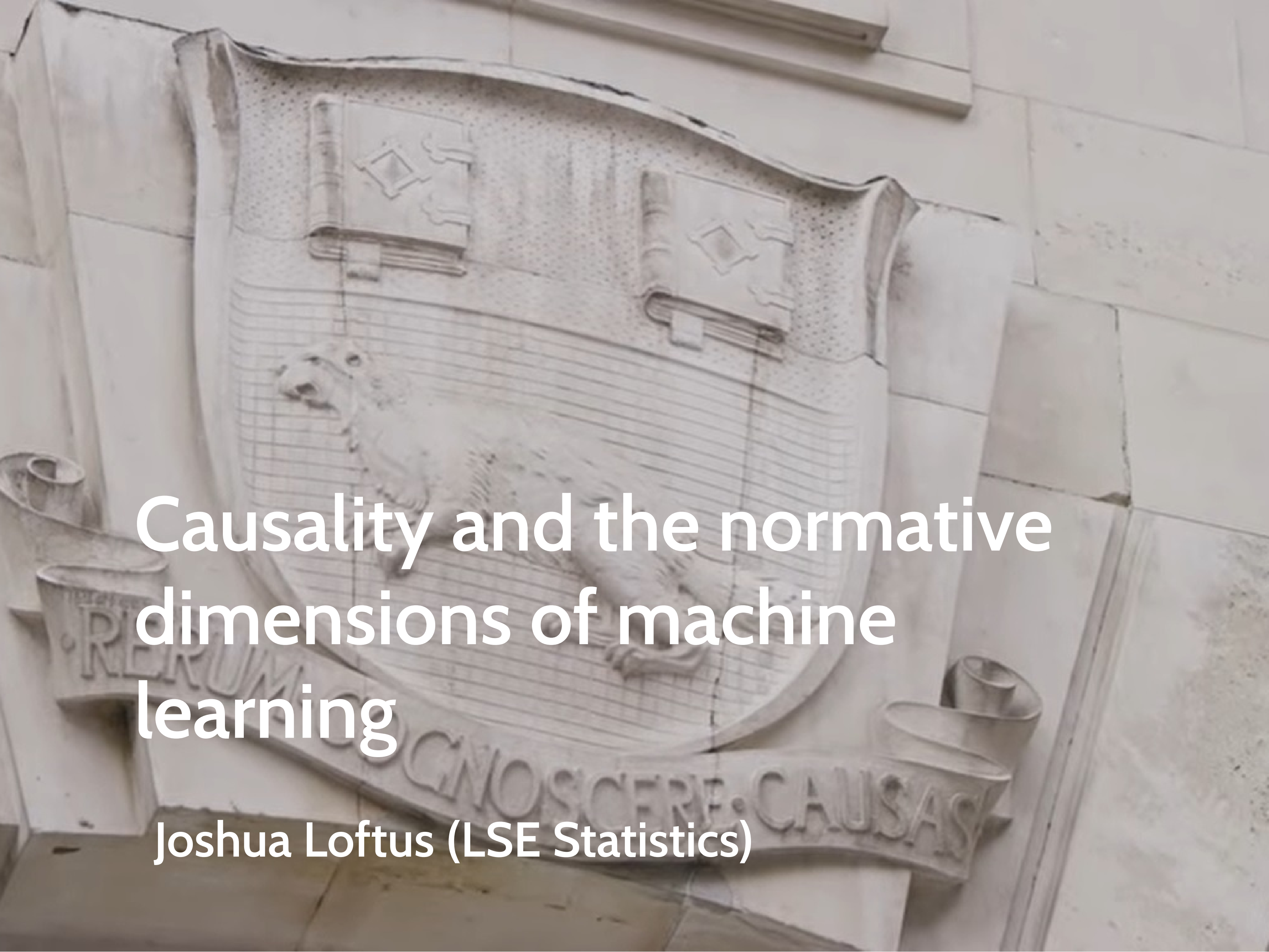


Causality and the normative dimensions of machine learning

Joshua Loftus (LSE Statistics)

High level intro

Causality, what is it good for?

Causal fairness

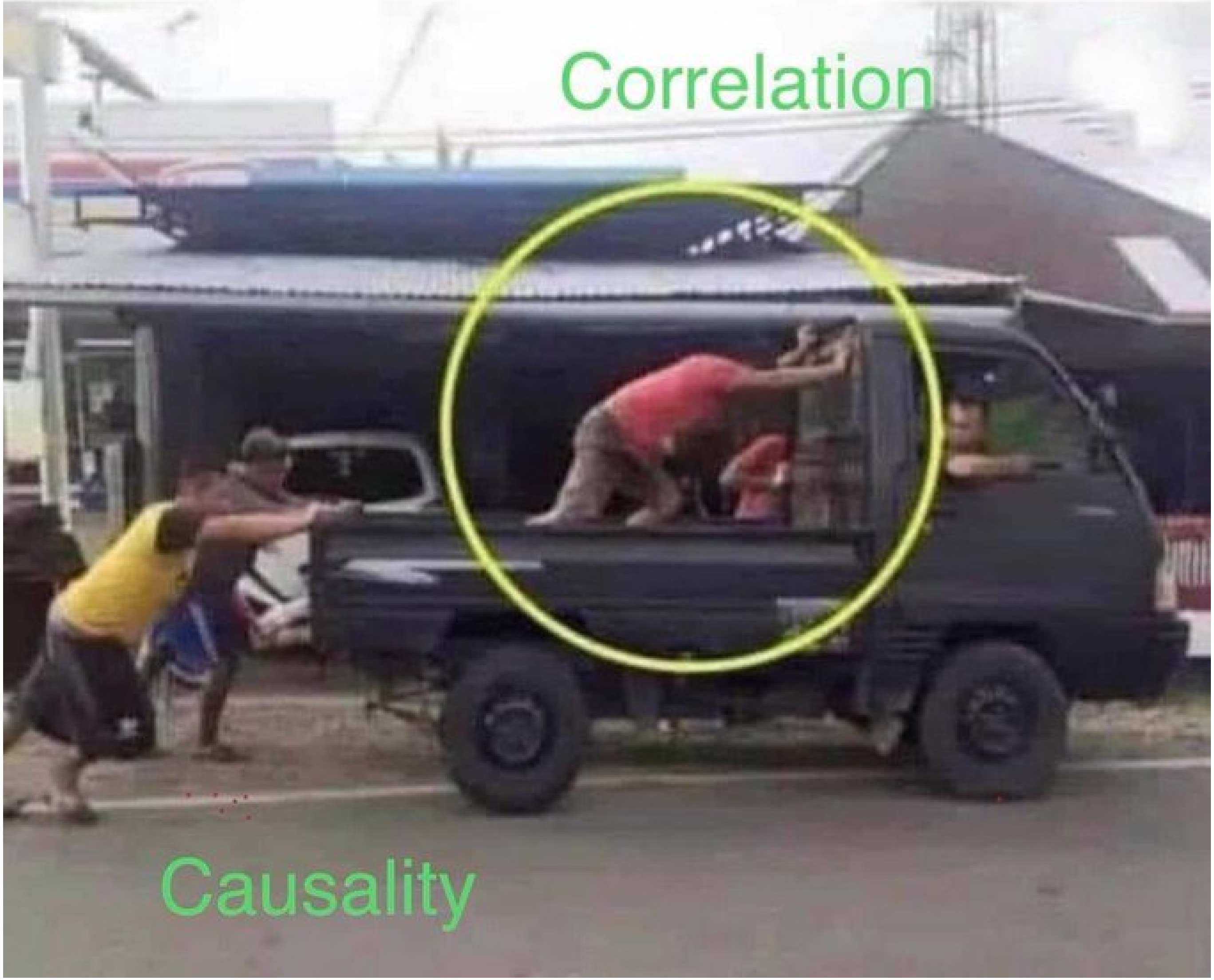
In prediction and ranking tasks, and with **intersectionality**

Designing interventions

Optimal fair policies, causal **interference**

Concluding thoughts

Tech solutionism, using ML/AI in every situation



Imagination

Albert Einstein:

> Imagination is more important than knowledge. For knowledge is limited, whereas imagination [...] stimulat[es] progress, giving birth to evolution.

Stephen Jay Gould:

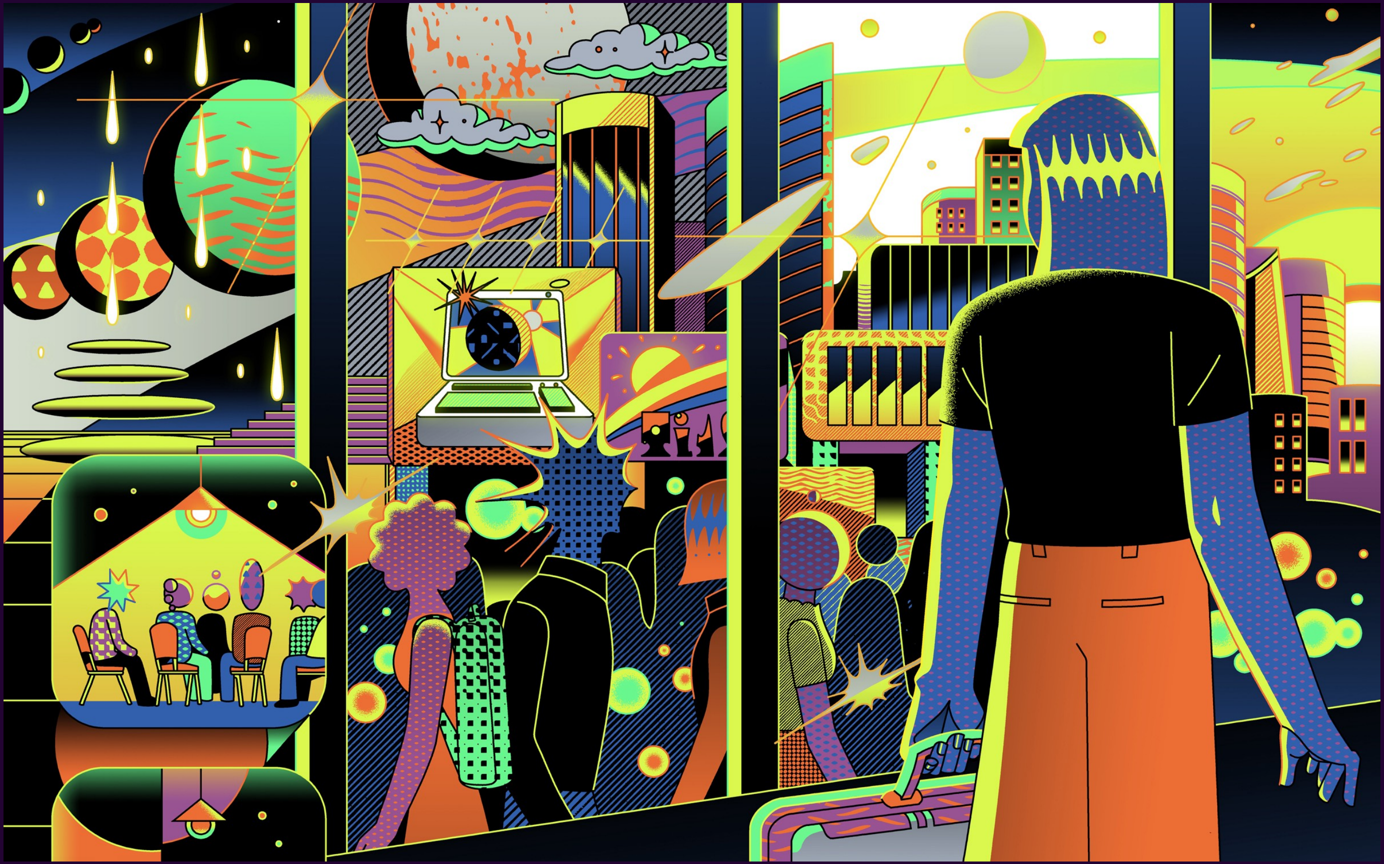
> I am, somehow, less interested in [...] Einstein’s brain than [that] people of equal talent have lived and died in cotton fields and sweatshops.

David Graeber:

> the ultimate, hidden truth of the world is that it is something that we make, and could just as easily make differently

Science fiction

Anxiety Is the Dizziness of Freedom, story by Ted Chiang



(Art: Jinhwa Jang)

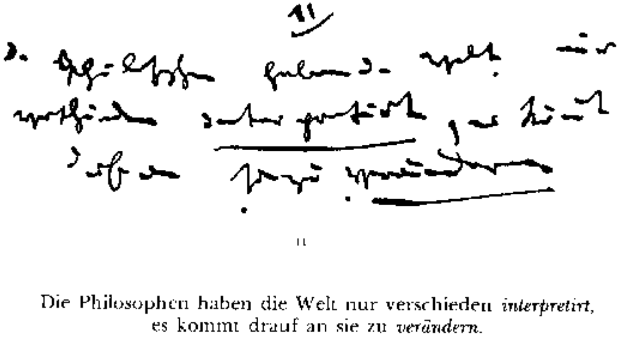
Fictional technology, like the "universe splitter" app combined with a messenger

> prism created two newly divergent timelines, one in which the red LED lit up and one in which the blue one did, and it allowed communication between the two

Experiments with both potential outcomes observed?! No...

> Every branch was of paramount importance to its inhabitants; no one was willing to act as a guinea pig for anyone else.
>
> What prisms did offer was a way to study the mechanisms of historical change. Researchers began comparing news headlines across branches, looking for discrepancies and then investigating their causes.

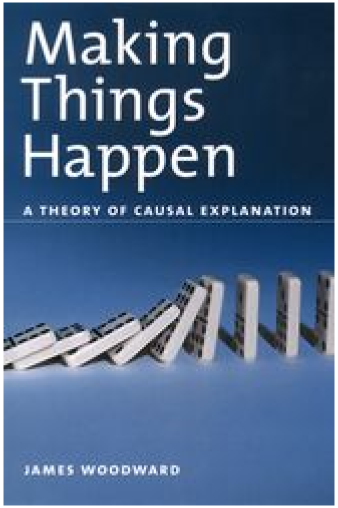
Why is causality important?



11

Die Philosophen haben die Welt nur verschieden *interpretirt*, es kommt drauf an sie zu *verändern*.

ML models have hitherto only predicted the world in various ways; the point is to change it



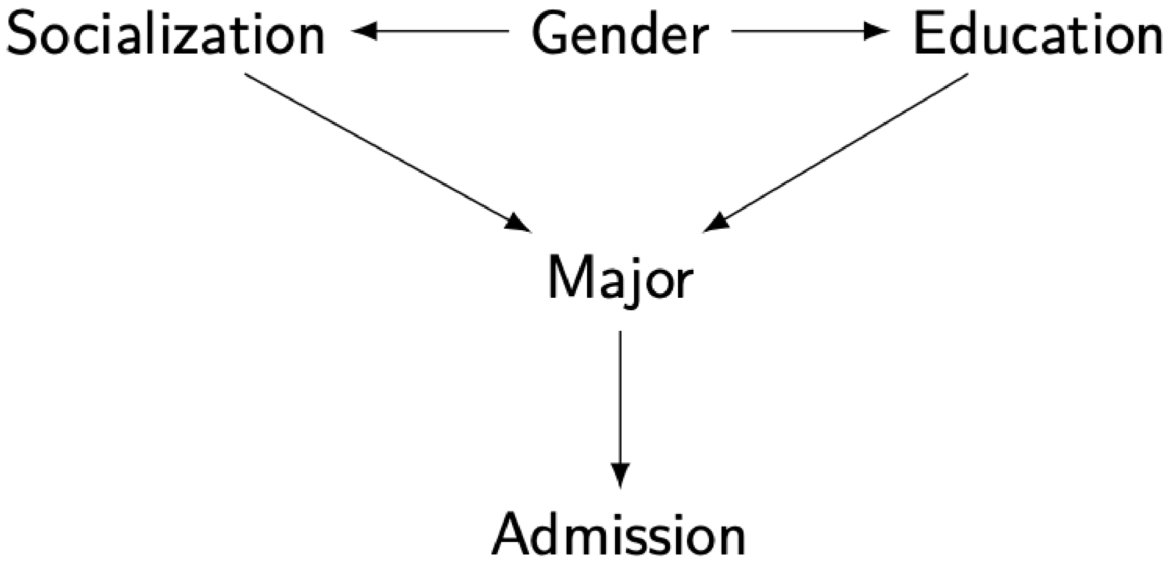
(Evolutionary explanation?)

Berkeley graduate admissions example

> 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 completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.

From the final paragraph of Bickel et al (1975)

Directed Acyclic Graph (DAG) models



- Assumption: directed paths show conditional dependence
- Assumption: intervention to change one variable also affects all variables on paths away from it

Two interpretations

Interventions

Can we change socialization/education, for example including more women role models in STEM curricula?

Counterfactual fairness

I was admitted to grad school. *If I had been a different gender*, maybe I would have applied to a different department and not been admitted... *Is that fair*?

Bet on causality

My claim: **fairness**, combined with a **focus on causality**, orients us toward discovering and mitigating the **root causes**

In particular, it will:

- expose flaws in the *status quo* application of fairness concepts like **merit** / just deserts
- show how to better operationalize those concepts
- bridge the "merit" and consequentialist approaches

Interventions, counterfactuals, thought experiments

- Pearl's ladder of causation level 2

Intervention, action, manipulation, policy change

What will (or is more likely to) happen if...

- But we can also go to level 3

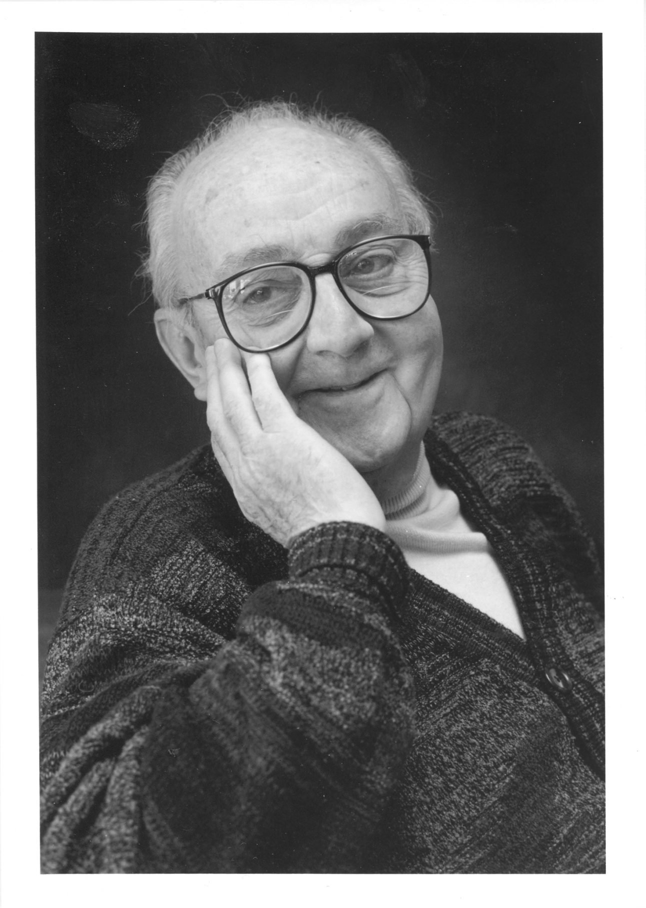
Counterfactuals, potential histories(?)

What would have (been more likely to have) happened if...

- Thought experiments

Models as thinking tools, *over-simplified models as diagnostics for more realistic ones*

Statistical wisdom: models as (thinking) tools



George Box

> All models are wrong but some are useful

therefore,

> ... the scientist must be alert to what is **importantly wrong**
>
> ... **cannot obtain a "correct" one** by excessive elaboration

Imbens (2019) on DAGs, reviewing The Book of Why:

> TBOW and the DAG approach fully deserve the attention of all researchers and users of causal inference as one of its leading methodologies.

DAGs conveniently express some things but not others

> In economics the endogeneity often arises from agents actively making choices regarding the causal variable on the basis of anticipated effects of those choices

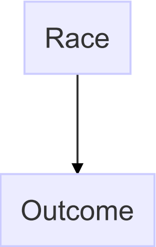
DAGs and Potential Outcomes are both useful

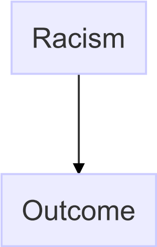
> identification results are also easier to derive in the PO framework

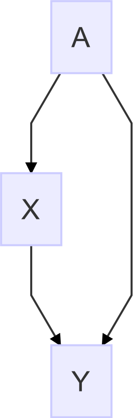
- What's **importantly wrong** about a given DAG?

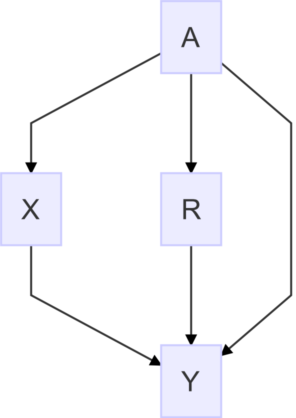
Every(?) DAG is importantly wrong

Transparency: we can say specifically what we disagree on









Kusner et al (2017), Kilbertus et al (2017), Nabi and Shpitser (2018), Zhang and Bareinboim (2018), Chiappa (2019), and a growing list of others

Formalism vs interpretation

As an analogy, consider the **axioms of probability**, and various **interpretations**

- Frequentist
- Bayesian
- Subjectivist
- Logical
- Epistemic
- etc

"... some are useful" -- **useful to who, and for what**?

Pragmatism: judge by practical effects, iterate, progress(?)

Making (DAG) models less wrong

Intersectional fairness

- Multiple sensitive attributes, e.g. race and gender
- Variety of relationships with other mediating variables
- Some of these mediators may be resolving/non-resolving for different sensitive attributes

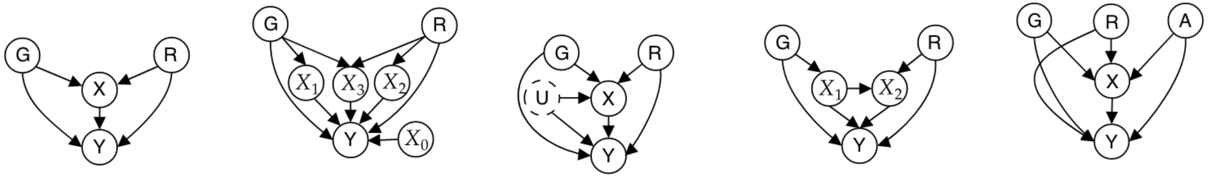
Lots of scholarship, not much using formal mathematical models. But see Bright et al (2016), O'Connor et al (2019), and a few other references in our paper

Yang, Stoyanovich, Loftus, Causal Intersectionality and Fair Ranking (FORC 2021, to appear)

"Moving company" example

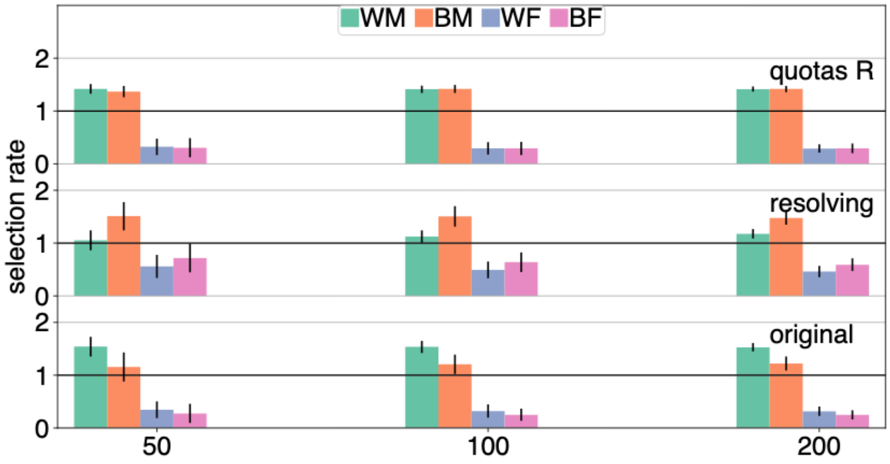
Race, gender, weightlifting test, application score

Weightlifting considered a resolving variable (company argues it is a necessary qualification)



(a) $\mathcal{M}_1$ **(b)** $\mathcal{M}_2$ **(c)** $\mathcal{M}_3$ **(d)** $\mathcal{M}_4$ **(e)** $\mathcal{M}_5$

Figure 2 Causal models that include sensitive attributes G (gender), R (race), and A (age), utility score Y, other covariates $\mathbf{X}$, and a latent (unobserved) variable U.



Untangling intersectional relationships

Causal models are a useful formal language

Empirically, causal mediation with **multiple mediators** and **multiple causes** is a very hard problem, limiting realistic application until better methods/data are available

(I don't think the difficulty is an artifact of our approach, it reflects that fairness/justice/equity are fundamentally hard problems)

Making (DAG) models less wrong

Interventions under interference

- Designing an optimal policy / intervention / allocation
- Relaxing common assumption that intervention on individual/unit i does not effect other individuals/units

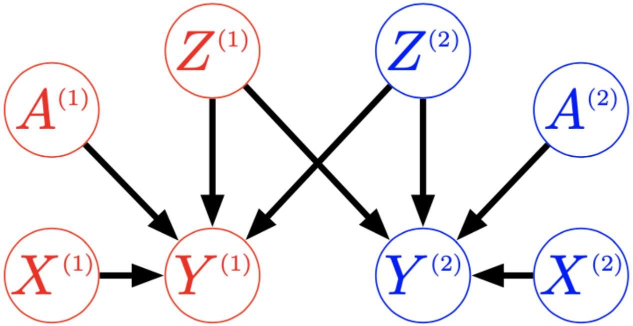
(in fairness/justice applications that will usually be **importantly wrong**)

- Constraint: bound **counterfactual privilege**, preventing "rich get richer" effect

Kusner, Russell, Loftus, Silva, Making Decisions that Reduce Discriminatory Impacts (ICML 2019)

Optimal fair policies under interference

Intervention $\mathbf{Z}$ trying to increase $\mathbf{Y}$



Privilege constraint, for $\tau \geq 0$

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

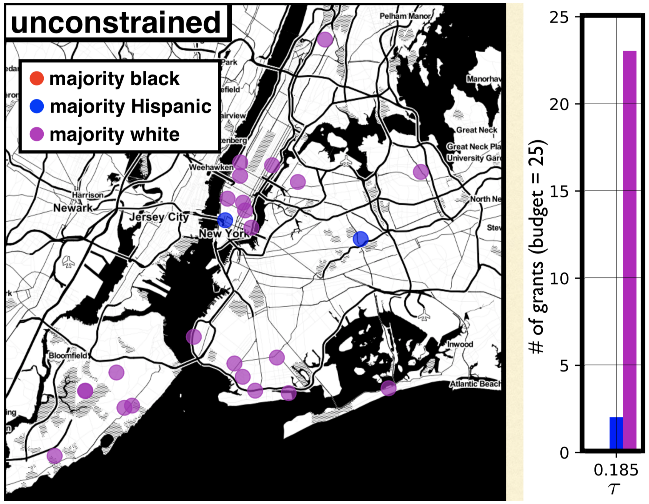
Optimization problem (with budget constraint b)

$$\mathbf{Z} = \arg\max \sum_i \mathbb{E}\left[\hat{\mathbf{Y}}^{(i)}(a^{(i)}, \mathbf{Z}) | \mathbf{A}^{(i)}, \mathbf{X}^{(i)}\right]$$

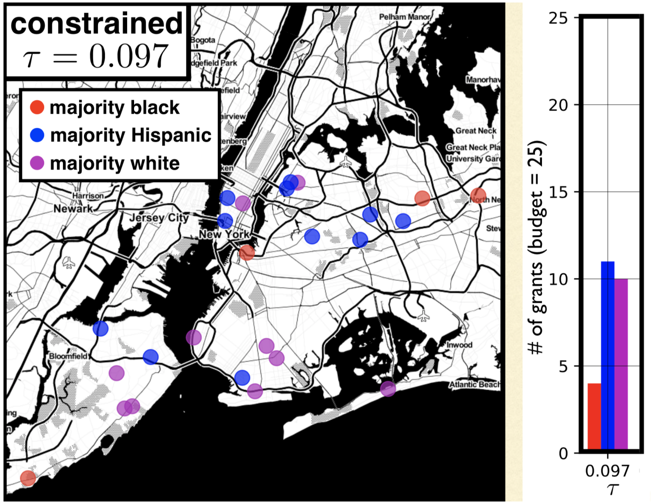
$$s.t. \quad \sum_i \mathbf{Z}^{(i)} \leq b$$

Allocating resources to (NYC) schools

Without constraint



With constraint



Future work

What about discriminatory impacts that already exist?

- Measuring existing inequality
- Intervening earlier in the ML pipeline

When are social categories oversimplified?

- Race as a multi-dimensional construct
- Race as a cause vs. race as an outcome
- Social categories as constructs created over time

Can we modify the status quo?

- Interventions on graphs
- Measuring inequality before and after intervention
- Breaking cycles of disadvantage

Closing thoughts

Merit, just desert, qualification, utility...

The things we really care about are usually not measured

And probably cannot be--or at least not for long--due to Goodhart-Campbell laws

Causal models can point us in directions of building consensus and converging toward truth/justice

Turing group

Matt Kusner (UCL)

Chris Russell (AWS/ELLIS)

Ricardo Silva (UCL)

NYU

Julia Stoyanovich

Ke Yang

Lucius Bynum

Margarita Boyarskaya

Questions?

Thank you for listening!

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

joshualoftus.com