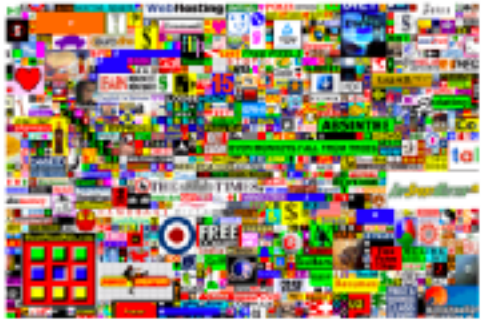
ON THE RISE

Investment in technologies that use artificial intelligence has climbed in recent years.

PURDUE UNIVERSITY | College of Science

Source: Richard Johansson, University of Gothenburg

Example systems in society



Web advertising

Lending

Ridesharing

Public policy

Source: Richard Johansson, University of Gothenburg

What could be the problem?

- Data are (about) people, can cause harm
- Algorithmic outcomes often not explainable
- A lot of data incidentally produced by daily life:
    - Social media
    - Ubiquitous cameras, microphones, location tracking
    - Medical treatment
- Important new uses
    - Legal: Surveillance, “predictive” policing, sentencing, fraud detection, military applications
    - Economic: School admissions, hiring/promotion, loans, insurance, accounting controls, advertising
    - Medical: Health insurance, diagnosis, decision making

Source: Lawrence Hunter, Colorado School of Mines

Ethical concerns

- Preserving privacy
  - Methods for handling sensitive data
  - Uses of data science that undermine privacy
- Avoiding bias
  - Data selection and unintentional red-lining
  - Re-inscription of existing biases
- Mitigating malicious attacks
  - Intentional subversion of machine learning systems
  - Hazards of learning from the open internet

Source: Lawrence Hunter, Colorado School of Mines

Anonymized data isn’t always safe

- In 1997, Latanya Sweeney identified the Governor of Massachusetts’ medical records.
    - Massachusetts released hospital records anonymized by removing names, addresses and SSNs
    - Voter records have name, address, ZIP code, birth date, and sex of every voter
    - Sweeney used zip code, birthdate and gender to uniquely identify Weld’s records
    - 87% of US identified by zip, birthdate & gender
- Similar with Netflix (using IMDB) and search logs

Source: Lawrence Hunter, Colorado School of Mines

Privacy != Security

- Some data cannot be anonymized
  - Genome sequences are inherently identifying
  - Even a few hundred well-picked SNPs...
- Often, people's desires about their data involve questions of trust
  - Willingness to share medical data with academic researchers, but not pharmaceutical companies
- Privacy is not a binary value
  - Different sorts of exposure to different sorts of people evoke different responses

Source: Lawrence Hunter, Colorado School of Mines

Bias in data science

- “Objective” algorithms are thought to be free of the biases that plague people.
- Algorithms, especially ones that learn, can inadvertently re-inscribe those biases
- Algorithms are opaque, hard to interrogate
- Increasingly widespread

Source: Lawrence Hunter, Colorado School of Mines

Proxies for discrimination

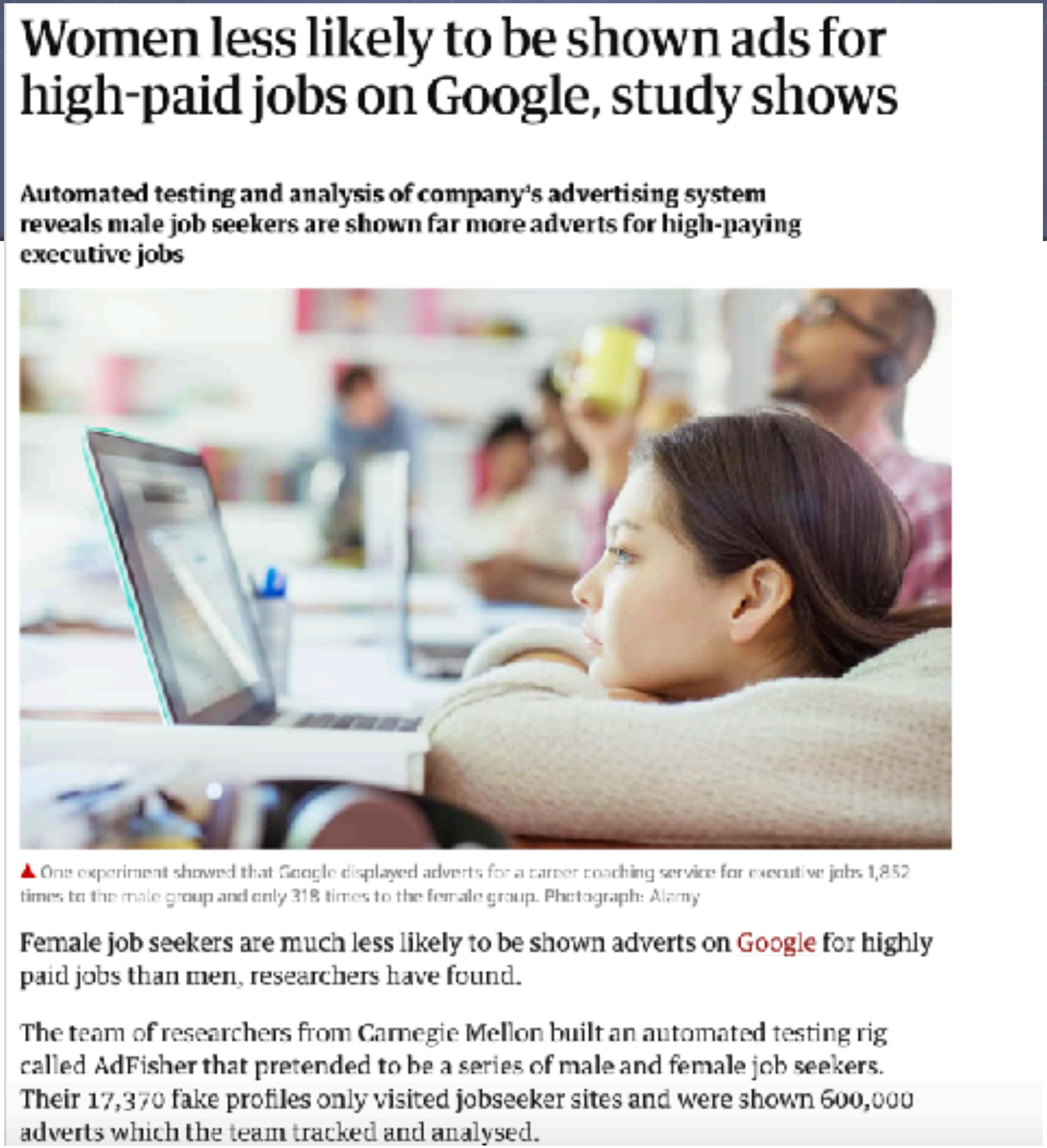
- Illegal, and generally perceived as wrong to make choices based on race, gender, religion, national origin, etc.
- However, proxies for these are everywhere:
  - Zip codes
  - Names (gender, race, national origin)
  - Purchase histories (including movies or tv shows)
- Machine learning that uses biased historical record + any proxy is likely to re-inscribe bias

Source: Lawrence Hunter, Colorado School of Mines

Discrimination in online ad delivery

- Sweeney observed in 2013 that black-identifying names turned out to be much more likely than white-identifying names to generate ads that including the word “arrest” (60 per cent versus 48 per cent).
- Google uses a learning algorithm to place ads that are most often clicked on.
- Likely to be a reflection of people clicking on those ads more for ‘black’ names

Source: Lawrence Hunter, Colorado School of Mines



Women less likely to be shown ads for high-paid jobs on Google, study shows

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs

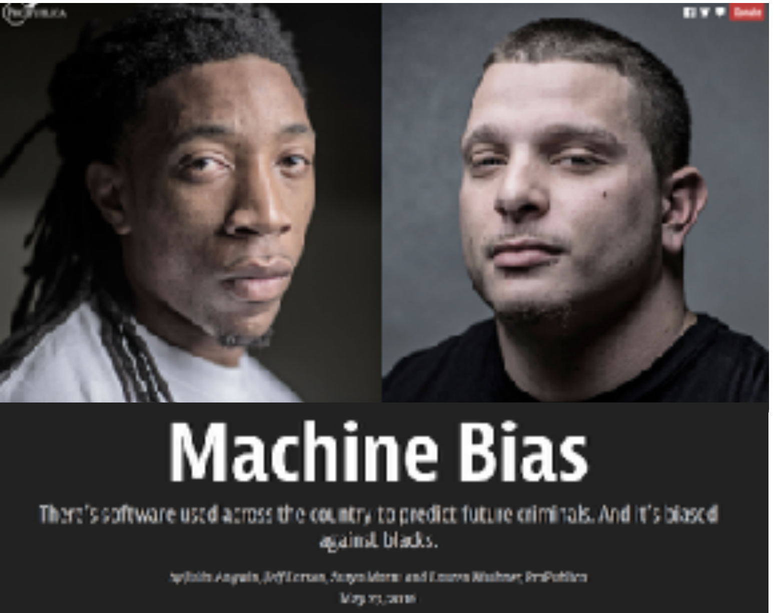
▲ One experiment showed that Google displayed adverts for a career coaching service for executive jobs 1,852 times to the male group and only 318 times to the female group. Photograph: Alamy

Female job seekers are much less likely to be shown adverts on Google for highly paid jobs than men, researchers have found.

The team of researchers from Carnegie Mellon built an automated testing rig called AdFisher that pretended to be a series of male and female job seekers. Their 17,370 fake profiles only visited jobseeker sites and were shown 600,000 adverts which the team tracked and analysed.

PURDUE UNIVERSITY | College of Science

Example: COMPAS system



Machine Bias

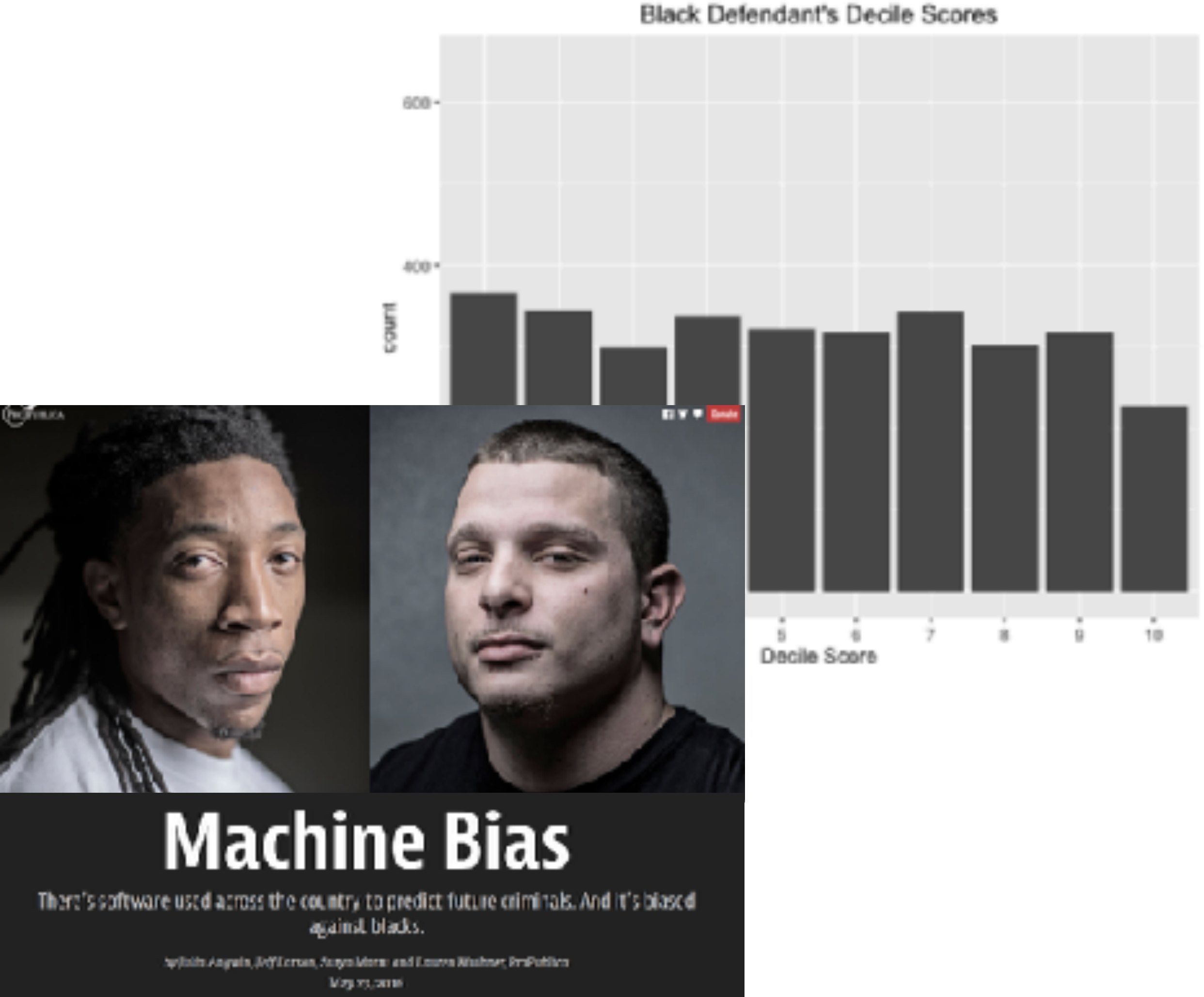
There's software used across the country to predict future criminals. And it's biased against blacks.

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

May 23, 2016

Source: Richard Johansson, University of Gothenburg

Example: COMPAS system

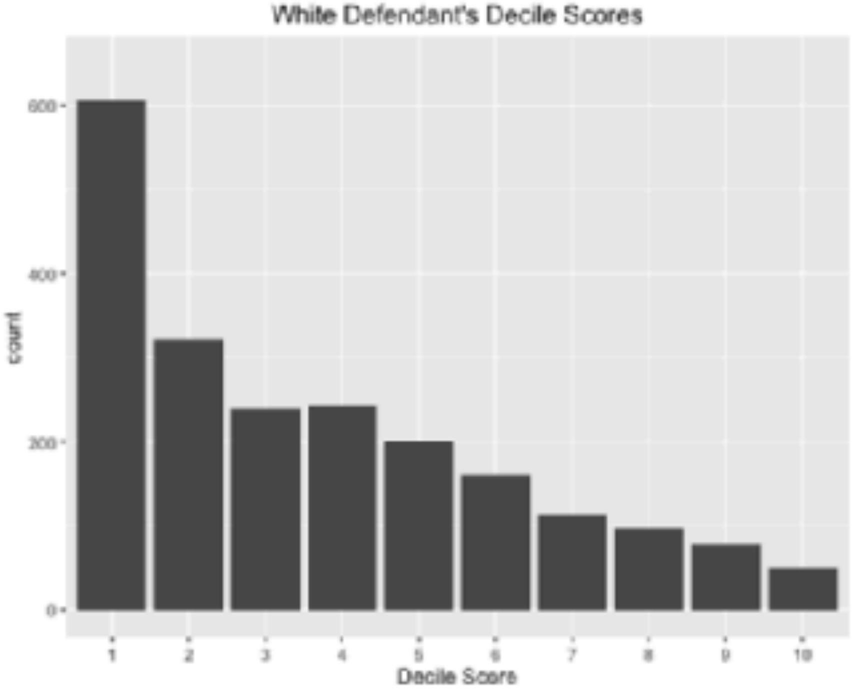


Machine Bias

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



PURDUE UNIVERSITY | College of Science

Source: Richard Johansson, University of Gothenburg

Can predictive system discriminate unfairly?

- When predictive systems are more widely applied in situations where they affect people's lives, they may run into troubles concerning anti-discrimination laws
- There is no consensus about how to define and less how to deal with the problem

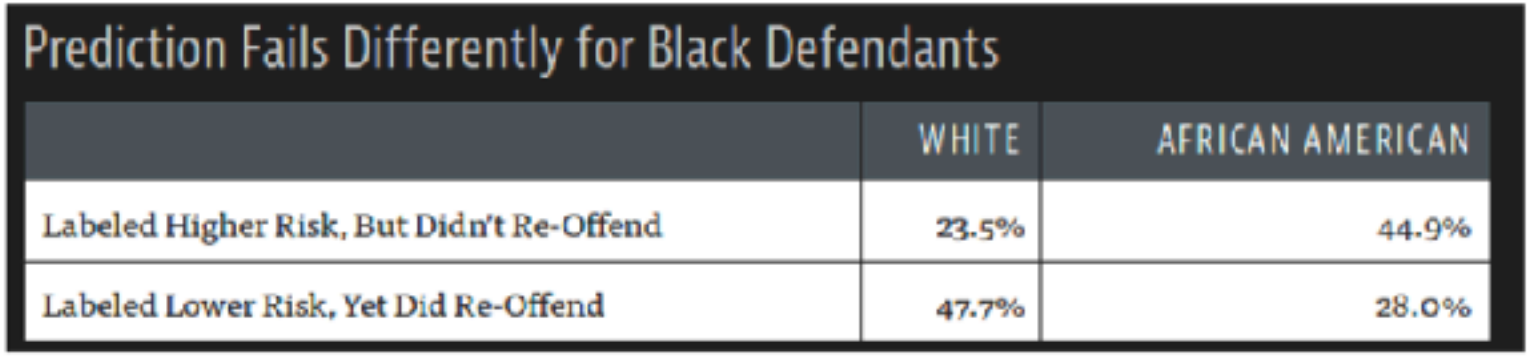
Source: Richard Johansson, University of Gothenburg

The legal situation

- Anti-discrimination laws in several countries prohibit unfair treatment of people based on protected attributes (e.g., gender, race)
- These laws often evaluate the fairness of a decision making process by mean of two distinct notions
  - Disparate treatment: if its decision are (partly) based on the subject's protected attribute information
  - Disparate impact: if its outcomes disproportionately hurt (or benefit) people with certain sensitive attribute values
- Are these ideas in conflict? How can they be operationalized?

Source: Richard Johansson, University of Gothenburg

Example: COMPAS system



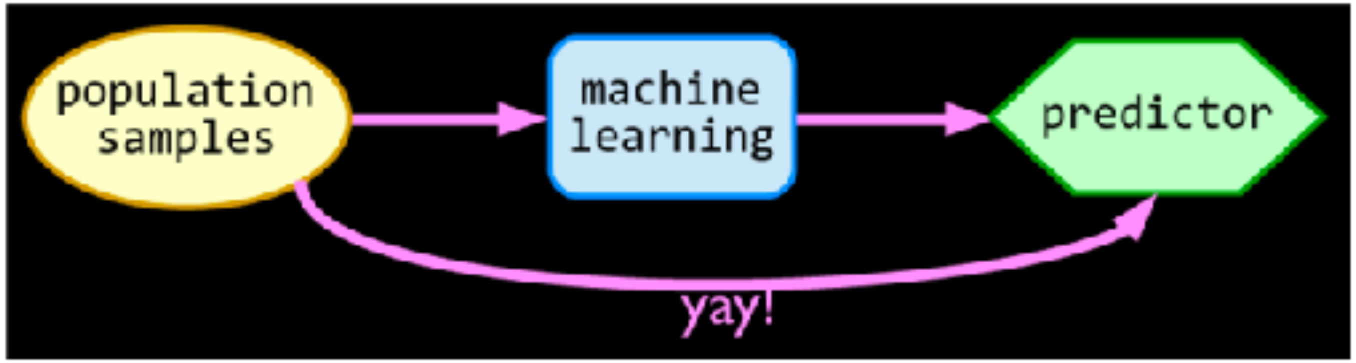
Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

- Non-offending blacks get higher scores
  - Note that race is not a feature in the model
- ML models can pick up subtle biases from the training set!

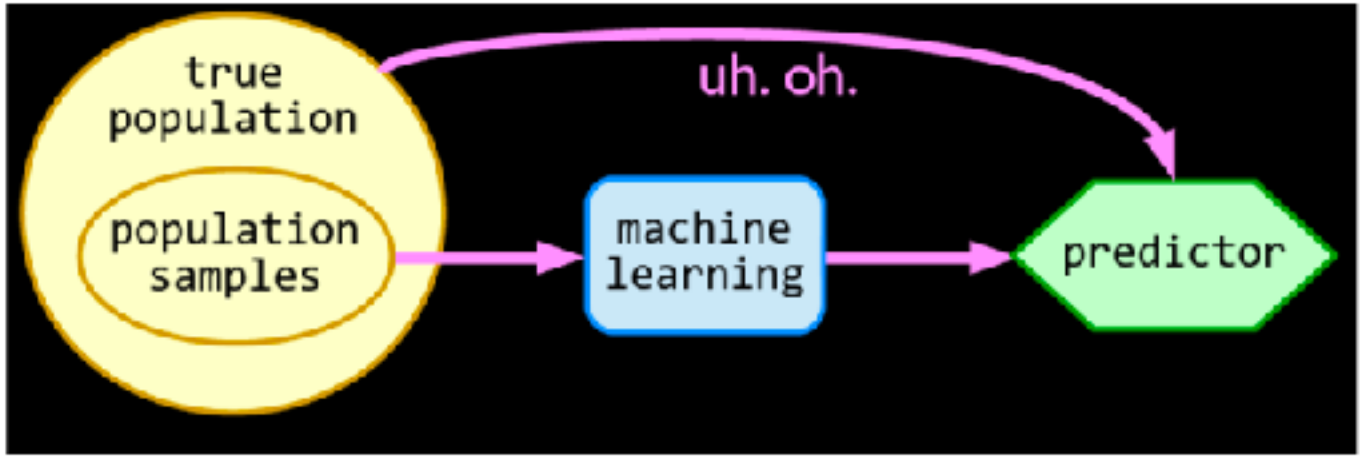
Source: Richard Johansson, University of Gothenburg

Collecting training data



Source: Richard Johansson, University of Gothenburg

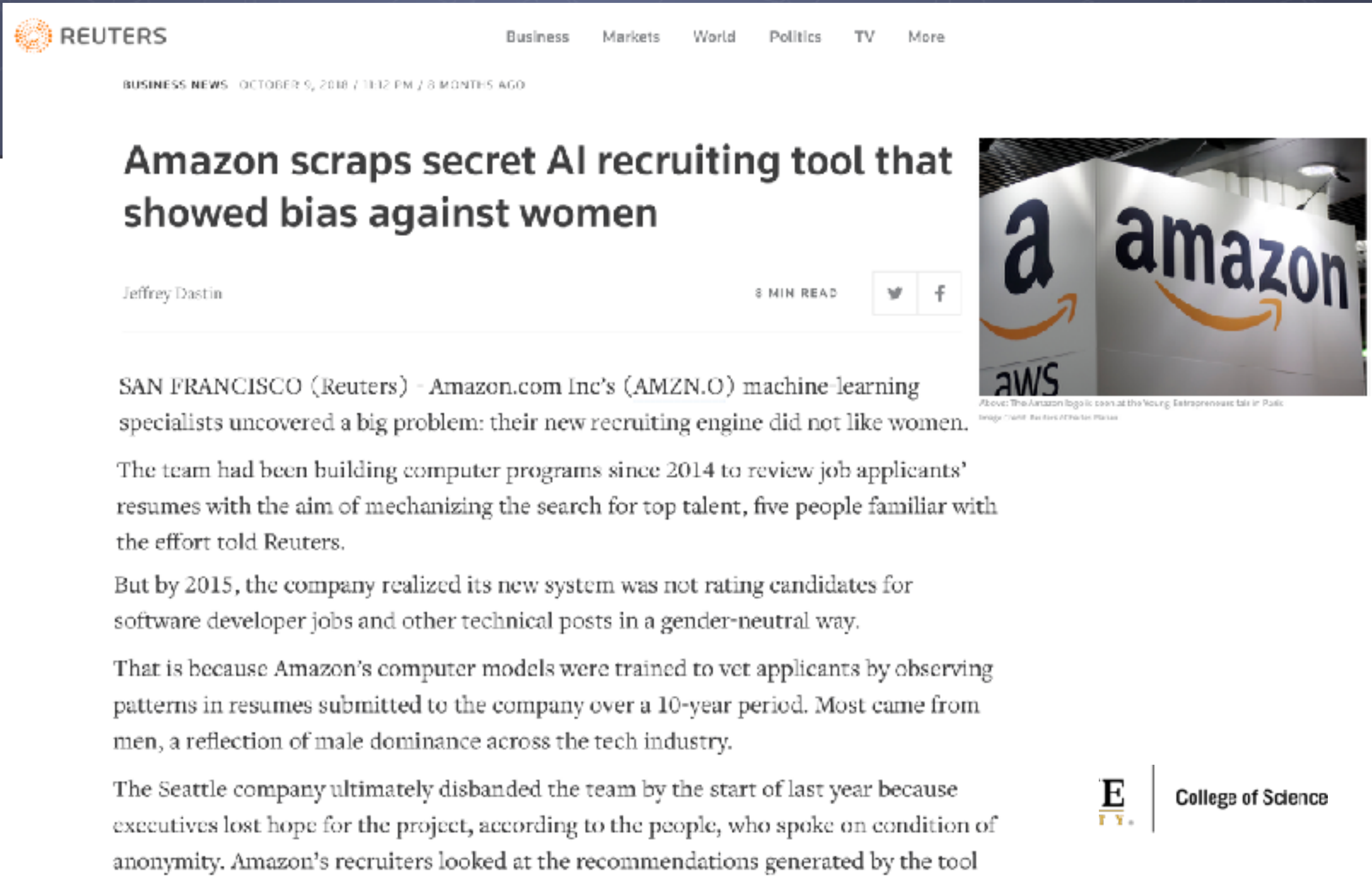
Collecting training data



Source: Richard Johansson, University of Gothenburg

Bias in models and patterns

- Bias is in the data, not the learning algorithm
- Models trained on data will learn any biases in the data
    - ML for recidivism prediction will learn race bias if data has a race bias
    - ML for resume processing will learn gender bias if data has a gender bias



REUTERS

Business Markets World Politics TV More

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / 8 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

Above: The Amazon logo is seen at the Young Entrepreneurs fair in Paris

Image Credit: Reuters/Charles Platiau

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

But by 2015, the company realized its new system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way.

That is because Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.

The Seattle company ultimately disbanded the team by the start of last year because executives lost hope for the project, according to the people, who spoke on condition of anonymity. Amazon's recruiters looked at the recommendations generated by the tool

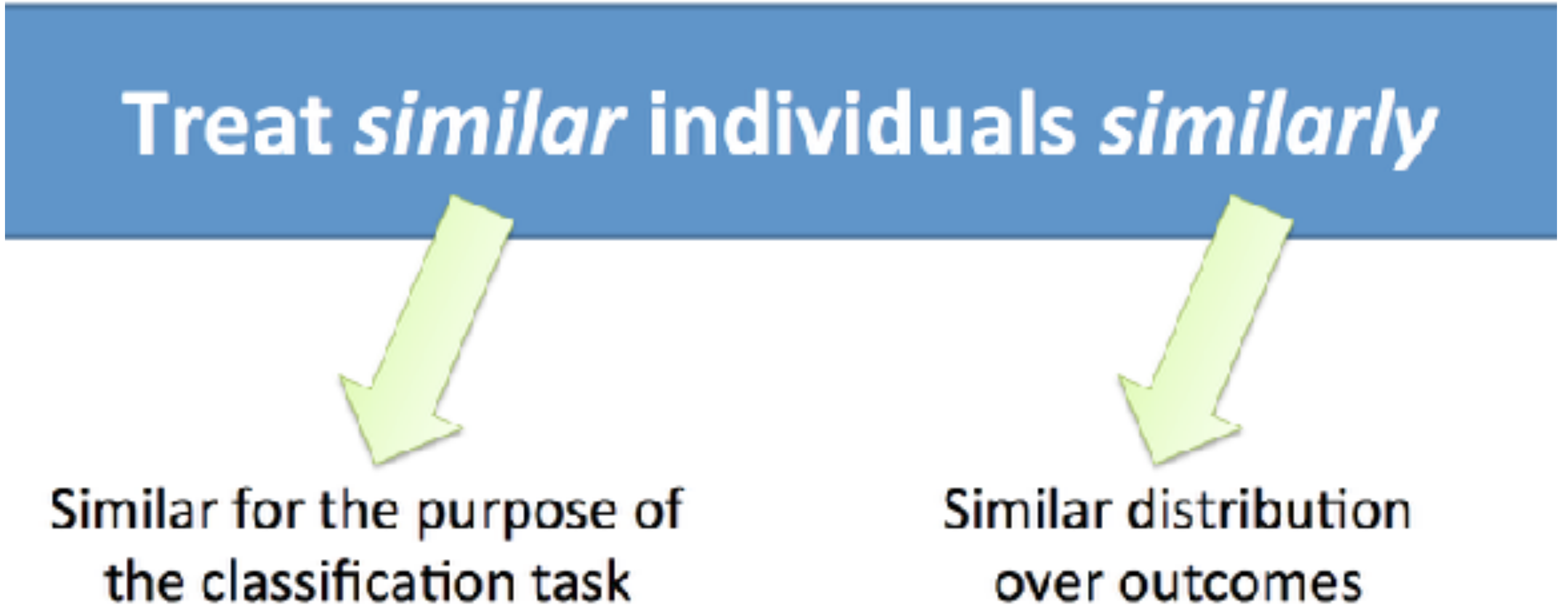
Potential solution: Fairness via blindness

- Ignore all irrelevant/protected attributes
- But you don’t need to see an attribute to be able to predict it with high accuracy
  - E.g. User visits artofmaniless.com … has 90% chance of being male

Potential solution: Group fairness

- Equalize two groups S, T at the level of outcomes
  - For example, S=minority, T=everyone else
  - P(outcome o | S) = P(outcome o | T)
- This is not strong enough as a notion of fairness
  - Sometimes desirable, but can be abused
  - Self-fulfilling prophecy: select smartest students in T, random students in S, then students in T will perform better

Potential solution: Individual fairness

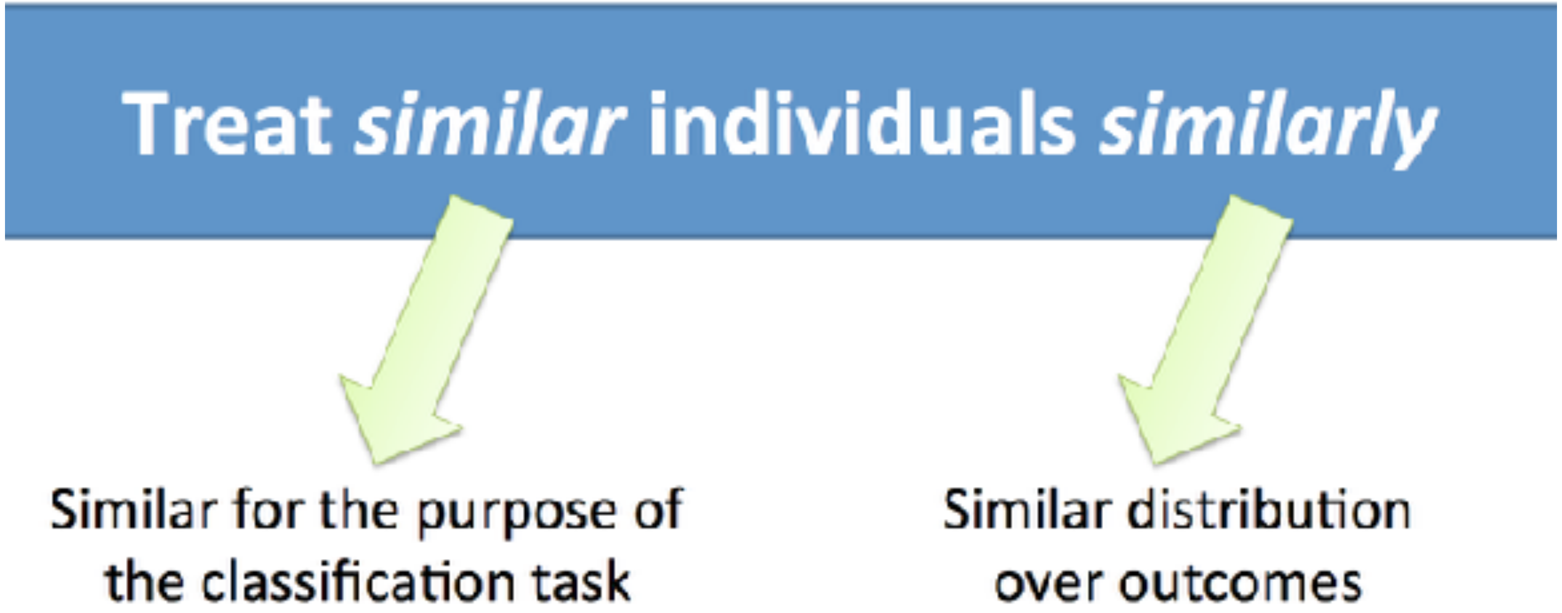


Treat *similar* individuals *similarly*

Similar for the purpose of the classification task

Similar distribution over outcomes

Potential solution: Individual fairness



Treat *similar* individuals *similarly*

Similar for the purpose of the classification task

Similar distribution over outcomes

How to do this? Current active area of research