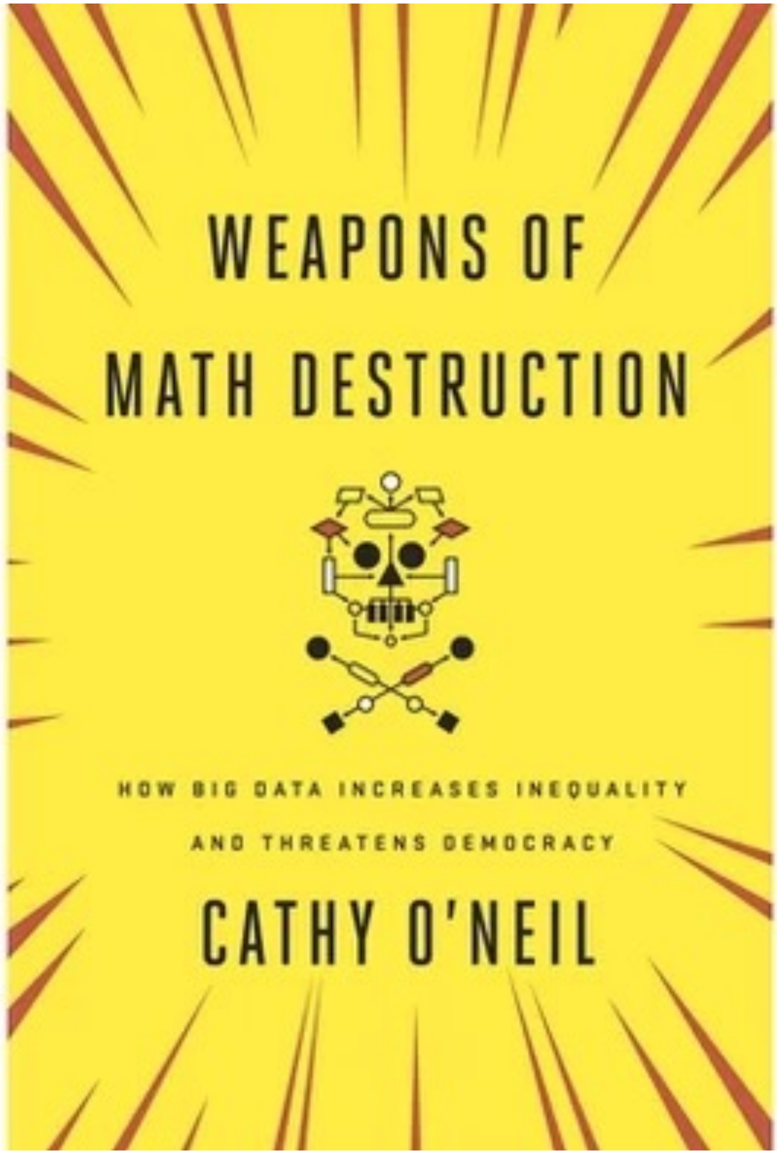
CS440/ECE448 Lecture 5: Fairness

Mark Hasegawa-Johnson

Outline

- Fairness Problems
  - Opacity; Scale; Damage
- Conditional versus Unconditional Fairness
  - Demograpic Parity vs. Equal Odds vs. Predictive Parity
- Proxy Variables
  - Redlining


WEAPONS OF
MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY
AND THREATENS DEMOCRACY
CATHY O'NEIL

Benefits of Statistical Models

- Before statistical models, many decisions were blatantly unfair
  - College admissions: Who were your parents?
  - Housing loans: Does the loan officer like the way you look?
- In many cases, statistical models are provably more accurate and more fair
  - College admissions: Weighted sum of grades, SAT, essay, interview
  - Housing loan: Weighted sum of income, debt, education

Problems with Statistical Models

- Opacity
  - If you knew the formula, you could game it, therefore decision-makers keep their formulas secret
  - Since you don’t know the formula, you don’t know when it is giving undue weight to something that happened to you in an unfortunate accident
- Scale
  - A successful statistical model gets adopted by every decision-maker
  - If they’re all making the same decision, they all make the same mistake
- Damage
  - On average, a statistical model is better than a biased human
  - … but the one person for whom the model fails might have their life destroyed, especially if every decision-maker uses the same model

Examples of the problem

- **Opacity**: The “Level of Service Inventory-Revised” (LSI-R) was used to decide who gets parole in at least two states, and many counties/precincts.
  - It did not ask about race.
  - It did ask “when was your first encounter with police” and other questions that are highly correlated with race.
- **Scale**: The collapse of the world economy in 2008 was caused by a statistical model with a bug. Most large banks used the Gaussian copula model to decide who got home loans; it failed to correctly model the risk of multiple simultaneous defaults.
- **Damage**: Companies can’t use medical tests to determine hiring, but they are allowed to use personality tests. In 2016, a lawsuit found that all the employers in one metro area were using the same “personality test” to screen applicants, so people with ”undesirable” personalities could not work.

AI Decision-makers

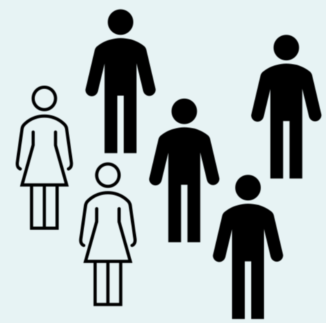
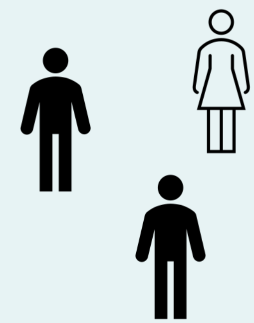
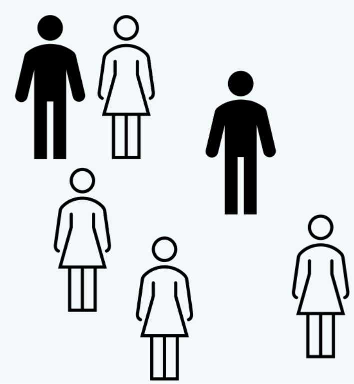
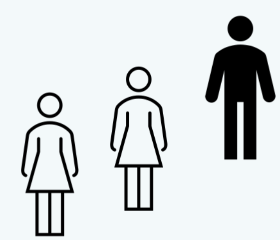
- At most large companies, the only job applications read by a human are those that are first approved by an AI
- Is this fair? Why or why not?

Outline

- Fairness Problems
  - Opacity; Scale; Damage
- Conditional versus Unconditional Fairness
  - Demograpic Parity vs. Equal Odds vs. Predictive Parity
- Proxy Variables
  - Redlining

Is your AI decision-maker fair?

- $f(X)$ = the decision your AI makes
- Y = the decision a human would make
- A = some attribute that shouldn't matter (e.g., gender)

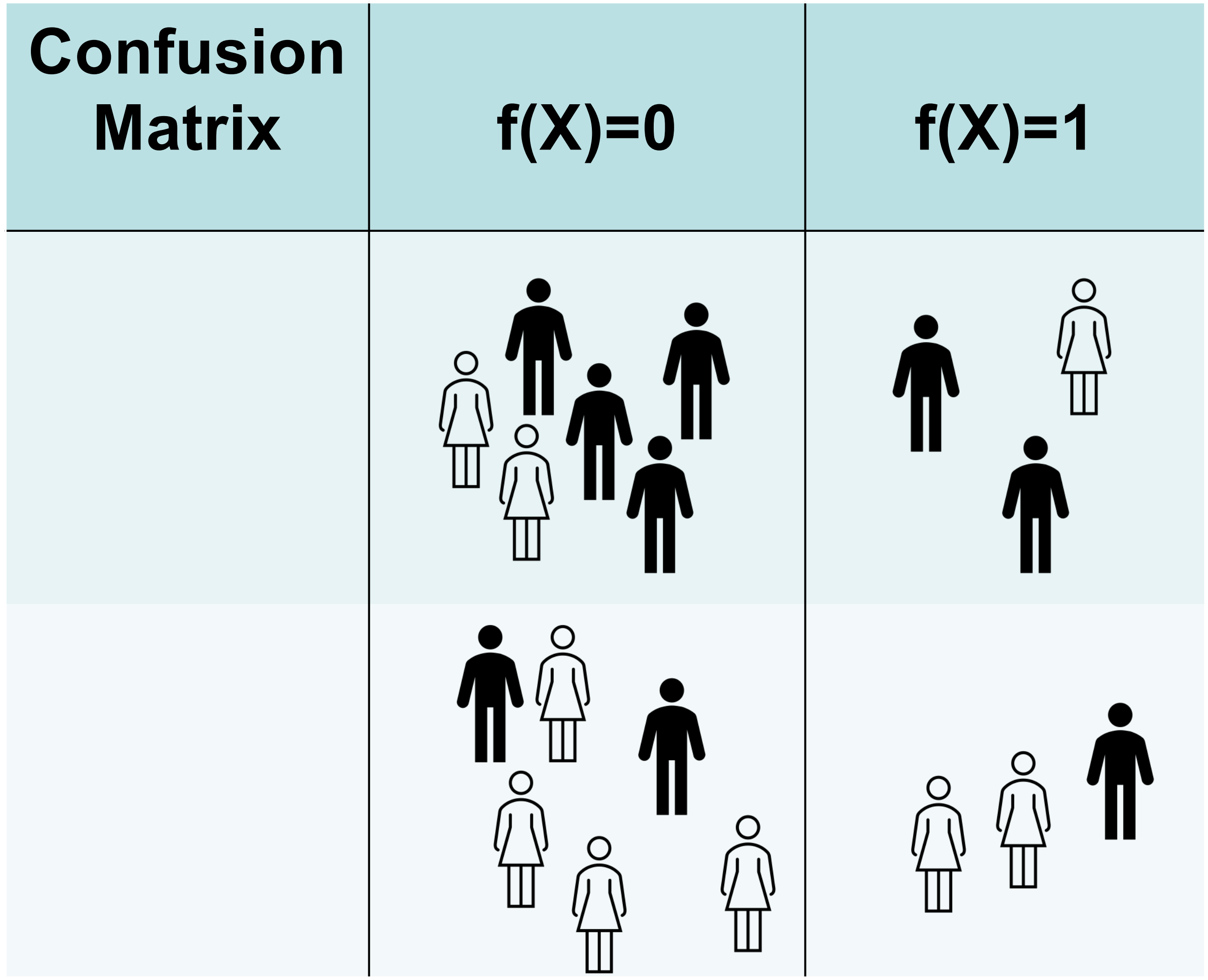
Confusion Matrix	**f(X)=0**	**f(X)=1**
Y=0		
Y=1		

Is your AI decision-maker fair?

- Demographic parity: Do equal fractions of all groups succeed?
- Equal opportunity: Do well-qualified people succeed at equal rates?
- Predictive parity: Are the people who succeed, from all groups, equally qualified?

Demographic parity:
Do equal fractions of all groups succeed?

$$P(f(X) = 1 | A = 1) = P(f(X) = 1 | A = 0)?$$

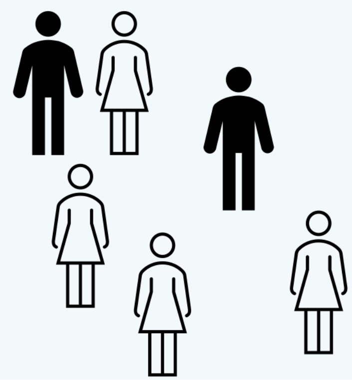
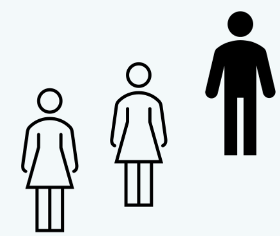


Why it matters

- (Generational justice): If a group is not represented in positions of power, then children in that group will be less likely to seek positions of power

Equal opportunity: Do well-qualified people succeed at equal rates?

$$P(f(X) = 1 | Y = 1, A = 1) = P(f(X) = 1 | Y = 1, A = 0)?$$

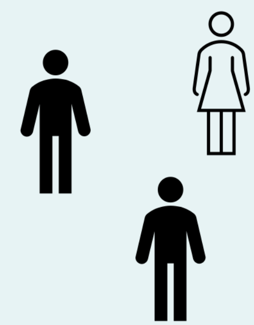
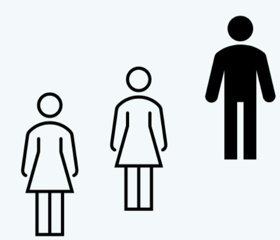
Confusion Matrix	**f(X)=0**	**f(X)=1**
Y=1		

Why it matters

- (Individual justice): Your chance of success should only depend on your qualifications. It should not depend on irrelevant attributes.

Predictive parity: Are the people who succeed, from all groups, equally qualified?

$$P(Y=1|f(X)=1, A=1) = P(Y=1|f(X)=1, A=0)?$$

Confusion Matrix		**f(X)=1**
Y=0		
Y=1		

Why it matters

- (Perceived justice): If people who succeed from group A are perceived to be unqualified more often than those from group B, people will believe you are discriminating against group B.

All three types of fairness are possible only if you define "qualified" in a group-independent manner

- Definition of conditional probability:
$$P(Y = 1|f(X) = 1) = \frac{P(f(X) = 1|Y = 1)P(Y = 1)}{P(f(X) = 1)}$$
- Demographic parity: $P(f(X) = 1)$
- Equal opportunity: $P(f(X) = 1|Y = 1)$
- Predictive parity: $P(Y = 1|f(X) = 1)$
- … all three are simultaneously independent of A if and only if $P(Y = 1)$ is independent of A, i.e., if and only if you can define a "well-qualified person" using a group-independent definition

Quiz

Go to PrairieLearn, try the quiz!

Why do we have to care?

- Fairness might mean lowering the GPA cutoff for job interviews for the student who's living out of their car (though privacy probably means we can't know where they live…)
- … or it might mean making sure they have a better place to live.
- AI is used to schedule job interviews more often than it's used to solve homelessness.
- …therefore, AI should be designed subject to society's overall fairness strategy.
  - DP, EO, or PP are often undesirable as hard constraints. Better to print the statistics and then evaluate in a larger context.

Outline

- Fairness Problems
  - Opacity; Scale; Damage
- Conditional versus Unconditional Fairness
  - Demograpic Parity vs. Equal Odds vs. Predictive Parity
- Proxy Variables
  - Redlining

Other Useful Definitions of Fairness in AI

Individual Fairness:

The dissimilarity between two outcomes should be less than the dissimilarity between the people.

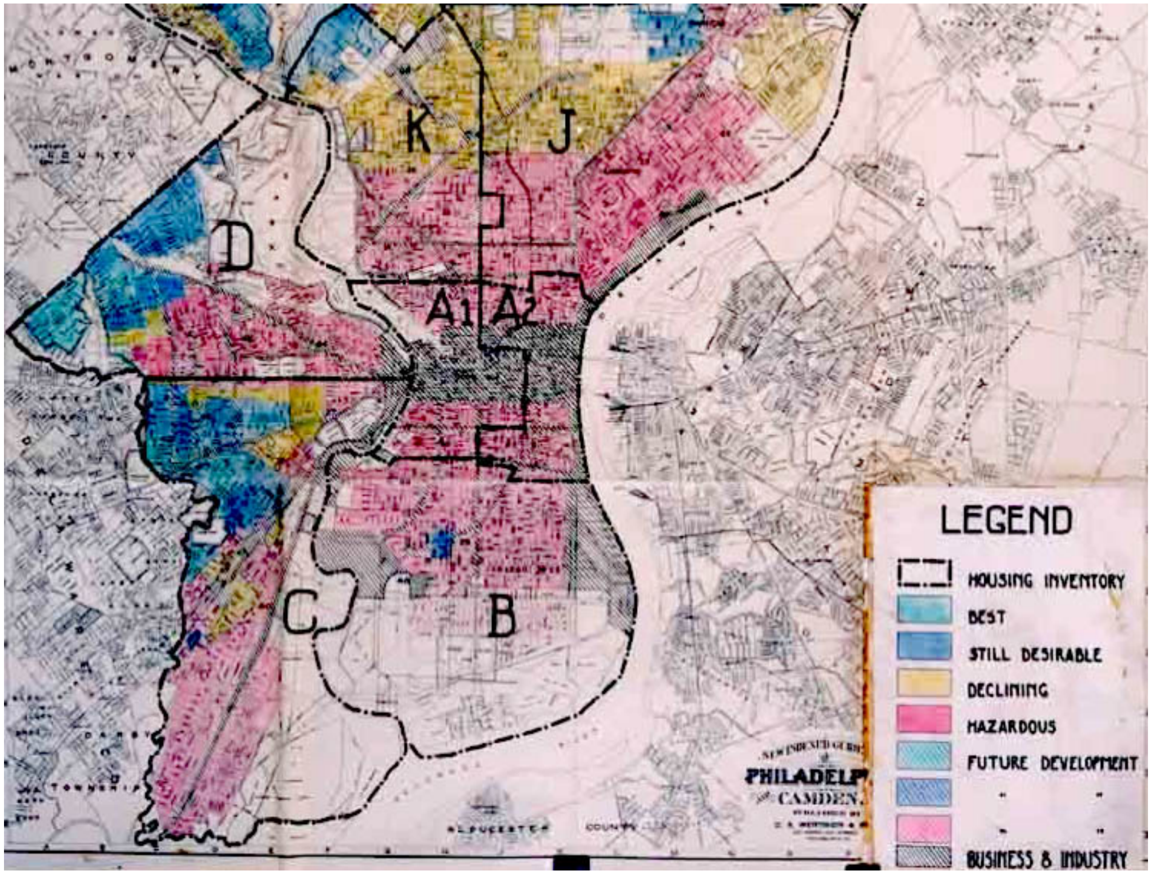
Counterfactual Fairness:

If a person's protected attribute were changed (and all their other attributes were possibly changed, according to their dependence on the protected attribute), then the outcome should not change.

A hard real-world problem that I want to make you aware of, and that has no black-box solution: How do we define "relevant" and "irrelevant" attributes?

Redlining

- “Redlining” is the practice of withholding home loans or investment from people who live in “bad neighborhoods”
- Traditionally, “bad neighborhood” meant that most people who lived there were racial minorities



https://commons.wikimedia.org/wiki/File:Home_Owners%27_Loan_Corporation_Philadelphia_redlining_map.jpg

Redlining by AI

- Until recently, in many places, it was illegal for an AI to use race, gender, or ethnicity in its decision-making formula (still illegal in most of Europe)
- Many “proxy variables” correlate with race, gender, and ethnicity, e.g., home address, name, number of times you’ve had to speak to the police
- Widely-used AI decision-makers have been shown to make predictions, based on proxy variables, that are highly discriminatory in practice

Conclusion: What’s fair?

- Definition of conditional probability:

$$P(Y = 1|f(X) = 1) = \frac{P(f(X) = 1|Y = 1)P(Y = 1)}{P(f(X) = 1)}$$

- Demographic parity: $P(f(X) = 1)$
- Equal opportunity: $P(f(X) = 1|Y = 1)$
- Predictive parity: $P(Y = 1|f(X) = 1)$
- Balancing these three mutually incompatible definitions requires political decisions (like: What does it mean to be ”qualified”?), not just technology decisions