Statistical aspects of algorithmic fairness

Joshua Loftus

New York University

February 2020

Joint work with

Matt Kusner

UCL

Chris Russell

Surrey

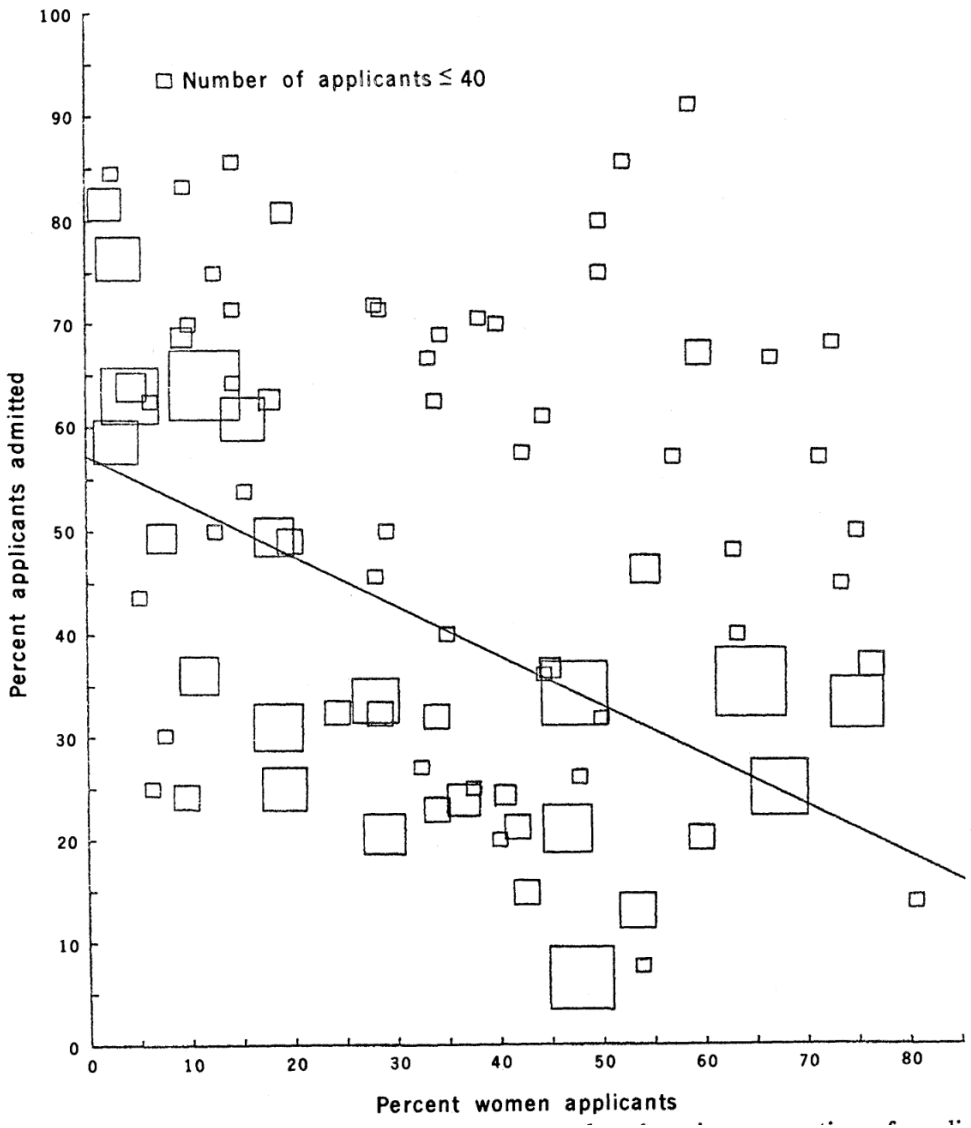
Ricardo Silva

UCL

The Alan Turing Institute

And my current student @ NYU Stern, Margarita Boyarskaya

Using data to detect discrimination



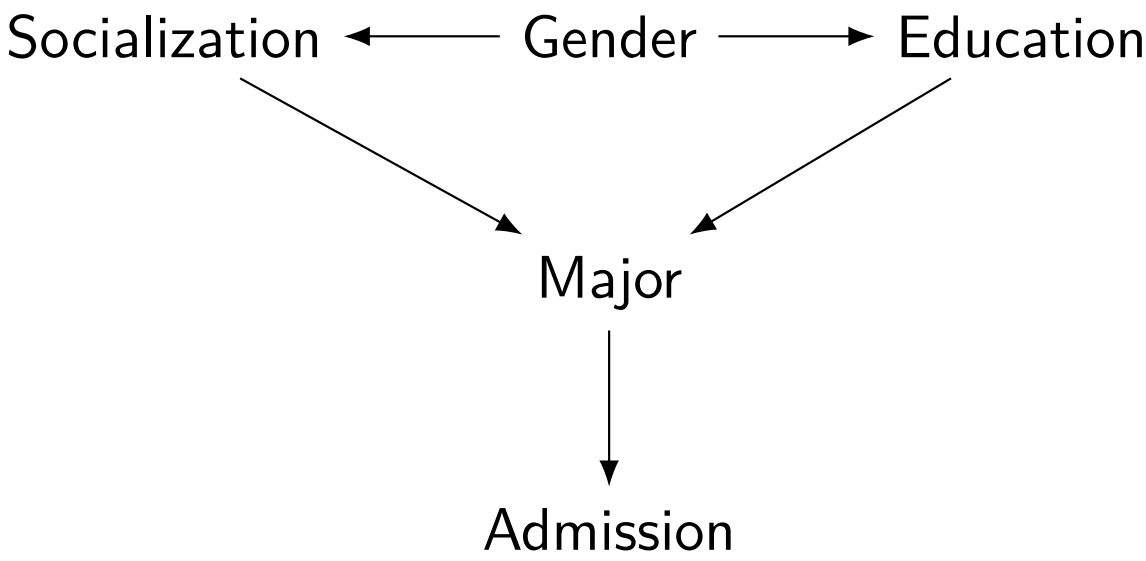
Sex Bias in Graduate Admissions: Data from Berkeley (P. J. Bickel, E. A. Hammel, J. W. O'Connell, 1975)

Lesson: we should condition on covariates?

The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of com- pleted degrees, and less well funded, and that frequently offer poorer professional employment prospects.

- from the final paragraph of Bickel et al (1975)

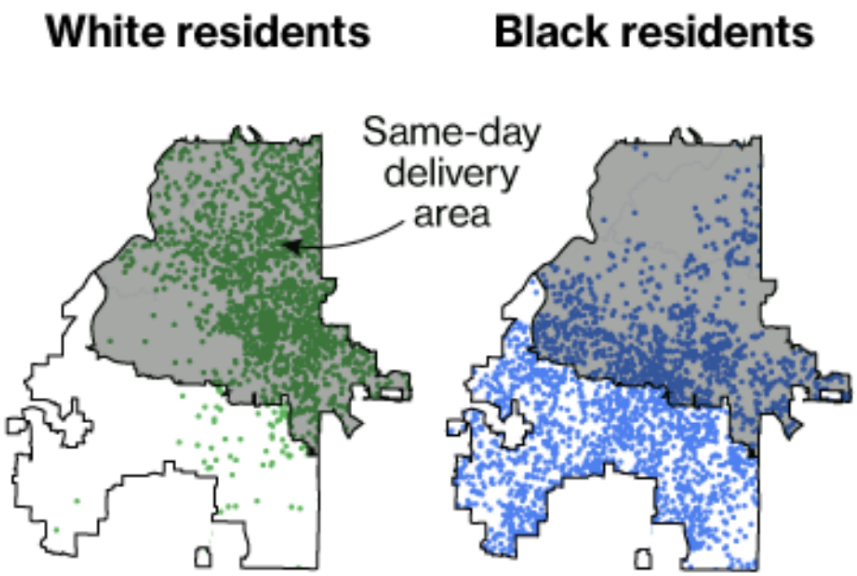
Conditional dependence relations

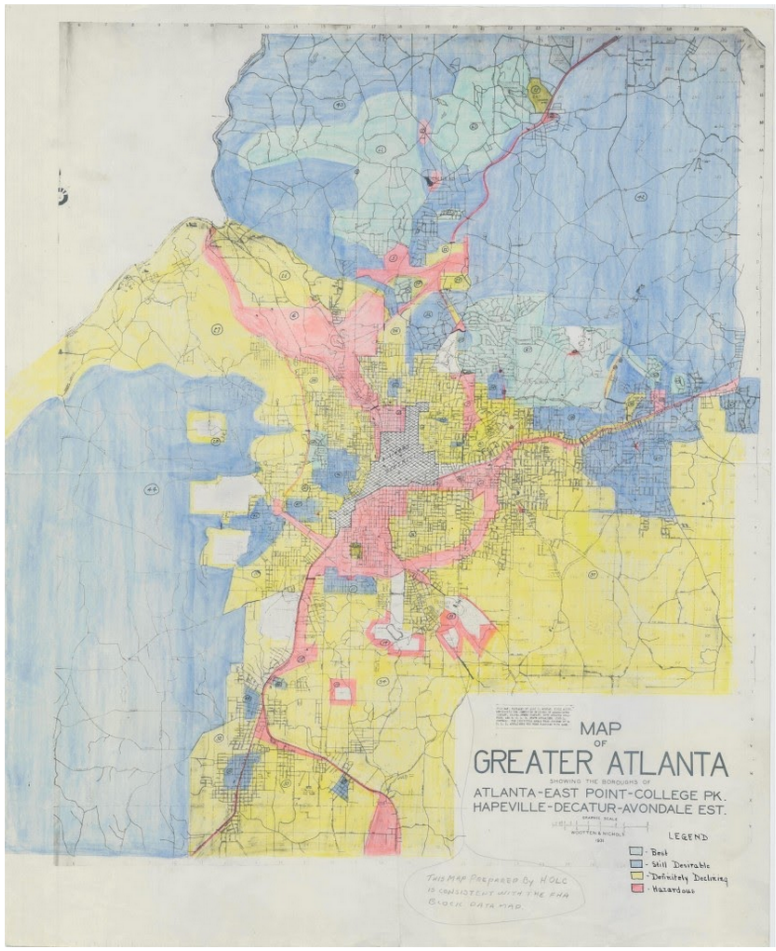


When Joshua showed enjoyment of math as a student, what if his teachers, family, role models, etc, had encouraged him toward a career in nursing or teaching?

Redlining: we should NOT condition on covariates?

The northern half of Atlanta, home to 96% of the city's white residents, has same-day delivery. The southern half, where 90% of the residents are black, is excluded.





Policing: we should condition on labels?

Officer characteristics and racial disparities in fatal officer-involved shootings (David J. Johnson, Trevor Tress, Nicole Burkel, Carley Taylor, and Joseph Cesario, PNAS 2019)

> *A persistent point of debate in studying police use of force concerns how to calculate racial disparities. Racial disparities in fatal shootings have traditionally been tested by asking whether officers fatally shoot a racial group more than some benchmark, such as that group's population proportion in the United States. [. . .]*
>
> *However, using population as a benchmark makes the strong assumption that White and Black civilians have equal exposure to situations that result in FOIS. [. . .] When violent crime is used as a benchmark, anti-Black disparities in FOIS* ***disappear or even reverse****.*

Policing: we should NOT condition on labels?

Knox and Mummolo:

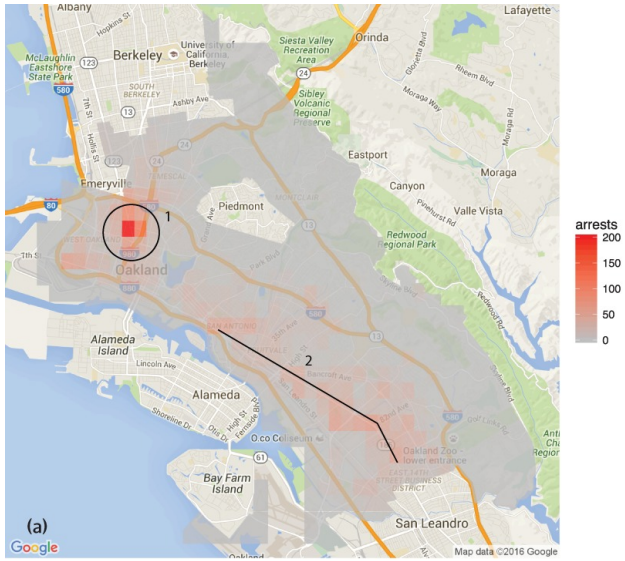
> *Johnson et al.'s (1) analysis cannot inform the original claims without accounting for Bayes' rule*

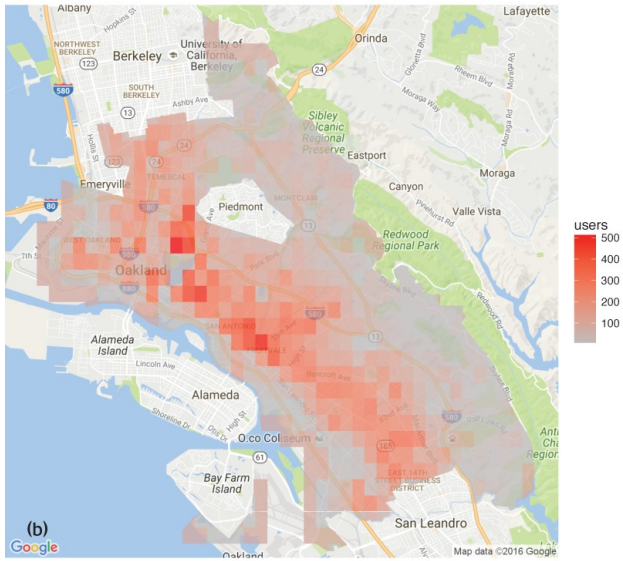
Response:

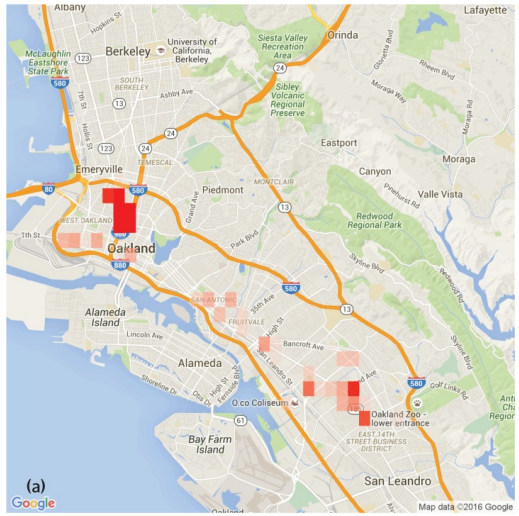
> *. . . the analyses in our report account for racial differences in exposure by controlling for crime rates, a proxy for offending*

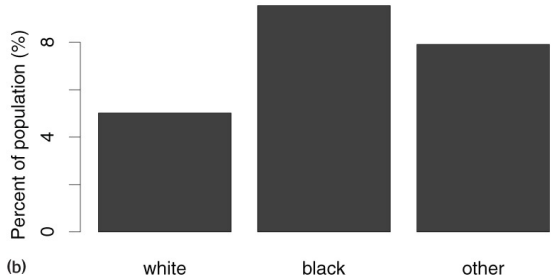
A debate that has been going on in policing for decades, perhaps even since Quetelet in the 19th century (*despite one side being obviously wrong*)

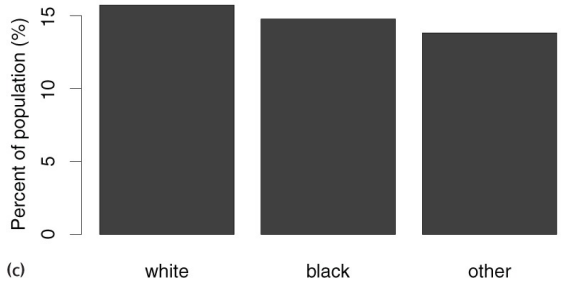
PredPol: To predict and serve, Lum and Isaac (2016)











Bloomberg interlude

You want to spend the money on a lot of cops in the streets. Put those cops where the crime is, which means in minority neighborhoods, So, [inaudible] unintended consequences is people say, "Oh, my God, you are arresting kids for marijuana that are all minorities." Yes, that's true. Why? Because we put all the cops in the minority neighborhoods. Yes, that's true. Why do we do it? Because that's where all the crime is.

- Mike Bloomberg, arguing in 2015 for **targeted policing of minorities**

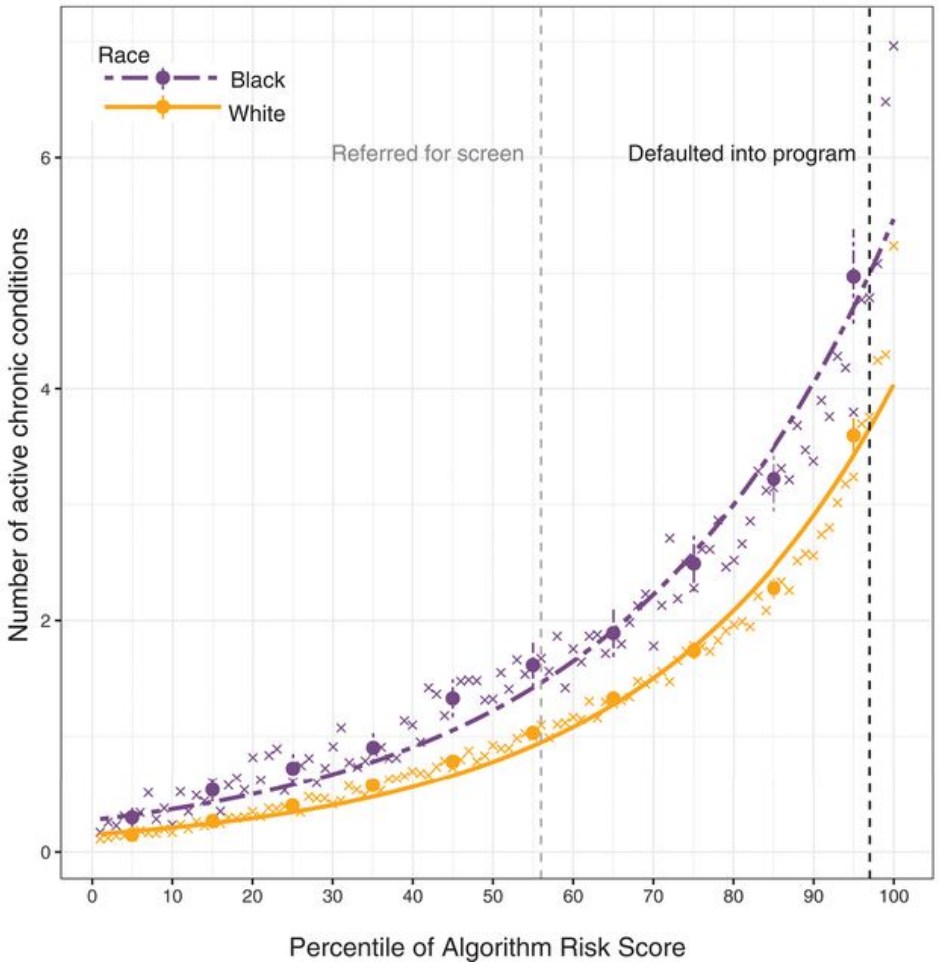
What makes these examples confusing?

Statistical issues

- How data/models (fail to) relate to the "real world"
- Causation vs association
- Sampling bias

These are urgent, important, first order problems, not just curiosities–in some cases they're the whole "game"

Sampling bias in healthcare patient data



Obermeyer et al (2019)

- Algorithm assigns risk scores by predicting healthcare **costs** from patient records
- Underestimates risk of health conditions for black patients compared to white patients
- Adjusting algorithm to close the gap results in 2.5x black patients receiving more care

"New products and services, including those that incorporate or utilize artificial intelligence and machine learning, can raise new or exacerbate existing ethical, technological, legal, and other challenges, which may negatively affect our brands and demand for our products and services and adversely affect our revenues and operating results"

TOM SIMONITE BUSINESS 02.11.19 07:00 AM

GOOGLE AND MICROSOFT WARN THAT AI MAY DO DUMB THINGS

ALYSSA FOOTE

Source: WIRED article, Feb. 2019

Next: my research

- Pause for questions?

Next: my research

- Pause for questions?
- Now let's use math to help us understand some of these problems

The Alan Turing Institute

UNIVERSITY OF CAMBRIDGE

Statistical Laboratory

Basic idea

Model the relevant real-world causal relationships to separate discrimination on the basis of protected attributes from other, potentially correlated observables (and unobserved confounders)

(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 others)

Ethicists and social choice theorists have various notions about

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

which rely on causal reasoning. Roughly, it is unfair for individuals to experience different outcomes due to factors outside of their control.

A notation slide

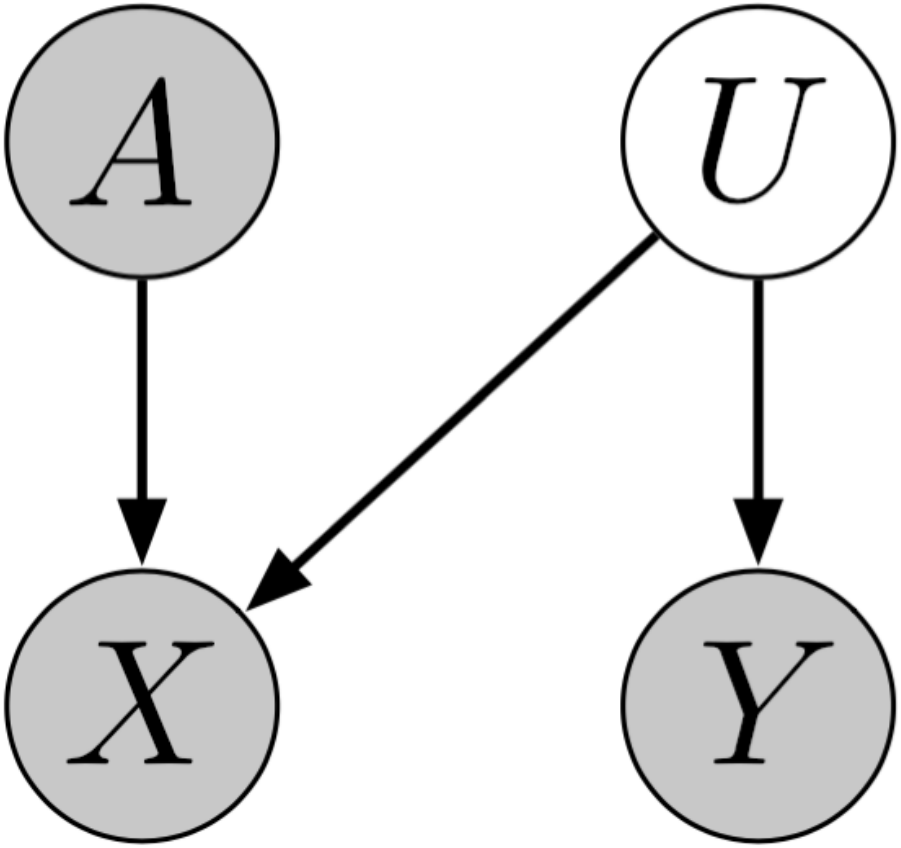
Supervised learning

- Outcome variable $\mathbf{Y}$ - consider as a score for decisions, or $\mathbf{Y} = 1$ as the desirable decision
- Sensitive/protected attribute(s) $\mathbf{A}$ - race, gender, ...
- Other predictors $\mathbf{X}$ - not sensitive (*prima facie*)

Machine learning task: learn a function $f(\mathbf{X}, \mathbf{A})$ from (labeled) training data to predict values of $\widehat{\mathbf{Y}} = f(\mathbf{X}, \mathbf{A})$ on (unlabeled/future) test data (by minimizing some loss function that measures closeness of $\widehat{\mathbf{Y}}$ to $\mathbf{Y}$ on the training data)

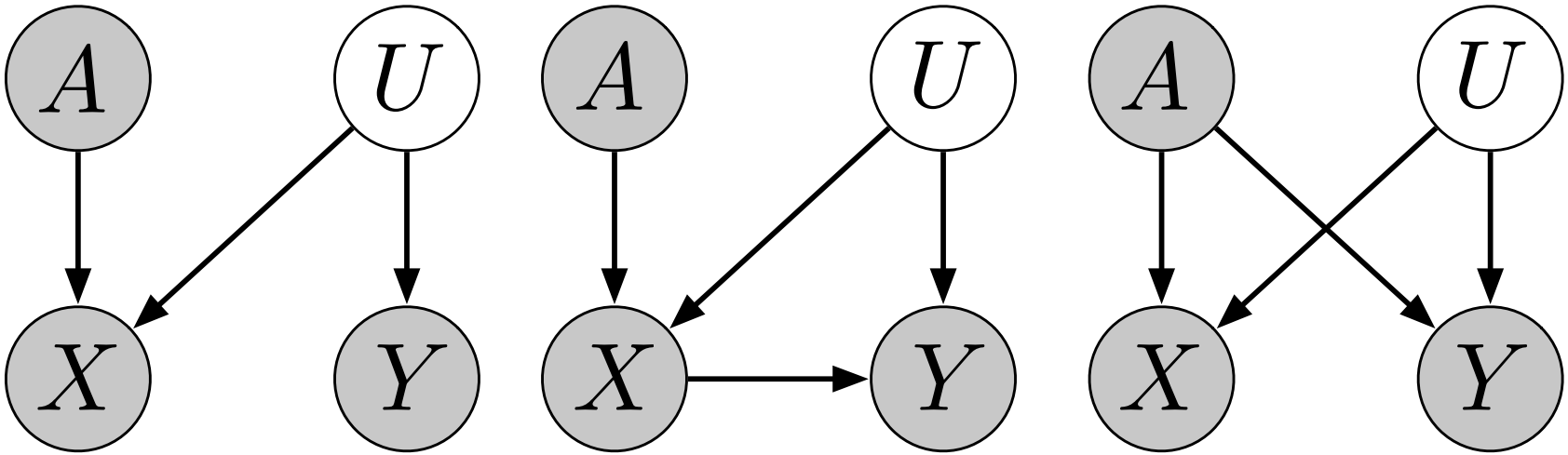
What would it mean for such function to be fair with respect to $\mathbf{A}$?

DAG/SEM causal model framework



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

Computing counterfactuals: follow the paths



Assume a probabilistic model for **U**, estimate parameters using observed data, change a to a' and propagate that change through the structural equations to all descendents of **A**

Counterfactual fairness

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

$$\mathbb{P}(\widehat{\mathbf{Y}}_{a}|\mathbf{X} = x, \mathbf{A} = a) = \mathbb{P}(\widehat{\mathbf{Y}}_{a'}|\mathbf{X} = x, \mathbf{A} = a)$$

for all a'.

Proposition: structural counterfactual fairness

Any estimator $\widehat{\mathbf{Y}}$ which is a function of only non-descendents of $\mathbf{A}$ is counterfactually fair (sufficient condition, not necessary)

Intuitive appeal

- This definition captures the intuition that the outcome should not depend in a causal way on an individual's sensitive attributes *or* other causal consequences thereof.
- With correct model of the world* it addresses root sources of unfairness.
- Next: A few *observational* definitions of fairness, a simple example, and then back to causality...

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

(Likely violates "no causation without manipulation" unless we restrict to *perception* of the sensitive attribute)

Perhaps the most straightforward definition (and my favorite due to its elegant simplicity), often described as **equality of outcomes**

Demographic parity

Predictions (or decisions) are independent of $\mathbf{A}$:

$$\mathbb{P}(\widehat{\mathbf{Y}}|\mathbf{A}=0) = \mathbb{P}(\widehat{\mathbf{Y}}|\mathbf{A}=1)$$

Equality of opportunity (Hardt et al, 2016)

The accuracy of the algorithm does not depend on $\mathbf{A}$:

$$\mathbb{P}(\widehat{\mathbf{Y}} = 1 | \mathbf{A} = 0, \mathbf{Y} = 1) = \mathbb{P}(\widehat{\mathbf{Y}} = 1 | \mathbf{A} = 1, \mathbf{Y} = 1)$$

Demographic parity but only among individuals "qualified" for the desirable outcome

"Equal treatment," people tend to believe such treatment is fair.

Grgic-Hlaca et al. (2016)

Prediction does not *explicitly* use $\mathbf{A}$, i.e.

$$\widehat{\mathbf{Y}} = f(\mathbf{X})$$

Unfortunately encoded in US law (despite contradicting other laws)

Dwork et al. (2012)

Similar predictions for individuals who are similar (in their unprotected attributes). If $\mathbf{X}_i \approx \mathbf{X}_{i'}$ then

$$\widehat{\mathbf{Y}}(\mathbf{X}_i, \mathbf{A}_i) \approx \widehat{\mathbf{Y}}(\mathbf{X}_{i'}, \mathbf{A}_{i'})$$

Continuity condition in $\mathbf{X}$ but not $\mathbf{A}$. Can be related to matching approaches to causal inference

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. jail everyone)

Many versions of this can be proven with different sets of assumptions but basically the same conclusion: some fairness definitions are contradictory

Causal models as tools for understanding fairness

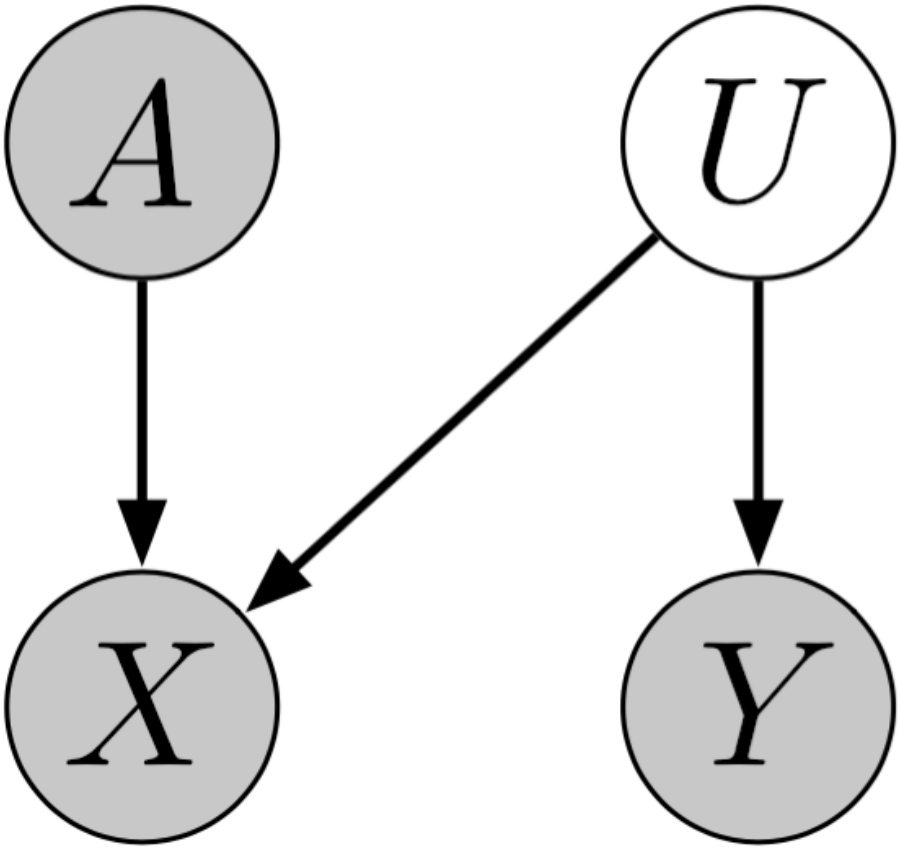
I do not advocate everyone use the counterfactual definition of fairness. Maybe there is no one "right" answer. This approach is useful for probing/understanding limitations.

Chief Justice Roberts: "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race" (PICS, 2007).

Equal treatment or fairness through unawareness

Is it actually any good?

Chief Justice Roberts is *mathematically* wrong

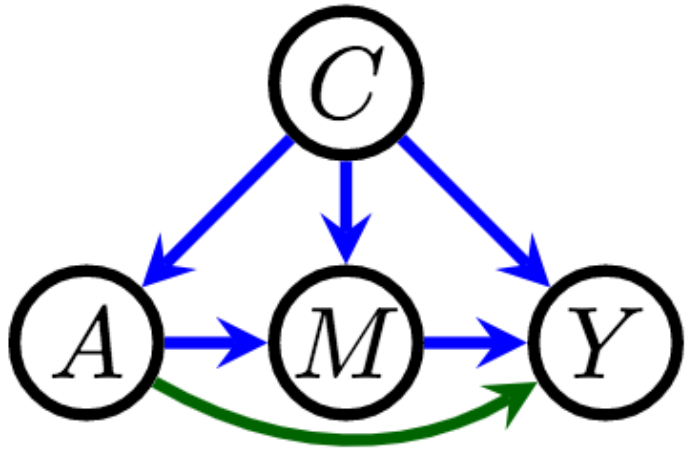


- Auto insurance risk **Y**
- Car color **X**
- Policyholder's gender **A**
- "Aggressiveness" **U**

"Equal treatment" actually *introduces* unfairness where there was none

Pathway analysis / decomposition

- Kusner et al (2017): path-dependent counterfactual fairness (supplement)
- Kilbertus et al (2017): proxies and resolving variables
- 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

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. People will disagree about pathways

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

- 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

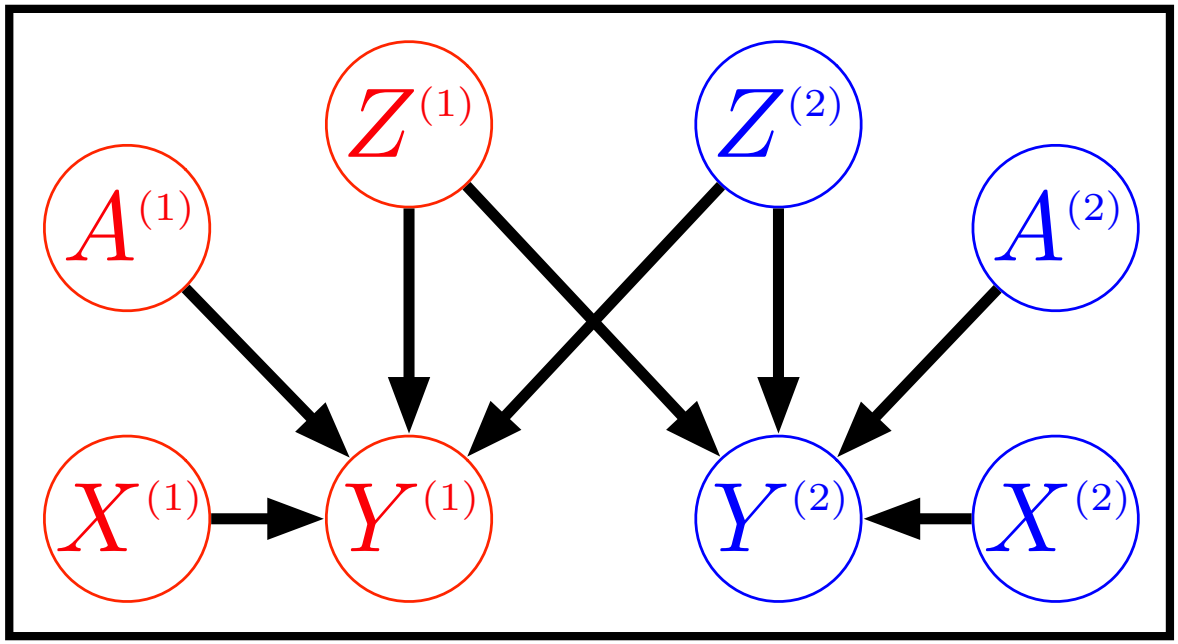
I think this is the right *path*. It's now about understanding the causes of unfairness well enough to reach consensus.

Recognizing separate design elements of predictive ML

- Model the **intervention** (that predictions will be used for) as its own separate variable **Z**
- Causal model including **Z**, explicitly model changes in the world due to use of ML
- Potentially multiple objectives to represent interests of different stakeholders

Recent work: Making Decisions that Reduce Discriminatory Impacts (Kusner et al, *ICML 2019*)

Example: fair intervention under interference



When designing an optimal, fair intervention $\mathbf{Z}$, instead of enforcing the equality in the definition of counterfactual fairness, we can also use an asymmetric bound on **counterfactual privilege**, for $\tau \geq 0$

$$\mathbb{E}[\widehat{\mathbf{Y}}(a, \mathbf{Z})] - \mathbb{E}[\widehat{\mathbf{Y}}(a', \mathbf{Z})] \leq \tau$$

- In practice these *asymmetric constraints* will only be active for privileged values of a (actual, left term), and inactive otherwise
- Intervention should not allocate resources in a way that helps people (in expectation) become more than τ (in terms of the outcome) units better than they would be if they were not privileged

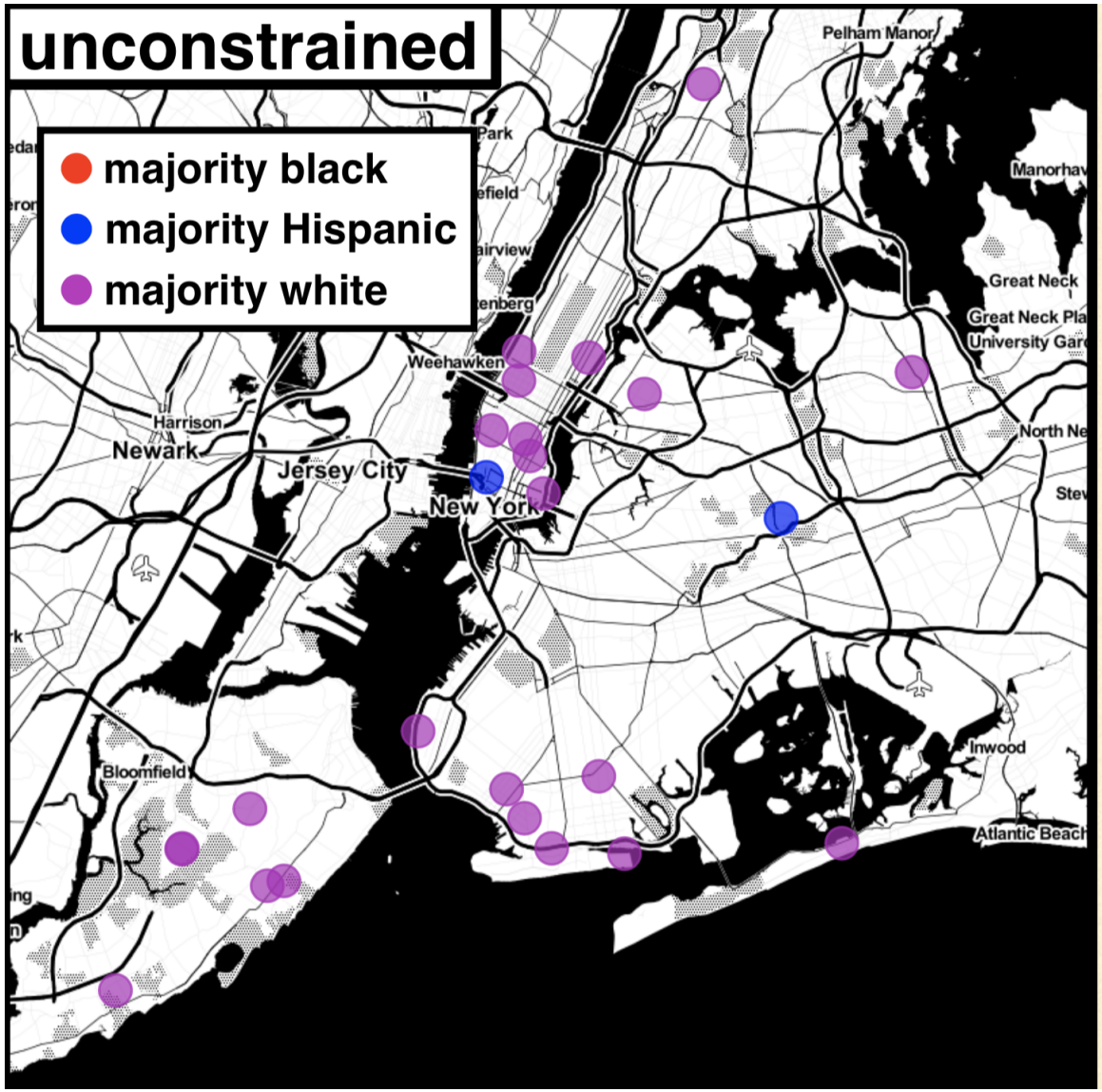
Optimal intervention under interference

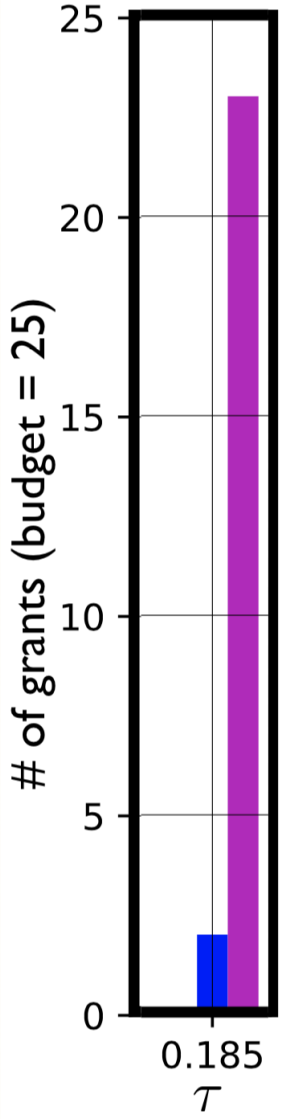
- Our goal is to design optimal interventions or policies $\mathbf{Z}$ subject to a budget constraint, e.g.

$$\mathbf{Z} = \arg\max \sum_i \mathbb{E}\left[\widehat{\mathbf{Y}}^{(i)}(a^{(i)}, \mathbf{Z}) | \mathbf{A}^{(i)}, \mathbf{X}^{(i)}\right] \quad s.t. \quad \sum_i \mathbf{Z}^{(i)} \leq b$$

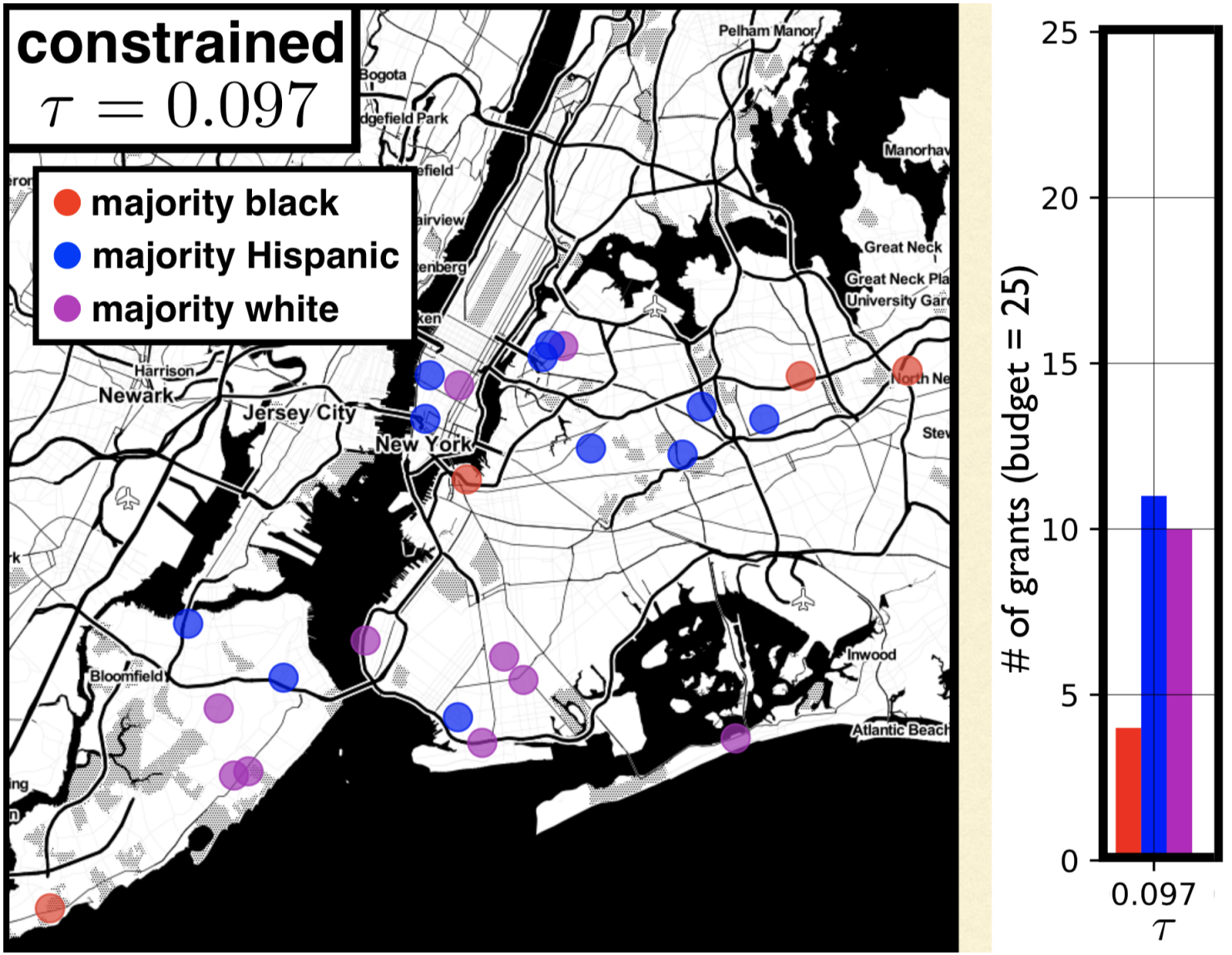
- Interference means $\mathbf{Y}^{(i)}$ is potentially a function of all of $\mathbf{Z}$ and not just $\mathbf{Z}^{(i)}$
- Next two slides: optimal interventions with and without counterfactual privilege constraint

School resource allocation without fairness constraint





School resource allocation bounded counterfactual privilege



Other directions

Things I'm working on

- Drawing attention to statistical aspects of fair ML, e.g. sampling bias in training data (equal opportunity?)
- Attacking bad ideas (from SCOTUS or elsewhere) like the "cost of fairness" or "trade-off" between "accuracy" and fairness (hint: **utility** may increase with fairness!)

Things it would be cool if other people worked on, in general

- Interdisciplinary/STS-informed tech/quantitative research
- Empirical ethics, "big data"/ML-informed social research
- Political/economy of whether any of this is actually headed in a good direction

Questions?

Thank you for listening!

Reading for a fairly general audience: *The long road to fairer algorithms*. Nature, 2020

joshualoftus.com