Limits and opportunities in algorithmic fairness

Columbia DSI Data For Good series

Joshua Loftus

10/5/2018

Unfair algorithms

Fairness: legal / observational perpective

Fairness: causal perspective

Selection bias

Conclusions

Objective of this talk

Highlight statistical challenges in fairness, particularly

- **Causation**
- **Selection bias**

Much of the material can be found in a survey (preprint) with my collaborators, Causal Reasoning for Algorithmic Fairness (2018)

I attempt to cite others as often as possible, and make (almost) no distinctions between preprints and peer reviewed citations

Opinions, both technical and non-technical, are my own, but if I say something you disagree with you should email Matt (just kidding!)

These slides can be found on my site joshualoftus.com, and I would be glad to hear from you at loftus@nyu.edu

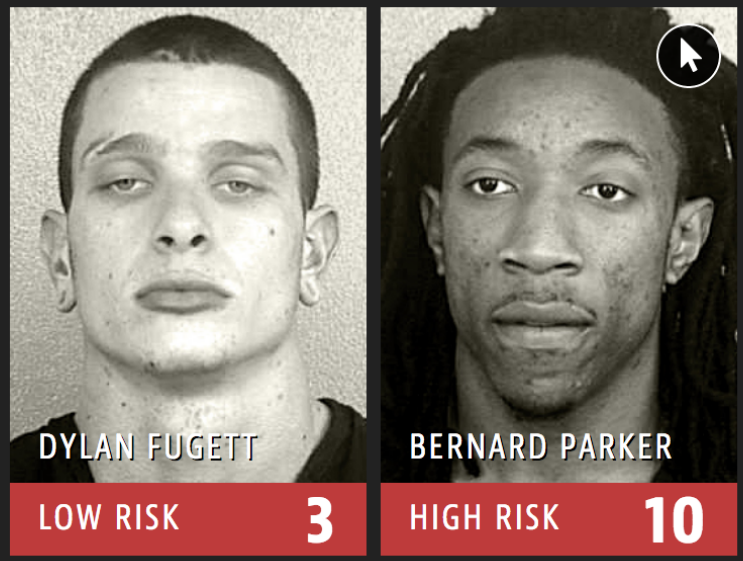
Unfair algorithms

Race: risk assessment and jail/bail

Dozens of risk assessments are being used across the nation — some created by for-profit companies such as Northpointe and others by nonprofit organizations. (One tool being used in states including Kentucky and Arizona, called the Public Safety Assessment, was developed by the Laura and John Arnold Foundation, which also is a funder of ProPublica.)

There have been few independent studies of these criminal risk assessments. In 2013, researchers Sarah Desmarais and Jay Singh examined 19 different risk methodologies used in the United States and found that "in most cases, validity had only been examined in one or two studies" and that "frequently, those investigations were completed by the same people who developed the instrument."

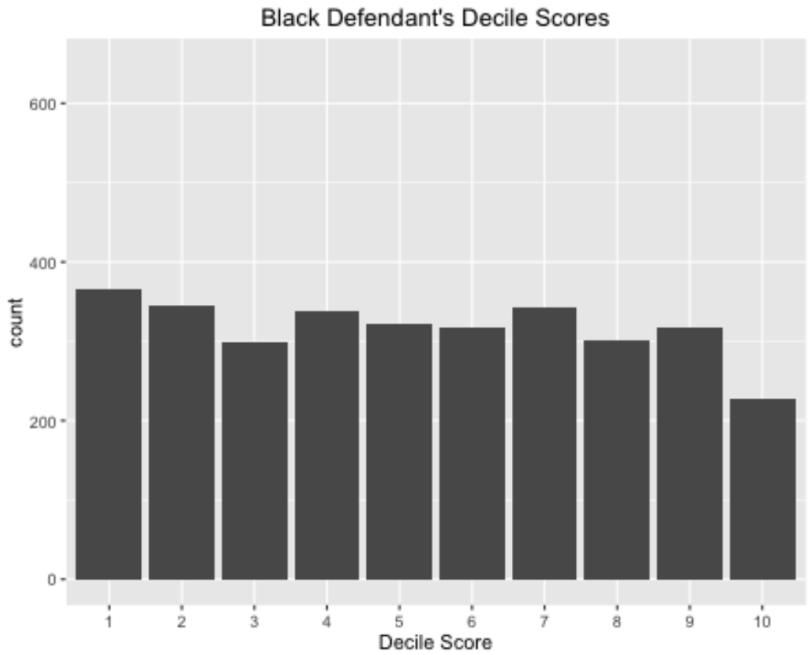
Two Drug Possession Arrests

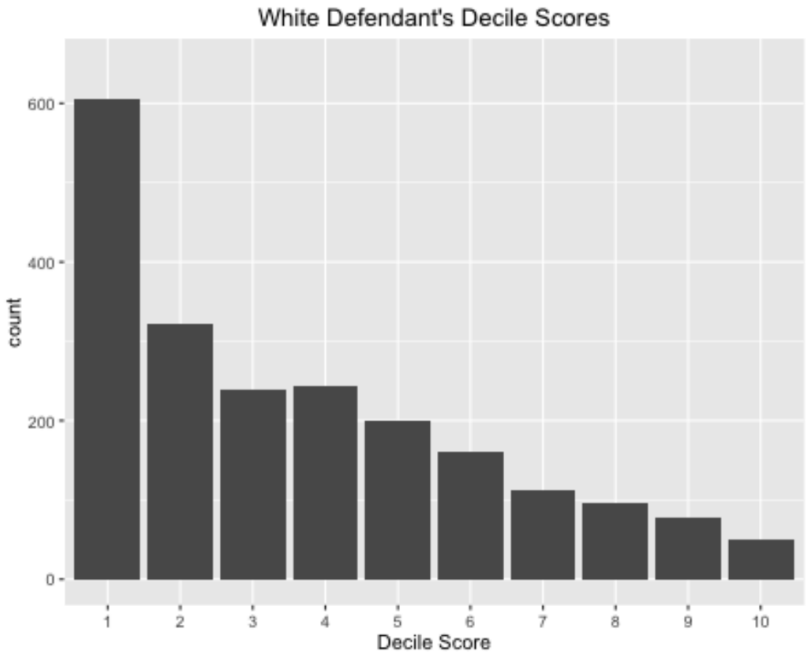


Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Prior offenses: left, attempted burglary; right, resisting arrest without violence

Race: risk assessment tool biased?





Race: biased after adjusting for actual re-offense

Risk of General Recidivism Logistic Model

	Dependent variable:
	Score (Low vs Medium and High)
Female	0.221*** (0.080)
Age: Greater than 45	-1.356*** (0.099)
Age: Less than 25	1.308*** (0.076)
Black	0.477*** (0.069)
Asian	-0.254 (0.478)
Hispanic	-0.428*** (0.128)
Native American	1.394* (0.766)
Other	-0.826*** (0.162)
Number of Priors	0.269*** (0.011)
Misdemeanor	-0.311*** (0.067)
Two year Recidivism	0.686*** (0.064)
Constant	-1.526*** (0.079)
Observations	6,172
Akaike Inf. Crit.	6,192.402

*Note: *p<0.1; **p<0.05; ***p<0.01*

PROPUBLICA

Machine Bias

Julia Angwin, Jeff Larson, Lauren Kirchner and Surya Mattu, May 23, 2016, 8 a.m. EDT

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

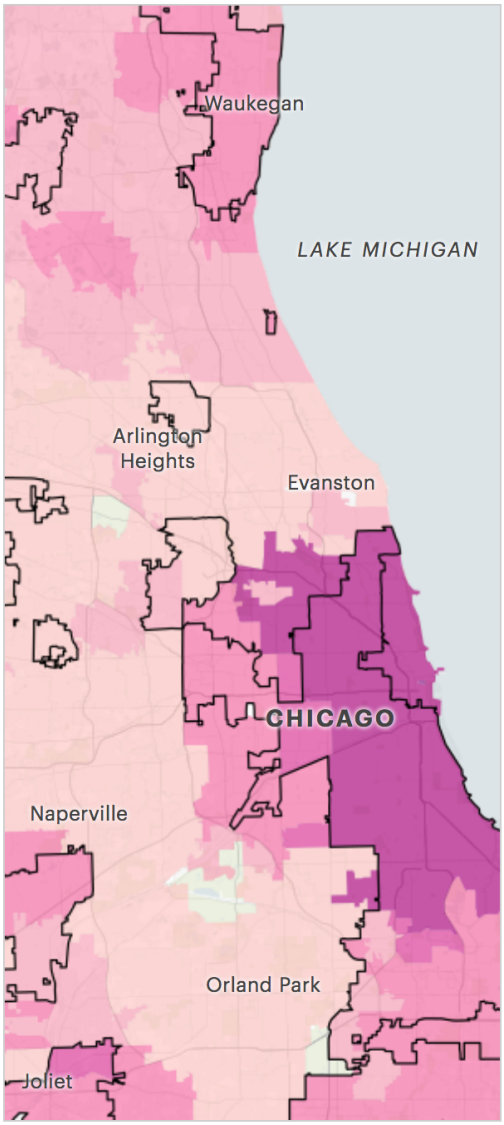
Minority Neighborhoods Pay Higher Car Insurance Premiums Than White Areas With the Same Risk

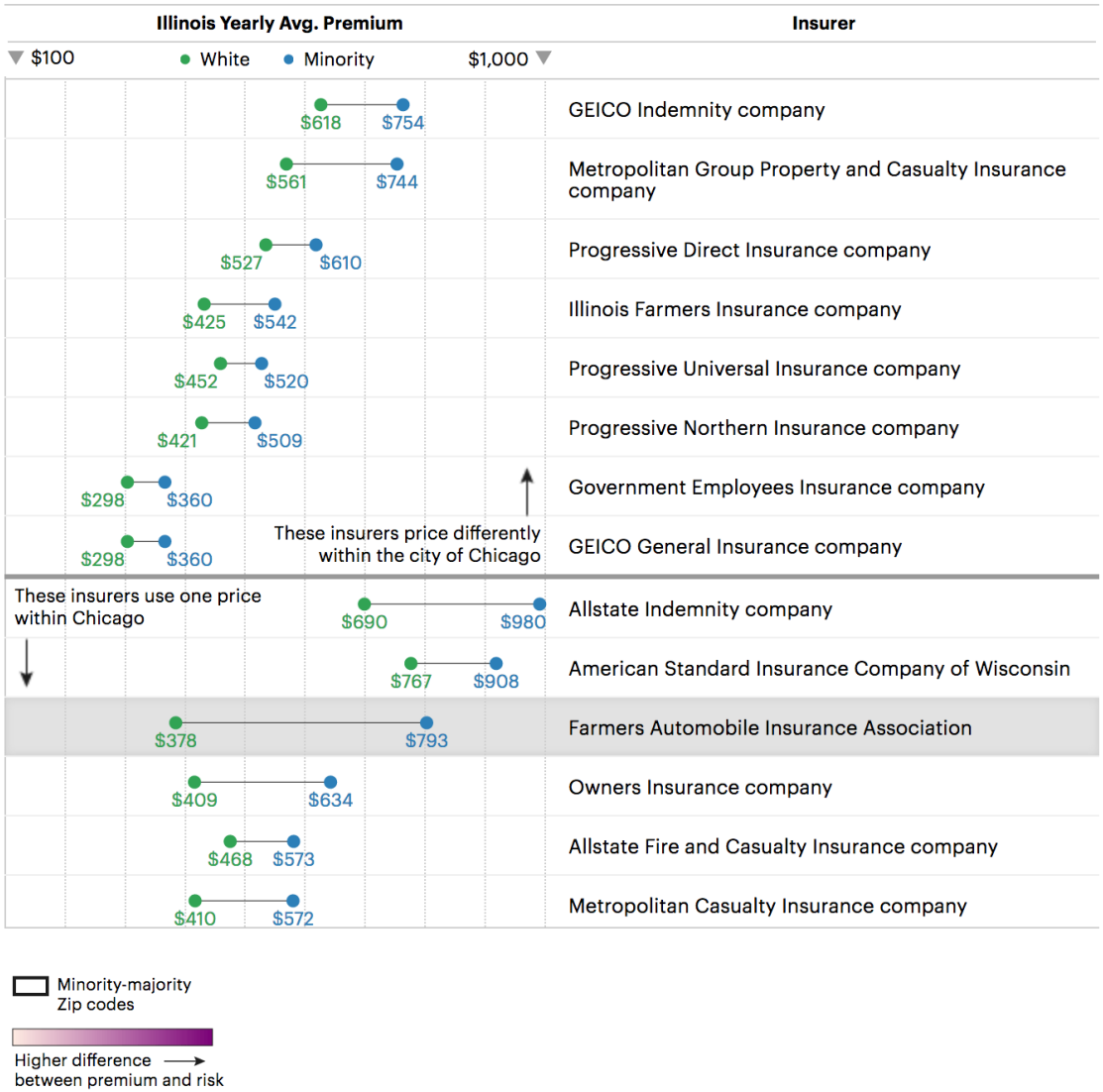
by Julia Angwin, Lauren Kirchner, Surya Mattu, April 4, 5 a.m. EDT

Facebook Says it Will Stop Allowing Some Advertisers to Exclude Users by Race

by Julia Angwin, Nov. 11, 2016, 10 a.m. EST

Location: charged a higher price to insure the same car





Age: excluded from job opportunities

AARP and the senators were reacting to a Dec. 20 report by ProPublica and The New York Times that dozens of the nation's leading employers, including Facebook itself, narrow their audience for job ads on Facebook and other platforms by age. The ability of advertisers to direct their messages at specific groups is a cornerstone of Facebook's business. But such micro-targeting becomes controversial when it fosters discrimination in legally protected categories such as race and age. ProPublica has reported that Facebook also accepted ads aimed at "Jew-haters" as well as housing ads that discriminated by race, gender, disability and other factors.

Read More

Dozens of Companies Are Using Facebook to Exclude Older Workers From Job Ads

Among the companies we found doing it: Amazon, Verizon, UPS and Facebook itself. "It's blatantly unlawful," said one employment law expert.

Also on facebook: housing ads and racial bias, lawsuit in SDNY

Gender: word associations enforce stereotypes

Here is a poem written by Google Translate on the topic of gender. It is the result of translating Turkish sentences using the gender-neutral "o" to English (and inspired by this Facebook post).

Gender

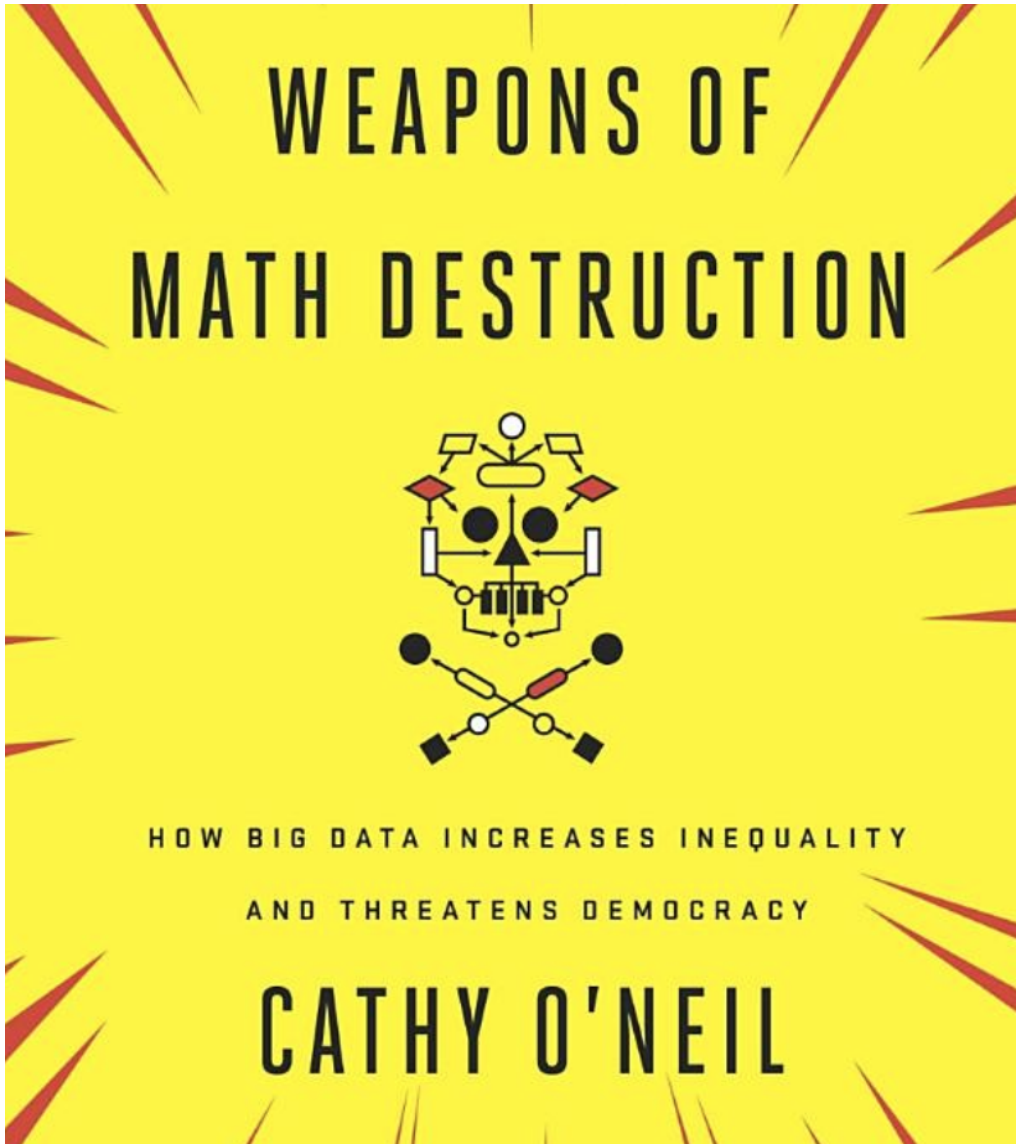
by Google Translate

he is a soldier
she's a teacher
he is a doctor
she is a nurse

he is a writer
he is a dog
she is a nanny
it is a cat

Credit: Nikhil Sonnad / Emre Sarbak

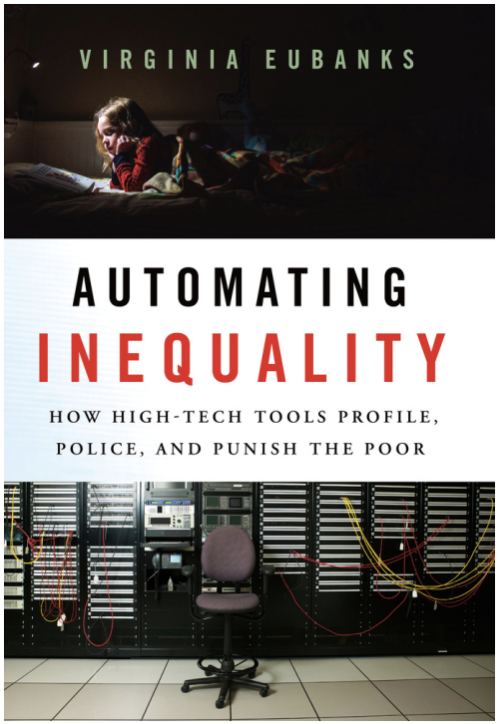
Book full of more examples by Cathy O'Neil

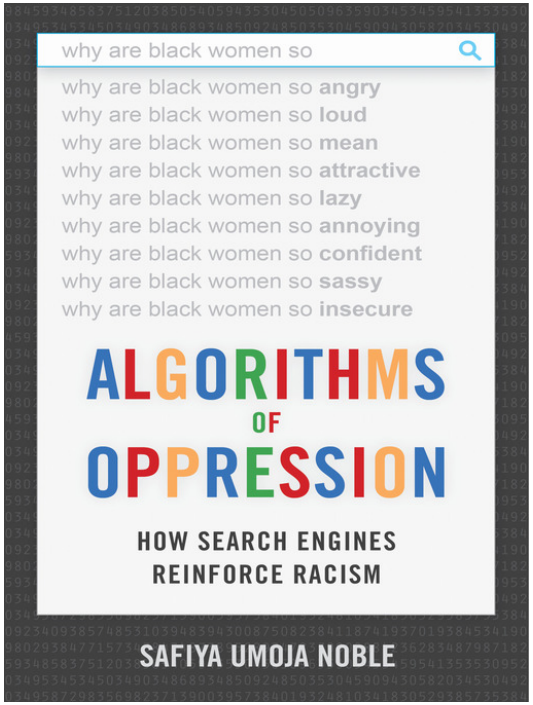


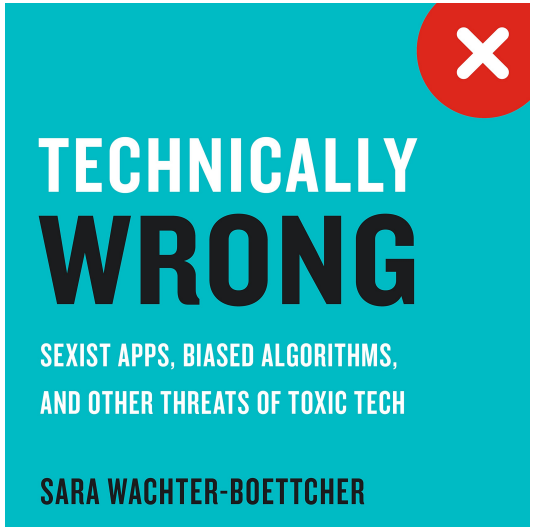
- Admissions
- Credit
- Employment
- Insurance
- Healthcare

Unique dangers: scale, opacity, proxies/wrong data, faux objectivity

More books full of examples: Eubanks, Noble, Wachter-Boettcher







The rise of big data policing: Surveillance, Race, and the Future of Law Enforcement by A. Ferguson

Discrimination

- Failure to apply relevant laws/regulations
- Poor design/incentives
- Ignorance, lack of representation in the field
- Bad people

"Unintentional" disparate impact

- Bias introduced by data collection/use
- Unfair world: many predictor variables **correlated** with both outcome and protected attributes (race, sex, etc)

Research to the rescue! (e.g. FAT/ML beginning at NIPS 2014)

And this seminar! FATES at Columbia DSI

Fairness: legal / observational perpective

Two kinds of unfairness (in US law)

Obvious: discrimination / disparate treatment

- Employment non-discrimination law covers sexual identity and gender *in only 20 states*
- USDOJ v. Fred C. Trump, Donald Trump, and Trump Management, Inc. (1973)

Less obvious: adverse/disparate impact, "unintentional"

- Educational requirements **unrelated** to actual job responsibilities
- Geography / zip codes: partisan gerrymandering **correlated** with racial gerrymandering

Demographic parity means that predictions are independent of A, i.e.

$$P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$$

Similar to **equality of outcomes**. Perhaps the most straightforward, but (surprisingly?) opposed by many...

Depending on causal structure of the world, may be counterfactually unfair (e.g. affirmative action may be necessary)

Hardt et al (2016)

In **equality of opportunity**, the *accuracy* of the algorithm does not depend on A:

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$$

Possibly unequal outcomes / counterfactually unfair. Does not address sources of unfairness affecting Y (i.e. implicitly assumes individuals have had equal opportunity in the past, or not responsibility of present decision to fix it)

Grgic-Hlaca et al (2016)

An algorithm yielding prediction $\hat{Y}$ satifies **fairness through unawareness** if it does not explicitly use A, i.e.

$$\hat{Y} = f(X)$$

Intuitively appealing, analogous to **equal treatment**, people tend to believe such treatment is fair, but we'll see an example that shows it can actually *introduce* counterfactual unfairness.

Chief Justice Roberts: "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race."

Dwork et al (2012)

Algorithm satisfies **individual fairness** if it makes similar predictions for individuals who are similar in their unprotected attributes. If $X_i \approx X_j$ then

$$\hat{Y}(X_i, A_i) \approx \hat{Y}(X_j, A_j)$$

Continuity-like condition for features other than A. Intuitively related to privacy and also *matching approaches to causal inference*. The flexibility/challenge in application is in knowing when $X_i \approx X_j$ (i.e. what is a good metric to measure similarity in the feature space)

The Fundamental (superficial?) Contradiction of Fairness

Various works showing impossibility of simultaneously satisfying several of the different fairness definitions at once: Kleinberg et al (2016), Chouldechova (2016)

(Simplified) impossibility theorem

Unless the world is already fair, the only solutions satisfying both equal treatment (or opportunity) and equal outcomes (demographic parity) are trivial ones
(e.g. put *everyone* in jail)

Fairness: causal perspective

Ethicists and social choice theorists have various notions about

- the role of agency in justice
- responsibility-sensitive egalitarianism
- luck egalitarianism

which involve or center on causal reasoning. Simplifying: it is unfair for individuals to experience different outcomes due to factors outside of their control.

Empirical studies by economists (Cappelen et al, 2013, and Mollerstrom et al, 2015) find most participants prefer redistribution to create fairer outcomes, and do so in ways that depend on how much control individuals have on their outcomes.

Oral presentation at NIPS (2017)

Joint work with Matt Kusner, Chris Russell, and Ricardo Silva

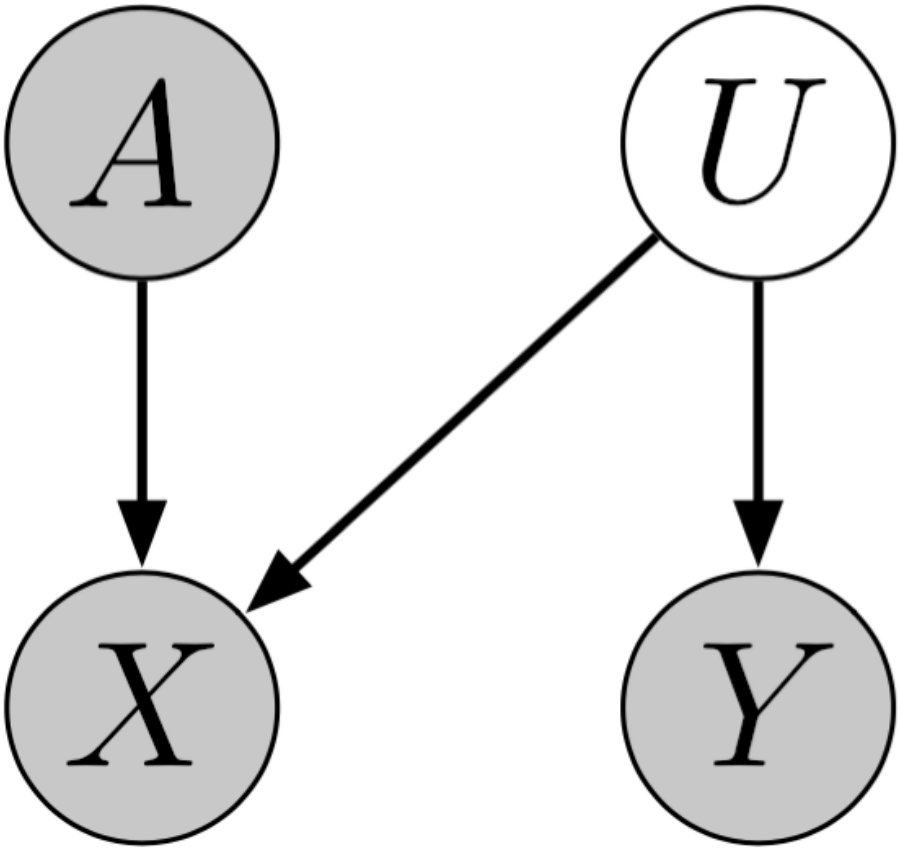
The Alan Turing Institute

Similar works: Pearl et al (2016) (textbook), DeDeo (2014), Kilbertus et al (2017), Johnson et al (2016), Nabi and Shpitser (2018), Zhang and Bareinboim (2017), Chiappa and Gillam (2018), and many others and more to come!

Idea

Explicitly model causal relationships between predictors and protected attributes (or other types of fairness-confounding)

DAG/SEM causal model framework



- Nodes: variables (U unobserved)
- Arrows: causal relationships / conditional (in)dependence
- Structural equations: functional forms of (arrow) relationships

Example: auto insurance risk Y, car colour X, "aggressiveness" U, policyholder's gender A

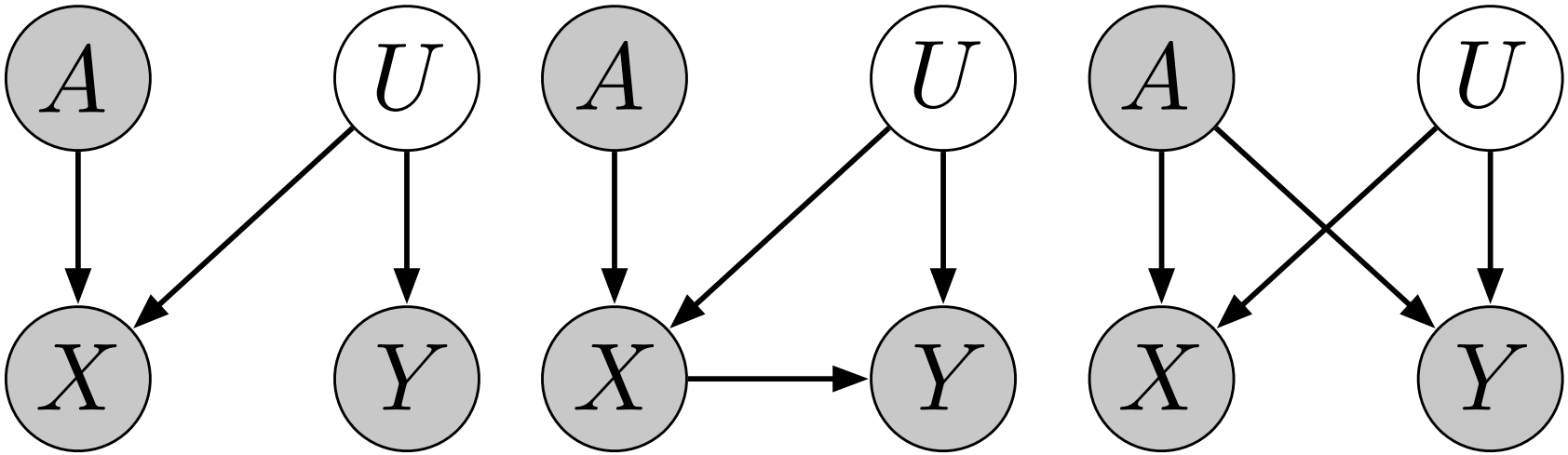
Counterfactuals

An individual has a red car. Perhaps it's because of a preference related to their gender a, or perhaps it's because of an aggressive personality u.

What if (counterfactually) the same individual had been a different gender a'? Perhaps they might not have chosen red.

Can we formalize and compute estimates of counterfactuals?

Computing counterfactuals: follow the paths



Assume a probabilistic model for U, change a to a' and propagate that change through the structural equations to all descendents of A

An estimator $\hat{Y}$ is **counterfactually fair** if

$$P(\hat{Y}_a|x, a) = P(\hat{Y}_{a'}|x, a)$$

Where counterfactuals are computed from a model of the world represented by a DAG. Determining a good model of the world is a subtle and important issue

Proposition

Any estimator $\hat{Y}$ which is a function of only non-descendents of A is counterfactually fair

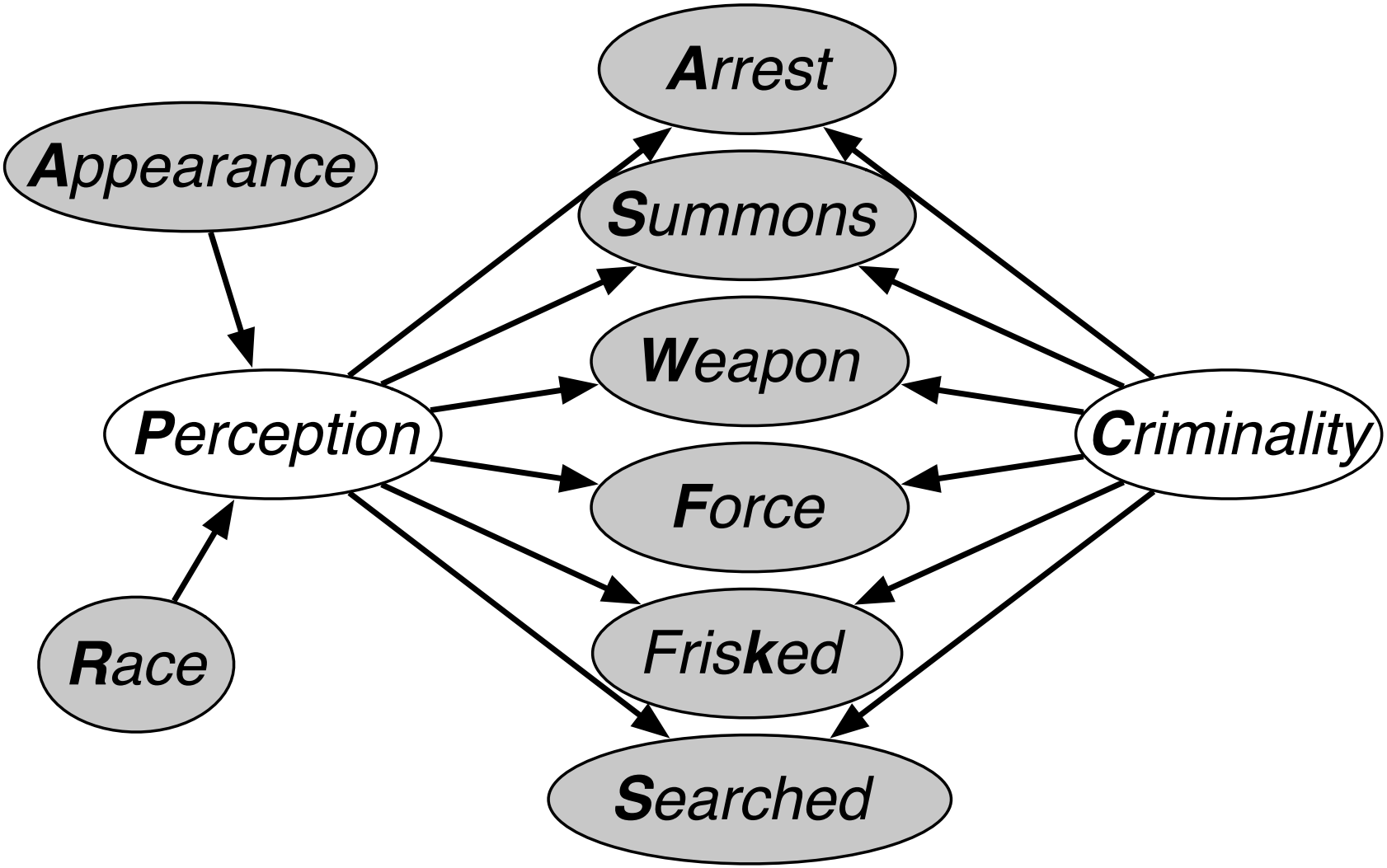
Counterfactual fairness captures the intuition that the outcome should not depend in a causal way on an individual's sensitive attributes *or* other (*irrelevant*) causal consequences thereof.

It is related to specific philosophical/legal conceptions of fairness.

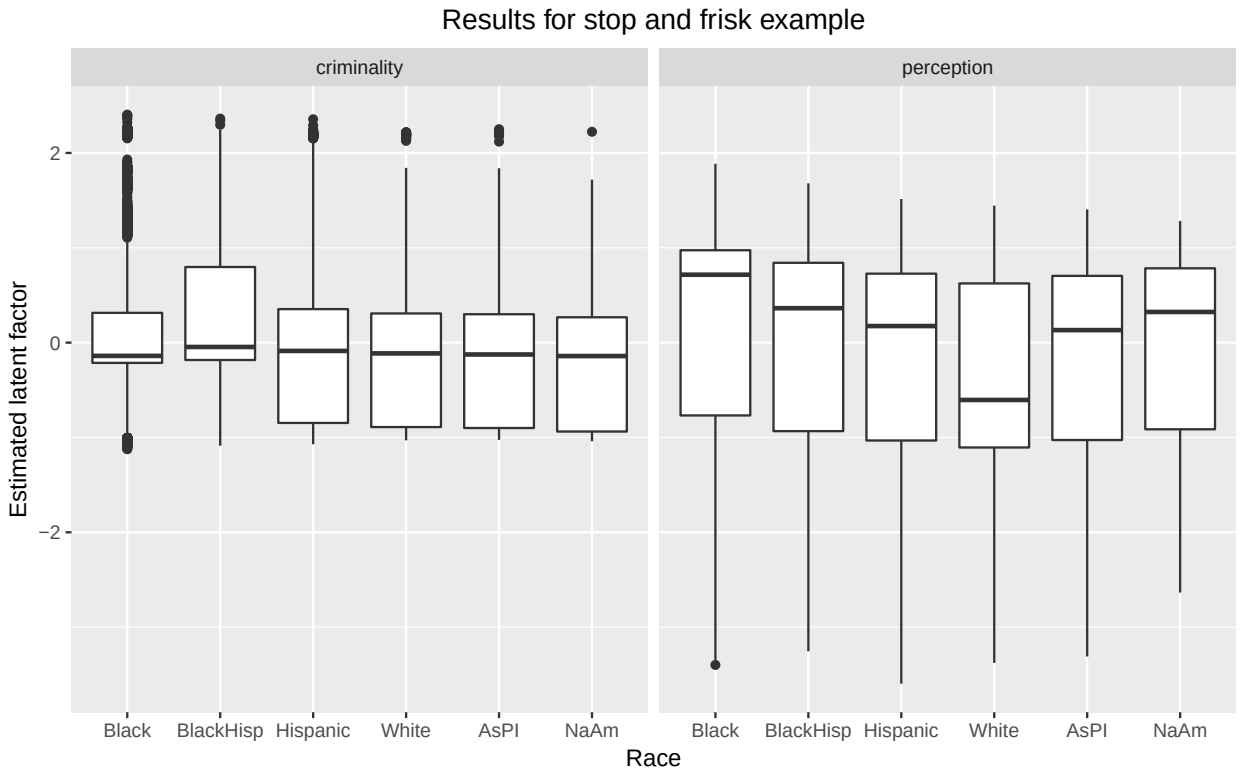
With correct model of the world* it addresses root sources of unfairness.

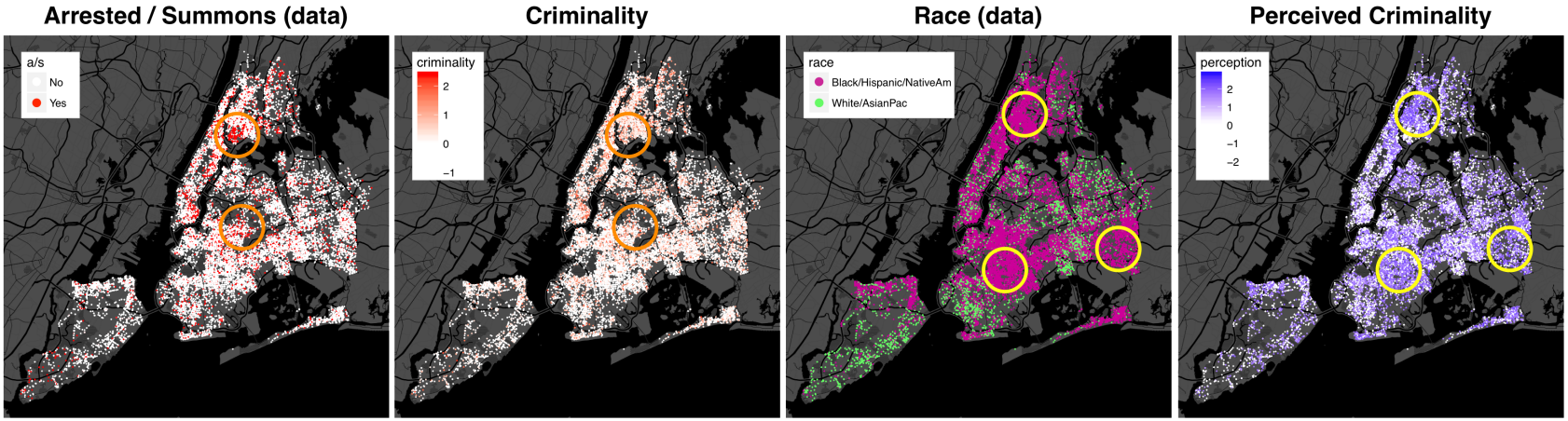
*Big assumption, but also **transparent** because the model is explicit.

Stop and frisk DAG



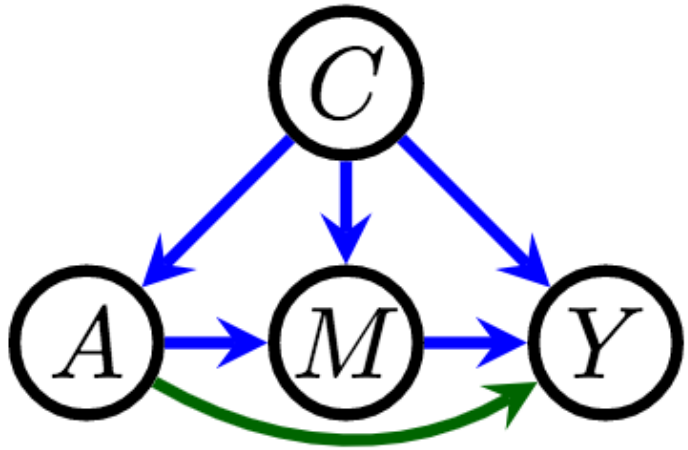
Stop and frisk results





Pathway analysis / decomposition

- Kusner et al (2017): path-dependent counterfactual fairness (supplement)
- Nabi and Shpitser (2018): path specific effects, mediators, constrain parameters
- Chiappa and Gillam (2018): more flexible modeling, modify features
- Zhang and Bareinboim (2017): (counterfactual) direct, indirect, and spurious



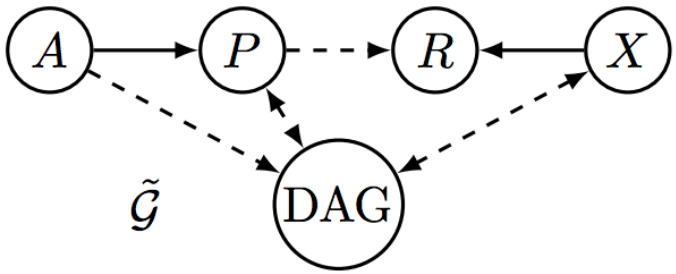
Model for crime data where the mediator can be, e.g., prior convictions

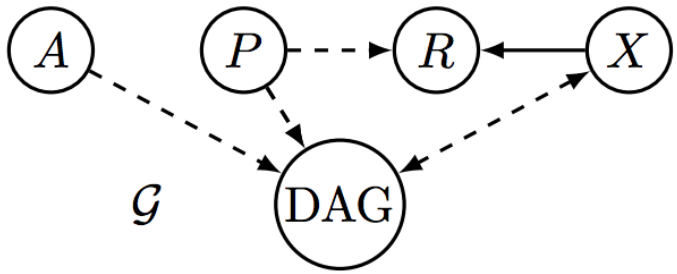
Resolving variable (mediator)

Allow certain pathways from A to $\hat{Y}$ if an intermediate variable X "resolves" discrimination–i.e. is in some sense an admissible source of inequality in Y.

Proxy discrimination

Counterfactuals more philosophically well-defined for proxies P which are highly correlated with A but may conceivably be changed with an intervention.





Pathways and proxies

- Enhanced capability of causal modeling to address fairness questions
- Capture *some aspects* of equal treatment, equal outcomes, equal opportunity
- Does this resolve the Fundamental Contradiction of Fairness?

In my opinion: no.

When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness (Russell et al, 2017, also NIPS)

- Competing causal models
- Approximate counterfactual fairness (relax equality constraint)
- Predictions approximately satisfy fairness across both (all) models
- Limitation: the more contradictory are the competing models, the more trivial the predictions (constant)
- Causal framing of fundamental contradiction

Resolving the contradiction

I think this is the right *path*. It's now about understanding the causes of unfairness well enough to reach consensus.

Subtleties of causal effects due to race/gender

What's in a name? A possible proxy for race or gender. Can conduct randomized experiments by sending the same resume but with different names, and looking for average causal effects due to race or gender of name.

- Bertrand and Mullainathan (2003)

  each resume is assigned either a very African American sounding name or a very White sounding name. [. . .] White names receive 50 percent more callbacks for interviews

Subtleties of causal effects due to race/gender

What's in a name? A possible proxy for race or gender. Can conduct randomized experiments by sending the same resume but with different names, and looking for average causal effects due to race or gender of name.

- Moss-Racusin et al (2012)

  science faculty from research-intensive universities rated the application materials of a student—who was randomly assigned either a male or female name—for a laboratory manager position. Faculty participants rated the male applicant as significantly more competent and hireable than the (identical) female applicant. These participants also selected a higher starting salary and offered more career mentoring to the male applicant.

But discrimination starts much earlier...

- Even if we make counterfactually fair hiring decisions based on the name proxy, what about all the unfair decisions affecting the same individuals preceding our most recent one?
- Given that this effect occurs at time t_2, is there any reason to believe it wasn't also occuring at time $t_1 < t_2$?
- Therefore, someone with an African-American/female sounding name *may have had to work harder* (or be more fortunate in other ways) in order to get an otherwise identical looking resume in the first place
- A preponderance of experimental (like the above), observational, and historical evidence suggests that this is how the world generally works. In light of this, how are we to understand concepts like "equal treatment" or "equal opportunity"?

Selection bias

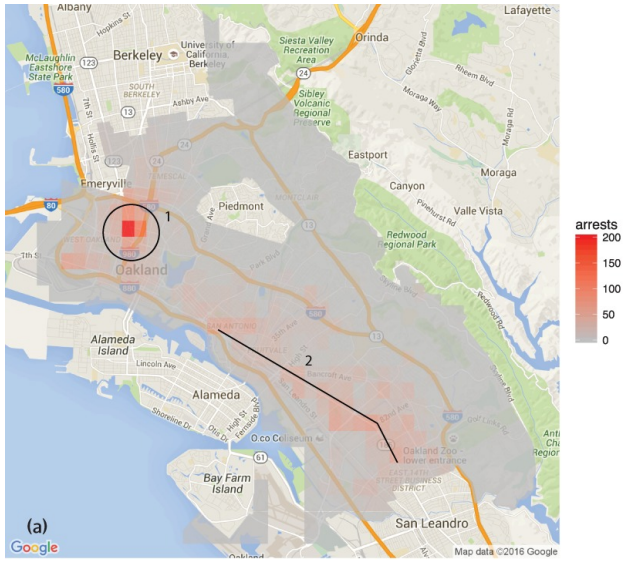
Do protected attributes affect whether someone is in our dataset?

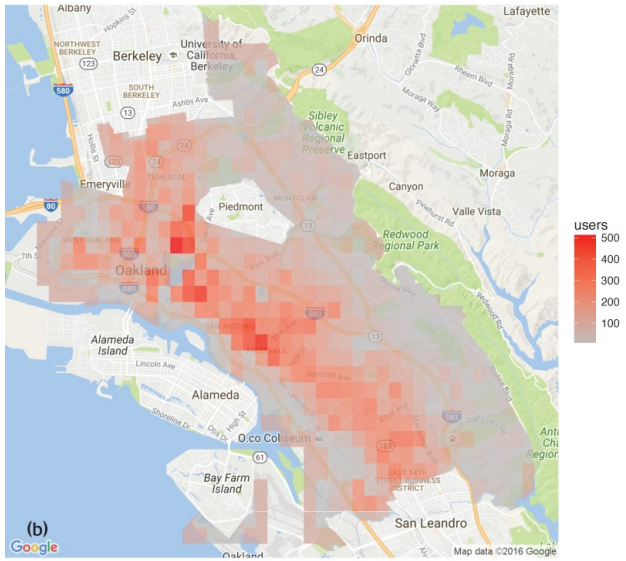
- Pervasive and underappreciated issue in fairness applications
- What are the baseline rates for demographic parity?
- e.g. what if a field has gender parity at the graduate student level, but only the top 10% of women graduates apply for faculty positions while 50% of the men do? What are the fairness responsibilities of hiring committees vs the graduate programs?

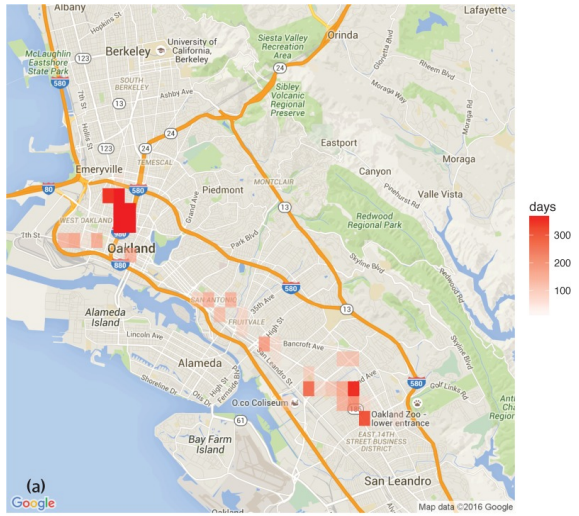
Widespread problem in police/criminal legal system data

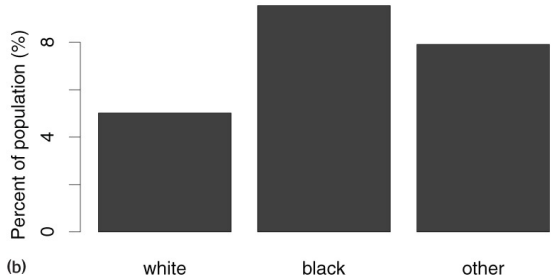
- PredPol locations: what about *unobserved* crimes?
- Stop and frisk: who *wasn't* stopped?
- COMPAS recidivism scores: do they appear biased *because* they are effective?

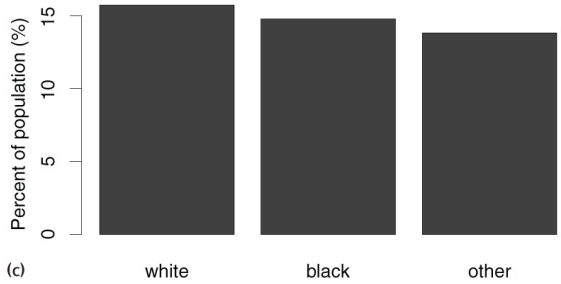
PredPol: To predict and serve, Lum and Isaac (2016)











Feedback loops?

- Arrest data $\neq$ crime data
- Send officers to a neighborhood $\rightarrow$ make an arrest there today
- Observe arrest in a neighborhood $\rightarrow$ send officers there tomorrow

See *Runaway Feedback Loops in Predictive Policing* (Ensign et al, 2017)

Pairs of precincts with same rate of marijuana calls/complaints:

> *In Brooklyn, officers in [. . .] Canarsie arrested people on marijuana possession charges at a rate more than* ***four times as high*** *as [. . .] Greenpoint. . . Canarsie precinct is 85% black. The Greenpoint precinct is 4% black.*
>
> *In Queens, the marijuana arrest rate is more than* ***10 times as high*** *in [. . .] Queens Village as it is in [. . .] Forest Hills. [. . .] the Queens Village precinct is just over half black, [. . .] Forest Hills has a tiny portion of black residents.*
>
> *And in Manhattan, [. . .] Harlem make marijuana arrests at* ***double the rate*** *of [. . .] northern part of the Upper West Side. [. . .] the precinct covering western Harlem has double the % of black residents . . .*

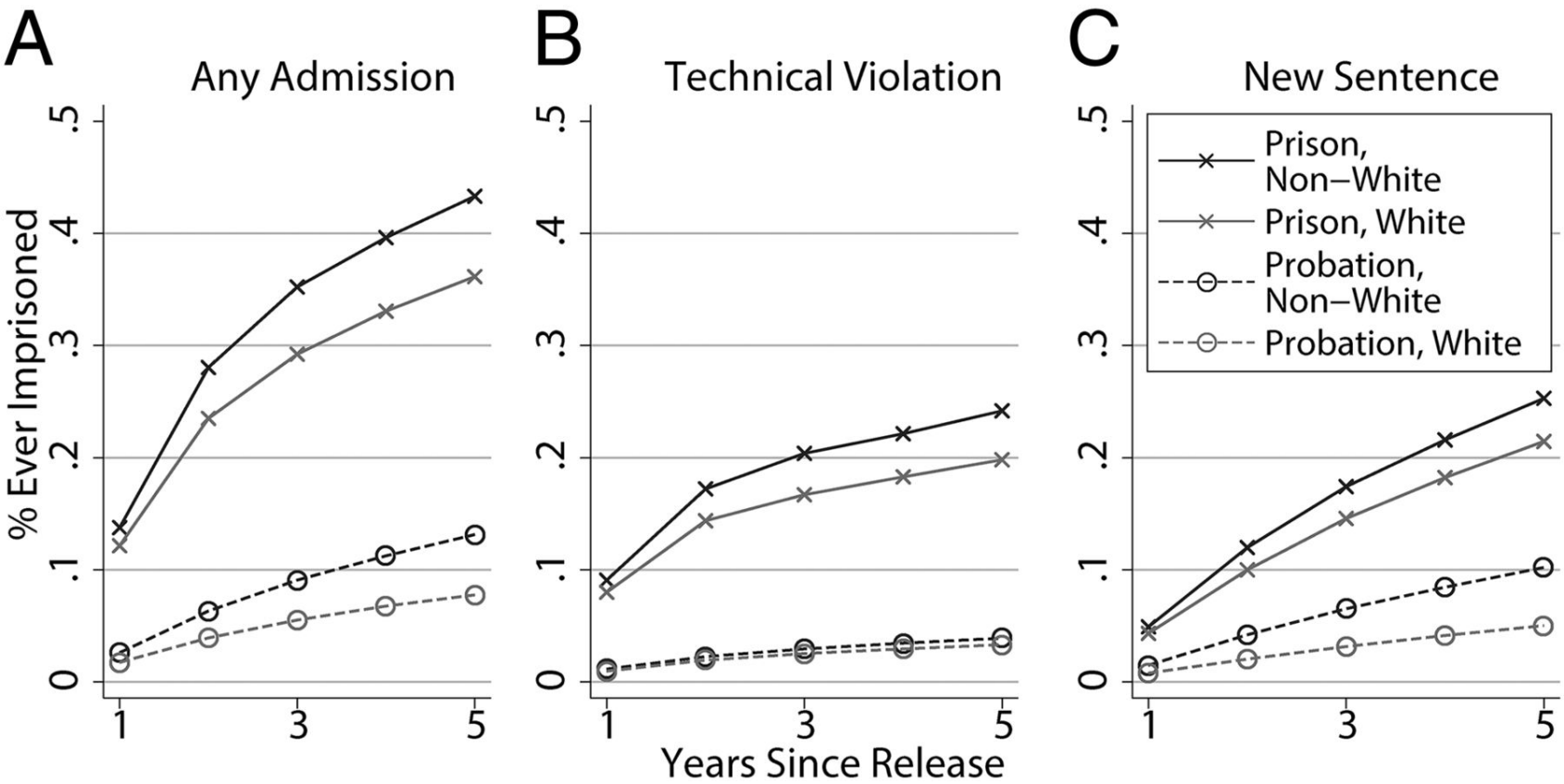
COMPAS scores: are they preventing recidivism?

Do prison sentences prevent recidivism?

Harding et al (2017) use variation in *judges* (instrumental variable) to estimate causal effects

> *. . .* ***imprisonment leads to future imprisonment****. In other words,* ***prison's figurative revolving door has real causal force****, rather than being the simple consequence of imprisonment of individuals at higher risk for future offending. For example,* ***being sentenced to prison rather than probation increases the probability of a future prison admission within 3 y after release by 18–19 percentage points****. [. . .] Although the estimated effect sizes are not different by race, substantially higher rates of imprisonment among nonwhites mean that* ***the consequences of these effects in the aggregate are much more severe for racial minorities****. . .*

Harding et al (2017)



Harding et al (2017)

Our results also demonstrate that the ***majority of this reimprisonment effect is generated by parole violations rather than prison admissions for new felony convictions. Postprison parole supervision surveilles and punishes, and in so doing increases incarceration****. In other words, the rise in incarceration was in part a self-perpetuating process resulting from the workings of the criminal justice system itself.*

Causal inference despite selection bias: Bareinboim et al (2014)

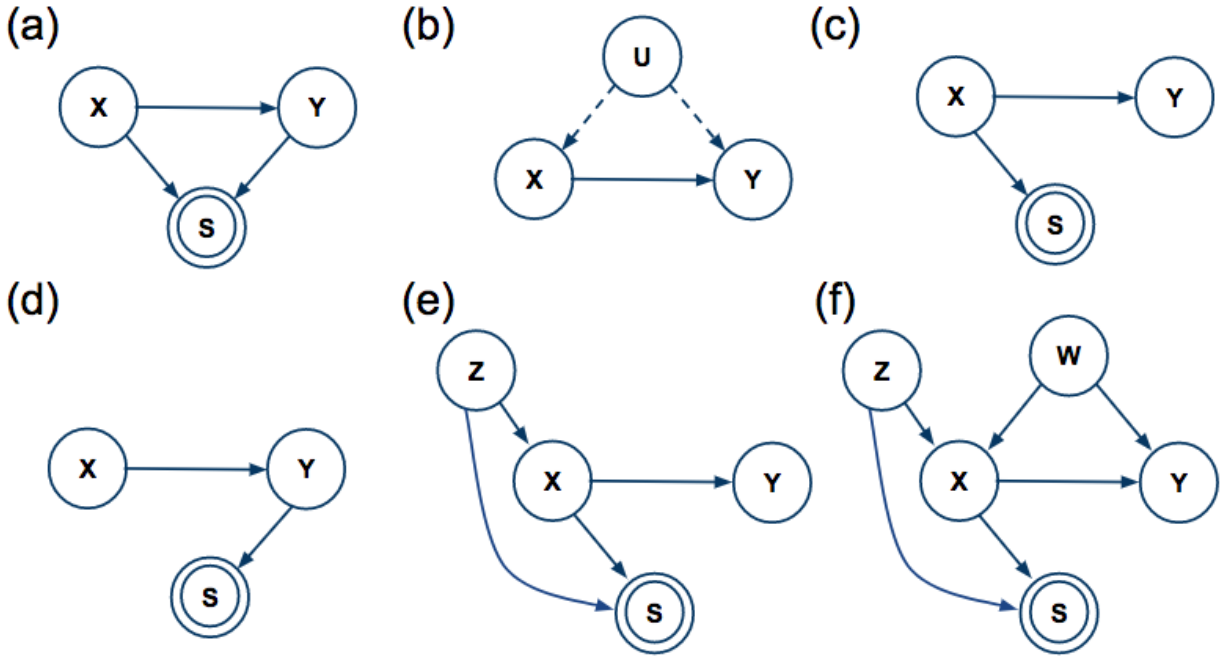


Figure 1: (a,b) Simplest examples of selection and confounding bias, respectively. (c,d) Treatment-dependent and outcome-dependent studies under selection, $Q = P(y|x)$ is recoverable in (c) but not in (d). (e,f) Treatment-dependent study where selection is also affected by driver of treatment Z (e.g., age); Q is recoverable in (e) but not in (f).

Selection bias

Keep an eye out for selection bias and you'll start to see it everywhere... Why?

The (many) causes are topics of study in some social sciences.

Short answer: all of history. (My answer: it's political)

Goodhart's law (Charles Goodhart, economist)

Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes. See also: Manheim and Garrabrant (2018)

Campbell's law (Donald Campbell, psychologist)

The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.

Conclusions

Takeaway messages and directions

Fairness requires understanding the real world

Even if someone doesn't want to use causal fairness definitions, it's important to try to understand how the data is generated/measured, including causal relationships and selection biases. Not doing so can even result in *increasing* unfairness as in the red car example

Fairness requires affirmative action

Affirmative action, also known as positive discrimination: is it really *necessary*, or just a great idea? (My answer: both)

Shift focus from individuals to organizations

- *Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment* Barabas et al (2018)
- *Causal Interventions for Fairness* Kusner et al (2018)

Trying to be humble

Fairness, like data science is interdisciplinary. STEM fields have been privileged by a lot of investment from both the government and industry, but social sciences and humanities are *extremely important* for understanding fairness.

Improvements in algorithms and methodology can only get us so far. Learning even one fairly simple causal graph from observational data alone can be challenging (or impossible), never mind reaching consensus on a complex causal graph with contentious (possibly even adversarial) parties.

Thanks for listening!