Responsible Data Science

Fairness and Causality

Prof. George Wood

Center for Data Science
New York University

data RESPONSIBLY

NYU

Defining causal inference

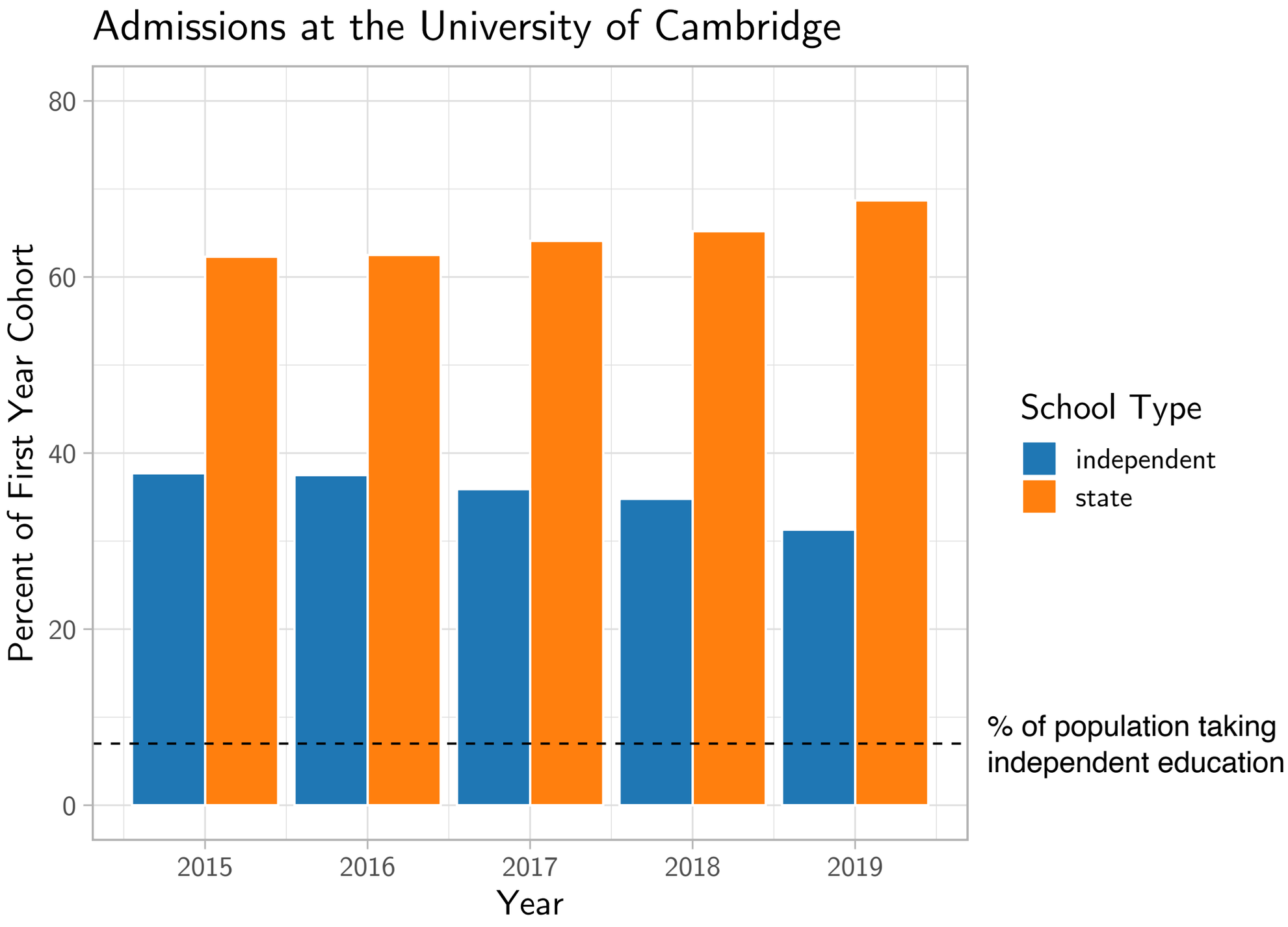
Causal model + statistical inference

Focus ① today

Focus ②

Causal reasoning and fairness

What is a causal model?



What is a causal model?

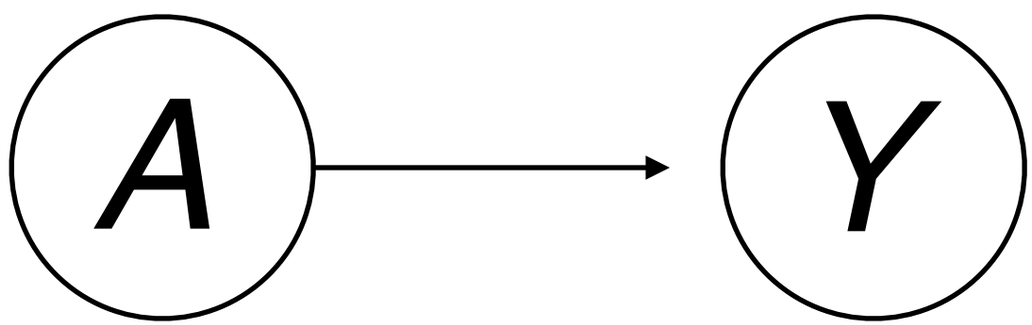
***A*, Treatment:**

Student attends an independent school

→

***Y*, Outcome:**

Student gets a place at Cambridge



We can represent this causal structure using a Directed Acyclic Graph (DAG)

Nodes: variables

Arrows: causal relationships

What is a causal model?

A causal model presupposes a counterfactual

***A*, Treatment:**

Student attends an independent school

→

***Y*, Outcome:**

Student gets a place at Cambridge

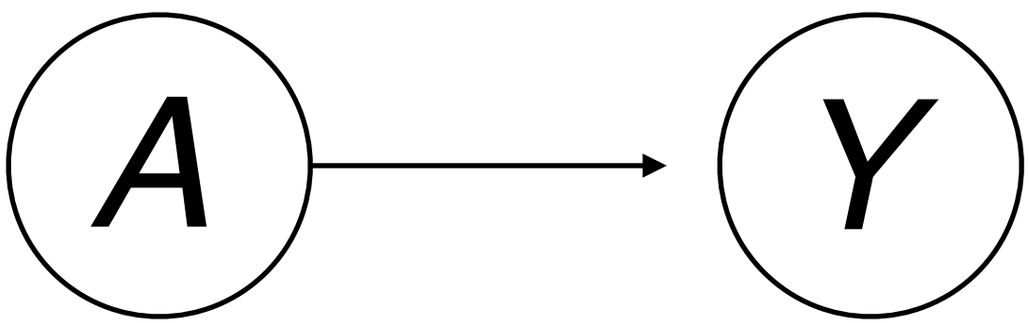
***A′*, Treatment:**

Student does not attend an independent school

→

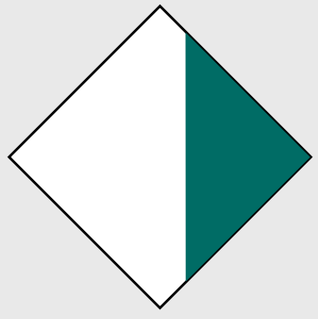
***Y*, Outcome:**

Student does not get a place at Cambridge



Association and causation

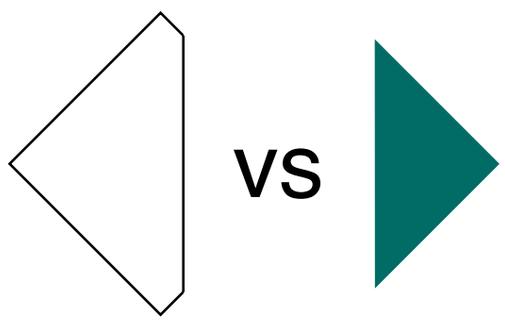
Population of units



Untreated

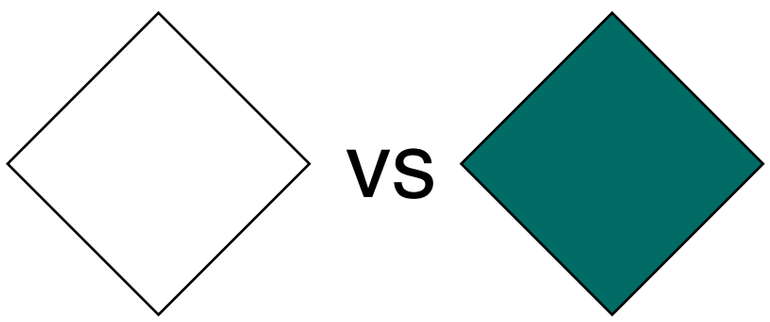
Treated

Association



$E[Y^{a=0}]$ $E[Y^{a=1}]$

Causation

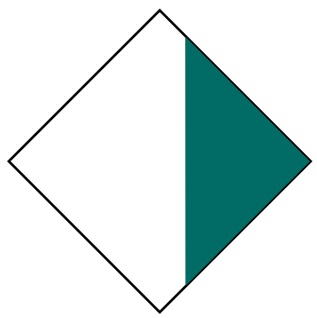


$E[Y|A=0]$ $E[Y|A=1]$

Adapted from **Hernán and Robins (2020),** ***Causal Inference: What If***

Association and causation

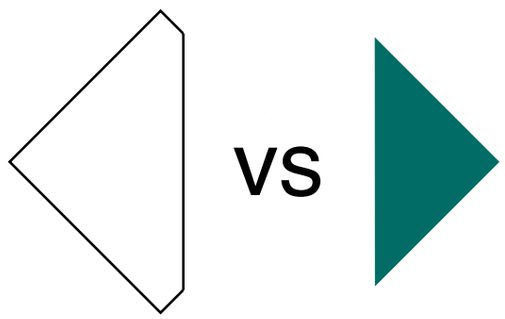
Population of units



What is the probability of going to Cambridge for students at state schools

What is the probability of going to Cambridge for students at independent schools

Association



$E[Y^{a=0}]$ $E[Y^{a=1}]$

⇒ The world as it is

Adapted from **Hernán and Robins (2020), *Causal Inference: What If***

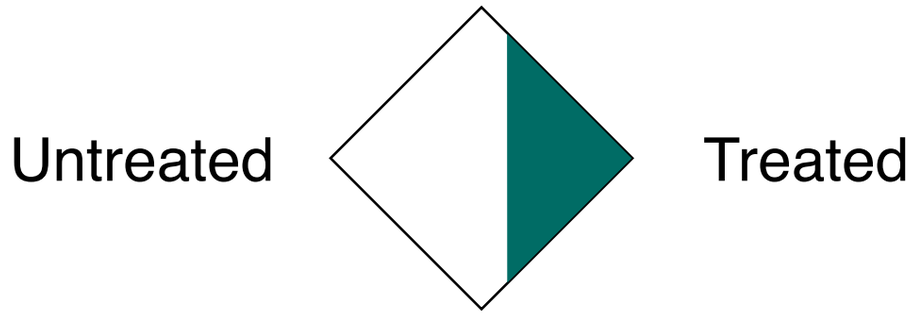
Association and causation

What if a student at the independent school had attended a state school?

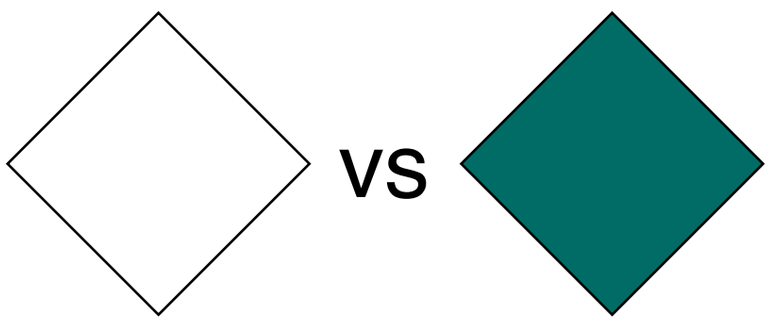
What if a student at the state school had attended an independent school?

⇒ A counterfactual world

Population of units



Causation



$\mathrm{E}[Y|A=0]$ $\mathrm{E}[Y|A=1]$

Adapted from **Hernán and Robins (2020),** ***Causal Inference: What If***

Fundamental problem of causal inference

We cannot observe the counterfactual!

Building blocks of causal modeling

- Units, *i (e.g.* individuals, countries)
- Treatments (also called actions, interventions), ***A***
- Outcomes, ***Y***

*For simplicity, we'll consider the case where **A** is binary, i.e. the unit receives treatment (**A** = a) or it does not (**A** = a′)*

We may also observe other features of the unit *i*. In the machine learning context, these features are typically represented by a matrix **X**

A basic definition of causal effects

- Causal effect for individual *i:* $Y_i^{a=1} \neq Y_i^{a=0}$
- Average causal effect in population: $\mathrm{E}[Y^{a=1}] \neq \mathrm{E}[Y^{a=0}]$

Capital letters represent random variables. Lower case letters denote particular values of a random variable.

E represents the expected value.

Association and causation

Typical supervised learning task:

Predict *Y* given ***A*** and a matrix of features ***X***

$$\mathrm{p}(y \mid \mathrm{A} = a, \mathbf{X} = x)$$

If we observe some value of *a* and *x*, what would we observe about *y*?

In causal inference, we want to know

$$\mathrm{p}(y \mid \mathbf{A} \leftarrow a, \mathbf{X} = x)$$

What happens to *Y* when we intervene on ***A***?

Sometimes $p(y \mid A \leftarrow a)$ is written as $p(y \mid do(a))$

This is really, really difficult

Complicating our casual model

***U*, Confounding variable:**

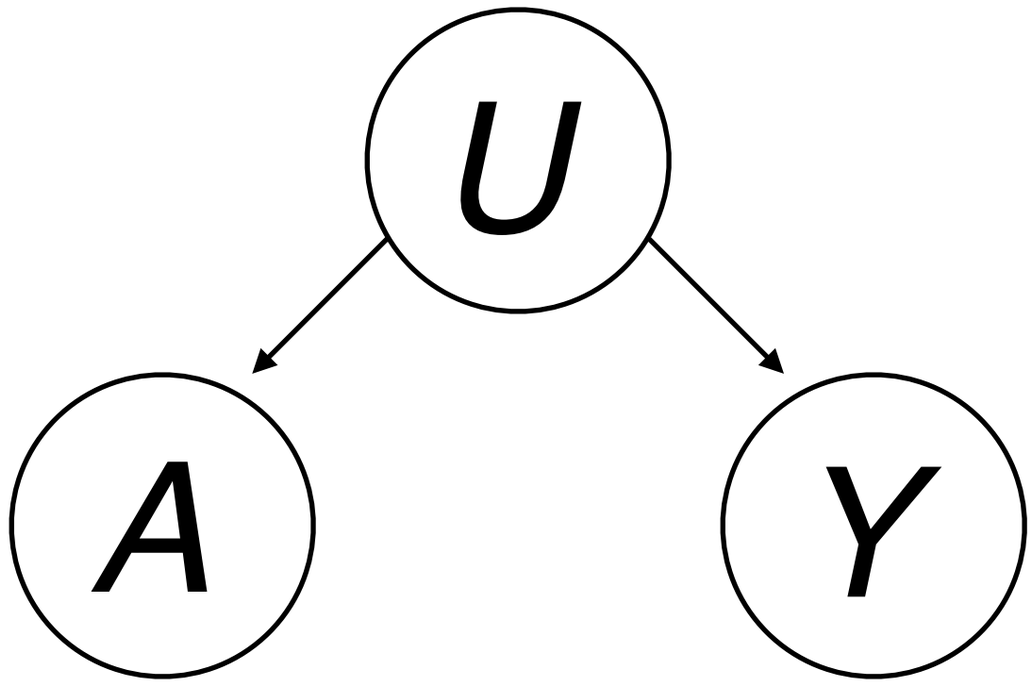
?

***A*, Treatment**:

Student attends an independent school

***Y*, Outcome**:

Student admitted to Cambridge



Confounders

***U*, Confounding variable:**

Student's family is wealthy

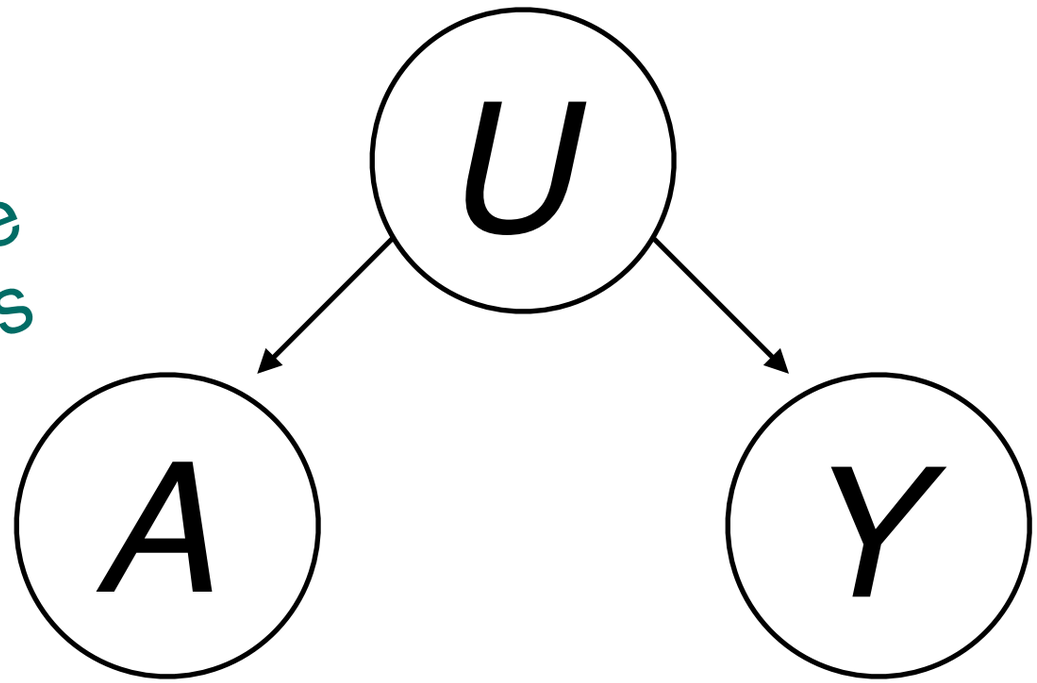
***A*, Treatment**:

Student attends an independent school

***Y*, Outcome**:

Student admitted to Cambridge

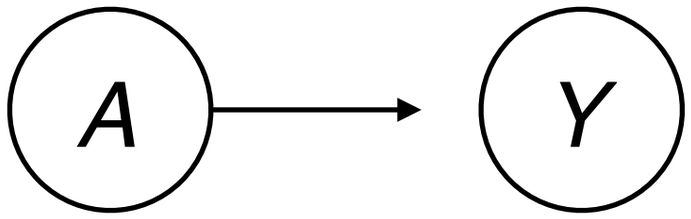
We could postulate many confounders



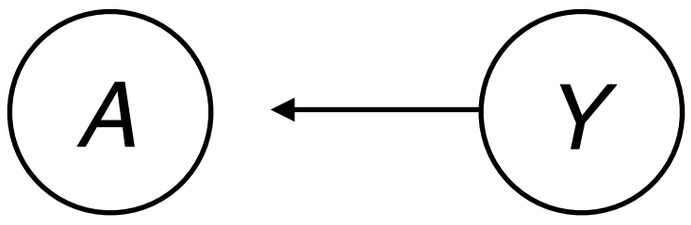
Aside: Edges in DAGs

(1)

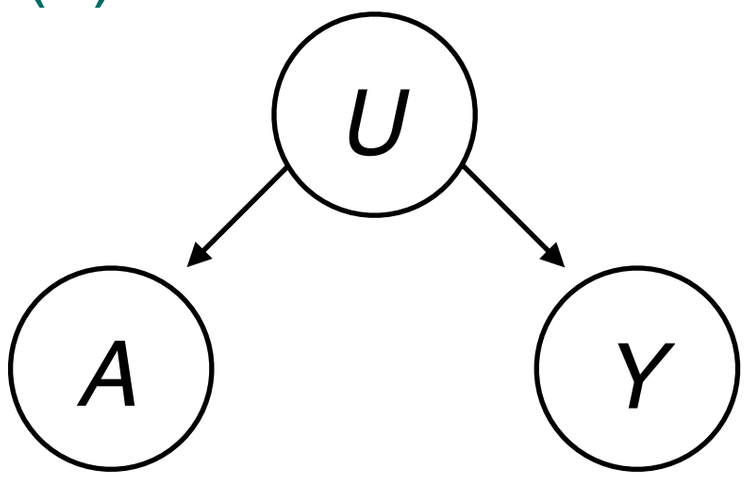
(2)



(3)



(4)

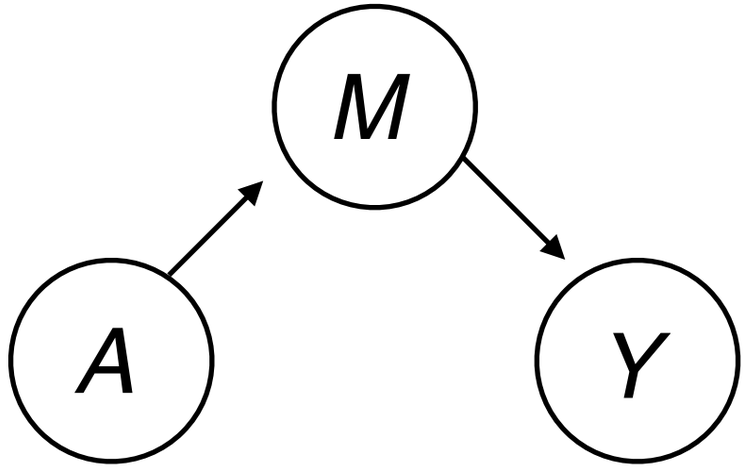


Which variables would change if we intervene on A?

Aside: Edges in DAGs as functions

- We can represent edges as functions and simulate or compute the consequences of an intervention

$$m = f(a) + \epsilon_m, \qquad y = g(m) + \epsilon_y$$



$$a \leftarrow a', \qquad m \leftarrow f(a') + \epsilon_m, \qquad y \leftarrow g(f(a') + \epsilon_m) + \epsilon_y$$

We can postulate many causal worlds

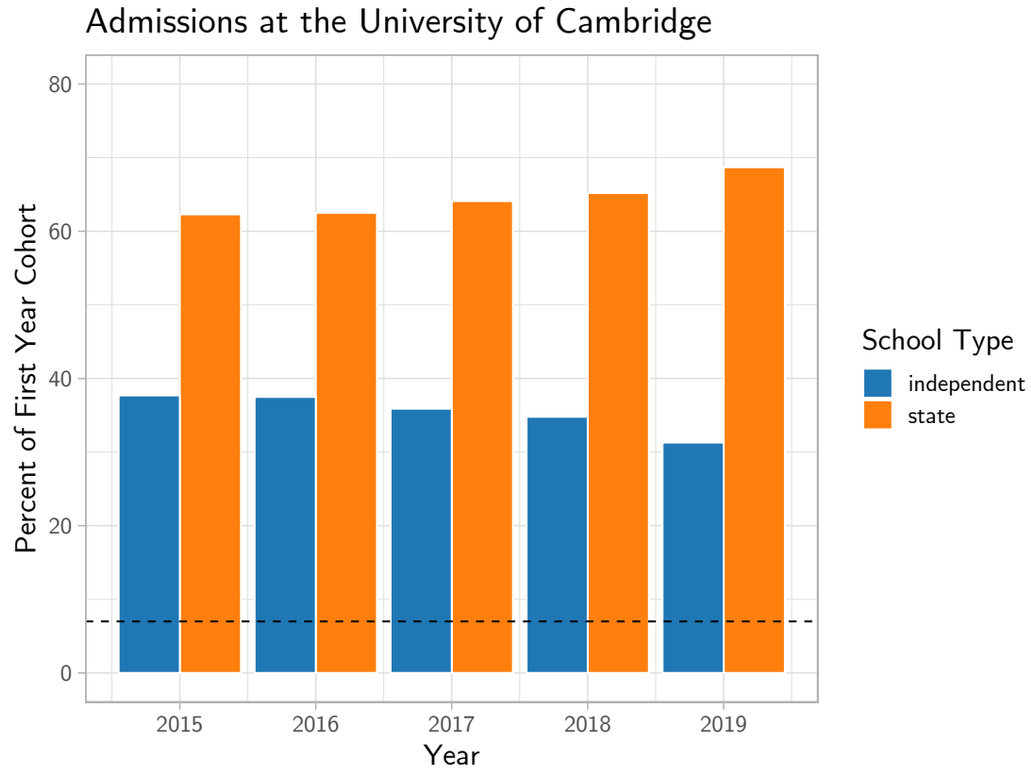
NASA

We can postulate many causal worlds

$A_{school} \rightarrow Y_{admission}$

$A_{wealth} \rightarrow Y_{admission}$

$A_{wealth} \rightarrow Y_{admission} \leftarrow A_{school}$



Among others…

Russell, Kusner, Loftus, and Silva (2017) ***When Worlds Collide***

Two important questions

1. How does the world postulated in a causal model relate to the "real world"?
   - Association vs causation
   - Omitted variables
   - Confounders, mediators
2. How does the data represent the "real world"?
   - Sampling bias
   - Measurement
   - Unobserved or unobservable features

These issues are omnipresent in causal inference. They cannot be ignored.

Aside: Where does ML fit in?

In broad terms:

- Once we have postulated a causal model(s) and we have sufficient data, we need an estimation procedure

- In a simple setting, this could be a difference in means $E[Y^{a=1}] - E[Y^{a=0}]$

- Regression, ML, etc can be useful procedures for estimating a treatment effect in more complex settings, e.g.
  - Suppose the probability of receiving a treatment is related to disease severity; we need to account for disease severity in our causal model and in our estimation procedure

- Importantly, ML (no matter how complex) does not guarantee better causal inference

How do causal models relate to fairness?

Causation and fairness

- Many ideas in (algorithmic) fairness rely on causal reasoning
- Consider admissions to Cambridge. Why might we consider it unfair?
- Student's chance of admission is lower if they attend a state school, and these circumstances are *morally arbitrary* (outside of a student's control)
- We often invoke a counterfactual when discussing fairness, e.g. COMPAS; what if the defendant had been White…

A causal framework for fairness

Typical supervised learning task:

Predict *Y* given *A* and a matrix of features *X*

$$\mathrm{p}(y \mid \mathrm{A} = a, \mathbf{X} = x)$$

- Learn a function $f(\boldsymbol{X}, \boldsymbol{A})$ from training data to predict values $\hat{Y}$
- Tune for accuracy, i.e. minimize loss function $L(Y, \hat{Y})$

Y: outcome
A: sensitive/protected attributes(s), e.g. race
X: predictors that are not sensitive/protected

We want the function to be fair with respect to A

Counterfactual fairness

Kusner, M.J., Loftus, J., Russell, C., and Silva, R., *arXiv:1703.06856v3* (2018)

A predictor $\hat{\mathbf{Y}}$ is counterfactually fair if under any context X = *x* and A = *a*,

$$P(\hat{\mathbf{Y}}_{a} \mid \mathbf{X} = x, \mathbf{A} = a) = P(\hat{\mathbf{Y}}_{a'} \mid \mathbf{X} = x, \mathbf{A} = a)$$

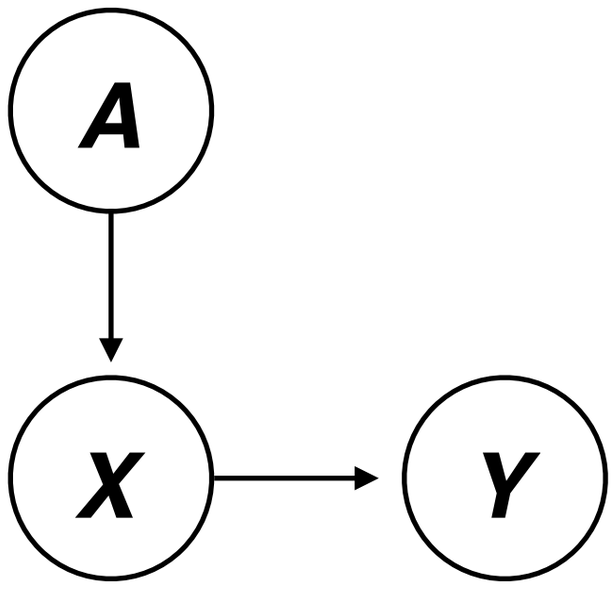
for all a'

Capital letters represent random variables. Lower case letters denote particular values of a random variable.

Counterfactual fairness

Kusner, M.J., Loftus, J., Russell, C., and Silva, R., *arXiv:1703.06856v3* (2018)

Is COMPAS counterfactually fair?

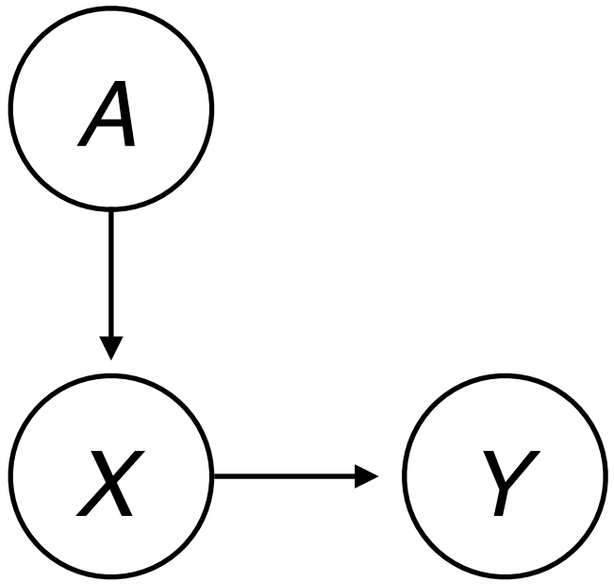


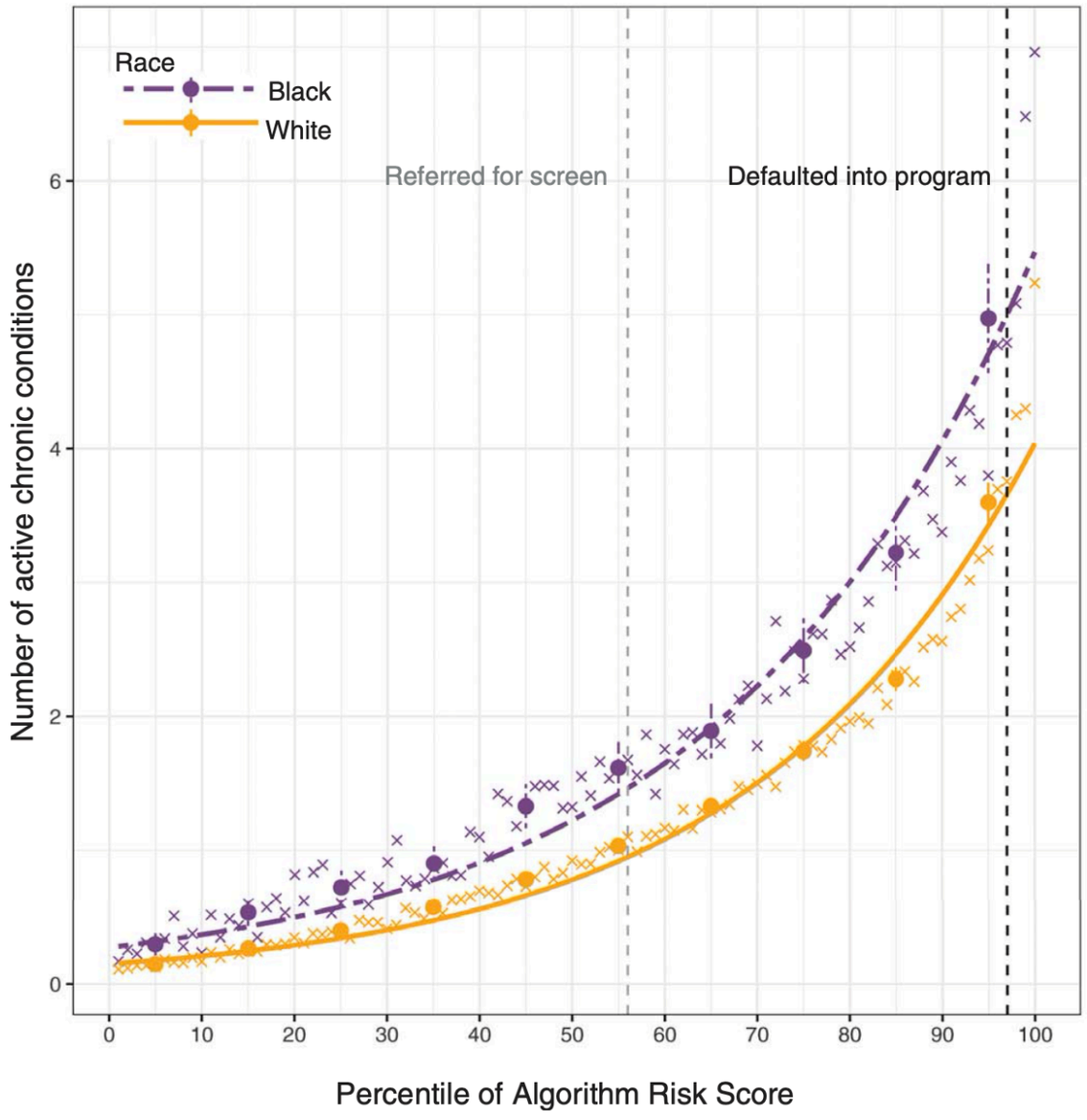
- A: protected attribute, race
- X: predictors, e.g. previous charges, contact with criminal justice system
- Y: recidivism

X is descendant (downstream) of **A**; **Y** = f(**X**, **A**)

Counterfactual fairness

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S., *Science* (2019)





- A: protected attribute, race
- X: medical expenditures
- Y: future healthcare needs

$$P(\hat{\mathbf{Y}}_{a} \mid \mathbf{X} = \$50{,}000) \neq P(\hat{\mathbf{Y}}_{a'} \mid \mathbf{X} = \$50{,}000)$$

Counterfactual fairness

Russell, C., Kusner, M.J., Loftus, J.R., and Silva, R., *NeurIPS* (2017)

- The prediction/outcome should not be a causal descendant of an individual's protected attribute*
- This is contingent on the postulated causal model representing the world as it is; what if the model is a poor representation?
- Promotes transparency: causal model must be postulated
- Idea: many (competing) worlds can be postulated

* we'll revisit this in a moment

Bloomberg's world

So 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.

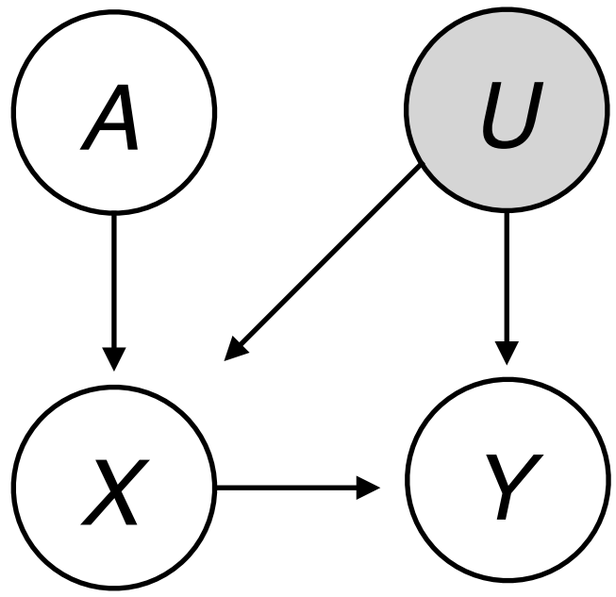
So one of the unintended consequences is people say, "Oh my God, you are arresting kids for marijuana that are all minorities." Yes, that's true. Why? Because we put all the cops in minority neighborhoods. Yes, that's true. Why do we do it? Because that's where all the crime is.

Mike Bloomberg (2015)

Bloomberg's world

To make a thief, make an owner; to create crime, create laws.

Ursula K. Le Guin, The Dispossessed

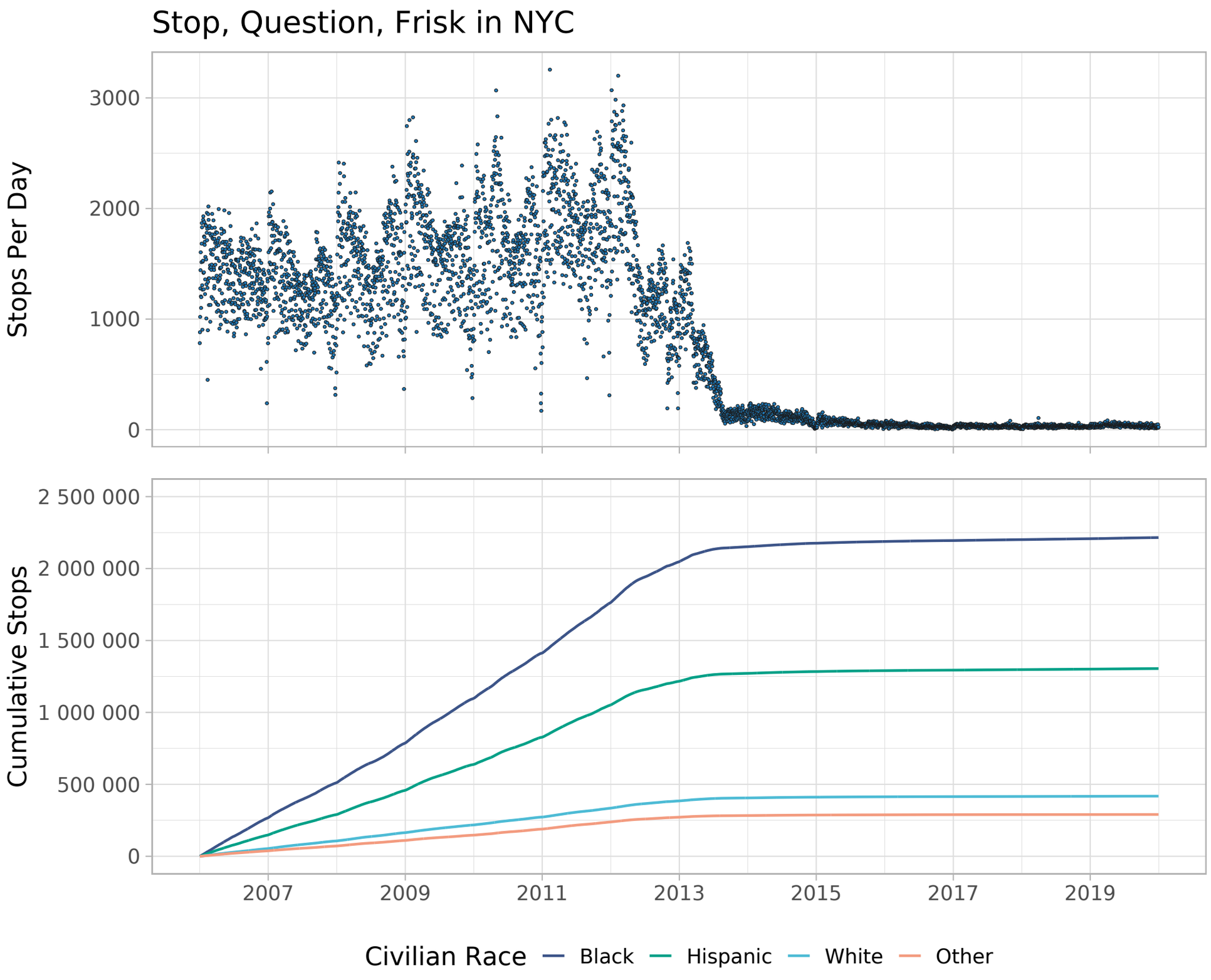


- A: racial composition of neighborhood
- X: police deployment rate
- U: other factors influencing enforcement patterns and charge rate
- Y: criminal charge

Bloomberg argued the city should determine X based on Y, encoding the targeted policing of minorities.

The strategic subjects list applied the same logic in Chicago and conspicuously excluded race from the algorithm.

The consequences



What does it mean for race to be a causal factor?

Sen, M. and Wasow, O. *Annual Review of Political Science* (2016)

- "No causation without manipulation" (Holland, 1986)
- What does it mean to "intervene" on race, age or other "immutable characteristics"?
- "Race as a bundle of sticks;" a composite of different factors
- E.g. Sweeney, L. "Discrimination in Online Ad Delivery" (2013) studied the effects of a racial cue (person's name) on ad delivery

What does it mean for race to be a causal factor?

Kohler-Hausmann. I., *Northwestern University Law Review* (2019)

But central to all counterfactual causal accounts of racial discrimination is the notion that there is a solid state race in units (individuals, neighborhoods, etc.), an objective fact about the units that can be isolated after stripping away all confounders. For something to be a treatment, there must be a way to pick out what *the-treatment* is—distinct and apart from all of the things that are *not-the-treatment* so that we are sure we are talking about identical units that differ only on the-treatment. If we cannot pick apart *the-treatment* from *not-the-treatment*, then we are not estimating a treatment effect of race and race alone when we compare the outcomes of candidates with some list of similar credentials and signals for different racial categories. We are doing something else.

Isla Kohler-Hausmann [emphasis added]

Causal descendants; the case of Berkeley admissions

Bickel, E. A., Hammel, J., and O'Connell, W., *Science* (1975)

- An early paper on fairness studied graduate admissions at Berkeley
- Women applicants were admitted at lower rates
- However, women applied to more competitive departments, on average
- At the department-level, women were slightly favored in admissions

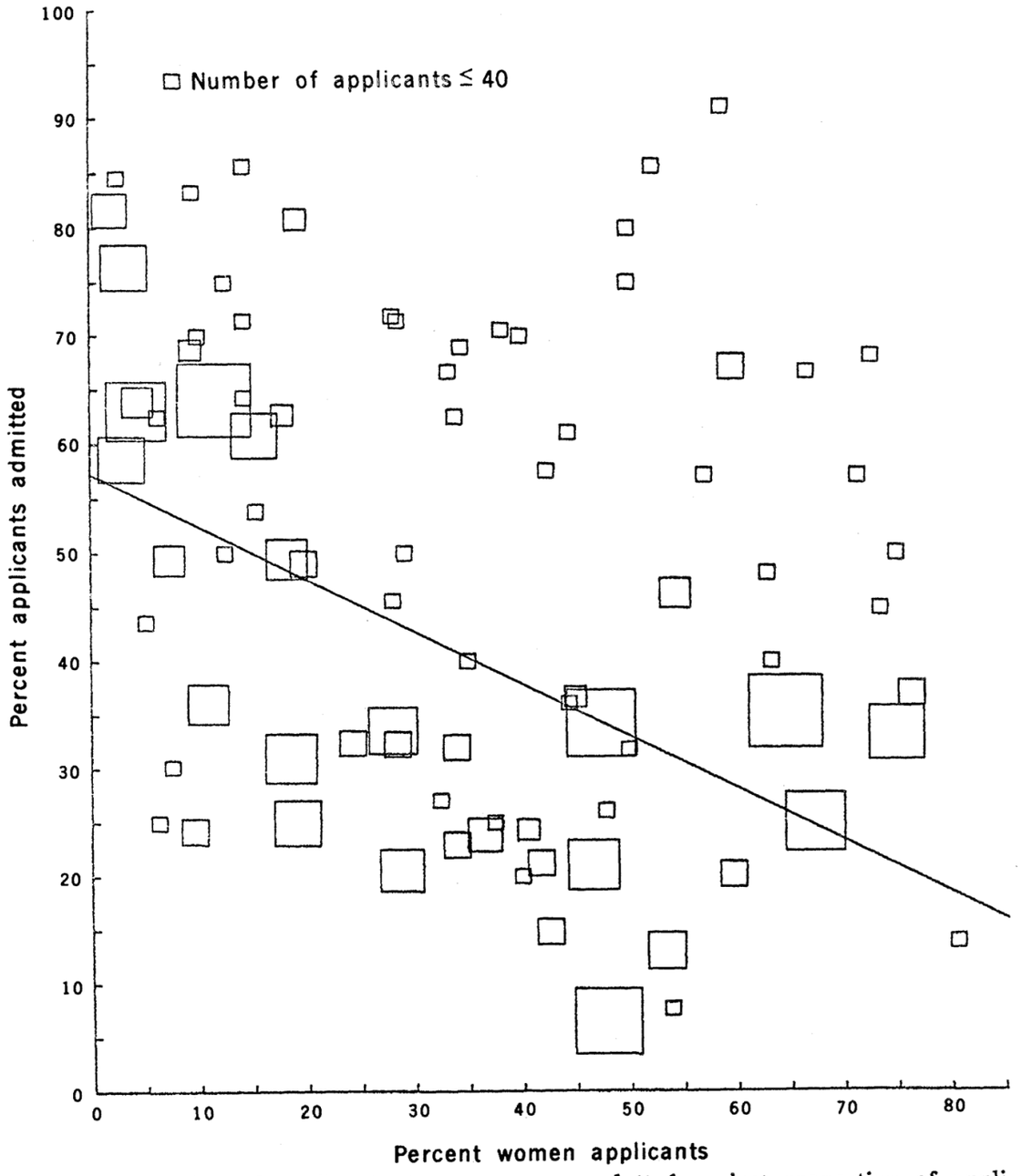


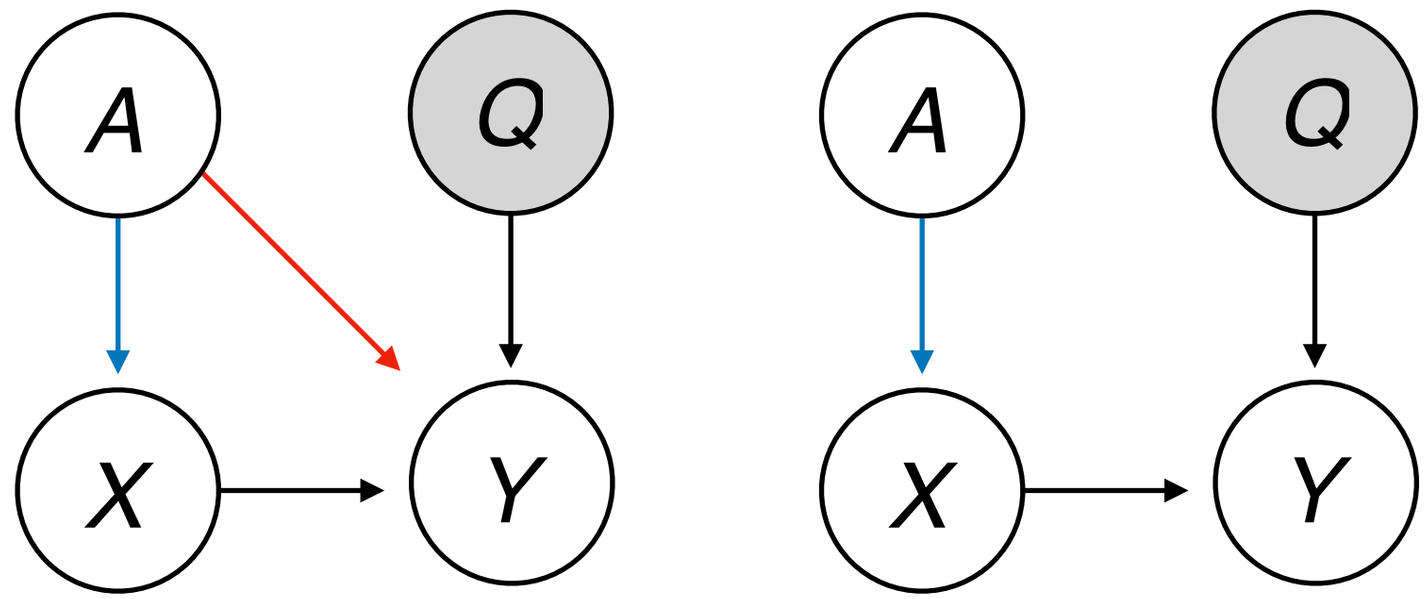
Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

Causal descendants; the case of Berkeley admissions

Chiappa, S. and Gillam, T.P.S. *arXiv:1802.08139* (2018)

Developed idea of “path-specific counterfactual fairness”

- A: gender
- X: department choice
- Q: qualification
- Y: admission



Unfair

Fair?

- Fair at what decision point? For which decision maker?
- Berkeley (the vendor) might say “you can’t expect us to resolve sexism in broader society!”

tem. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.

τ-Controlled counterfactual privilege

Suppose a vendor wants to implement an intervention z.
Proposes constraining "counterfactual privilege" such that:

$$E[\hat{Y}_i(a_i, \mathbf{z})] - E[\hat{Y}_i(a_i', \mathbf{z})] \leq \tau$$

- Exclude intervention assignments that allow an individual *i* to become more than τ units better off in expectation due to the interaction of z and A
- Anything $\geq \tau$ is considered unfair privilege

Revisiting other fairness definitions under a causal framework

Kusner, M.J., Loftus, J., Russell, C., and Silva, R. arXiv:1703.06856v3 (2018)

Demographic parity

A predictor $\hat{Y}$ satisfies demographic parity (i.e. equality of outcomes) if:

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

- Predictions are independent of **A**

Equality of opportunity

A predictor $\hat{Y}$ satisfies equality of opportunity if:

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

- Predictions are independent of **A**
- But only among individuals above “threshold” for desirable outcome

Fairness through unawareness

A predictor $\hat{Y}$ satisfies fairness through unawareness (i.e. equal treatment) if:

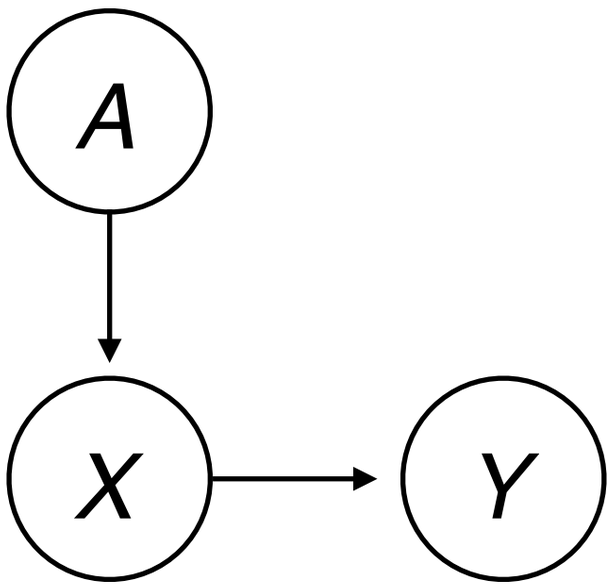
$$\mathrm{Y} = f(\mathbf{X})$$

- Predictions do not use **A** at all

Chief Justice Roberts

The way to stop discrimination on the basis of race is to stop discriminating on the basis of race.

Chief Justice John Roberts (2007)

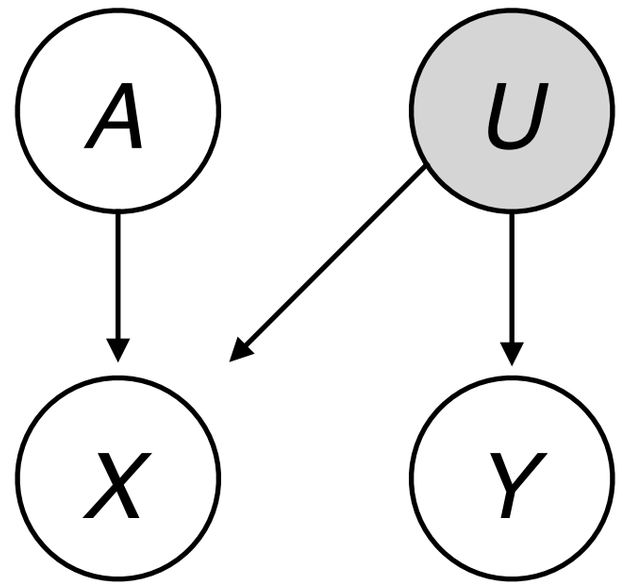


- Fairness through unawareness; “equal treatment”
- But A cannot be disentangled from X
- This is a common pattern of counterfactual unfairness

Credit: **Josh Loftus**

Chief Justice Roberts' view can introduce unfairness

Kusner, M.J., Loftus, J., Russell, C., and Silva, R. arXiv:1703.06856v3 (2018)



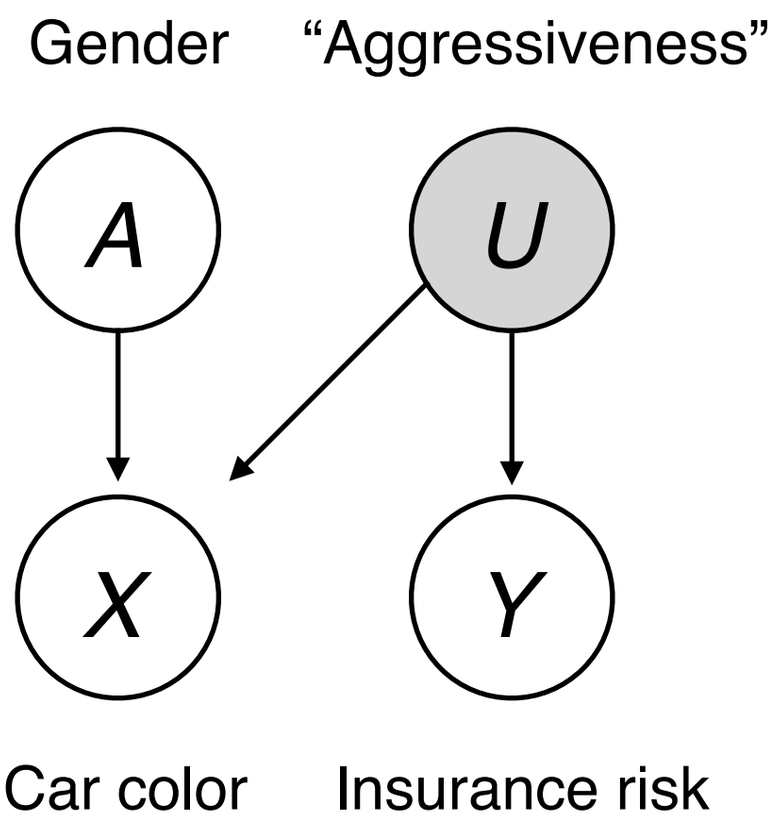
Note that X doesn't cause Y in this model!

- The variable **X** is a descendant of **U**
- **X** is also a descendant of **A**, i.e. **X** = *f*(**A**, **U**)
- If we use **X** to predict Y, we are using **U** and **A**

Credit: **Josh Loftus**

Chief Justice Roberts' view can introduce unfairness

Kusner, M.J., Loftus, J., Russell, C., and Silva, R. arXiv:1703.06856v3 (2018)



- "Fairness through unawareness" introduces unfairness
- **X** = *f*(**A**, **U**); by ignoring A we cannot adjust for its influence on X
- This would be counterfactually unfair
  - $P(\hat{Y}_{A\leftarrow a} \mid X = \text{red}) \neq P(\hat{Y}_{A\leftarrow a'} \mid X = \text{red})$

Individual fairness

A predictor $\hat{Y}$ satisfies individual fairness if:

$$Y(\mathbf{A}_i, \mathbf{X}_i) \approx Y(\mathbf{A}_j, \mathbf{X}_j)$$

Given a metric *d* that measures the similarity of individuals *i* and *j* in **X**.

Does this simply move the goalposts?

My view: no.

- Causal reasoning and counterfactual fairness clarifies what is at stake in a particular data science task
- Enhances transparency by requiring the specification of a causal model

However,

- It is a framework for assessing and enhancing fairness given causal model(s) + data, not a “solution to fairness”
- Underlying moral and ethical concerns around risk-assessment tools and other ML tasks do not go away, nor do problems of data bias
- We must think carefully about what a “counterfactual” means in terms of race and other sensitive/protected attributes