

CS440/ECE448

Lecture 7: Fairness

Mark Hasegawa-Johnson, 1/2024

Lecture slides: CC0



Some images may have other
license terms.



CC-SA 2.0, Kathy Simon, 2008

https://commons.wikimedia.org/wiki/File:Viola_and_Mina_share_food.jpg

Outline

- **Fairness Problems**
 - Weapons of Math Destruction
- **Conditional versus Unconditional Fairness**
 - Fair Action versus Fair Society
- **Mutually Incompatible Definitions of Fairness**
 - Demographic Parity vs. Equal Odds vs. Predictive Parity vs. Society
- **Proxy Variables**
 - Irrelevant proxy variables
 - Relevant proxy variables
 - Fairness as a multi-task objective

**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

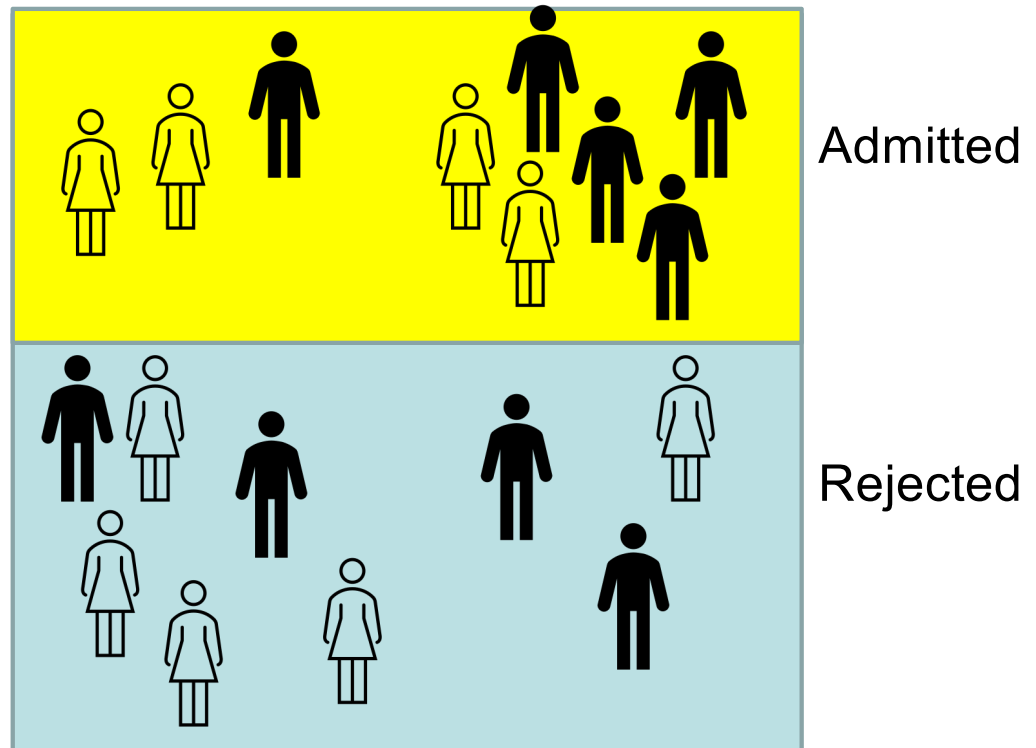
- **Opacity**: The “Level of Service Inventory-Revised” (LSI-R) was used to decide who gets parole in at least two states, and many counties/precincts.
 - It did not ask about race.
 - It did ask “when was your first encounter with police” and other questions that are highly correlated with race.
- **Scale**: 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 at least seven companies were using the same personality test, and therefore rejecting the same applicants, for the same frivolous reasons.

Outline

- Fairness Problems
 - Weapons of Math Destruction
- Conditional versus Unconditional Fairness
 - Fair Action versus Fair Society
- Mutually Incompatible Definitions of Fairness
 - Demographic Parity vs. Equal Odds vs. Predictive Parity vs. Society
- Proxy Variables
 - Irrelevant proxy variables
 - Relevant proxy variables
 - Fairness as a multi-task objective

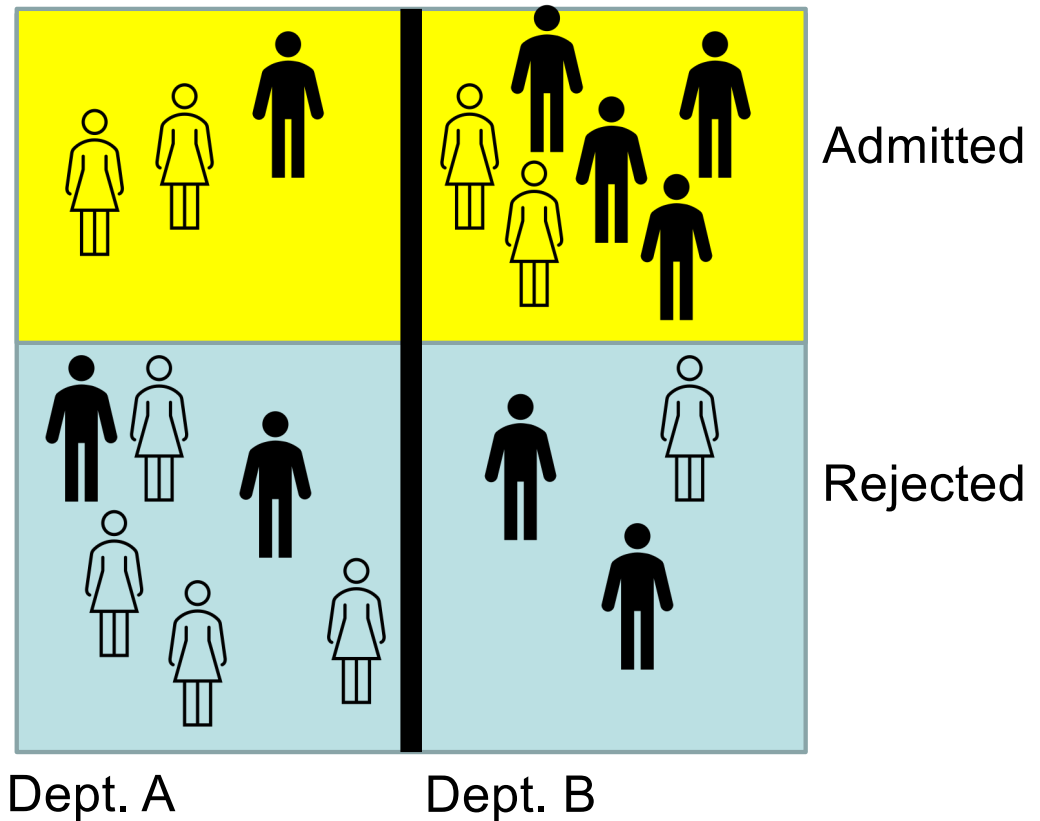
Are College Admissions Fair?

- Bickel, Hammel, and O'Connell, "Sex bias in graduate admissions: Data from Berkeley," *Science* 187(4175):398–404, 1975
- At that time, women were being admitted to the University of California at a far lower rate than men



Are College Admissions Fair?

Bickel, Hammel, and O'Connell showed that, *within each academic department*, $P(\text{admit}|\text{female})$ and $P(\text{admit}|\text{male})$ were the same.



Fair Action or Fair Society?

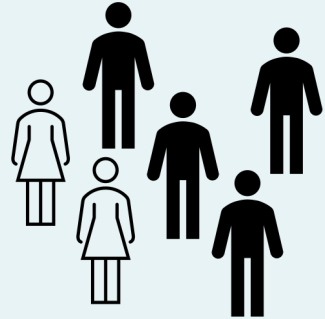
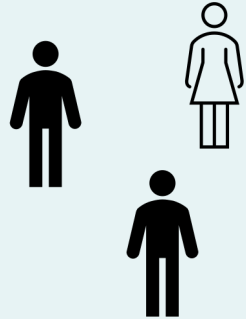
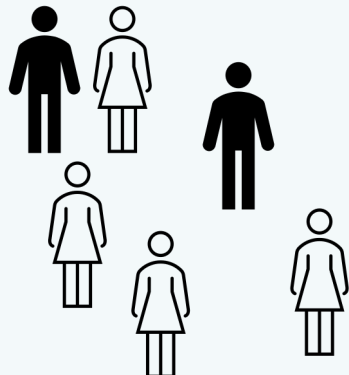
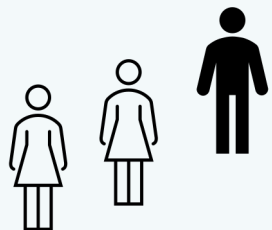
- Fair Action:
 - The admissions officers in each department were behaving in what seemed like a fair manner: admitting men and women in identical proportions
- Unfair Society:
 - The overall percentage of women admitted to college was lower than the percentage of men

Outline

- Fairness Problems
 - Weapons of Math Destruction
- Conditional versus Unconditional Fairness
 - Fair Action versus Fair Society
- Mutually Incompatible Definitions of Fairness
 - Demographic Parity vs. Equal Odds vs. Predictive Parity vs. Society
- Proxy Variables
 - Irrelevant proxy variables
 - Relevant proxy variables
 - Fairness as a multi-task objective

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)

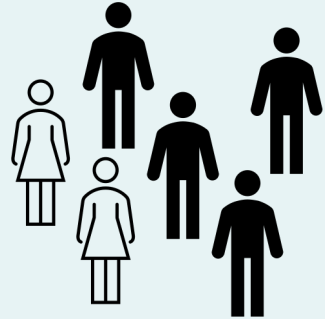
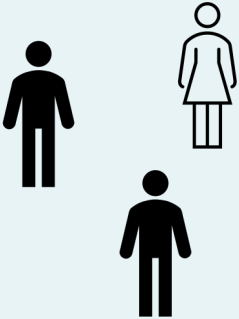
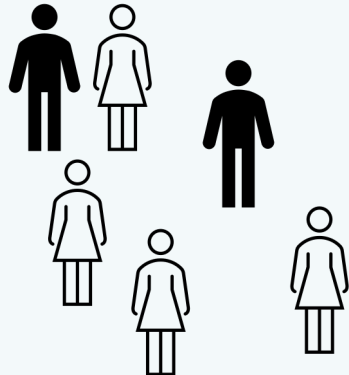
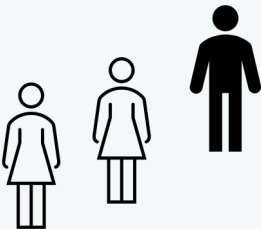
Confusion Matrix	$f(X)=0$	$f(X)=1$
$Y=0$		
$Y=1$		

Is your AI decision-maker fair?

- Demographic parity: Do equal fractions of all groups succeed?
- Equal odds: Do equal fractions of the “qualified” members of all groups succeed?
 - “qualified” = a human being would have chosen them?
- Non-discrimination: Are the people who succeed, from all groups, equally qualified?
 - “qualified” = a human being would have chosen them?

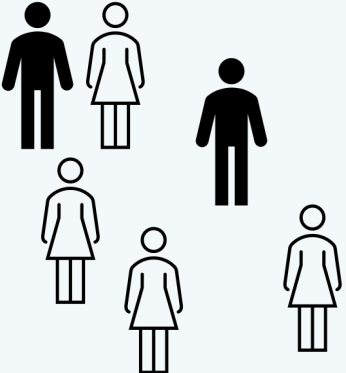
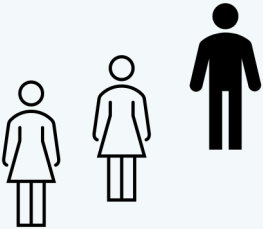
Demographic parity:
Do equal fractions of
all groups succeed?

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

Confusion Matrix	$f(X)=0$	$f(X)=1$
		
		

Equal odds: Do equal fractions of the “qualified” members of all groups succeed?

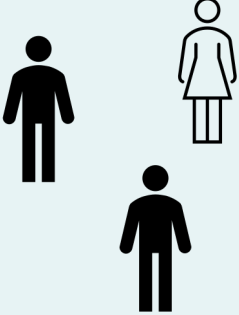
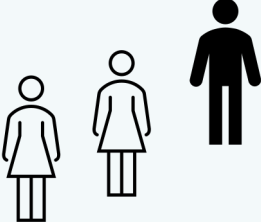
$$P(f(X) = 1 | Y = 1, A = 1) = P(f(X) = 1 | Y = 1, A = 0)?$$

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

Non-discrimination:

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		

Your AI can only be fair in all three ways if human judgment is fair

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

Definition of conditional
probability:

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

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

$$P(Y|f(X), A)$$
$$=$$

Those three things can only all be true,
all at the same time, if

- $P(Y|A = 1) = P(Y|A = 0)$

$$\frac{P(f(X)|Y, A)P(Y|A)}{P(f(X)|A)}$$

Quiz

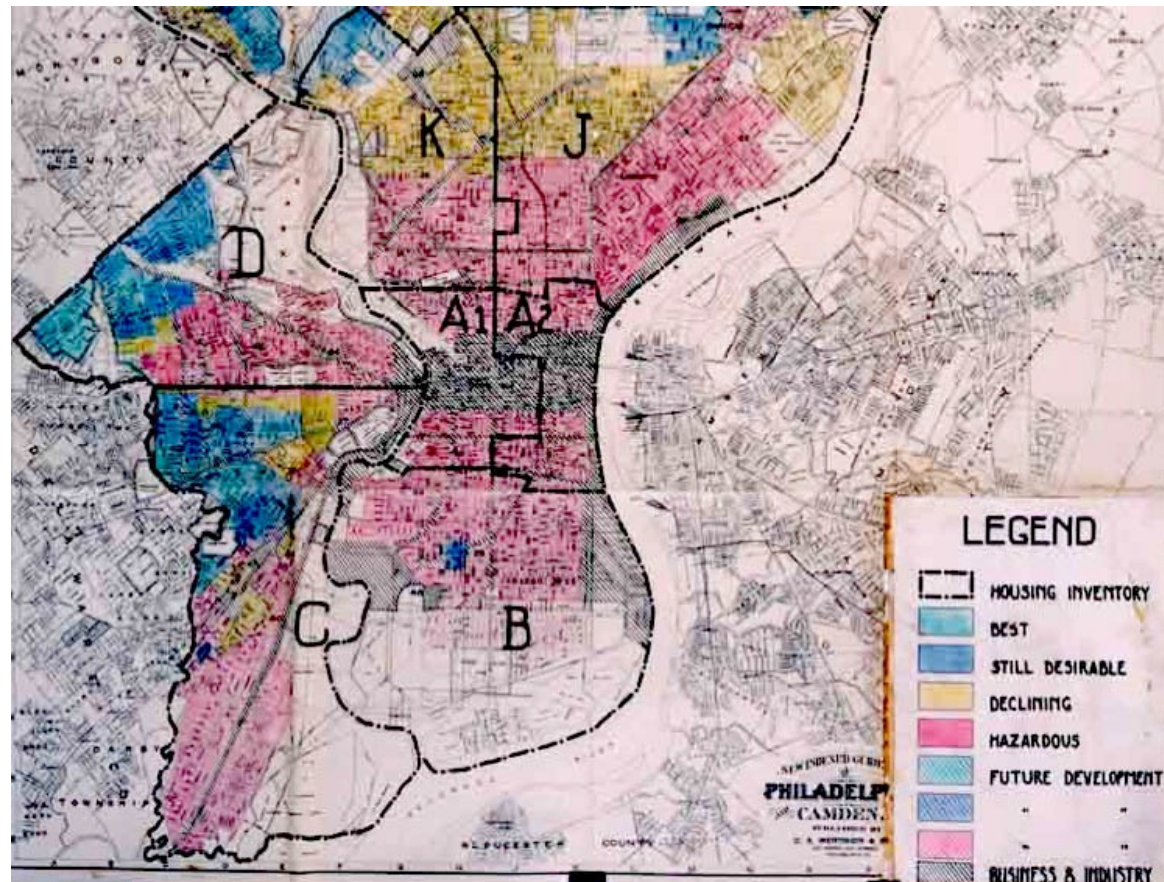
https://us.prairielearn.com/pl/course_instance/147925/assessment/2393723

Outline

- Fairness Problems
 - Weapons of Math Destruction
- Conditional versus Unconditional Fairness
 - Fair Action versus Fair Society
- Mutually Incompatible Definitions of Fairness
 - Demographic Parity vs. Equal Odds vs. Predictive Parity vs. Society
- **Proxy Variables**
 - Irrelevant proxy variables
 - Relevant proxy variables
 - Fairness as a multi-task objective

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](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 such as LSI-R were shown to make predictions, based on proxy variables, that were highly discriminatory in practice

Which observations should your algorithm use?

- Avoid using proxy variables correlated with race, gender, ethnicity, or wealth
- Try to only use variables that are relevant to the decision being made
- For example, college admission decisions might use only:
 - High school grades
 - Standardized test scores
 - Essays
 - Extracurriculars

Are any observations free of bias?

- High school grades
 - People with money can shop for a HS with grade inflation
- Standardized test scores
 - People with money can take the test many times
- Essays
 - People with money can hire tutors
- Extracurriculars
 - People with money can design an extracurricular portfolio

Fairness as a multi-task objective

No artificial intelligence is an island.

- John Donne, 1624 (paraphrased)

- If you're making decisions about somebody's qualifications, you need to have a good model of the way in which those qualifications were obtained
- No decision-maker can affect every step in a person's life, but fairness is the cumulative product of every step in a person's life. Therefore, decision-makers need to work together.

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.