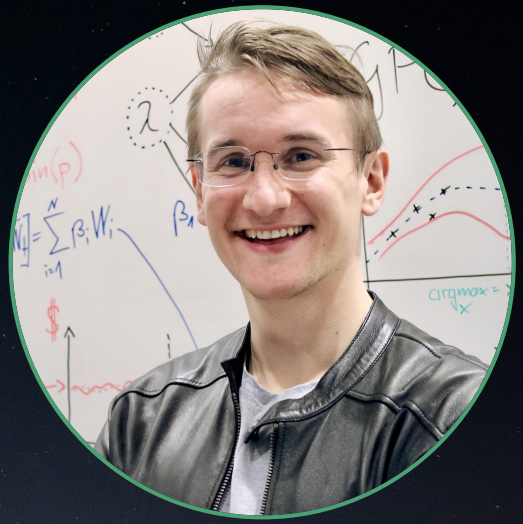
Causality, AI and Ethics

What works and what doesn't

Jakob Zeitler
Pioneer Fellow, Department of Statistics
Research Associate, Jesus College



Talk based on Bhadane et al. 2025 “Revisiting the Berkeley Admissions data: Statistical Tests for Causal Hypotheses”

AI + Decision Making


ETFLI
Panasonic
1 QZ
4 GHI
5 JKL
7 PRS
8 TUV
9 WXY


AI
AI
AI

1973 Berkeley Admissions Case

Is there sex-based discrimination?

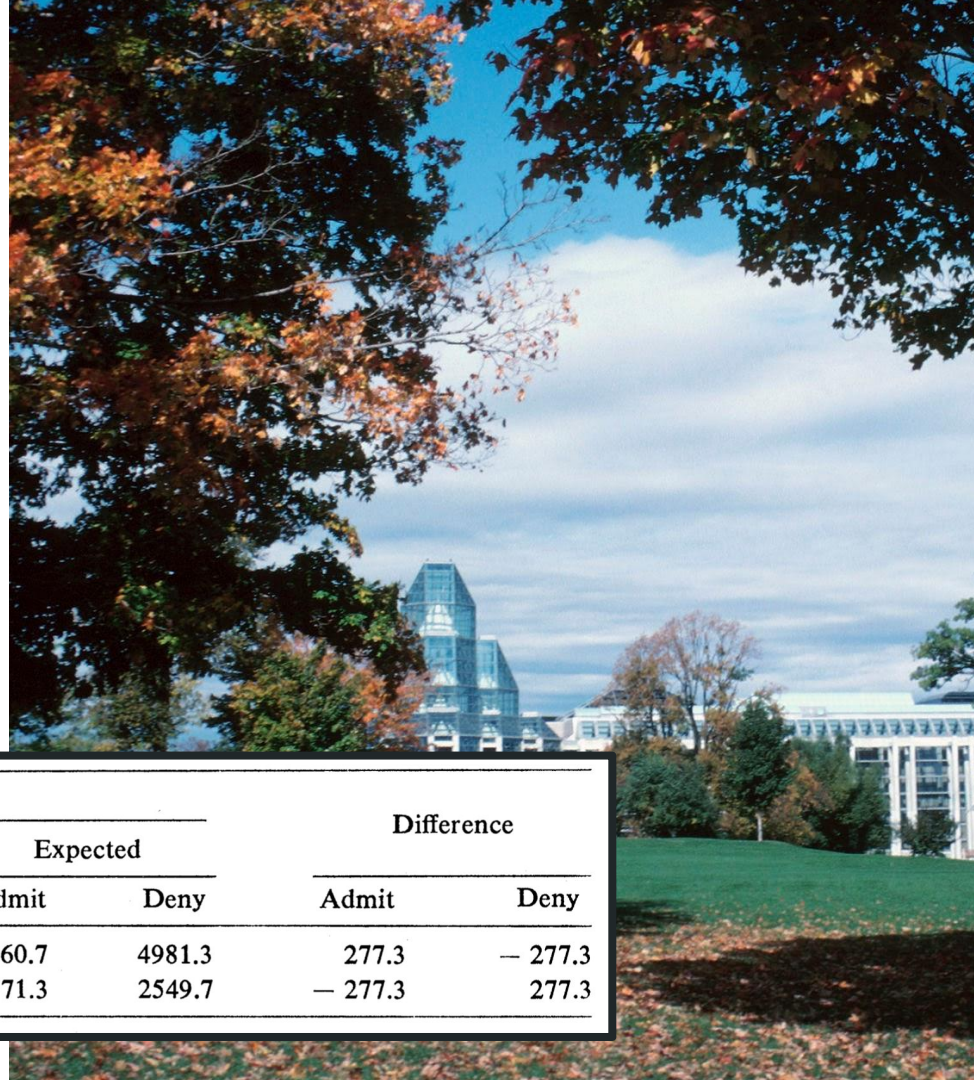
Applicants	Outcome	
	Expected	
	Admit	Deny
Men	3460.7	4981.3
Women	1771.3	2549.7

1973 Berkeley Admissions Case

Is there sex-based discrimination?

Male 44.2% Acceptance

Female: 34.6% Acceptance



Applicants	Outcome				Difference	
	Observed		Expected			
	Admit	Deny	Admit	Deny	Admit	Deny
Men	3738	4704	3460.7	4981.3	277.3	− 277.3
Women	1494	2827	1771.3	2549.7	− 277.3	277.3

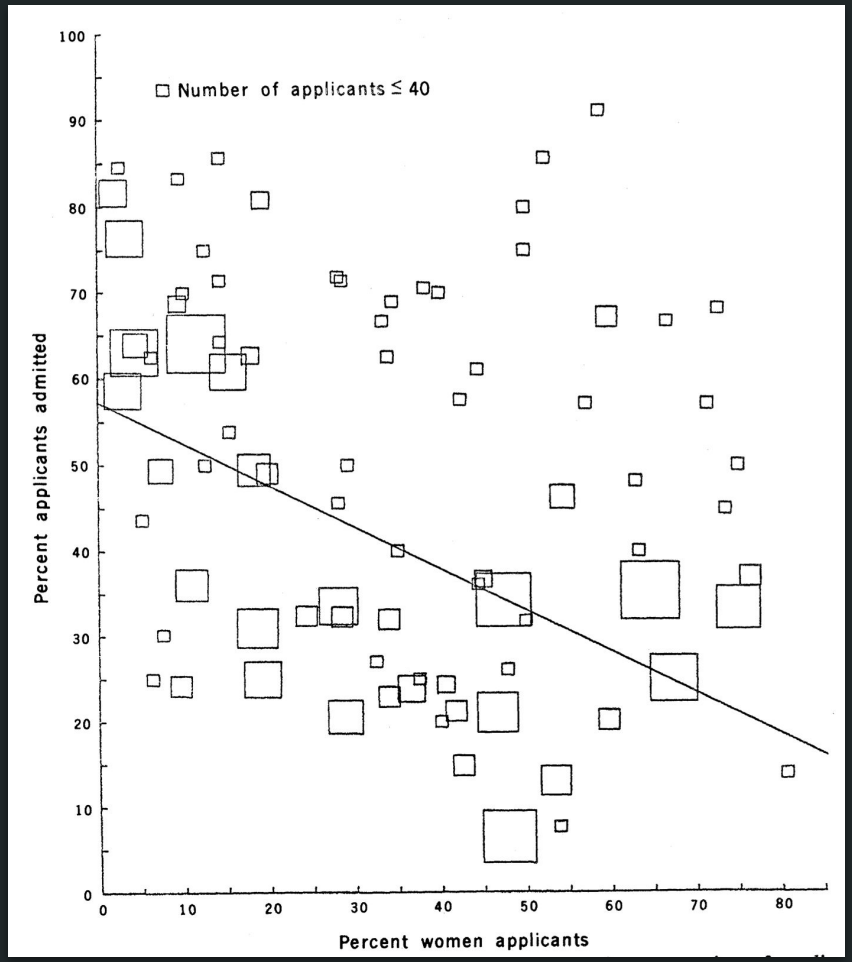
“The proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get into”

Bickel et al., 1974

“Picture a fishnet with two different mesh sizes. A school of fish, all of identical size, swim toward the net and seek to pass. The female fish all try to get through the small mesh, while the male fish all try to get through the large mesh. On the other side of the net all the fish are male.”

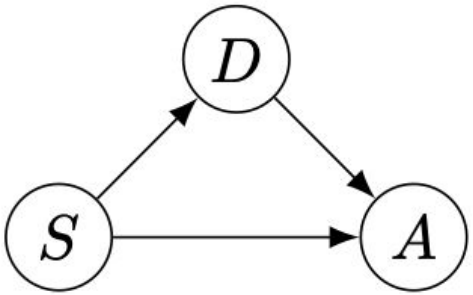
Bickel et al., 1974

Simpson's Paradox / Spurious Correlation

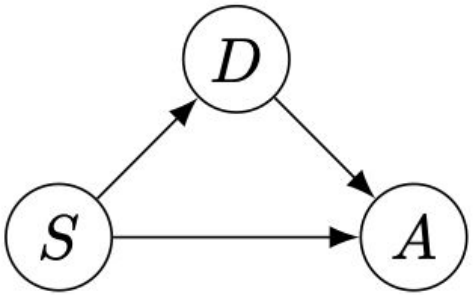


Did we miss anything?

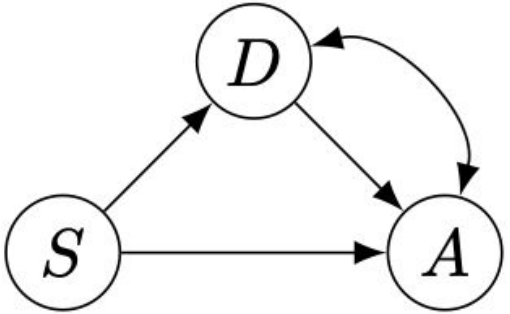
Problem: Unmeasured confounding



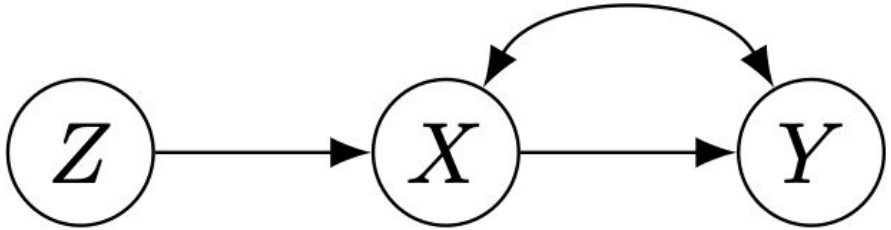
Problem: Unmeasured confounding



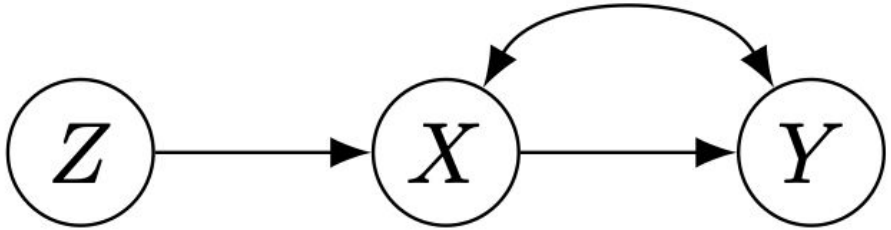
versus



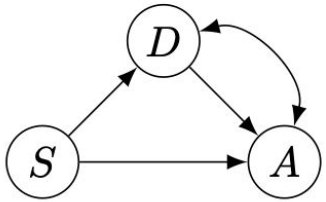
Instrumental Variable Inequalities



Instrumental Variable Inequalities

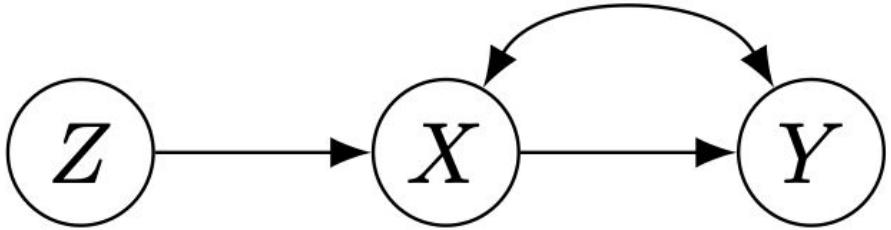


$$\max_x \sum_y \max_z P_M(X = x, Y = y \mid Z = z) \leq 1.$$



Example: Discrimination
If sex (S) impacts admissions (A), then IV inequalities are not satisfied.

Instrumental Variable Inequalities



$$\max_x \sum_y \max_z P_M(X = x, Y = y \mid Z = z) \leq 1.$$

Berkeley Data Inequalities *are* satisfied

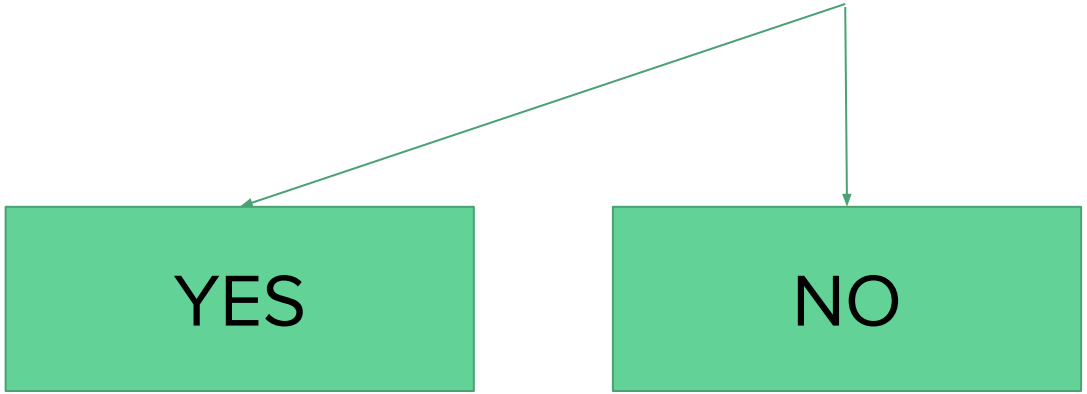
“ Therefore, satisfying the IV inequalities implies that fairness is ‘undecidable’ since both causal models [...] satisfy the IV inequalities.

Violating the IV inequalities, however, would imply that there is a direct effect of sex on admissions outcome.”

Bhadane et al., 2025

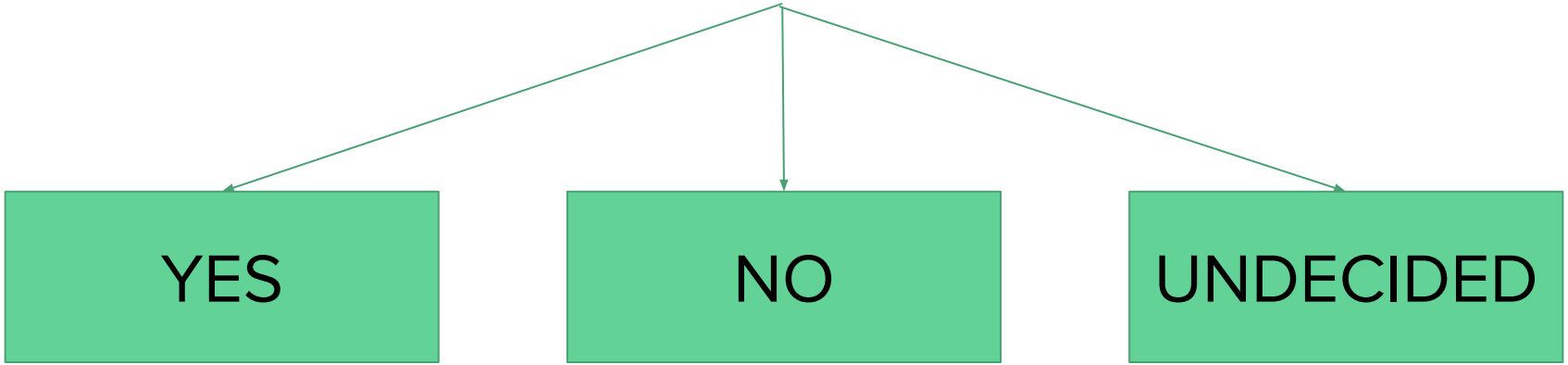
Workflow

Is Berkeley admissions discriminating based on sex?



Workflow

Is Berkeley admissions discriminating based on sex?



Conclusion

Assuming intellectual honesty about assumptions

Applying mathematical & statistical modeling

We often cannot conclude on fairness

Thank you to Bhadane et al. 2025 “Revisiting the Berkeley Admissions data: Statistical Tests for Causal Hypotheses”

What, then, are the **implications for our *Ethics* of AI?**

References

Talk based on Bhadane et al. 2025 “Revisiting the Berkeley Admissions data: Statistical Tests for Causal Hypotheses”

https://openreview.net/forum?id=5Sm8gC0Shb

https://homepage.stat.uiowa.edu/~mbognar/1030/Bickel-Berkeley.pdf