Fairness in Machine Learning and Its Causal Aspects

Ricardo Silva

University College London and The Alan Turing Institute

ricardo@stats.ucl.ac.uk

With joint work by Matt Kusner, Joshua Loftus and Chris Russell (Turing Institute)

Disclaimer

I am not a lawyer or philosopher. I am a data scientist by training.

I do not claim anything I say has a legal basis or a scholarly backing on ethics research.

Much of the work in this area is at the frontier of multidisciplinary research.

Machine Learning and Decision Making

- Machine learning is good old statistical science with a fancy hat.

- Granted, having a different motivation (Artificial Intelligence) does have a practical implication on how we do data analysis.

- In particular, machine learning does come with one major cultural baggage: an emphasis on (semi)autonomous decision making
    - What is a more down-to-earth name for AI? *Autonomous systems*.

computer says no

Implications

- The pipeline from data to decisions is less supervised by humans.
- There is a risk of paying less attention to biases *which may be in the data themselves*.
- There is a risk of making mistakes due to *lack of common sense* in a statistical/computer model.
- This has led to movements requesting more transparency from data-driven decisions.

Fairness

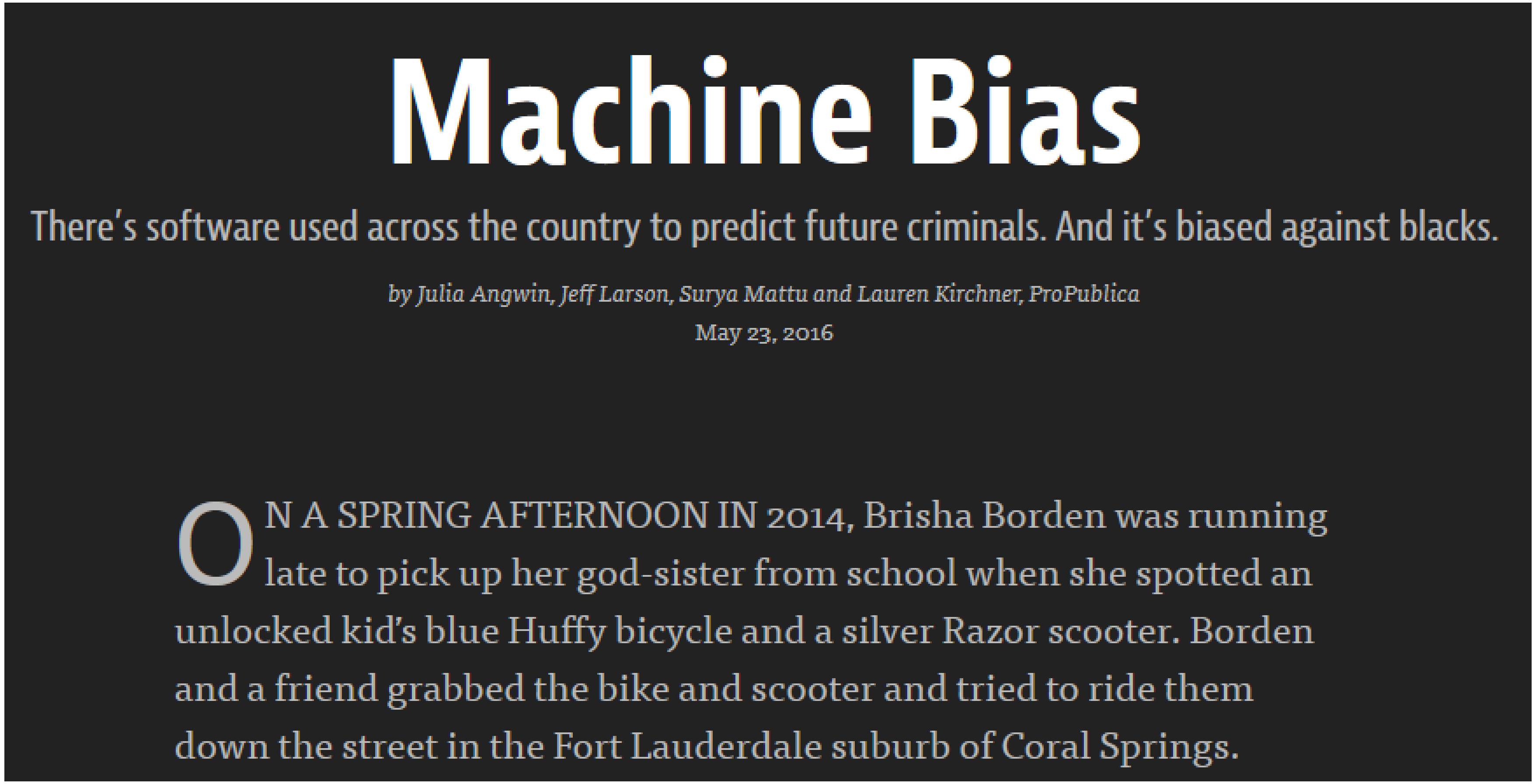
- I will not provide an universal definition of fairness.

- Instead, imagine we set in stone that some personal attributes are *protected*: that our decisions should not *discriminate* people on those attributes.

- But how do we define what discrimination (“unfairness”) is? The possibly confusing fact is that there is no unique way of defining it.

- I will throw my hat in the ring at some point in this talk.

Case Study:
The ProPublica/COMPAS Imbroglio



Machine Bias

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

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

What goes in COMPAS

“Correctional Offender Management Profiling for Alternative Sanctions”

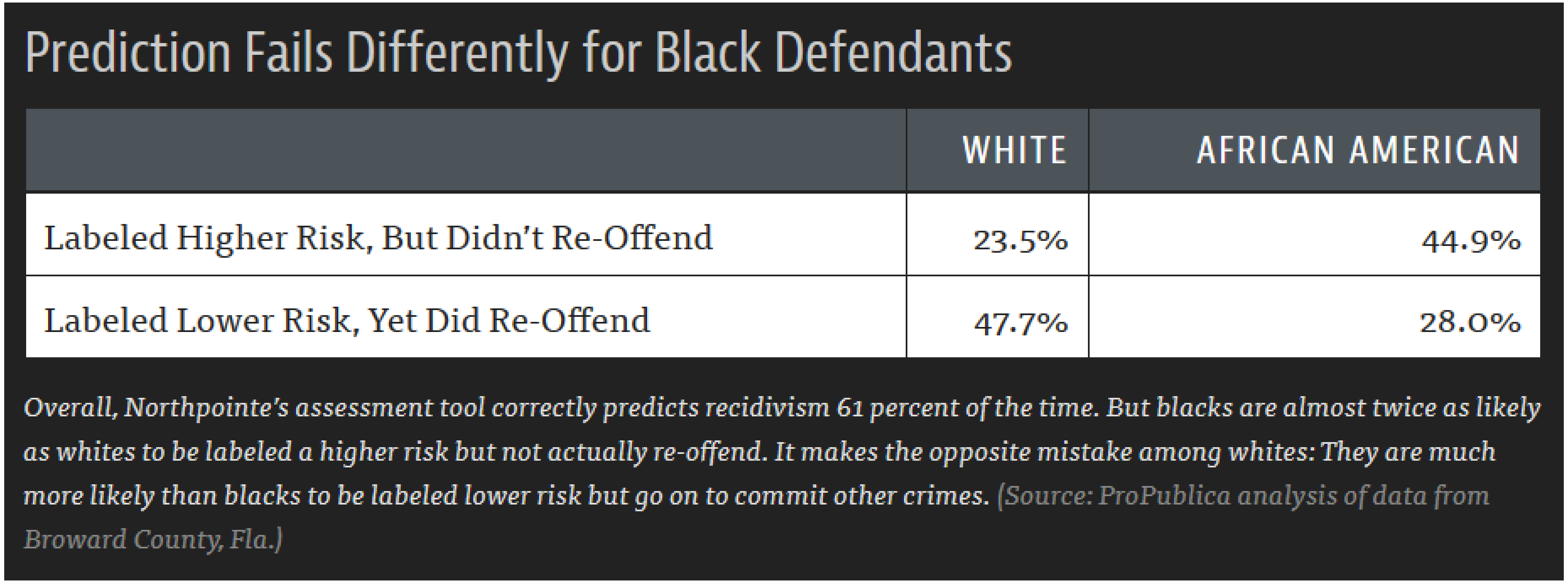
- Goal: risk assessment of recidivism.
- Means: a predictive model (“machine learning”, if you prefer the term) that gives you a score based on personal attributes such as number of previous offenses and substance abuse.
  - Its magic sauce is not public knowledge.
- This is *not* meant to be an autonomous system. However, there is evidence judges have used COMPAS scores in their sentences.

Validation

- Out-of-sample accuracy of 2-year recidivism events.
    - Overall: 68% of the time correct.
    - Among blacks: 67%.
    - Among whites: 69%.
    - No statistical evidence of differential treatment.

- However…

False Positives/False Negatives



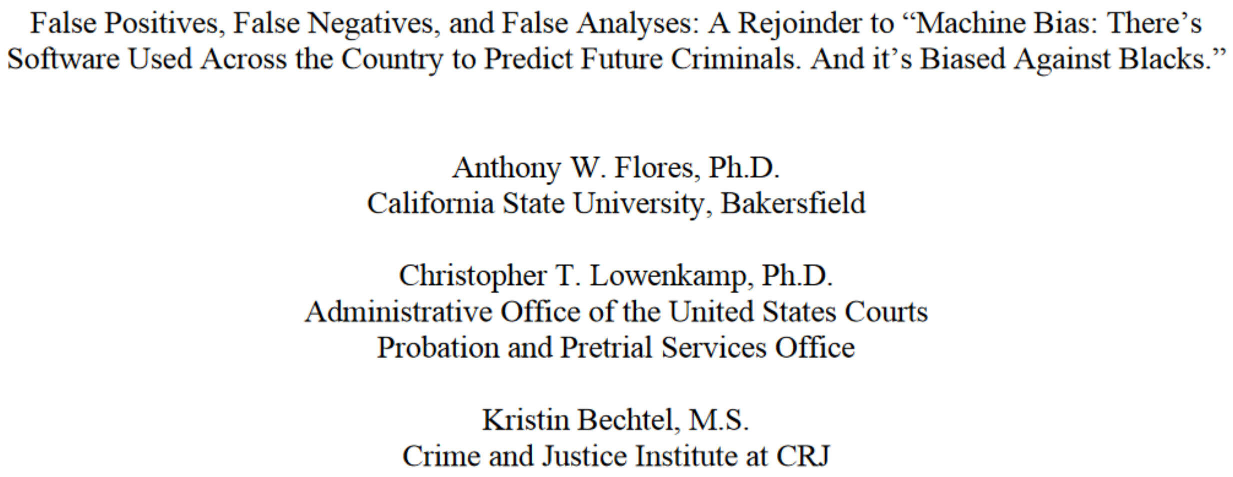
Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

The Rebuttal



False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks."

Anthony W. Flores, Ph.D.
California State University, Bakersfield

Christopher T. Lowenkamp, Ph.D.
Administrative Office of the United States Courts
Probation and Pretrial Services Office

Kristin Bechtel, M.S.
Crime and Justice Institute at CRJ

"The issue that is no longer up for debate is that (models) predict outcomes more strongly and accurately than professional judgment alone."

"(Models) are intended to inform objective decision-making."

http://www.crj.org/cji/entry/false-positives-false-negatives-and-false-analyses-a-rejoinder

“Test-fairness” in COMPAS

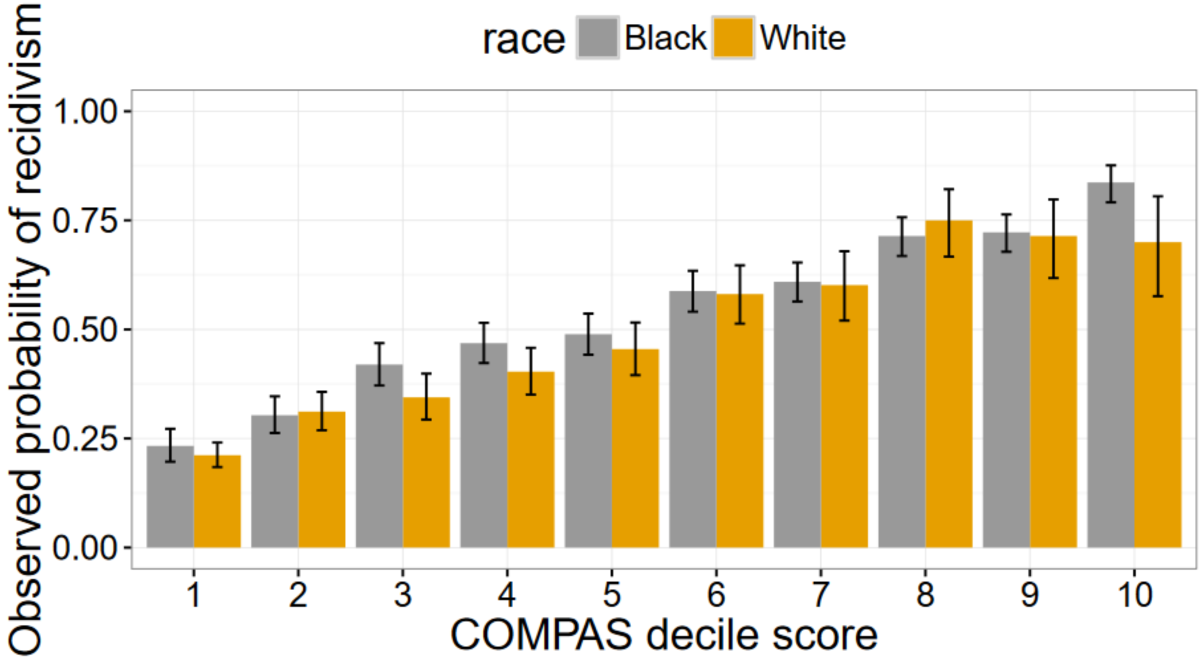


Figure 1: Plot shows $\mathbb{P}(Y = 1 \mid S = s, R)$ for the COMPAS decile score, with $R \in \{\text{Black}, \text{White}\}$. Error bars represent 95% confidence intervals.

A. Chouldechova, 2017. “Fair prediction with disparate impact”. https://arxiv.org/abs/1610.07524

Explaining the Disagreement

- ProPublica was looking at false positives/false negatives:

$$P(\textit{Low risk score} \mid \textit{Recidivism, White defendant}) > P(\textit{Low risk score} \mid \textit{Recidivism, Black defendant})$$

$$P(\textit{High risk score} \mid \textit{No recidivism, White}) < P(\textit{High risk score} \mid \textit{No recidivism, Black})$$

- The rebuttal was looking at “test fairness”:

$$P(\textit{Recidivism} \mid \textit{Low risk, White}) \cong P(\textit{Recidivism} \mid \textit{Low risk, Black})$$

$$P(\textit{Recidivism} \mid \textit{High risk, White}) \cong P(\textit{Recidivism} \mid \textit{High risk, Black})$$

They are both true in COMPAS!

“Doomed if we do, doomed if we don’t”

- Can’t we have a score with no discrepancy on all of these measures?

- NO! Except for pathological cases, *this will only be possible if recidivism does not vary by race* (regardless of score).
  - The argument is mathematical but follows from basic probability.

- But it is a matter of fact that, currently, recidivism *does* vary by race.

- So we cannot have test-fairness and the same rates of false positive/false negative regardless of race.

What is the Lesson?

- We can define “fairness” according to desirable particular statistical properties: test-fairness, balanced false positives/false negatives, etc.

- But these definitions may not be self-evident. In some contexts, they may sound “right”, and some other contexts, “wrong”, and not all satisfiable at the same time!

- What we propose is to make explicit causal assumptions about the world.

- The drawback of our proposal is that it relies on untestable causal assumptions. This is not a minor issue. However, I believe we need to bite this bullet.

Counterfactual Fairness

- If we have some protected attribute like race, and a decision such as length of sentence, then our decision satisfies *counterfactual fairness* if

  > “had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same”

- This explicitly asks for the causal concept of counterfactuals.

Wait, What?

- What do you mean by “had race been different”?

- It is indeed not obvious what a counterfactual such as “had I been a native American” means.

- For what follows, assume that counterfactuals are well-defined according to the context of each specific problem.
  - For instance, “race” meaning “race perception”, which can be controlled in some sense (information in a job application, including surrogates like name, cultural background, etc.)

A Toy Example: The Red Car

- Say you are in charge of pricing car insurance.

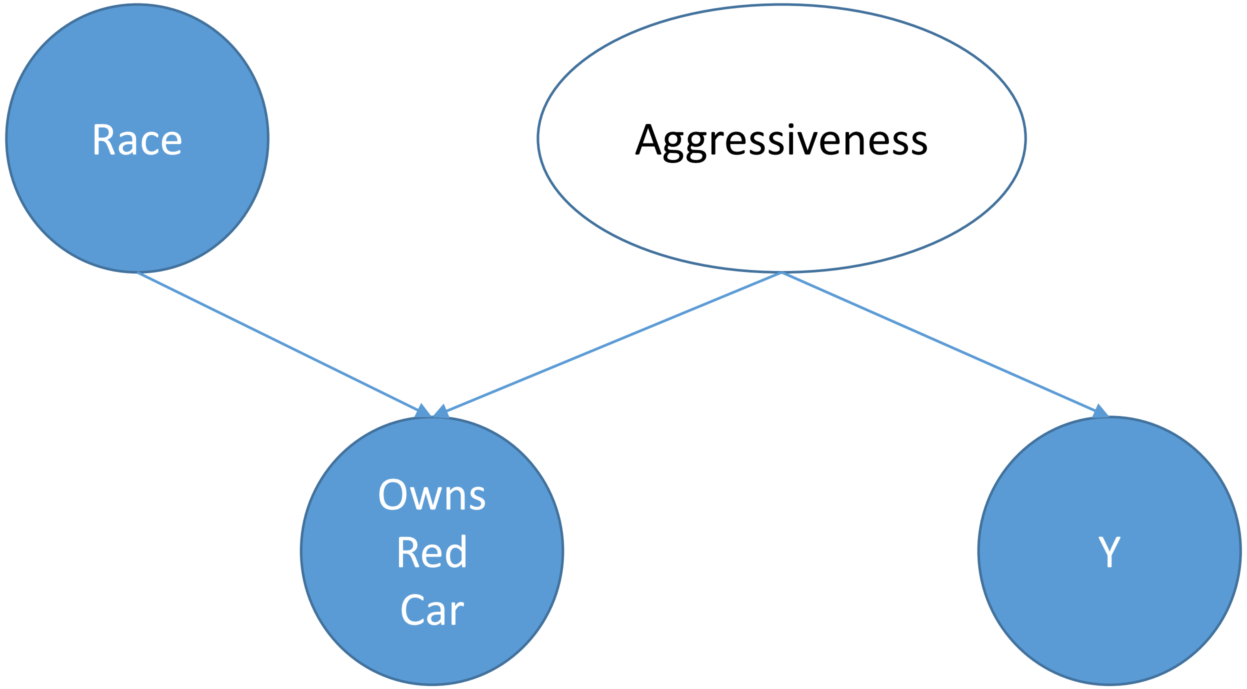
- Your database contains only three variables:
  - Outcome (“*Y*”): the rate by which a person provokes traffic accidents
  - Attributes: race, and whether the person owns or not a red car

- A concept called *fairness through unawareness* says: as long as I don’t use race in my predictions, I am being far.

- What could go wrong here?

A Toy Example: The Red Car

- Postulated causal structure



“Agressiveness” here is a *latent variable*, data we cannot observe directly.

Pearl, Glymour and Jewell (2016) *Causal Inference in Statistics: a Primer*. Cambridge University Press.

Fairness through Unawareness?

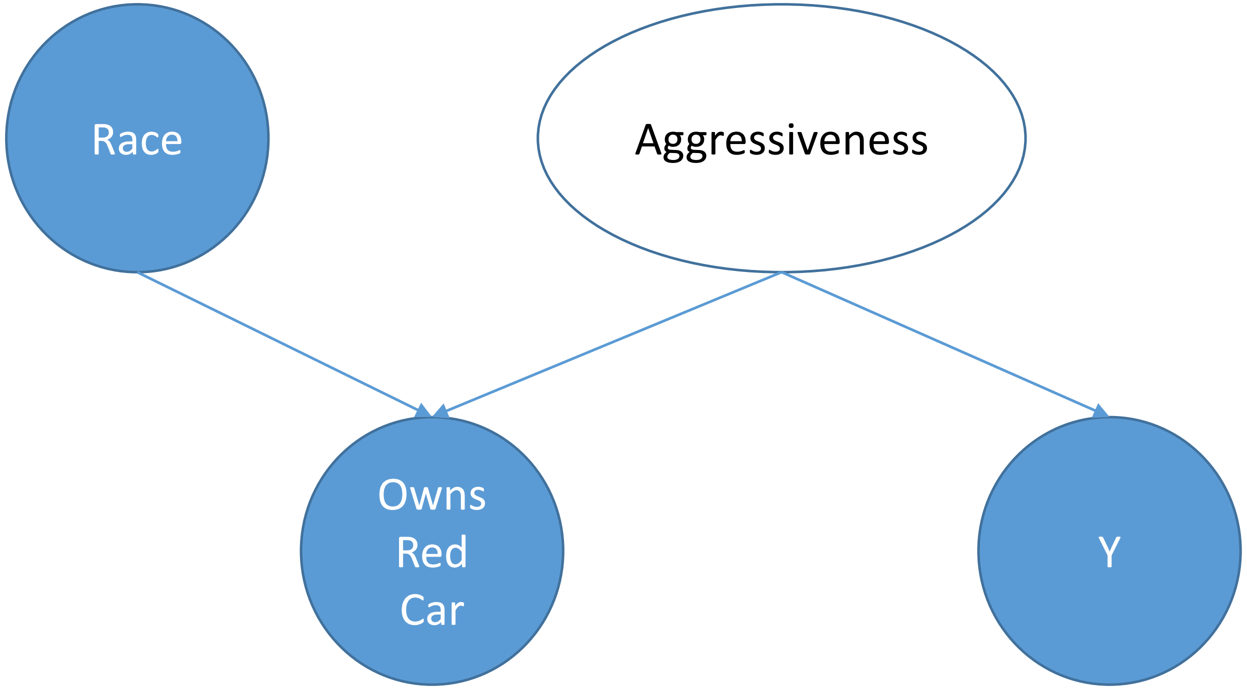
- Let’s say the model behind this structure is linear.

- What do you think the linear regression of *Y* on “Owns Red Car” will give to you? Remember, this is motivated by fairness through unawareness.

- What do you think the linear regression of Y on both “Red Car” and race will give to you?

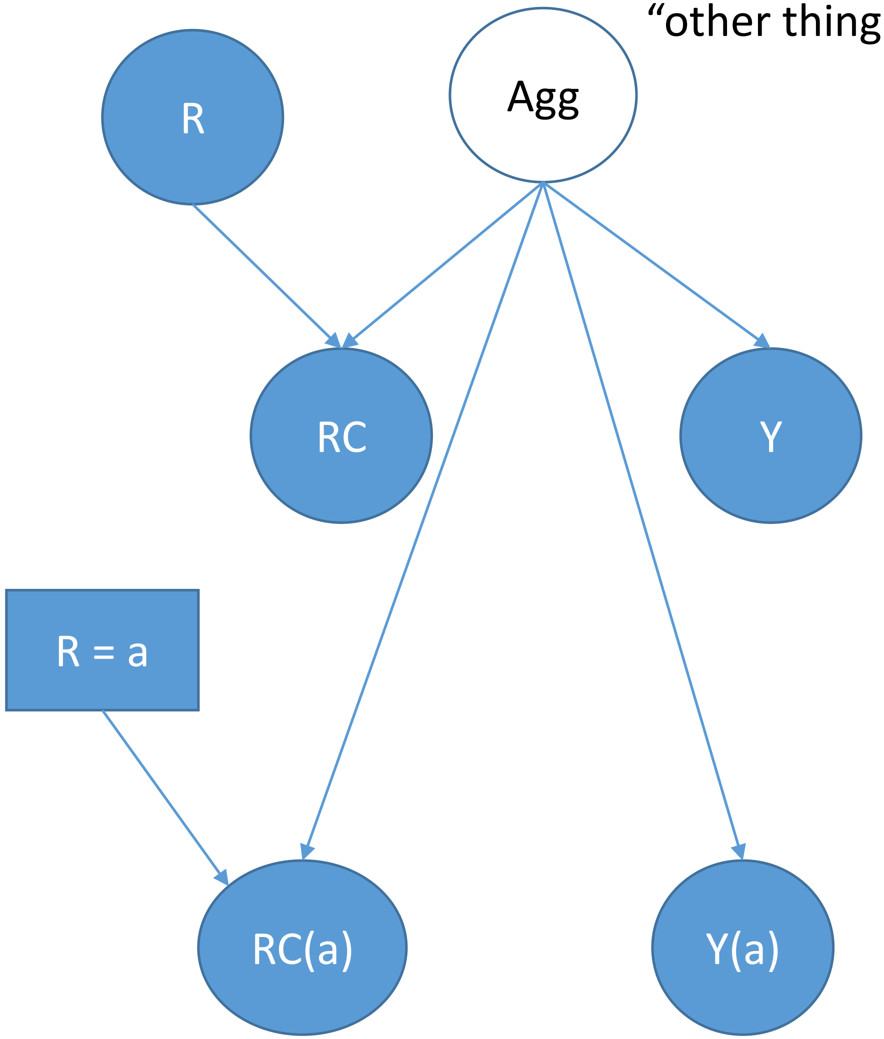
Counterfactual Analysis, in a Picture

- The “factual” world for a given individual



Pearl, Glymour and Jewell (2016) *Causal Inference in Statistics: a Primer*. Cambridge University Press.

Counterfactual Analysis, in a Picture



“other things being equal”

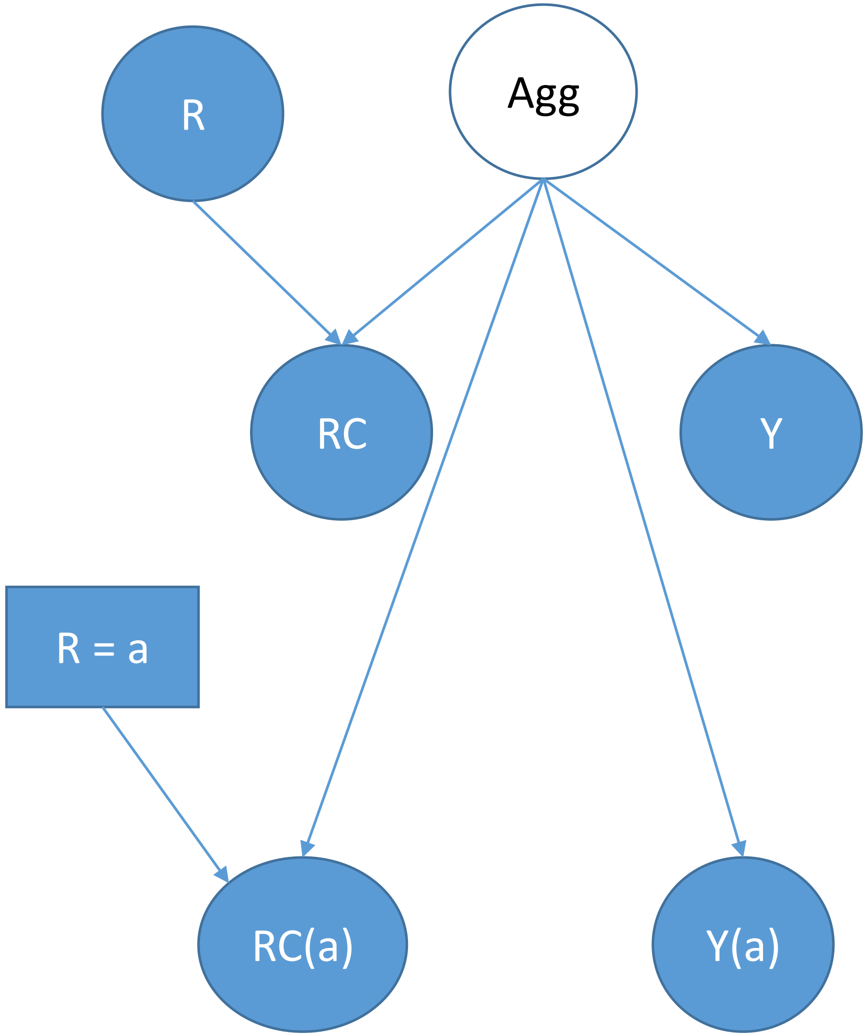
“Factual world”

“Counterfactual world”

- Aggressiveness is “exogenous”: no causes among the observed variables.
- Setting “R = a” is contrary to the fact (a counterfactual).
- The new information propagates to all observed variables.
- Y(a) = Y, but this is not important for the moment.

Counterfactual Analysis, in (Toy) Equations

- Say, *Red Car* = *Race* + *Aggressiveness*
  - *Race* as either 0 or 1, for the sake of an example.
  - *Red Car* as the proportion of cars owned which are red.
  - *Aggressiveness* as a psychometric score
- We can deduce aggressiveness levels, *Aggressiveness* = *Red Car* – *Race*.
- Now, keep it constant, change *Race* to new values like *a* or *a'*.
- Then *Red Car*(*a*) = *a* + *Aggressiveness* etc.

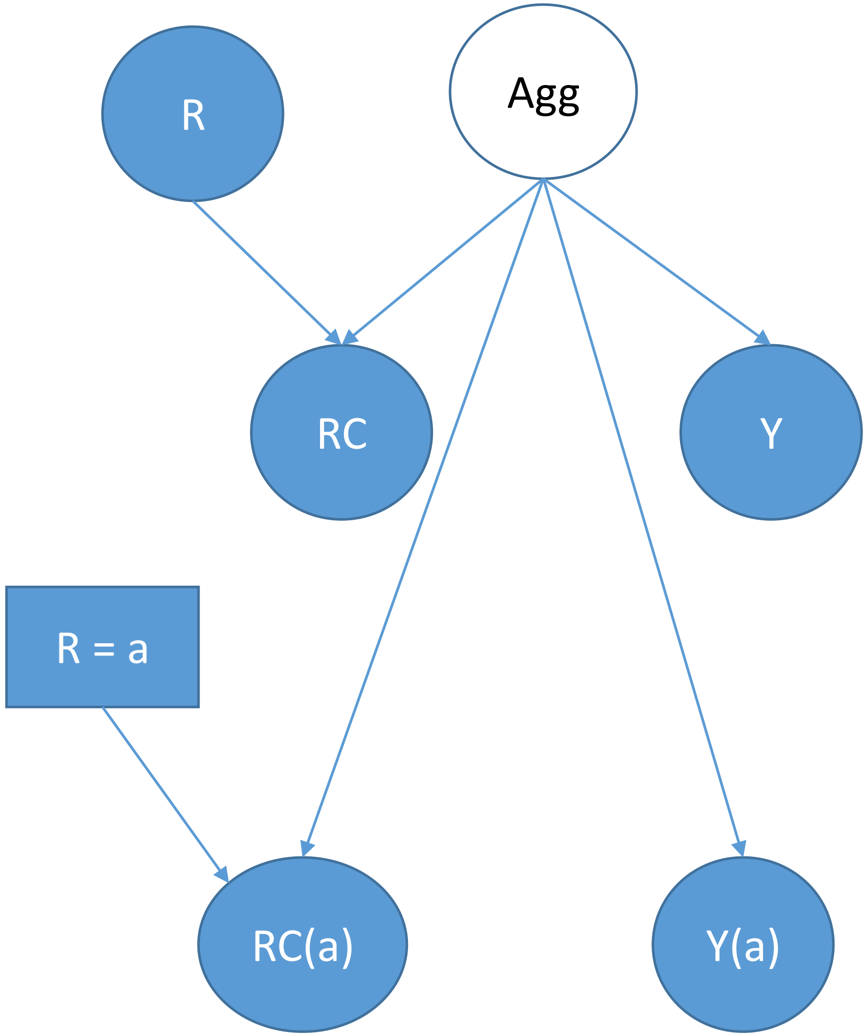


Counterfactual Analysis, in (Toy) Equations

- Finally, if we regress our outcome *Y* on *Red Car*, we get a *predictor S*.

- But *S* depends on *Red Car*, so *S*(*a*) and *S*(*a’*) are different.

- This is not counterfactually fair: had race been different, other things being equal, decisions based on *S* would be different.

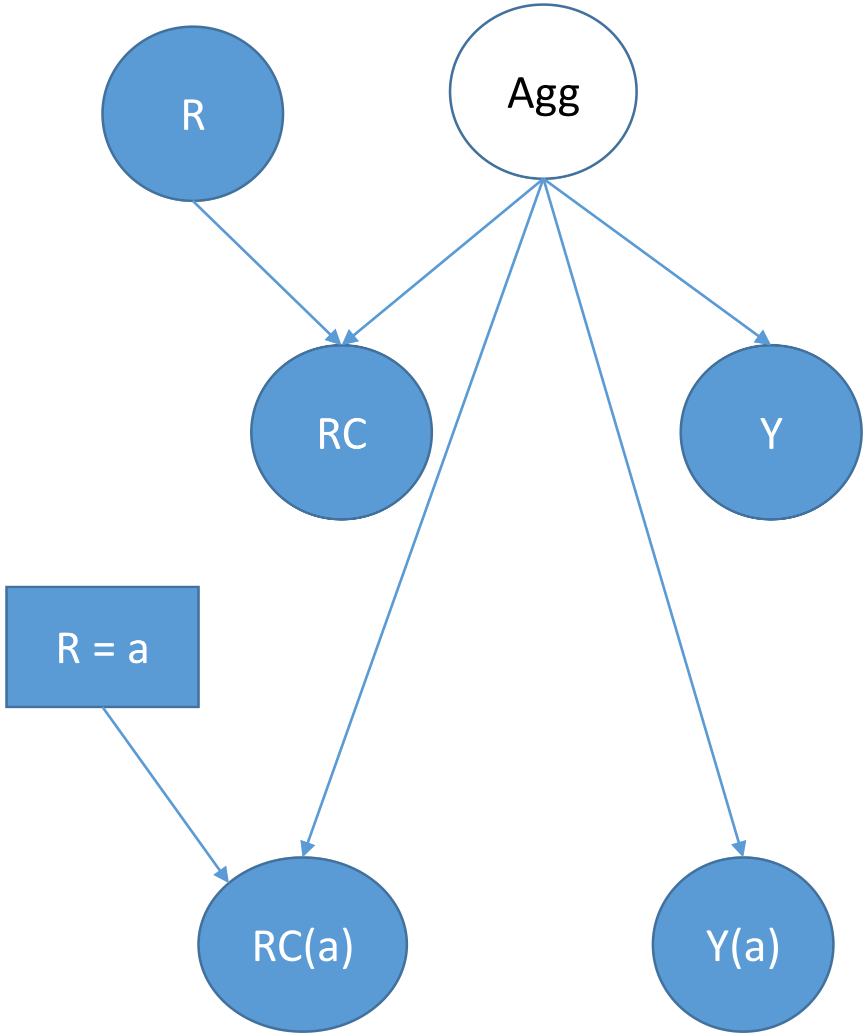


Counterfactual Analysis, in (Toy) Equations

- What to do instead?

- If we regress *Y* on *Agressiveness* only, than our resulting *S* is invariant across different levels of *Race*.

- Our evidence for *Agressiveness* is based on the “factuals”. So, contrary to fairness through unawareness, we *are* using *Race* information here.

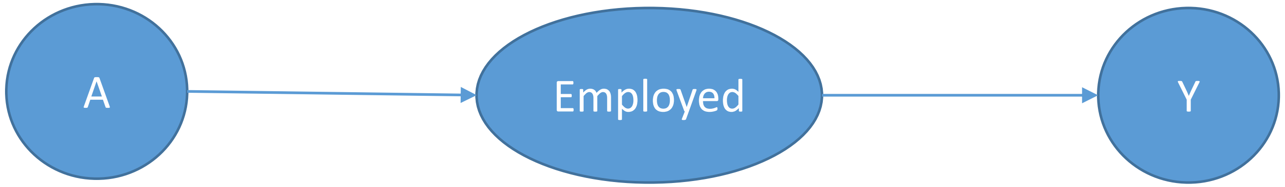


Hold On

- So defining *S* this way still depends on *Race*? In one sense, yes.
  - Because *Aggressiveness* is deduced using *Race* and *Red Car*.
- How is this fair? The comparison is done *within* counterfactual versions of the same individual, *not* across individuals.
- Contrast:
  - Look at people with the same value for *Red Car*. Fixing this, compare different people with *Race = a* and *Race = a'*. *S will be different* for these two groups of people, since *Aggressiveness* is different. **Here, *Red Car remains the same, Aggressiveness changes*.**
  - Look at a single person: deduce *Agressiveness* from *Race* and *Red Car*. Do the thought experiment of what would have happened to *S* had *Race* been different, other things being equal. *S* remains unchanged. **Here, *Aggressiveness remains the same, Red Car changes*.**

Another Toy Example:
Credit in an Unfair World

- Say *A* is an indicator of a demographic factor (race, etc.) and *Y* is whether a person will default on a loan.
- Postulate the following model of the world:



- Suppose I am able to have *S* = *Y* by whatever means. Perfect prediction. Is this counterfactually fair if *A* is a “protected attribute”?

Another Toy Example: Credit in an Unfair World

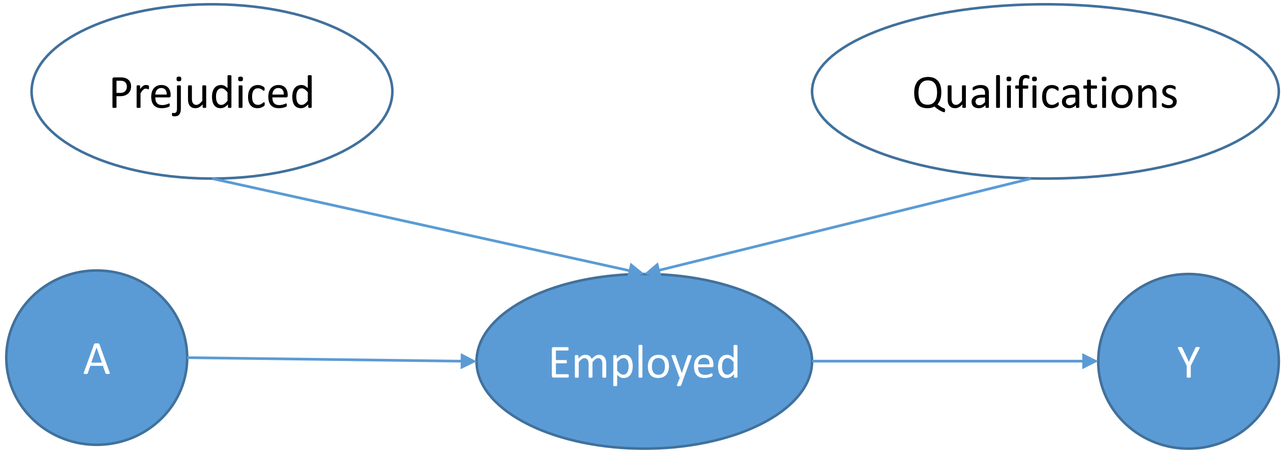
- No, it is not counterfactually fair!

- Does this make sense? Don’t be afraid of circularities:
  - If it doesn’t, then maybe you should not consider *A* to be protected.
  - If it does, then the world that generates *Y* is itself unfair. You must not aim at perfect predictions of *Y*!
  - Choose your primitives, but causal assumptions can help you with that.

- How would this be possible? Postulate finer-grained causal model.

Another Toy Example:
Credit in an Unfair World, Take 2

- *Prejudiced* is a latent variable indicating whether person offering job is employed.
- *Qualifications* is a latent variable representing the qualifications of the individual.
- What about basing our decisions on *Qualifications* only?



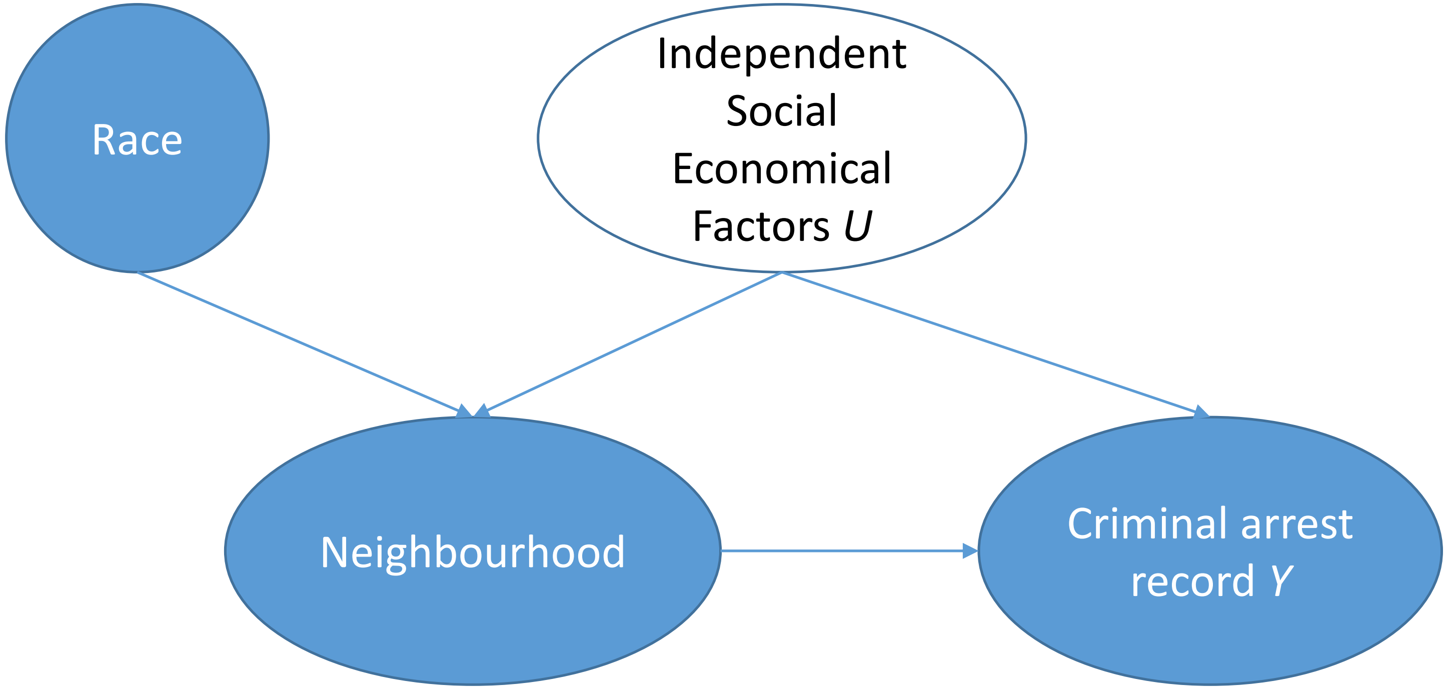
Fairness vs. Accuracy

- We can’t have both, unconstrained.

- We can build a predictor *S* that explain the “fair” components of defaulting on a loan. Accuracy will suffer, but perhaps this should be a price that needs to be paid.

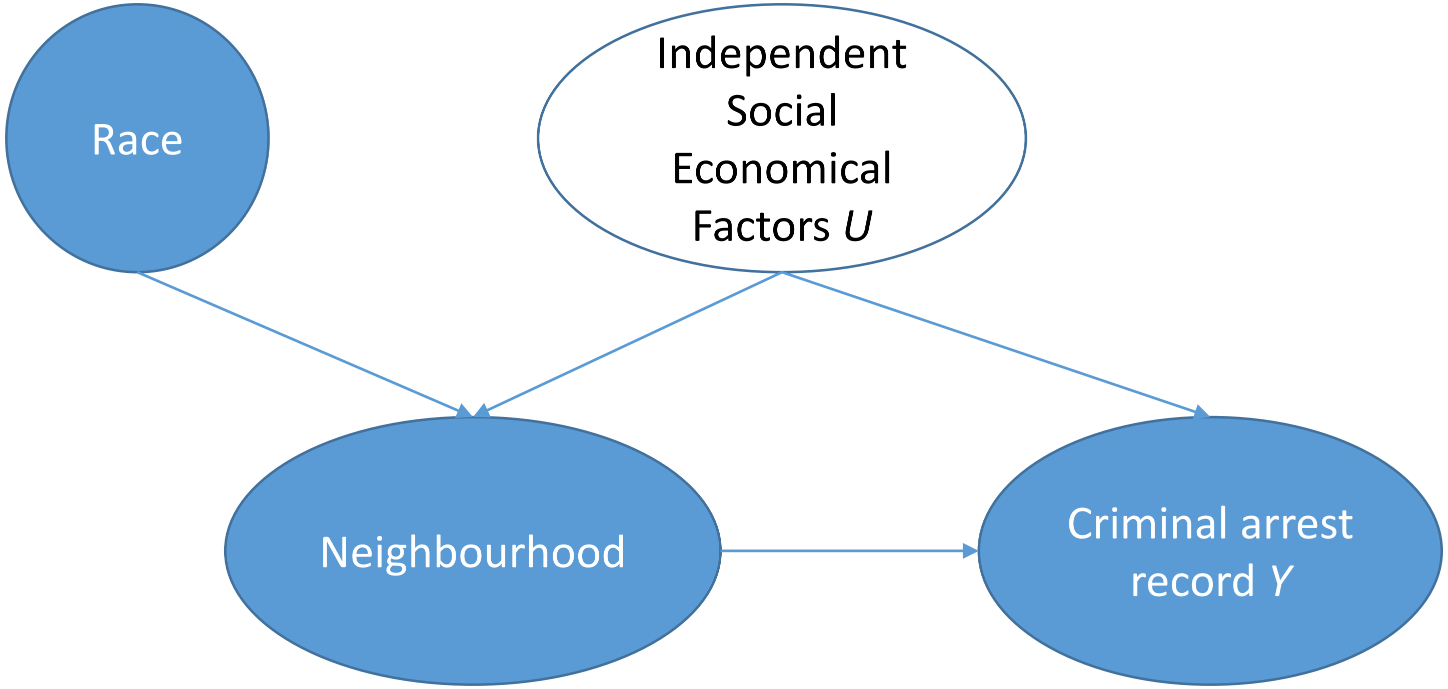
- Otherwise, the risk is to perpetuate unfairness in the world.
    - Bad credit limiting education which feeds prejudice and so on.

Another Toy Example: High Crime Regions


Race
Independent Social Economical Factors U
Neighbourhood
Criminal arrest record Y

Another Toy Example: High Crime Regions

- Sample selection bias embedded in the *Neighbourhood* → *Arrest* link.


Race
Independent Social Economical Factors U
Neighbourhood
Criminal arrest record Y

Another Toy Example: High Crime Regions

- Contrast this to *Equalized Odds*, which says that if *S* is independent of *Race* given *Y*, then all is well.
    - That is, if false positives/false negatives are the same by race.

- If our outcome *Y* is itself contaminated by unfair mechanisms, then anything that relies on calibration with respect to the “true” *Y* may be on shaky grounds according to counterfactual fairness.

The (High) Price to be Paid

- We need causal assumptions to be able to say whether predictor *S* is counterfactually fair or not.

- This will typically involve postulating latent variables and their relationship to the observed variables.
  - However, we do also highlight that postulating latent variables is not uncommon in social sciences.

- This is no easy task and should not be taken lightly. Designing it is however a separate task, independent of the construction of the predictor.
  - Essentially, we reduce the (possibly fuzzy) problem of defining fairness mostly to the problem of expressing causal assumptions.

Testing (some) Assumptions: Natural Experiments

"It turns out the department of correction's software was improperly giving some inmates credit for good behavior."

http://www.npr.org/2016/01/01/461700642/computer-glitch-leads-to-mistaken-early-release-of-prisoners-in-washington?utm_campaign=storyshare&utm_source=facebook.com&utm_medium=social

Testing (some) Assumptions: Controlled Changes of Practice

"How blind auditions help orchestras to eliminate gender bias"

https://www.theguardian.com/women-in-leadership/2013/oct/14/blind-auditions-orchestras-gender-bias

Back to COMPAS

- ProPublica was looking at false positives/false negatives:

$$P(\text{Low risk score} \mid \text{Recidivism, White defendant}) > P(\text{Low risk score} \mid \text{Recidivism, Black defendant})$$

$$P(\text{High risk score} \mid \text{No recidivism, White}) < P(\text{High risk score} \mid \text{No recidivism, Black})$$

- The rebuttal was looking at “test fairness”:

$$P(\text{Recidivism } \mid \text{Low risk, White}) \cong P(\text{Recidivism } \mid \text{Low risk, Black})$$

$$P(\text{Recidivism } \mid \text{High risk, White}) \cong P(\text{Recidivism } \mid \text{High risk, Black})$$

They are both true in COMPAS!

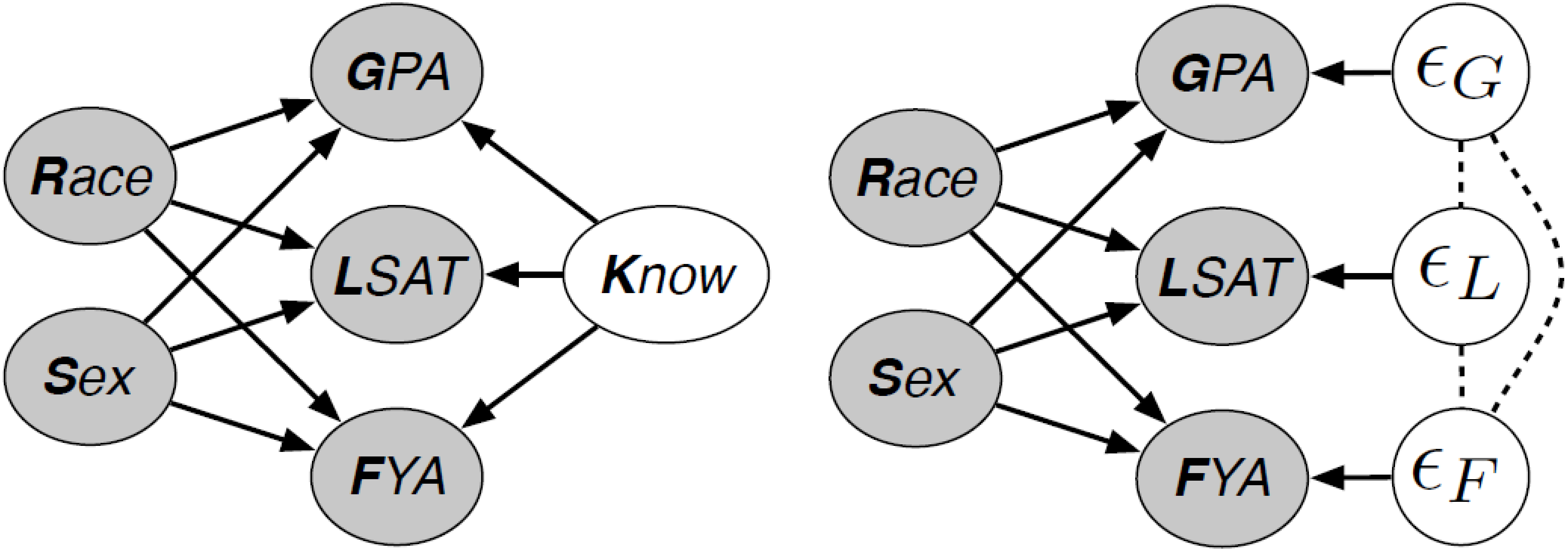
Recall

- *This disagreement happens because recidivism varies by race*.
  - The argument is mathematical but follows from basic probability.
- The disagreement seems undesirable but explainable if we believe that differences in recidivism by race are by construction unfair.
- Counterfactual fairness suggests that we should ignore “parts of” recidivism that comes from an unfair mechanism.
  - The COMPAS disagreement then becomes irrelevant.

Empirical Example: Law School Success

- Goal: to predict if an application to law school in the US will have a high first year average (FYA).
    - Law School Admission Council survey, 163 law schools, 21790 law students
- Postulate: predictions should not be biased by race and sex.
- In what follows,
    - We postulate two causal models as an illustration (we do not claim these are expert models)
    - We assess how fairness through unawareness (ignore race/gender) behaves under the assumption the causal model(s) is(are) correct.

Two Illustrative Models



Accuracy Results

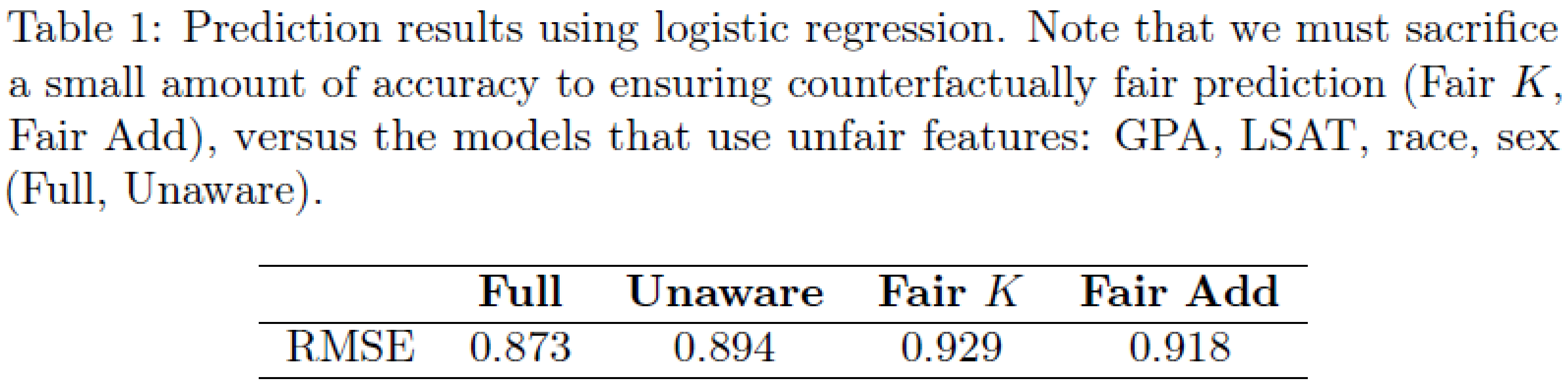
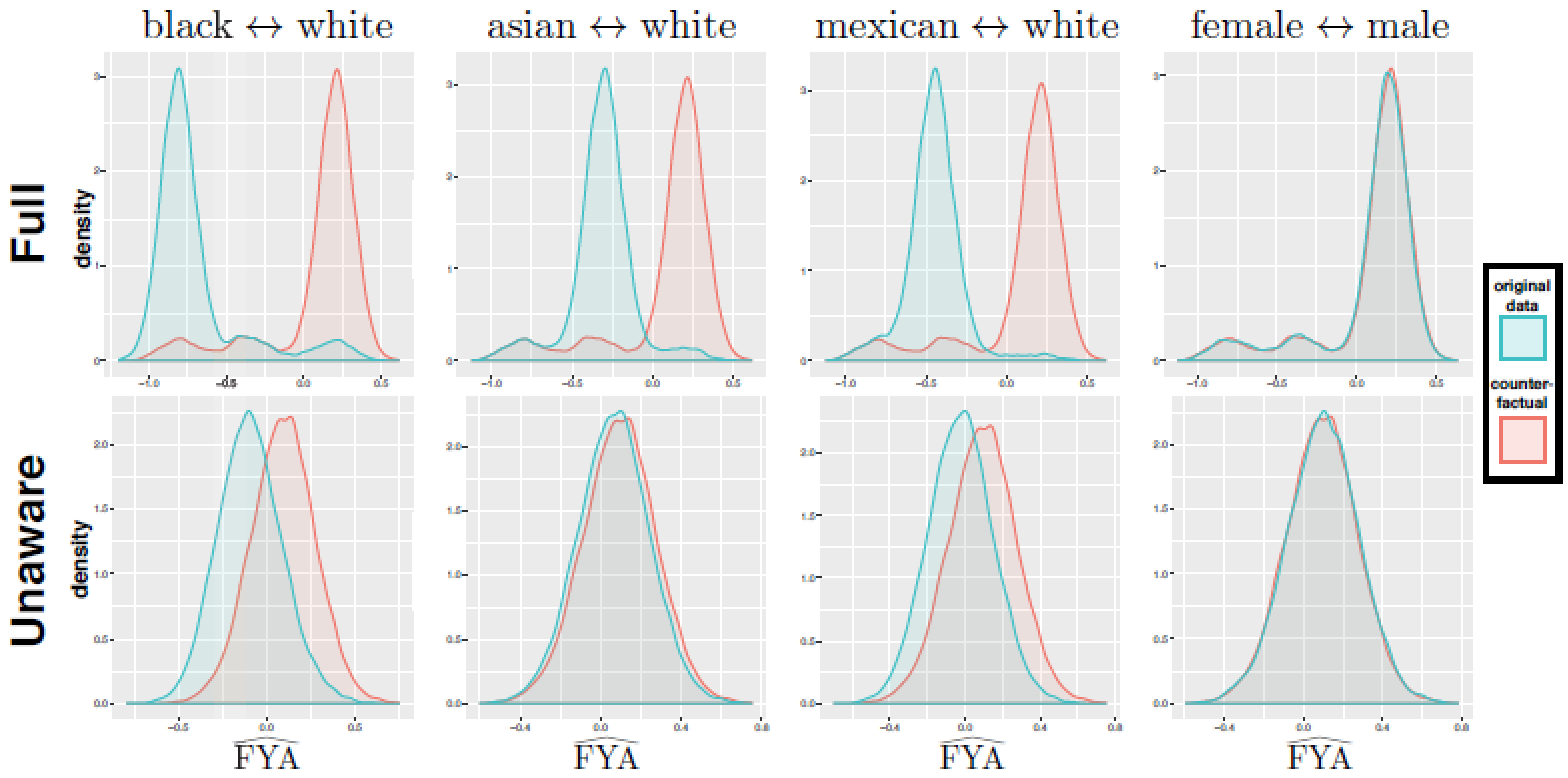


Table 1: Prediction results using logistic regression. Note that we must sacrifice a small amount of accuracy to ensuring counterfactually fair prediction (Fair K, Fair Add), versus the models that use unfair features: GPA, LSAT, race, sex (Full, Unaware).

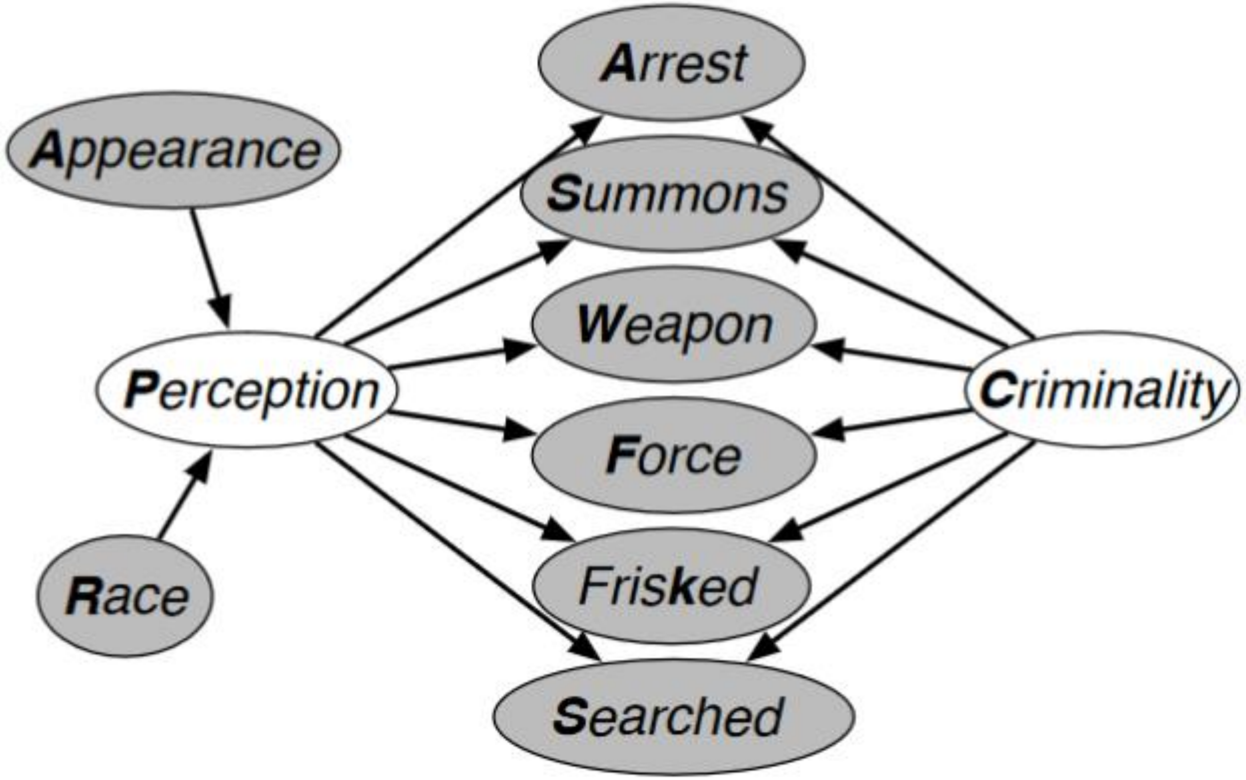
	Full	**Unaware**	**Fair K**	**Fair Add**
RMSE	0.873	0.894	0.929	0.918

Counterfactual Distribution of Predictions



Empirical Example: Criminality vs Perceived Criminality

- Separating actual and perceived criminality in police stops.
- Again: we are not experts, and the causal structure here is meant to be an illustration.
- NYPD records (since 2002) of stops.
  - Search or frisk
  - Males since 2014, 38609 records (90% of the data)



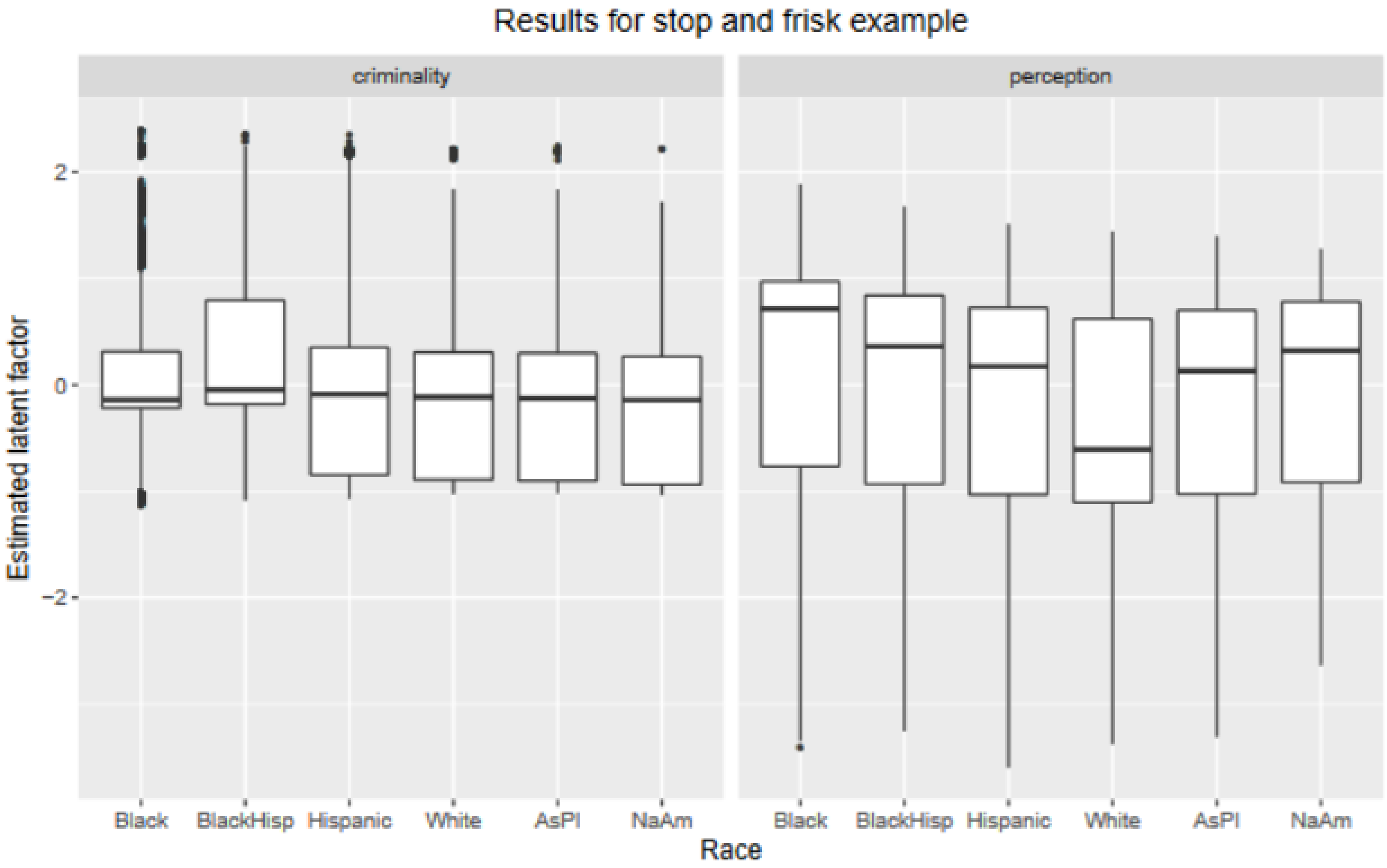
Analysing its Demographic Parity

- One criteria for fairness is whether our predictions are independent of protected attributes (e.g., are job offers independent of gender)

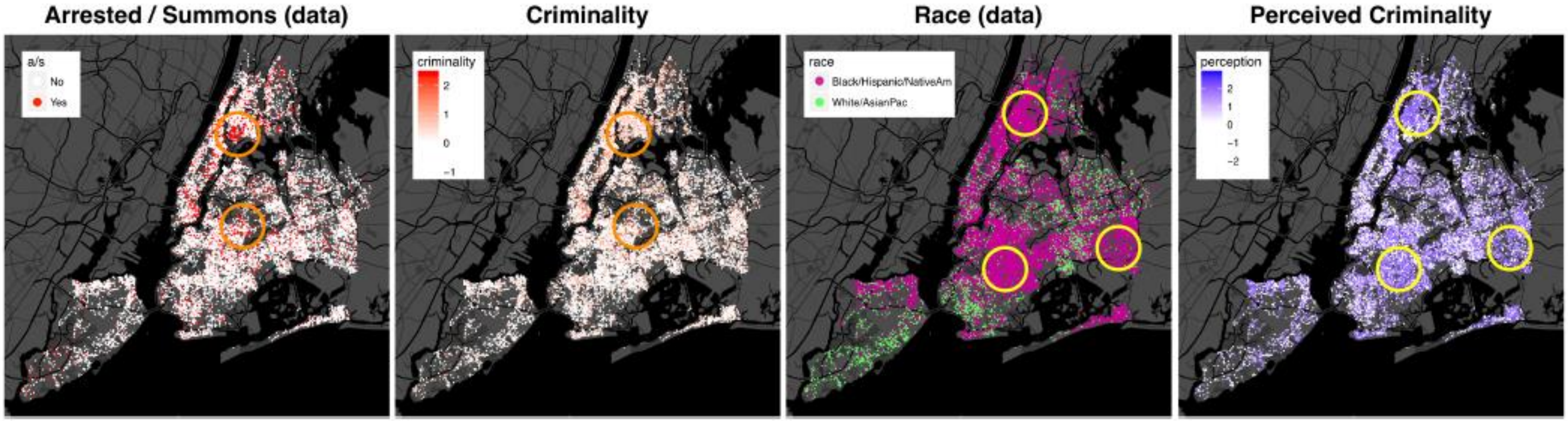
- In general this can hard to achieve.

- In the counterfactual inference scenario, this is in principle doable if the model is doing its job (e.g., *Criminality* independent of *Race*).

Race vs. Criminality/Perception



Spatial Arrangement



Arrested / Summons (data)

Criminality

Race (data)

Perceived Criminality

Conclusions

- With the rise of statistical models for human behaviour, we need to know in which ways these models are biased.

- The biases may be already in your data, and it would be irresponsible to propagate them.
  - See https://motherboard.vice.com/en_us/article/new-program-decides-criminality-from-facial-features for a ghastly example.

- The goal of counterfactual fairness is to put all assumptions on the table. Anything based on pure correlations is never going to suffice.

Thank You

Supplement

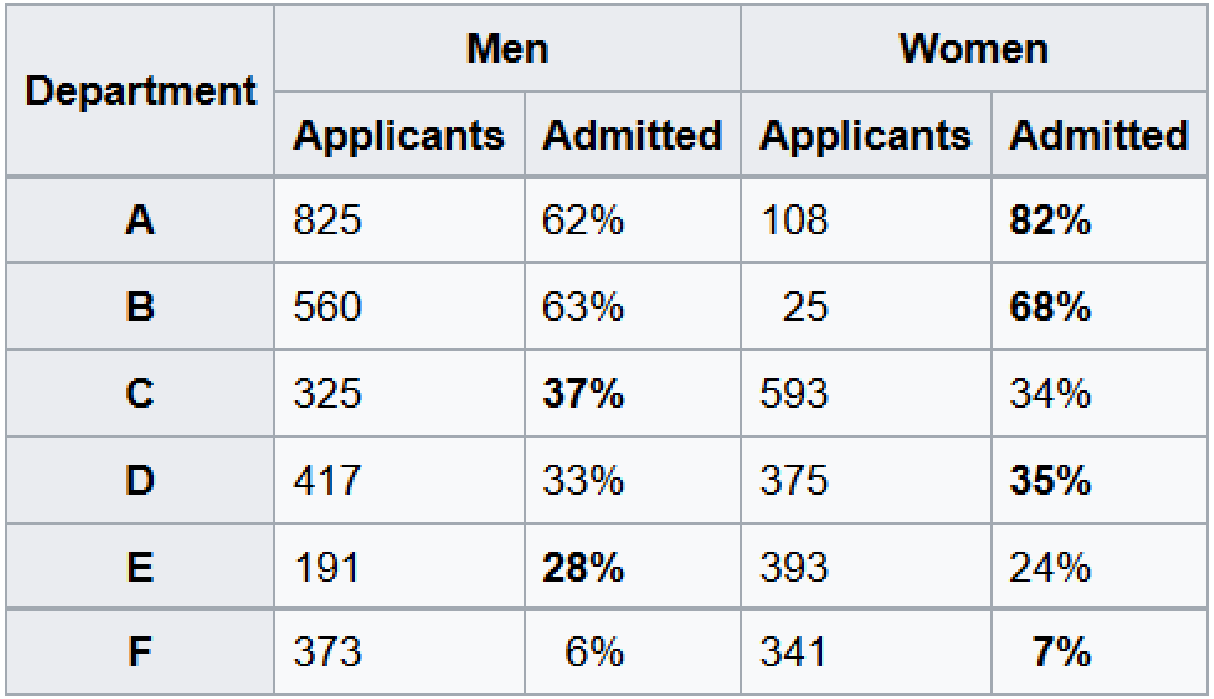
Stepping Back: a Historical Example

- In the early 1970s, UC Berkeley admission figures showed that 44% of male applications were admitted, against 35% of female applicants.

- Unfair decision making?

- Also, can we say something else about entanglements between the definition of what is *protected* and what is *fair*?
  - So far we have said that protected attributes are primitives: they are defined regardless of the definition of fairness.

Stepping Back: a Historical Example



Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

https://en.wikipedia.org/wiki/Simpson%27s_paradox

Another Example

- A drug has been observed to have a success rate higher than the alternative of not taking it.

- However, among males is performs worse than placebo.

- And among females, it *also* performs worse than placebo.

- Would you recommend it as a treatment?

Another Example

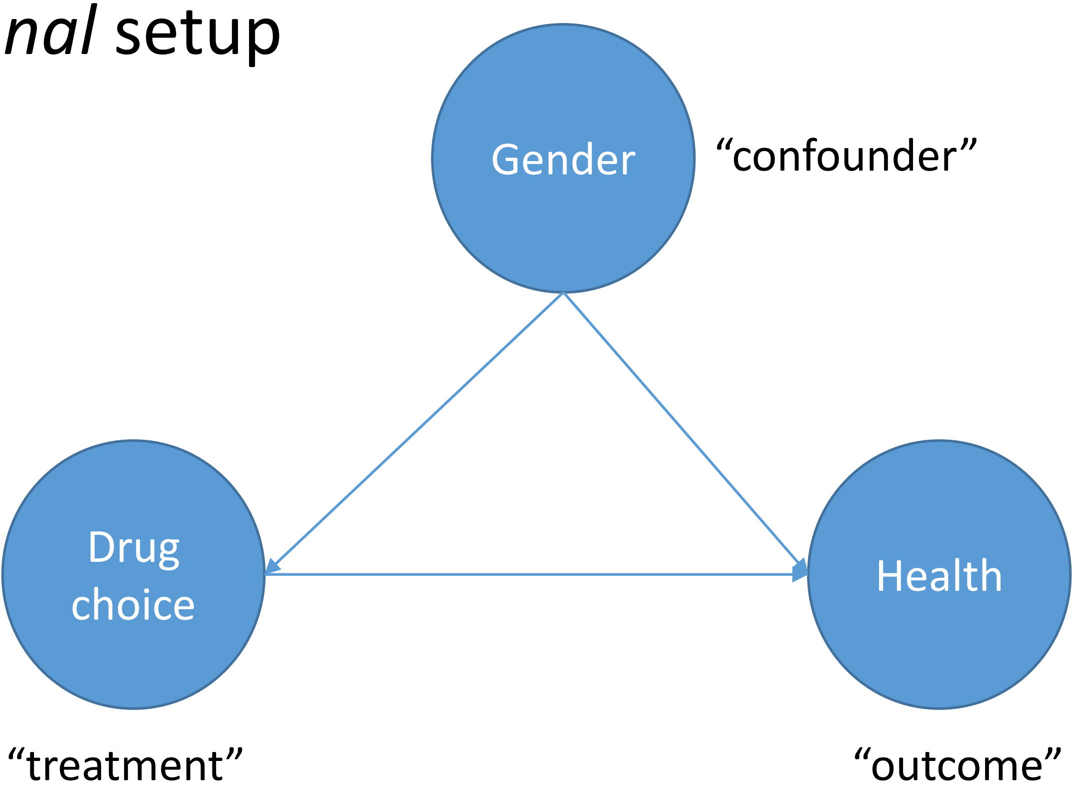
It looks like that if we stratify by gender, we should not give the drug.

But if we do *not* stratify, then we should.

Something is amiss. What is the “right” adjustment?

Disentangling It with Causal Diagrams

- An *observational* setup

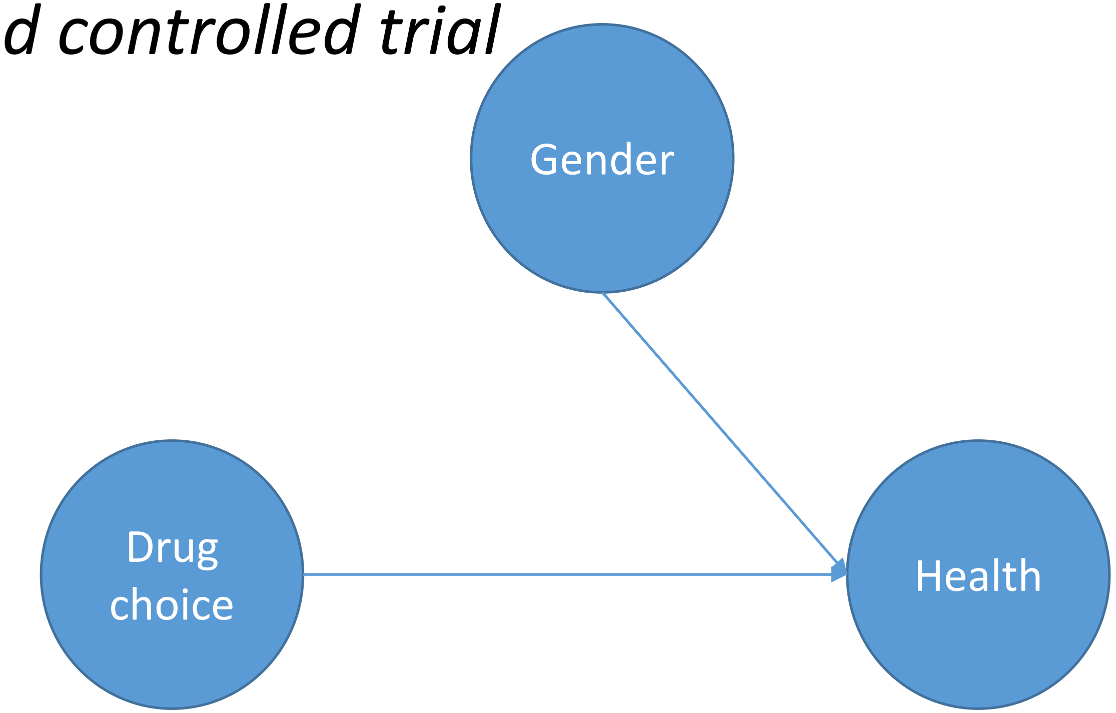


Males might be more likely to choose to take the drug.

They might also recover better than females, but still be better off without the drug.

Disentangling It with Causal Diagrams

- A *randomized controlled trial*



Link between gender and drug choice broken.

The drug *is* harmful in this hypothetical setup.

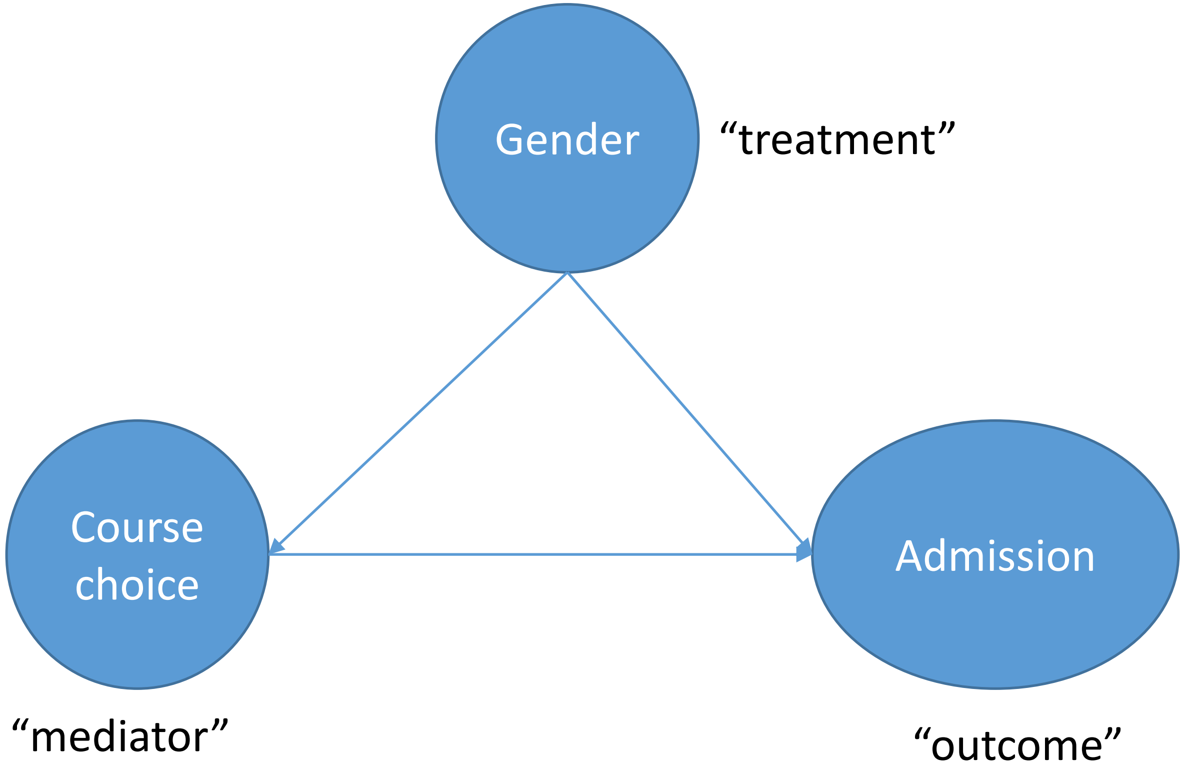
What is a Cause?

In this case, we say drug choice is a cause of health status because, if we control the choice directly, then as we vary the choice the distribution of health status changes.

This is the idea behind randomized controlled trials, but in socials sciences we can’t resort to them, and assumptions about causal structures are necessary.

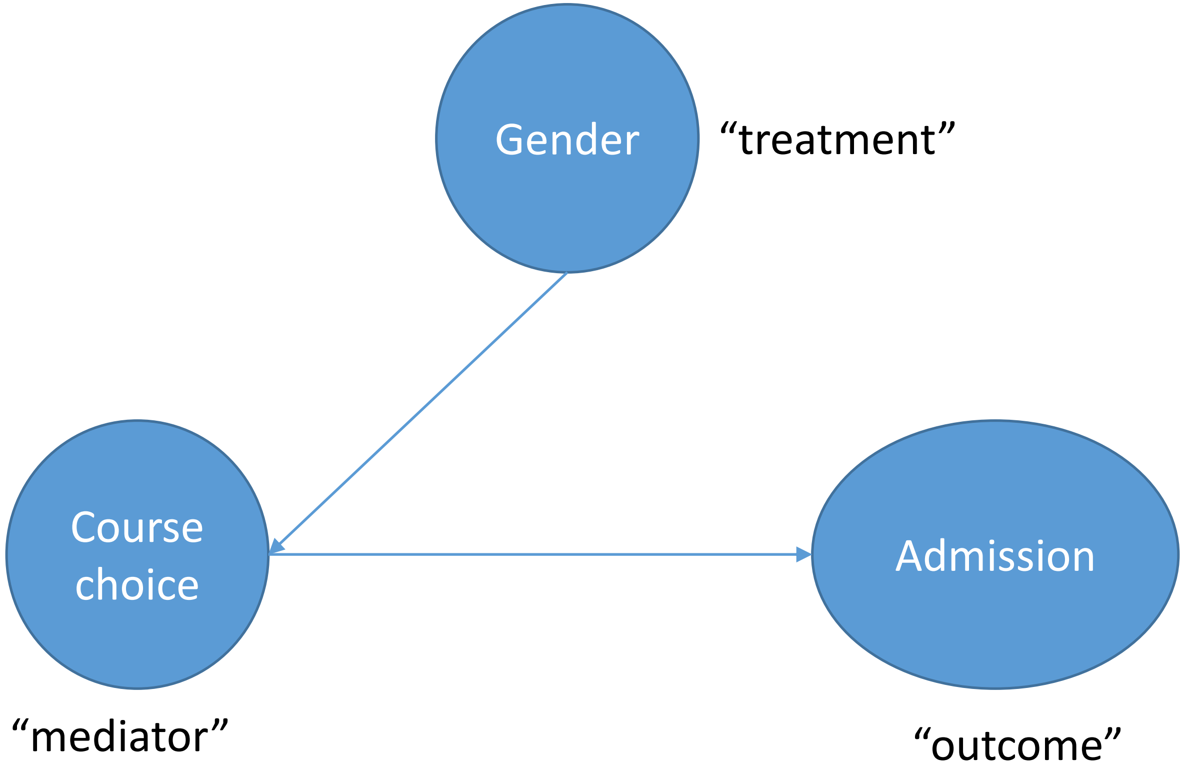
Back to Berkeley

- Should we stratify by course or not? Gender would be still a "cause" anyway if reality follows the structure below.



Back to Berkeley

- What if the world behaved like this? Controlling for course choice breaks the link, but gender is still a cause otherwise. Then what?



Under these Assumptions, Was Berkeley Unfair or Not?

- Gender *is* a cause of the admission decision according to these assumptions, but in some sense the competitiveness of a course should not be seen as an unfair factor in admissions.
- Feeding back to the definition of unfairness: if we hold that *Gender* should not be discriminated according to counterfactual fairness, then *Course Choice* should not be used.
- On the other hand, if we hold that *Course Choice* should be used along with counterfactual fairness, then *Gender* can be discriminated against.
- Knowing what to “hold true” while deducing what is unfair or not is part of the game. Don’t be afraid of changing initial assumptions of unfairness after eliciting causal structure.
- *The main lesson is that domain knowledge is required in order to judge whether a particular causal pathway is “sensitive” or not.*