CS 229br Lecture 6: Causality, Fairness, Privacy

Boaz Barak

Yamini Bansal
Official TF

Javin Pombra
Official TF

Dimitris Kalimeris
Unofficial TF

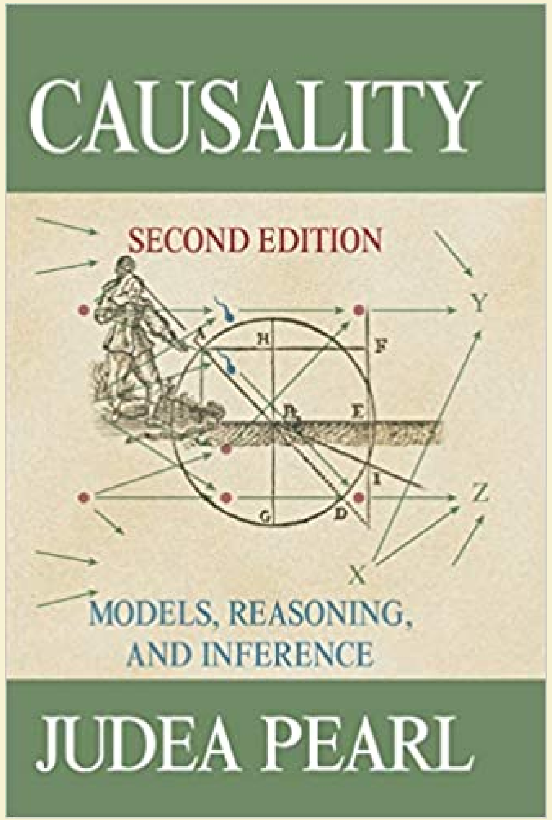
Gal Kaplun
Unofficial TF

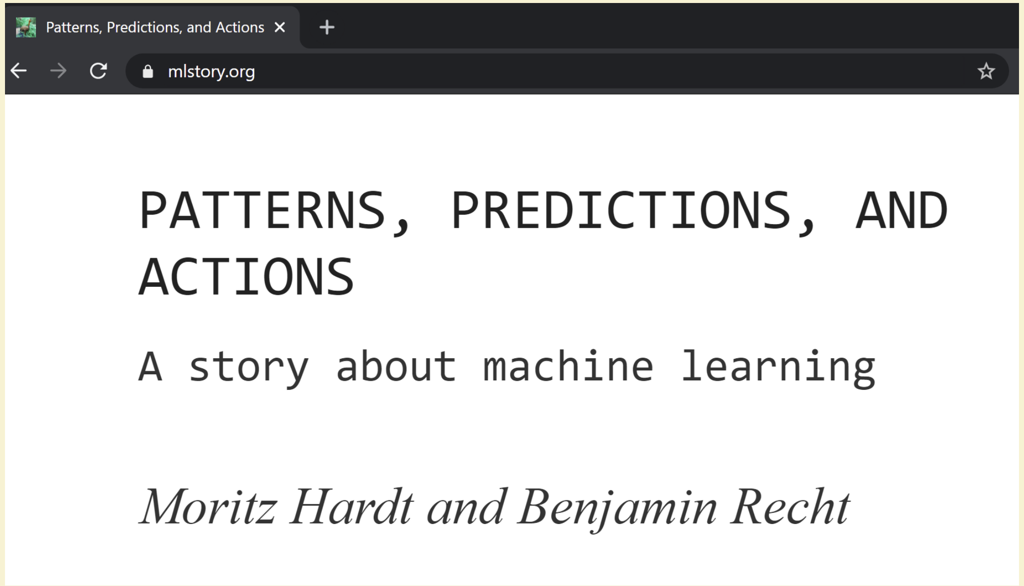
Preetum Nakkiran
Unofficial TF

Outline

- Part I: Causality
- Part II: Fairness

Causality


CAUSALITY
SECOND EDITION
MODELS, REASONING, AND INFERENCE
JUDEA PEARL


Patterns, Predictions, and Actions
mlstory.org
PATTERNS, PREDICTIONS, AND ACTIONS
A story about machine learning
Moritz Hardt and Benjamin Recht

Causality

Correlation ≠ Causation

But what is causation?

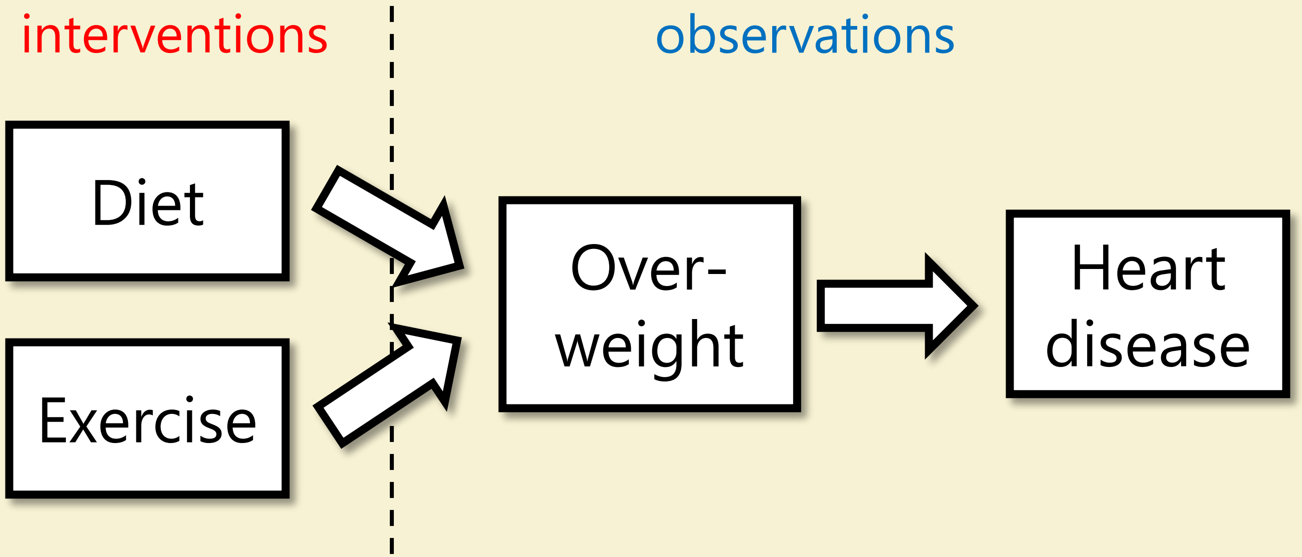
Intervention

Observation

"A causes B"

"Smoking causes cancer"

"Obesity causes heart disease"



Causality theory

Understand the conditions under which correlation = causation

Setup:

Observables: $A, B, C, D, \ldots$

Interventions: "$\text{do } A \leftarrow a$"

Correlation: $\Pr[\, B = b \mid A = a \,]$

Causation: $\Pr[\, B = b \mid \text{do } A \leftarrow a \,]$

Correlation: $\Pr[\, B = b \mid A = a \,]$

Causation: $\Pr[\, B = b \mid \mathrm{do}\ A \leftarrow a \,]$

e**X**ercise

over-**W**eight

Heart disease

Scenario 1:

$$X \leftarrow B(1/2)$$

$$W \leftarrow \begin{cases} 0, & X = 1 \\ B(1/2), & X = 0 \end{cases} \qquad H \leftarrow \begin{cases} 0, & X = 1 \\ B(1/2), & X = 0 \end{cases}$$

X	W	H	Prob
1	0	0	1/2
0	0	0	1/8
0	0	1	1/8
0	1	0	1/8
0	1	1	1/8

Scenario 2:

$$W \leftarrow B(1/4)$$

$$X \leftarrow \begin{cases} 0, & W = 1 \\ B(1/3), & W = 0 \end{cases}$$

$$H \leftarrow \begin{cases} 0, & X = 1 \\ B(1/2), & X = 0 \end{cases}$$

	Scenario 1	Scenario 2
$\Pr[W = 1 \mid X = 0]$	1/2	1/2
$\Pr[W = 1 \mid \mathrm{do}\ X \leftarrow 0\,]$	1/2	1/4

Correlation: $\Pr[\, B = b \mid A = a \,]$

Causation: $\Pr[\, B = b \mid \text{do } A \leftarrow a \,]$

e**X**ercise

over-**W**eight

Heart disease

Scenario 1:

$X \leftarrow B(1/2)$

$$W \leftarrow \begin{cases} 0, & X = 1 \\ B(1/2), & X = 0 \end{cases} \qquad H \leftarrow \begin{cases} 0, & X = 1 \\ B(1/2), & X = 0 \end{cases}$$

X	W	H	Prob
1	0	0	1/2
0	0	0	1/8
0	0	1	1/8
0	1	0	1/8

Scenario 2:

$W \leftarrow B(1/4)$

$$X \leftarrow \begin{cases} 0, & W = 1 \\ B(1/3), & W = 0 \end{cases}$$

$$H \leftarrow \begin{cases} 0, & X = 1 \\ B(1/2), & X = 0 \end{cases}$$

	Scenario 1	Scenario 2
$\Pr[W = 1 \mid X = 0]$	1/2	1/2
$\Pr[W = 1 \mid \text{do } X \leftarrow 0\,]$	1/2	1/4

Cannot distinguish Scenario 1 and 2 from observations alone!

Estimating causal probabilities

$$
\begin{tikzcd}
 & X \leftarrow B(1/2) \arrow[dl] \arrow[dr] & \\
W \leftarrow \begin{cases} 0, & X = 1 \\ B(1/2), & X = 0 \end{cases} & & H \leftarrow \begin{cases} 0, & X = 1 \\ B(1/2), & X = 0 \end{cases}
\end{tikzcd}
$$

Assume: Know causal graph

Goal: Compute $\Pr[A = a \,|\mathrm{do}\, B \leftarrow b]$

$$\Pr[H = 1 | W = 0] = 1/6$$

$$\Pr[H = 1 | \mathrm{do}\, W \leftarrow 0] = 1/4$$

X	W	H	Prob
1	0	0	1/2
0	0	0	1/8
0	0	1	1/8
0	1	0	1/8
0	1	1	1/8

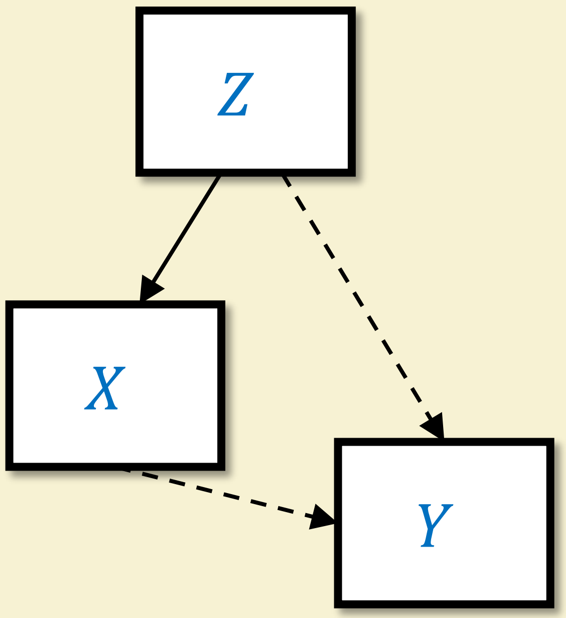
Known from observations

Controlling for X:

$$\Pr[H = 1 | \mathrm{do}\, W \leftarrow 0] = \Pr[H = 1 | W = 0, X = 0] \Pr[X = 0]$$
$$+ \Pr[H = 1 | W = 0, X = 1] \Pr[X = 1]$$

Apriori unknown

Adjustment formula



Known* from observations

$$\Pr[\, Y = y \mid \mathrm{do}\ X \leftarrow x \,] = \sum \Pr[Y = y \mid X = x, Z = z\,] \cdot \Pr[Z = z]$$

Apriori unknown

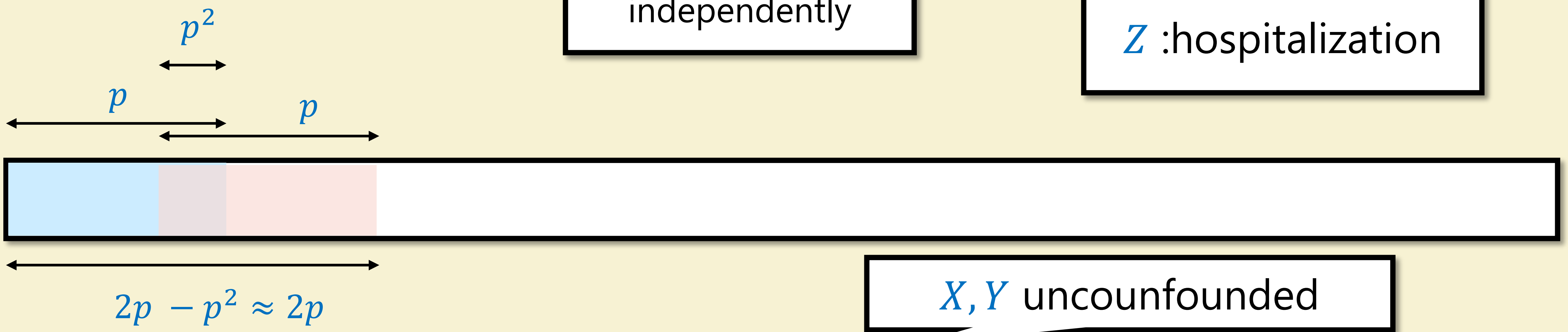
Control for wrong things

X :disease 1

Y :disease 2

Z :hospitalization

Both w prob $p \ll 1$ independently

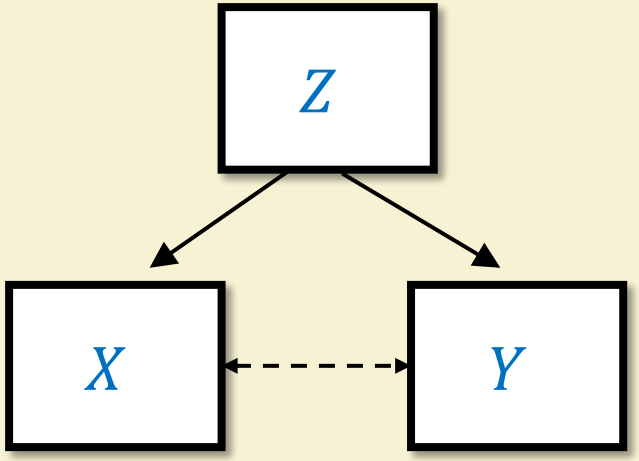


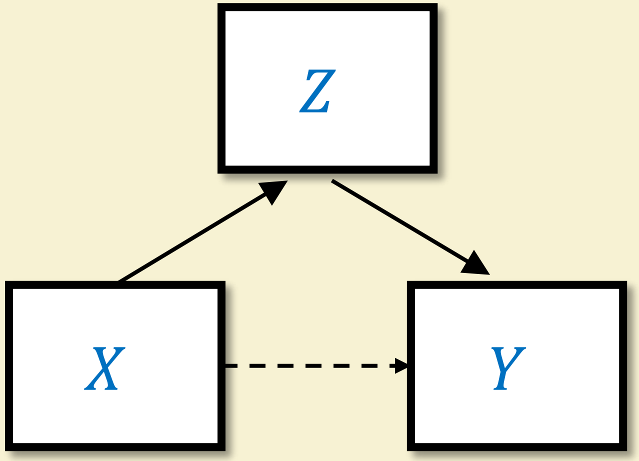
X, Y uncounfounded

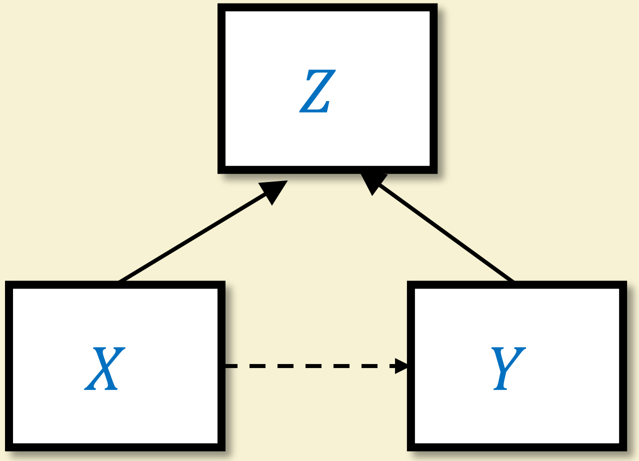
$$\Pr[X = 1 | Y = 1] = \Pr[X = 1 \,|\mathrm{do}\; Y \leftarrow 1] = p$$

Controlling for Z:

$$\underbrace{\Pr[X = 1 | Y = 1, Z = 1]}_{\approx \frac{p^2}{2p} = \frac{p}{2}} \cdot \underbrace{\Pr[Z = 1]}_{\approx 2p} + \underbrace{\Pr[X = 1 | Y = 1, Z = 0]}_{= 0} \cdot \Pr[Z = 0] \quad \approx p^2$$





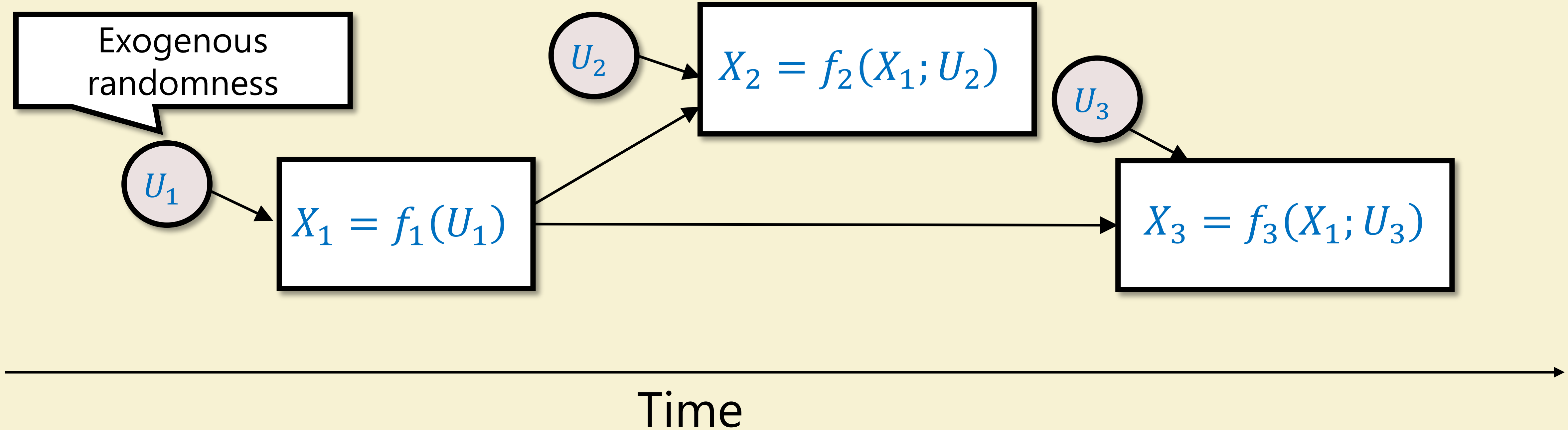


	Fork	Mediator	Collider
$\Pr[Y = y \mid do\ X \leftarrow x]$ vs $\Pr[Y = y \mid X = x]$	$\neq$	$=$	$=$
$\Pr[Y = y \mid do\ X \leftarrow x]$ vs $\sum \Pr[Y = y \mid X = x, Z] \Pr[Z]$	$=$	$\neq$	$\neq$

Casual Models

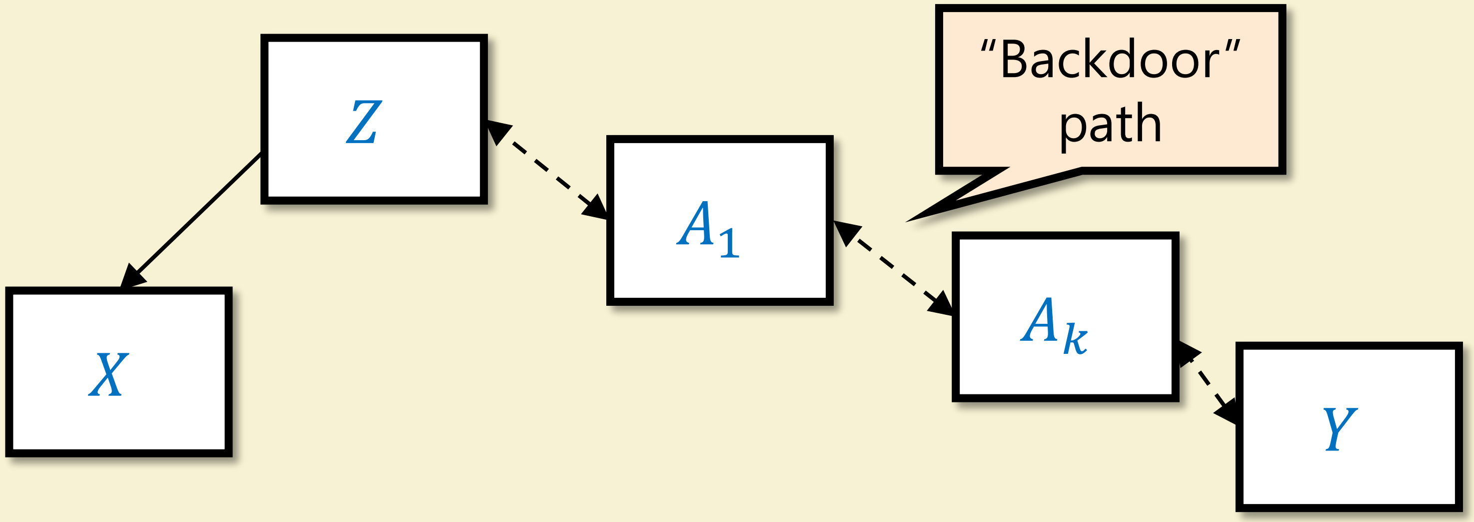
“Frequentist”:
$\Pr[\ A \mid do\ B\ ]$ is frequency of times that A occurs if we do B

“Bayesian”:
$\Pr[\ A \mid do\ B\ ]$ is probability A would have happened in “counter-factual” world where we did B



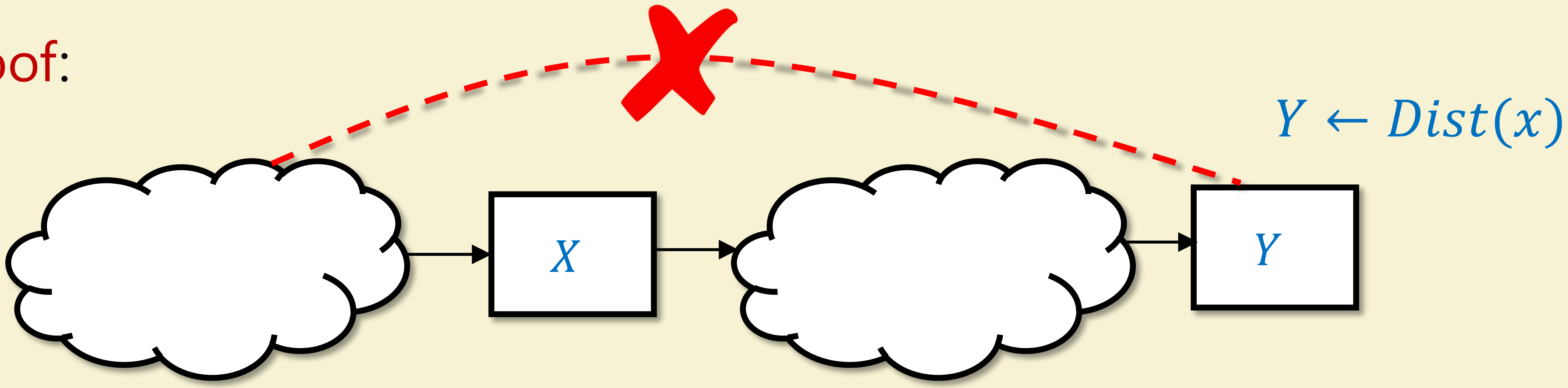
Backdoors

Def: X, Y are confounded if

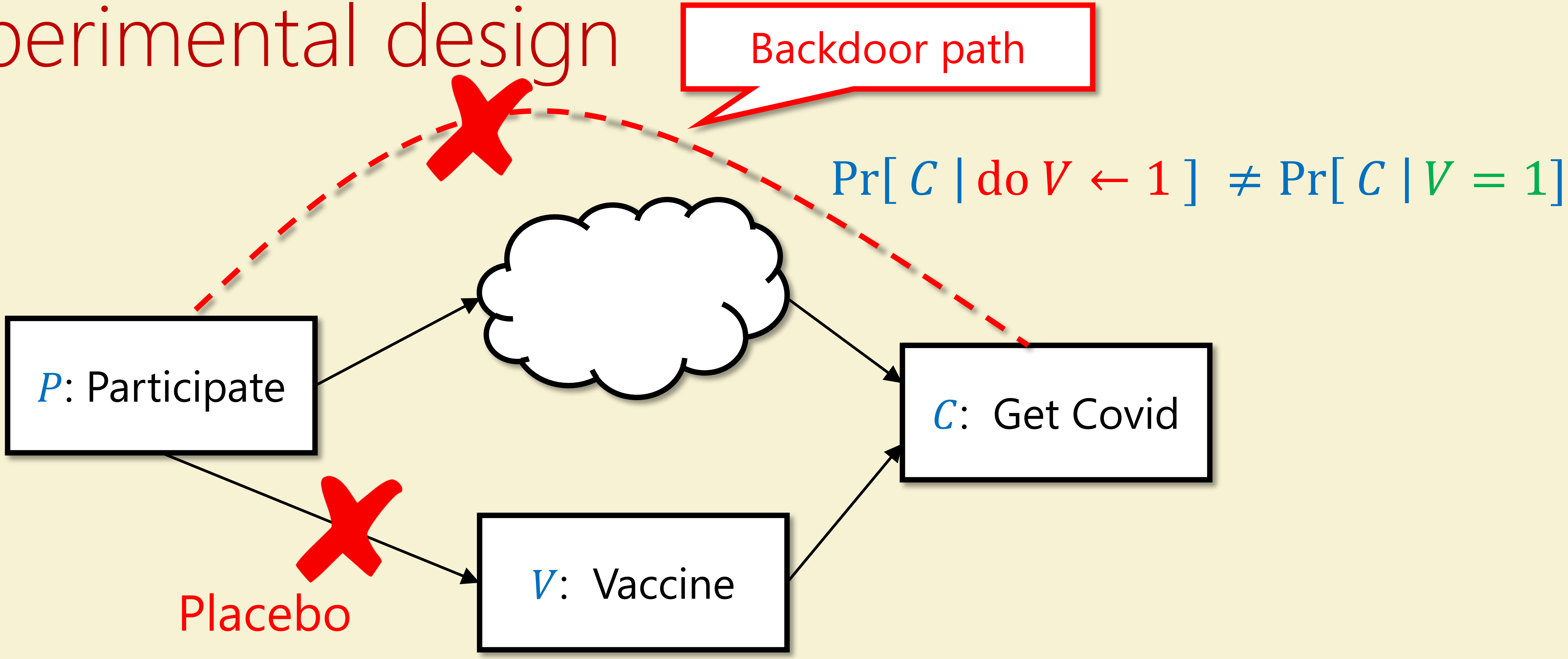


Thm: If X, Y not confounded then $\Pr[\,Y = y \mid \mathrm{do}\ X \leftarrow x] = \Pr[\,Y = y \mid X = x]$

Proof:

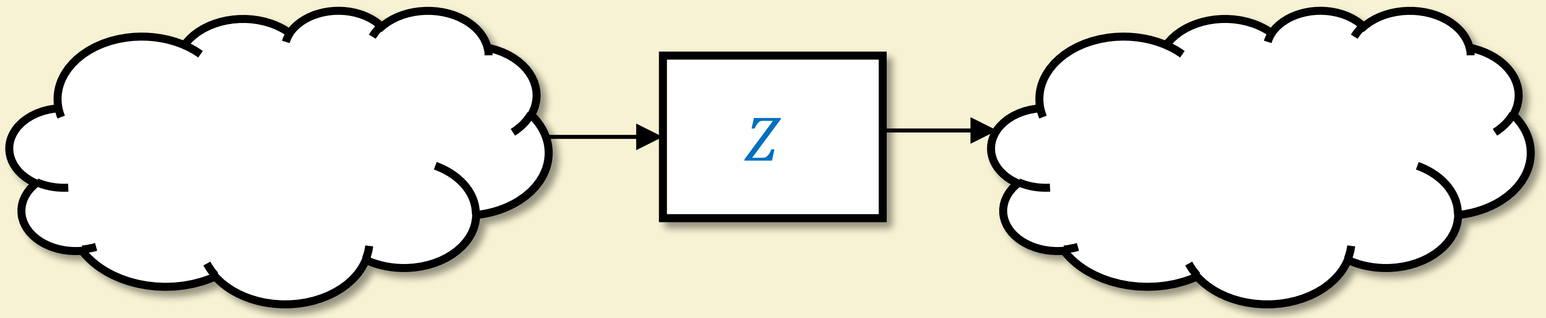


Experimental design

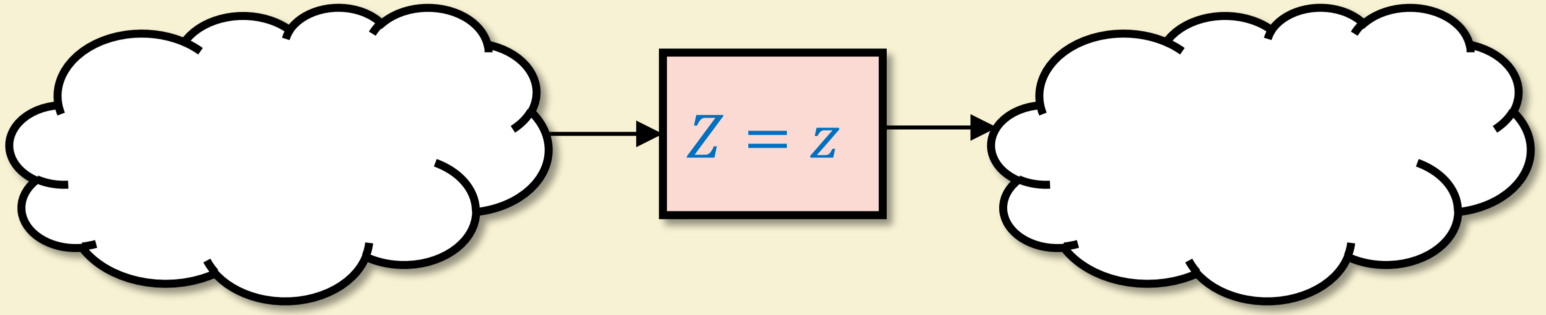


Treatment effect: $\Pr[C \mid \mathrm{do}\, V \leftarrow 1\,, P\,]$ vs $\Pr[\, C \mid \mathrm{do}\, V \leftarrow 0, P\,]$

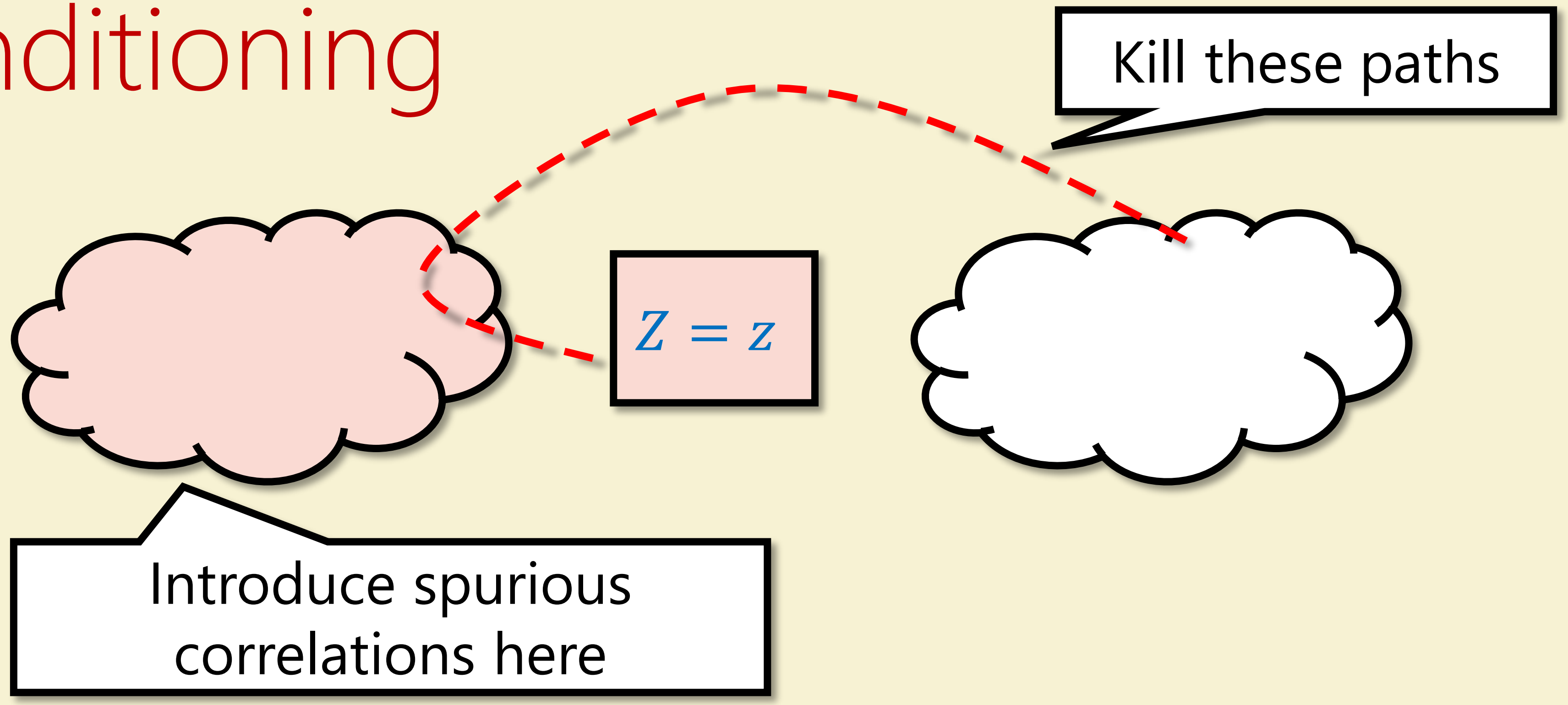
Conditioning



Conditioning



Conditioning



Average Treatment Effect

$T \in \{0,1\}$ – Treatment variable

$$Z \rightarrow T \rightarrow Y$$

$Y_t : Y \mid \mathrm{do}\ T \leftarrow t$

Goal: Estimate $\mathbb{E}[Y_1] - \mathbb{E}[Y_0]$

aka Z "admissable"

Def: T, Y "ignorable" controlling for Z if:

$T \perp (Y_0, Y_1) \mid Z$ i.e: choice of $T = 0,1$ independent of $Y|\mathrm{do}\ T \leftarrow t$

Average Treatment Effect

$T \in \{0,1\}$ – Treatment variable

Goal: Estimate $\mathbb{E}[Y_1] - \mathbb{E}[Y_0]$

Def: T, Y "ignorable" controlling for Z if:

$T \perp (Y_0, Y_1) \mid Z$ i.e: choice of $T = 0,1$ independent of $Y | \mathrm{do}\, T \leftarrow t$

Claim: If T, Y ignorable controlling for Z then

$$\Pr[\, Y = y \mid \mathrm{do}\, T \leftarrow t\,] = \sum \Pr[Y = y \mid T = t, Z = z] \Pr[Z = z]$$

Pf:

$$\sum \Pr[Y = y \mid T = 0, Z = z] \Pr[Z = z] = \sum \Pr[\, Y_0 = y \mid Z = z] \Pr[\, Z = z]$$

Propensity scores:

Learn model $e(z) \approx \mathbb{E}[T|Z = z\,]$

Let $e(z) = \mathbb{E}[T|Z = z]$

CLAIM: If Z admissible, $\mathbb{E}[\,Y \mid \text{do } T \leftarrow 1] = \mathbb{E}\left[\frac{Y \cdot T}{e(Z)}\right]$

Pf: $\Pr[Y = y \mid \text{do } T \leftarrow 1] = \sum_z \Pr[\,Y = y \mid T = 1, z]\Pr[z]$

For $y \neq 0$

$$= \sum_z \Pr[z] \frac{\Pr[Y=y, T=1|z]}{\Pr[\,T=1|z]} \quad = \mathbb{E}_Z\left[\frac{\Pr[Y=y, T=1\,|z]}{e(Z)}\right] \quad = \mathbb{E}_Z\left[\frac{\Pr[YT=y|z]}{e(Z)}\right]$$

$$\mathbb{E}\,[\,Y \mid \text{do } T \leftarrow 1] = \sum_y \Pr[Y = y \mid \text{do } T \leftarrow 1]\ \cdot y$$

$$= \sum_y \mathbb{E}_Z\left[\frac{\Pr[YT = y|z]\,y}{e(Z)}\right] \quad = \mathbb{E}_Z\left[\frac{Y \cdot T}{e(Z)}\right] \qquad \blacksquare$$

Double ML

Learn model $e(z) \approx \mathbb{E}[T|Z = z]$

Let $e(z) = \mathbb{E}[T|Z = z]$

Assume $Y = \psi(Z) + \tau \cdot T + Noise$

τ = treatment effect

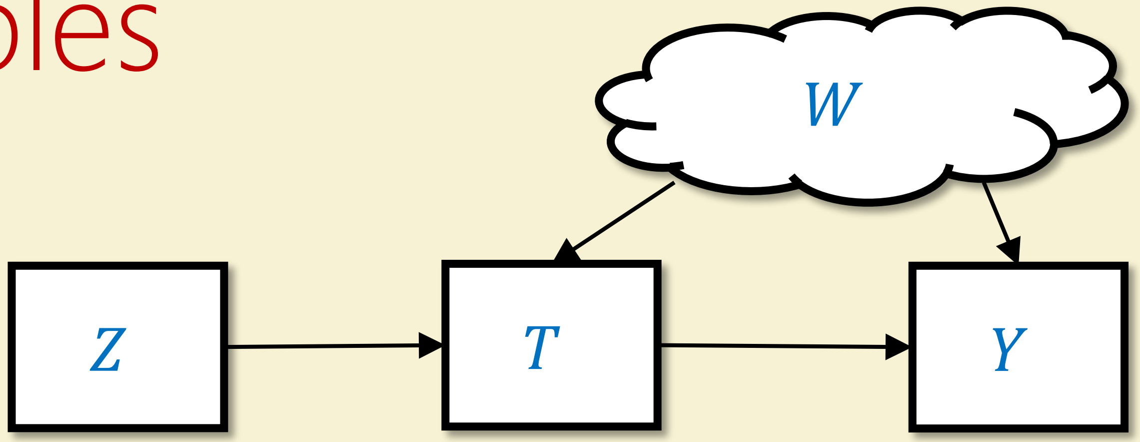
Observe (Z, T, Y) , learn model $f(z) \approx \mathbb{E}[Y|Z = z]$

$$f(z) \approx \psi(Z) + \tau \cdot e(z)$$

$$\Rightarrow \quad Y - f(z) \approx \tau \cdot (T - e(z))$$

Can estimate from data

Instrumental variables



W is unobserved: can't control for

Assume $Y = \tau \cdot T + f(W)$ $\quad$ $Cov\big(Z, f(W)\big) = 0$

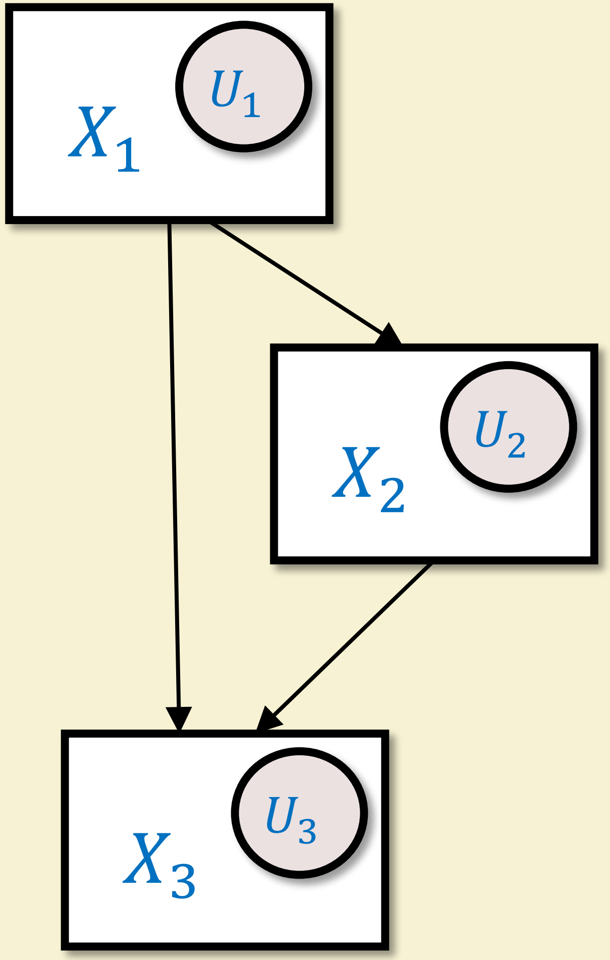
τ = treatment effect

$$\Rightarrow \qquad \tau = \frac{Cov(Z,Y)}{Cov(Z,T)}$$

Counterfactuals

Let u realization of $U_1 \ldots U_n$

$Y_{X \leftarrow x}(u)$ =output of Y if $U = u$ and $X = x$



Fairness

Fairness and machine learning

fairmlbook.org

Fairness and machine learning

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

RESEARCH-ARTICLE

The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making

Authors: Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian Authors Info & Affiliations

(Less)

Publication: Communications of the ACM • March 2021 • https://doi.org/10.1145/3433949

NIPS 2017 Tutorial on Fairness in Machine Learning

Solon Barocas, Moritz Hardt

Note: Focus on fairness in classification, not representation

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

NEWS

Google Algorithm Detects Lung Cancer Better Than Human Doctors

BY STEPHANIE MLOT 05.21.2019 :: 8:

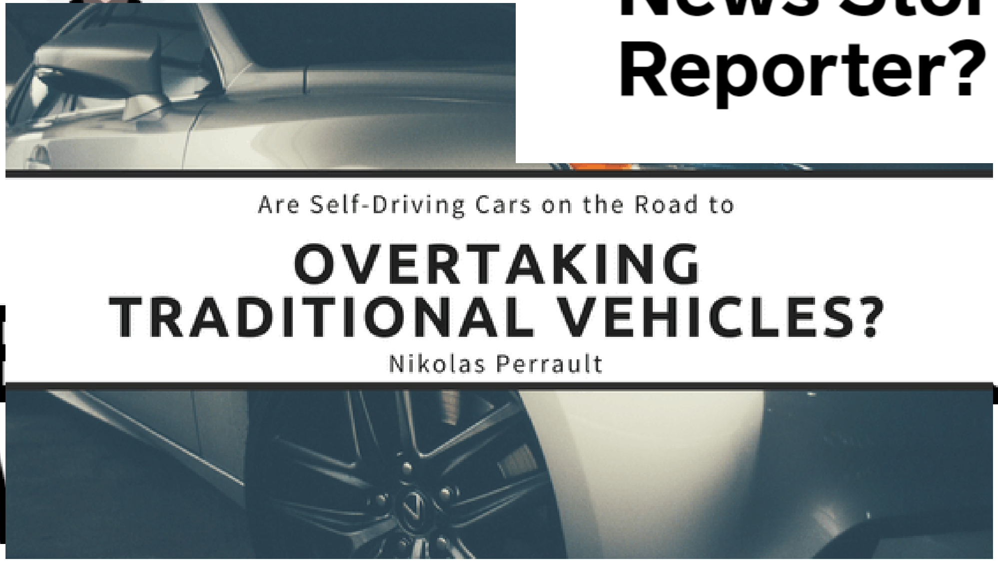
eplaced by Cheaper Softu

UR JOB?

STEVEN LEVY 04.24.12 04:46 PM

By Gary M

Can an Algorithm Write a Better News Story Than a Human Reporter?



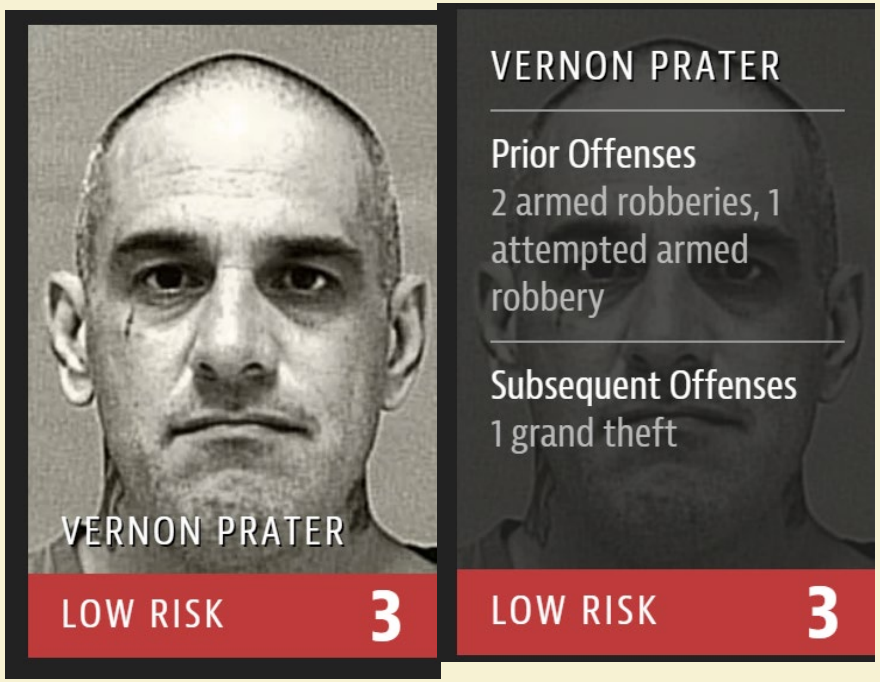
ROBO RECRUITING

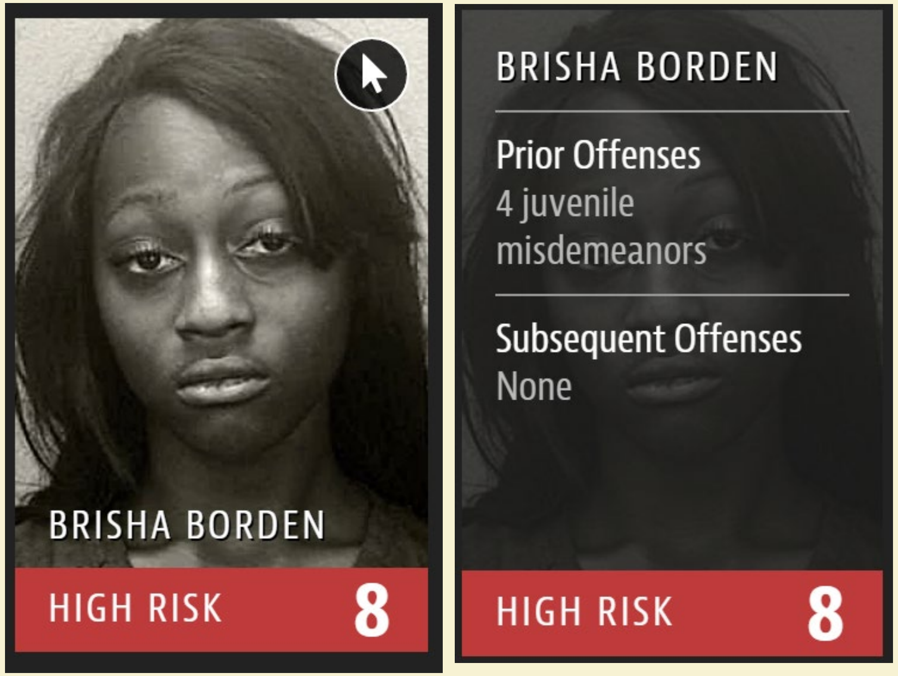
Can an Algorithm Hire Better Than a Human?

Claire Cain Miller @clairecm JUNE 25, 2015

Hiring and recruiting might seem like some of the least likely jobs to be automated. The whole process seems to need human skills that computers

Risk of Recidivism

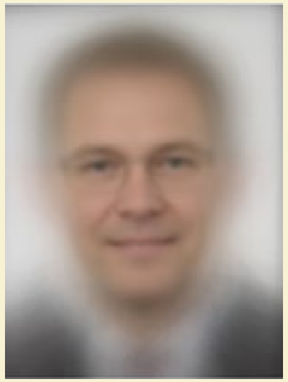




	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Angwin, Larson, Mattu, Kirchner 2016

Gender detection



99.7% correct

65.3% correct

Buolamwini, Gebru, 2018

Non-ML unfairness

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

Marianne Bertrand

Sendhil Mullainathan

AMERICAN ECONOMIC REVIEW
VOL. 94, NO. 4, SEPTEMBER 2004
(pp. 991-1013)

*"White names receive **50 percent more callbacks** for interviews. Callbacks are also more **responsive to resume quality** for White names than for African-American ones."*

Meta-analysis of field experiments shows no change in racial discrimination in hiring over time

Lincoln Quillian, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen
+ See all authors and affiliations

PNAS October 10, 2017 114 (41) 10870-10875; first published September 12, 2017;

Algorithms help?

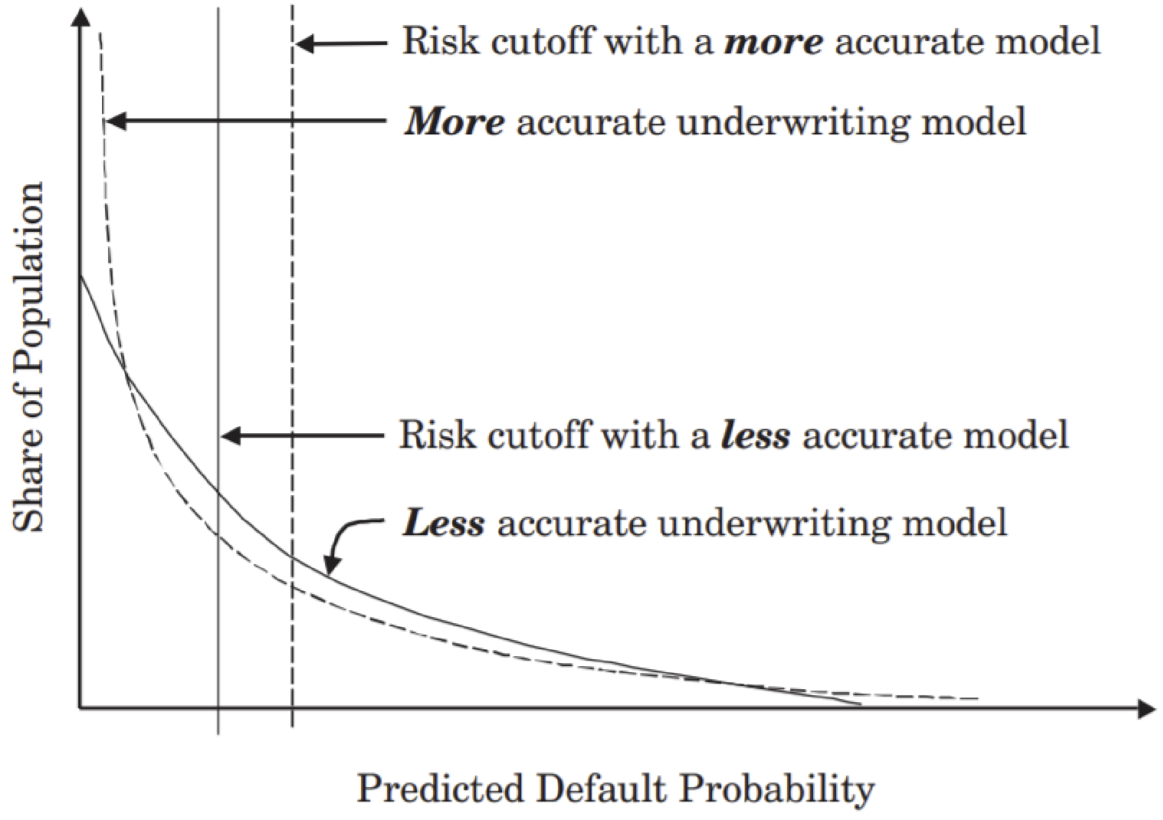
Original Articles

Automated underwriting in mortgage lending: Good news for the underserved?

Susan Wharton Gates, Vanessa Gail Perry & Peter M. Zorn

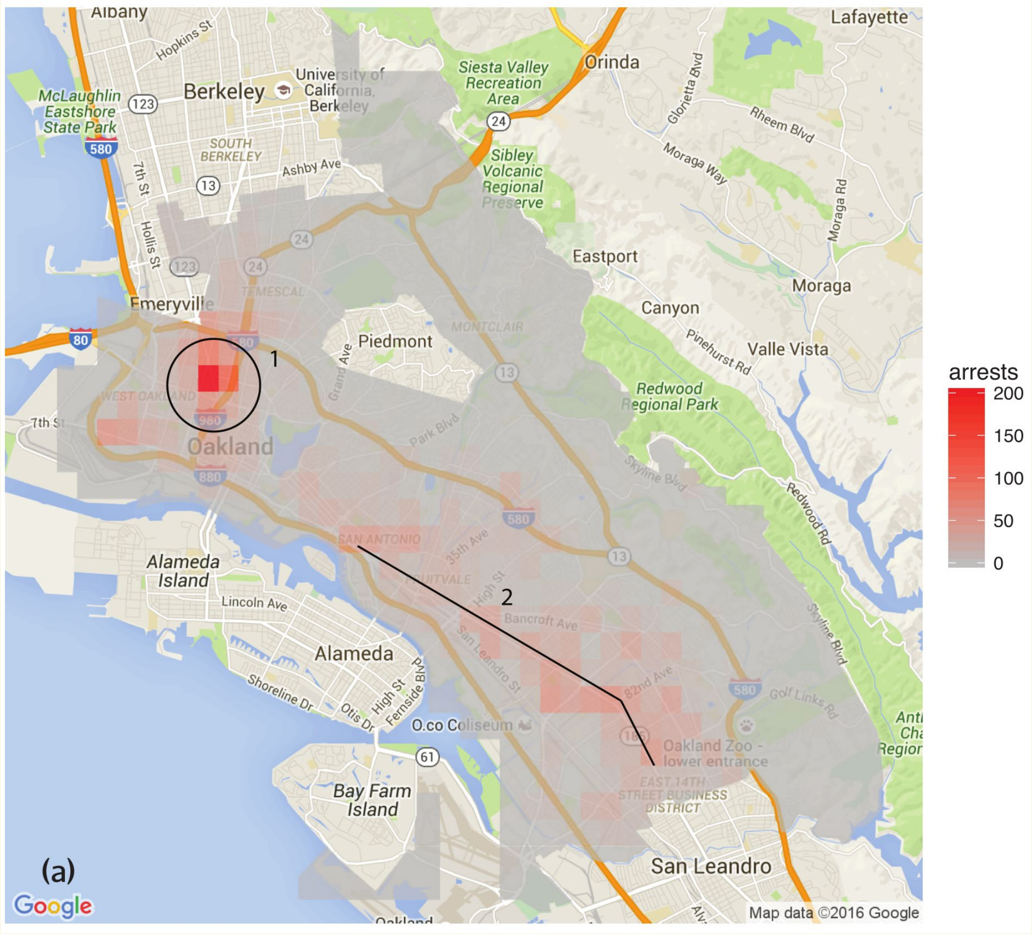
Pages 369-391 | Published online: 31 Mar 2010

Figure 6. **Effect of Introducing More Accurate Underwriting Models**

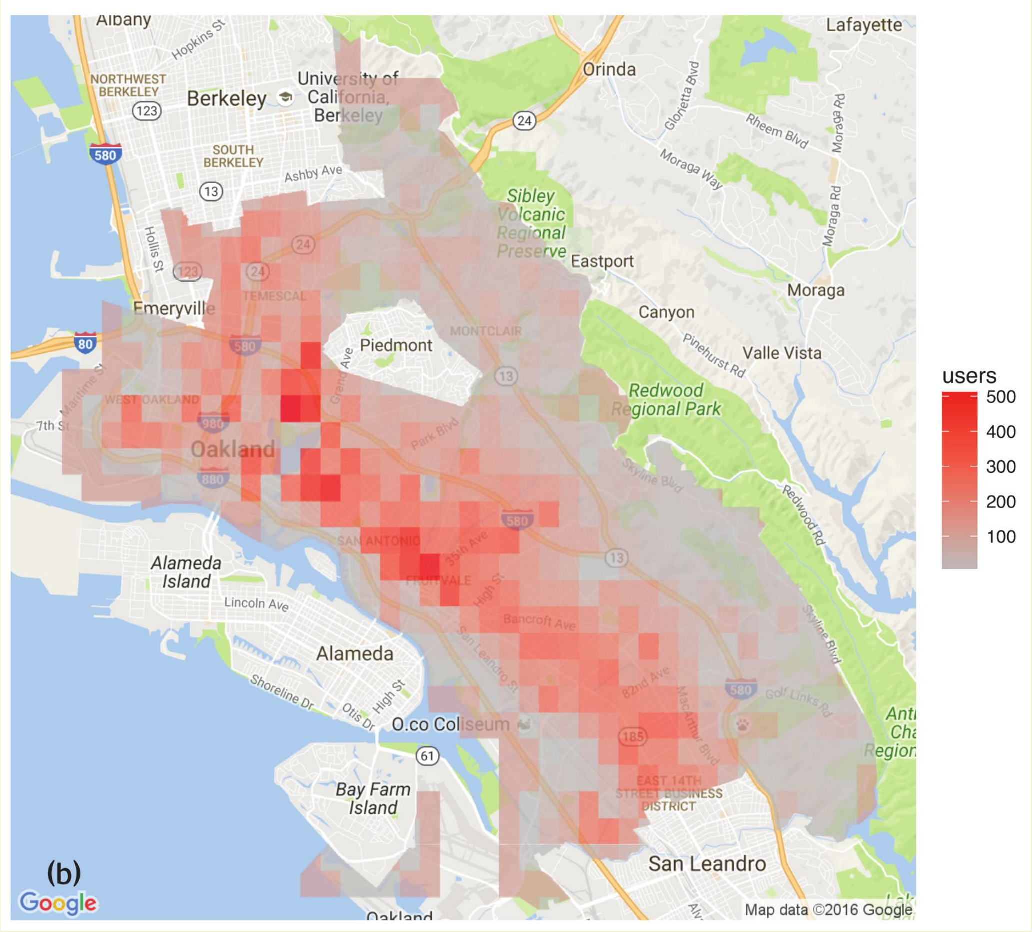


To predict and serve?

Predictive policing systems are used increasingly by law enforcement to try to prevent crime before it occurs. But what happens when these systems are trained using biased data? **Kristian Lum** and **William Isaac** consider the evidence – and the social consequences

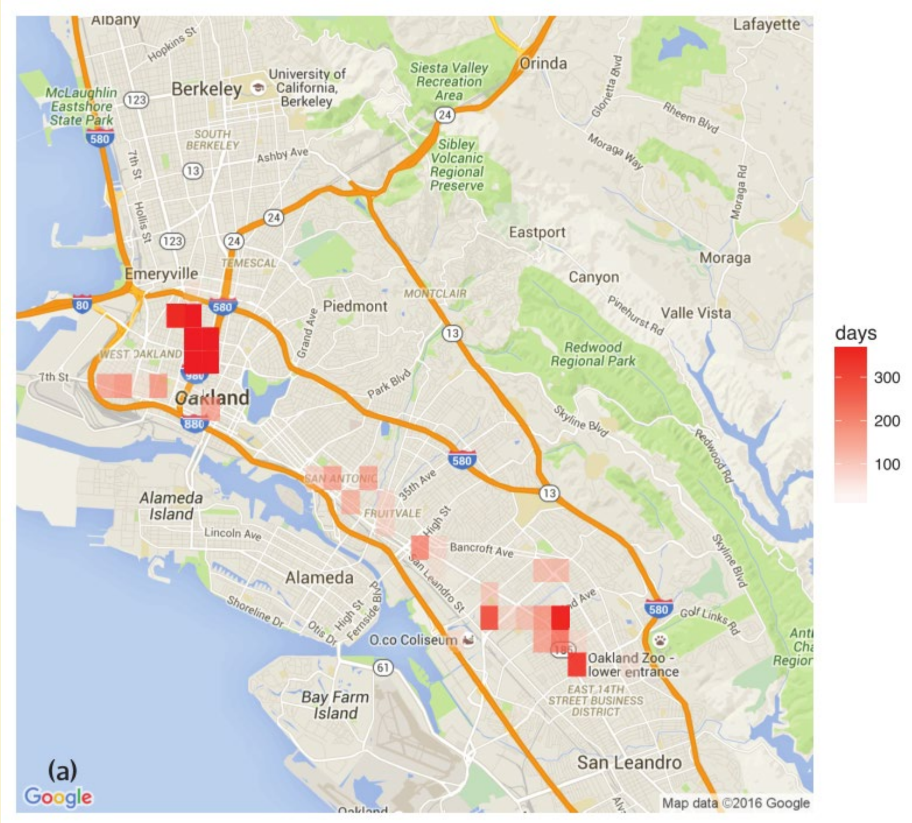


Arrests



Drug usage

Positive feedback loop



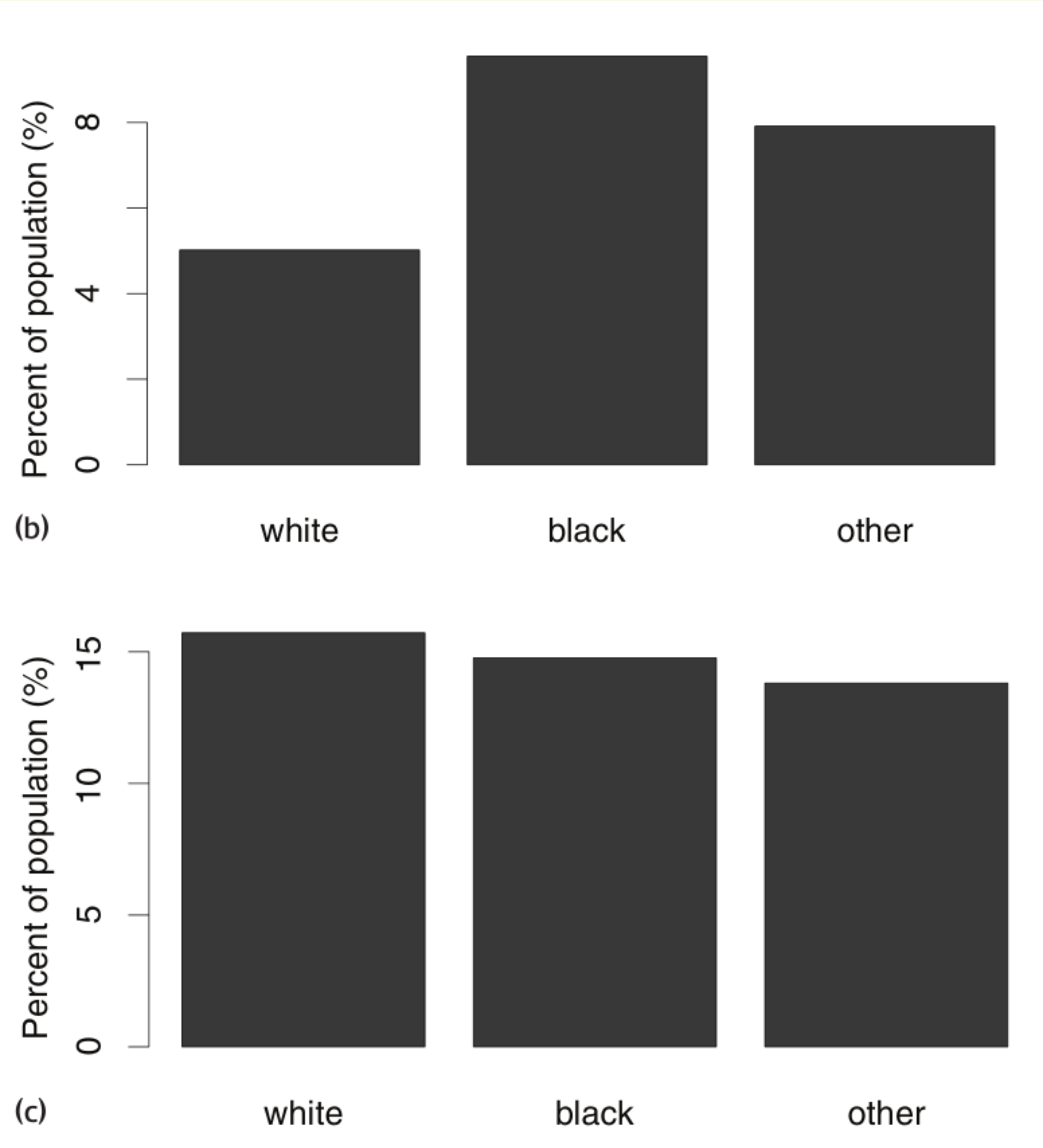


FIGURE 2 (a) Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland police data. (b) Targeted policing for drug crimes, by race. (c) Estimated drug use by race

Predicted crime

Making it formal

Unfairness definitions

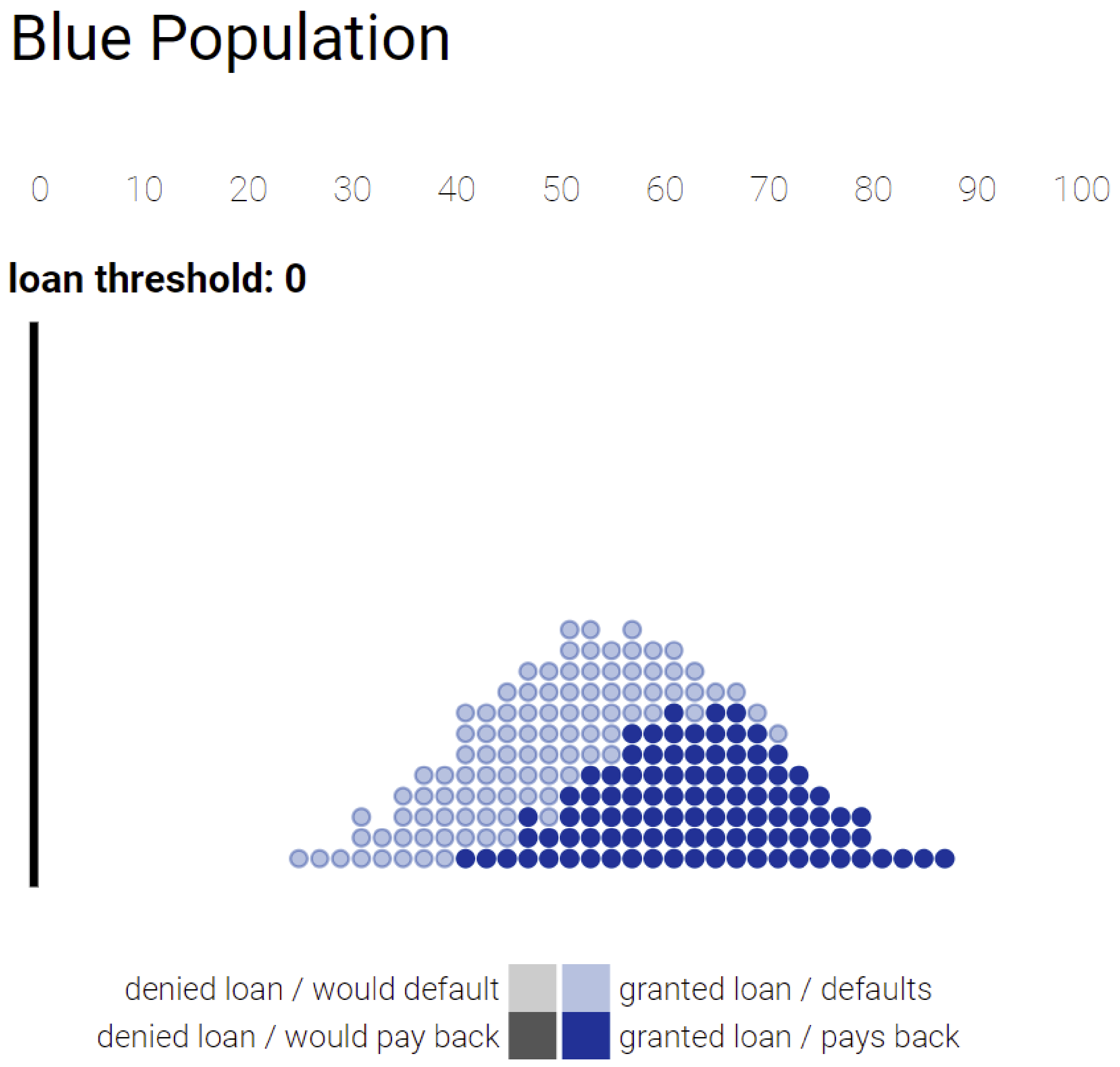
Components:

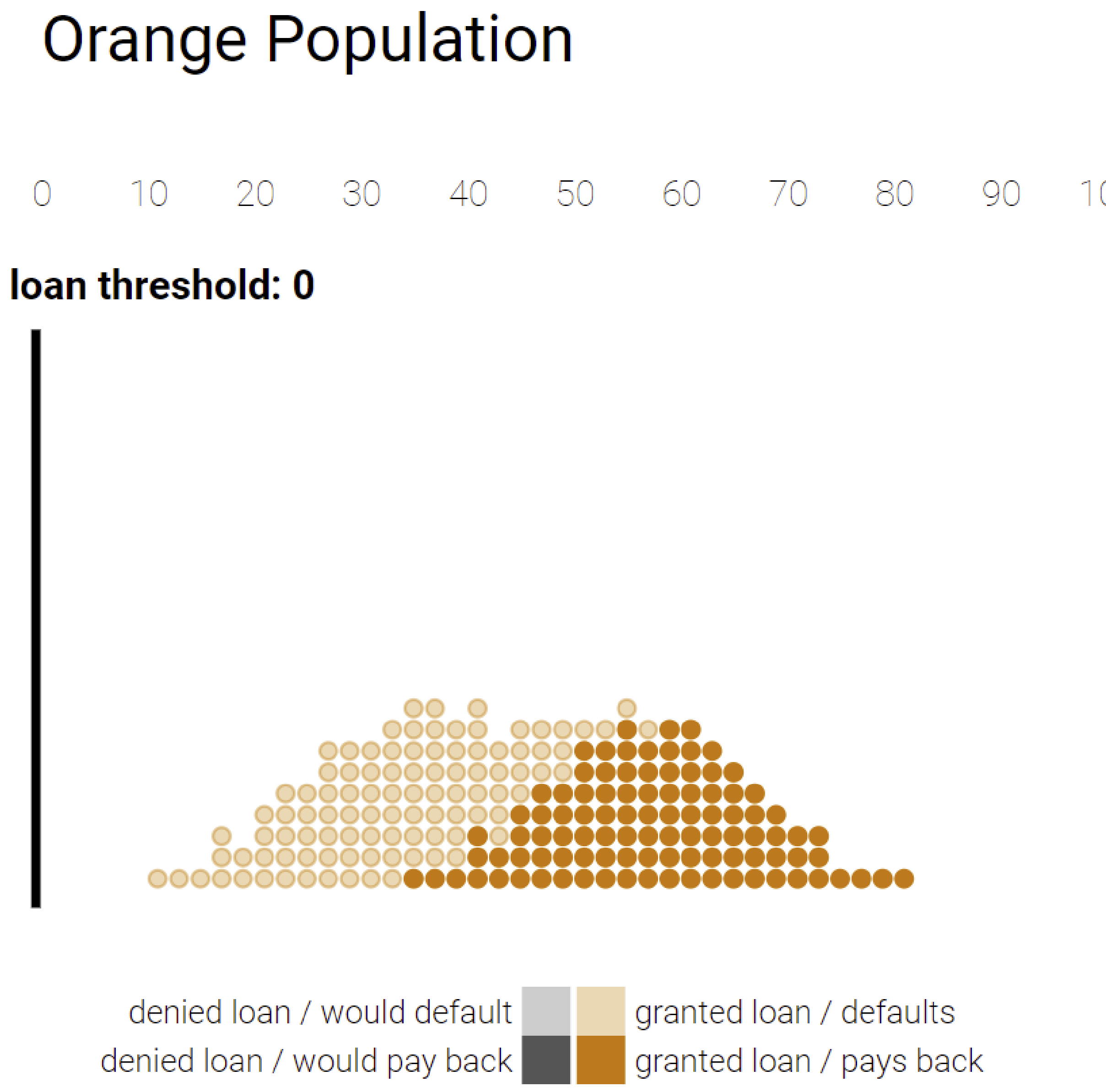
- Protected class*
- Unfairness measurement

Disparate treatment

Disparate impact

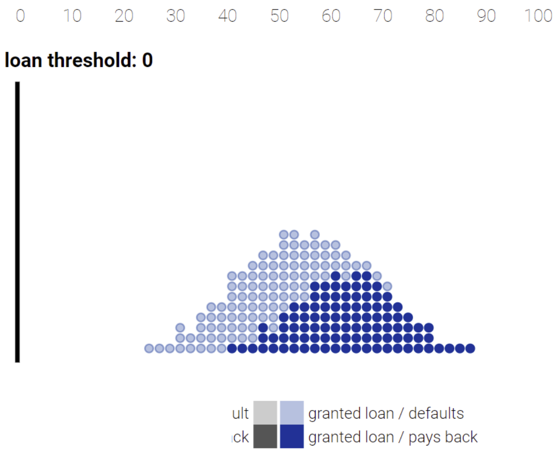
Race (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964); **National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967); **Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)



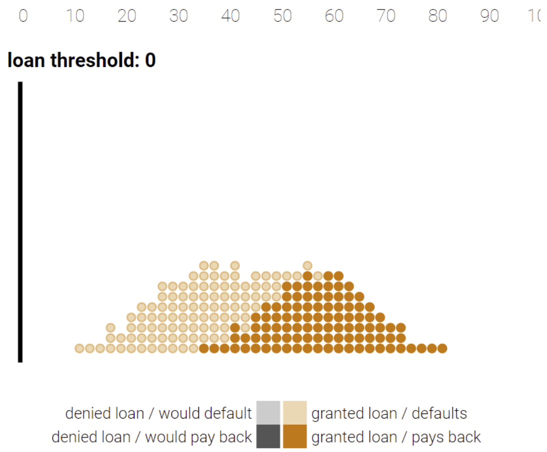


https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Blue Population

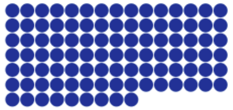


Orange Population

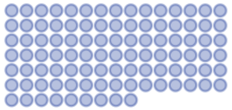


Total profit = -79200

Correct 50%
loans granted to paying applicants and denied to defaulters



Incorrect 50%
loans denied to paying applicants and granted to defaulters

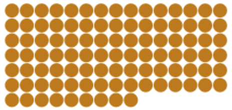


True Positive Rate 100%
percentage of paying applications getting loans

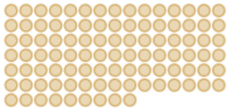
Positive Rate 100%
percentage of all applications getting loans

Profit: -39600

Correct 50%
loans granted to paying applicants and denied to defaulters



Incorrect 50%
loans denied to paying applicants and granted to defaulters



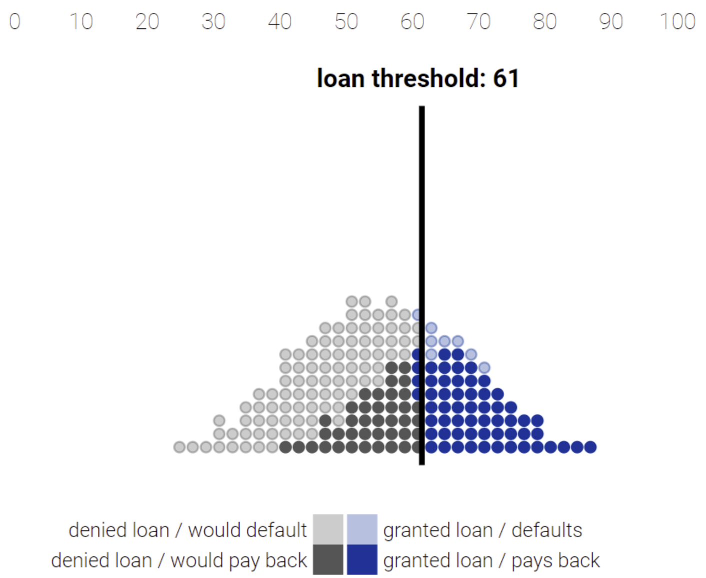
True Positive Rate 100%
percentage of paying applications getting loans

Positive Rate 100%
percentage of all applications getting loans

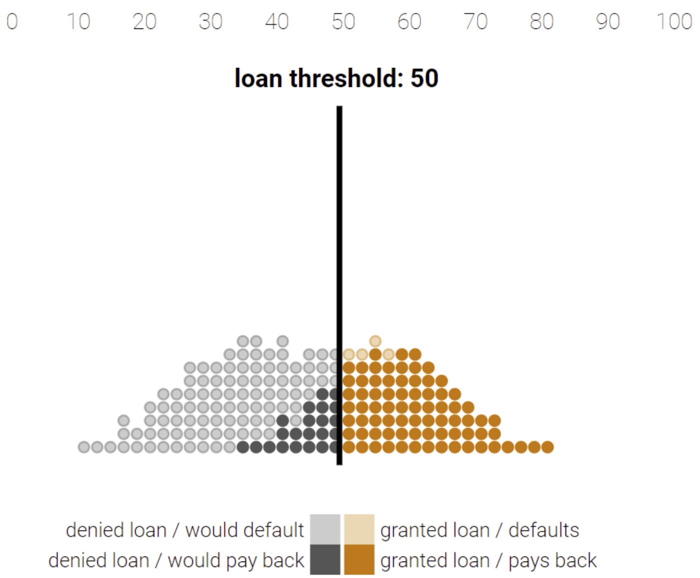
Profit: -39600

Maximize profit

Blue Population

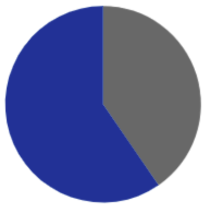


Orange Population



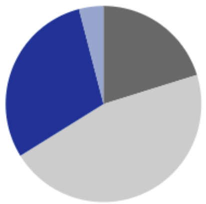
Total profit = **32400**

True Positive Rate 60%
percentage of paying applications getting loans

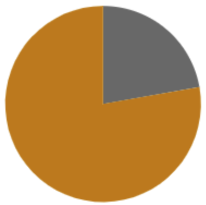


Profit: **12100**

Positive Rate 34%
percentage of all applications getting loans



True Positive Rate 78%
percentage of paying applications getting loans



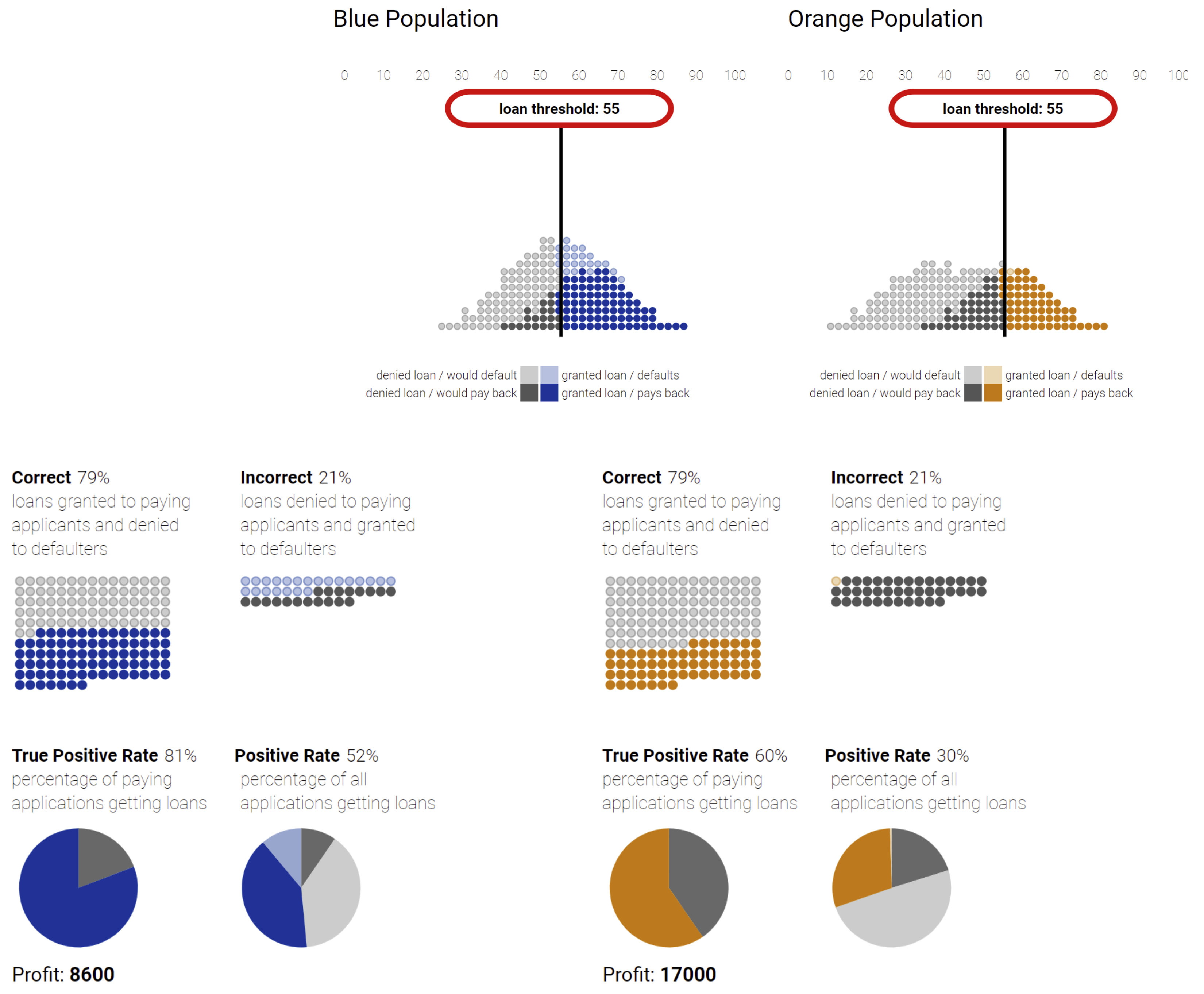
Profit: **20300**

Positive Rate 41%
percentage of all applications getting loans

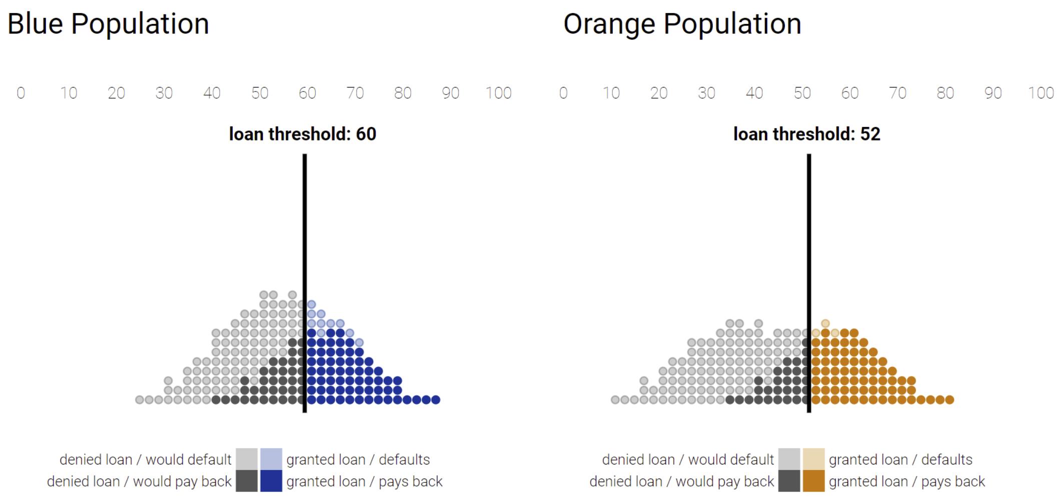
Ignore group

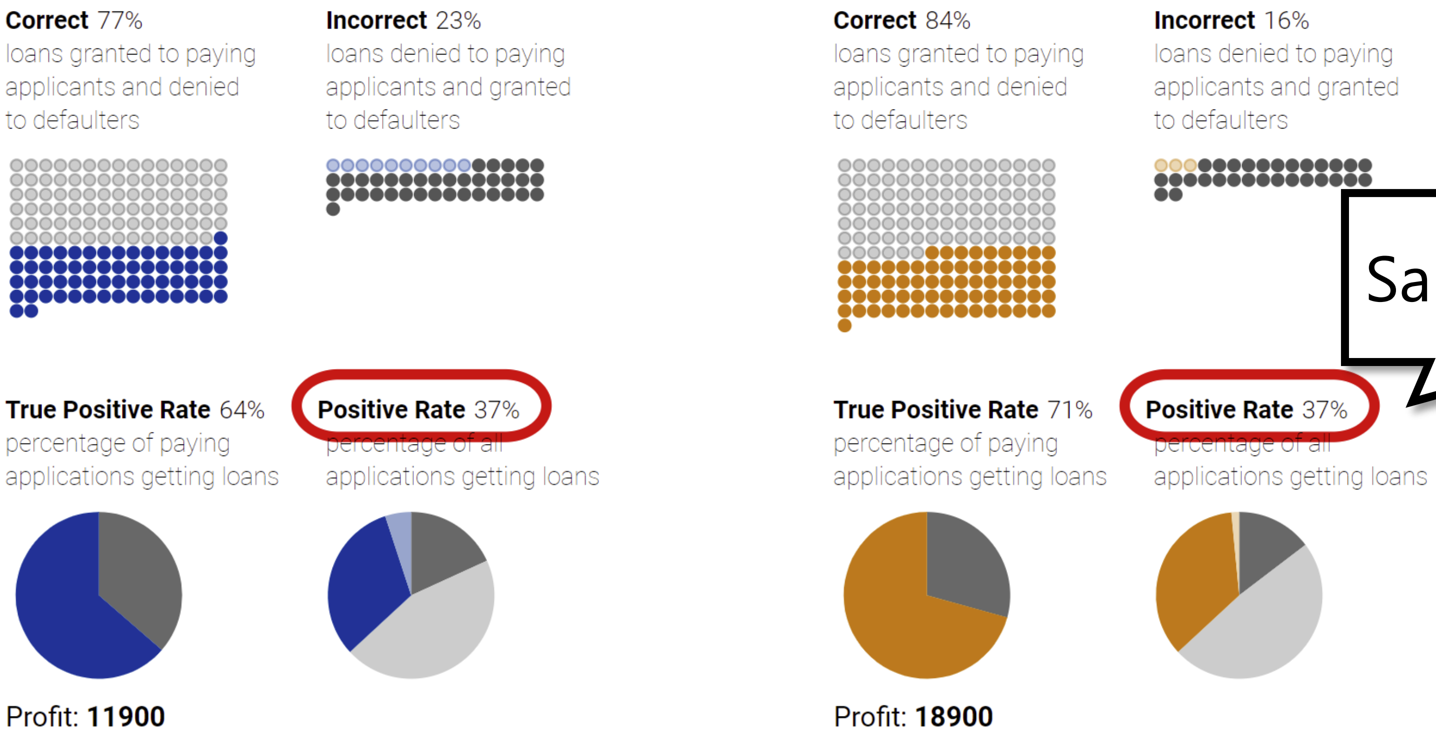
Calibrated from lender POV

Unfair from applicant POV



Demographic parity

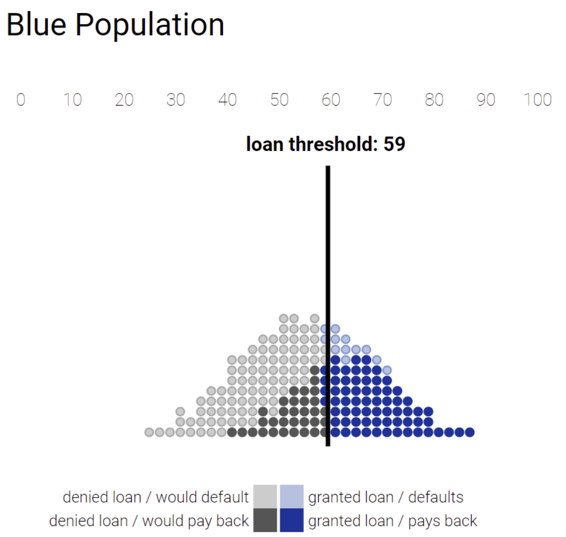


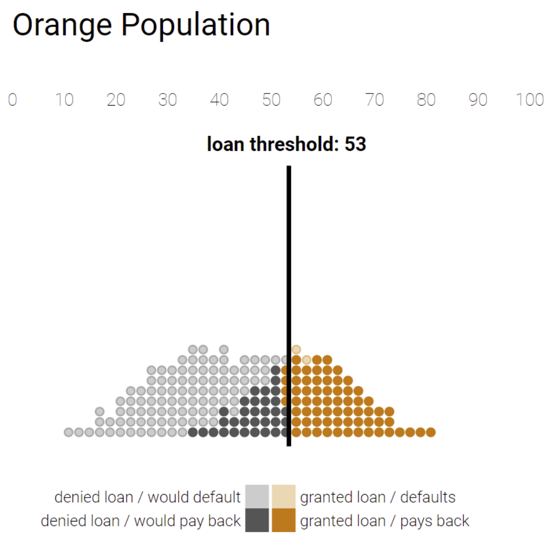


Same total loans

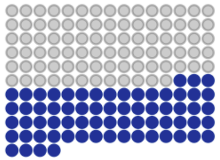
Accuracy advantage split between lender and applicant

Equal opportunity



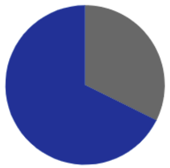


Correct 78%
loans granted to paying applicants and denied to defaulters

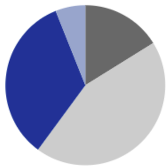


Incorrect 22%
loans denied to paying applicants and granted to defaulters

True Positive Rate 68%
percentage of paying applications getting loans

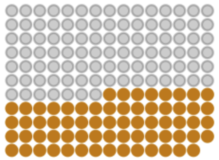


Positive Rate 40%
percentage of all applications getting loans



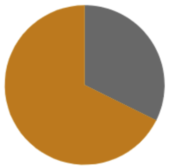
Profit: **11700**

Correct 83%
loans granted to paying applicants and denied to defaulters



Incorrect 17%
loans denied to paying applicants and granted to defaulters

True Positive Rate 68%
percentage of paying applications getting loans



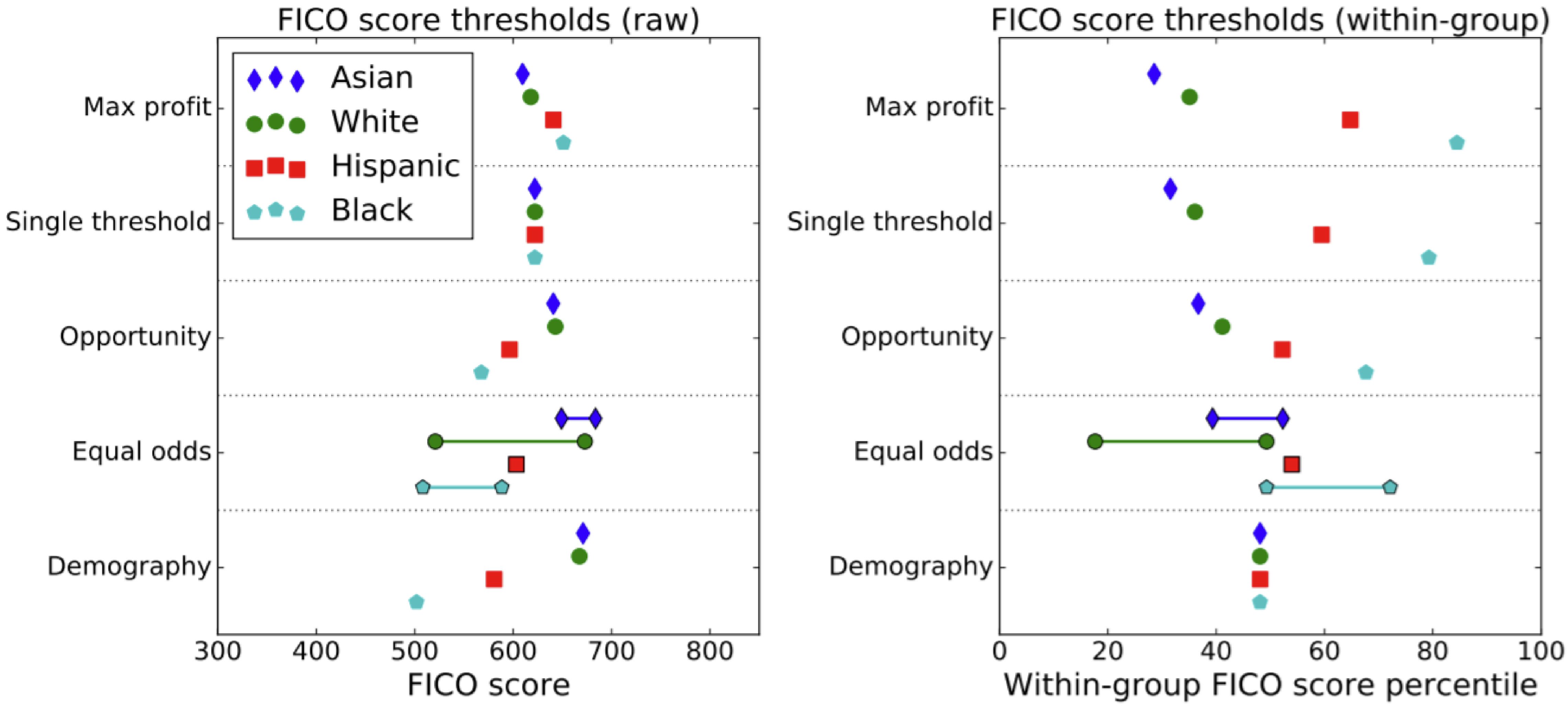
Positive Rate 35%
percentage of all applications getting loans

Profit: **18700**

Fair from applicant POV

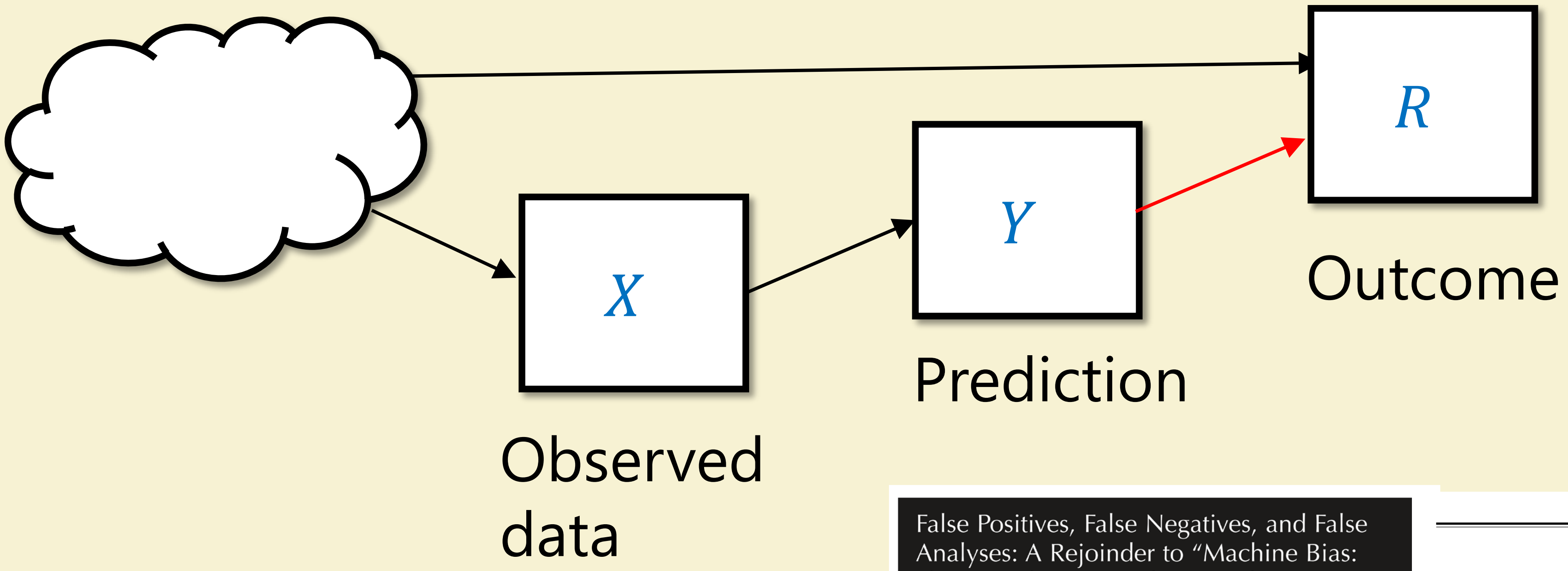
No demographic parity

Real world example: FICO scores



Hardt, Price, Srebro 2016

COMPAS Debate



	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%

Angwin, Larson, Mattu, Kirchner 2016

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. Flo
California State University, Bakersfi
Kristin Bech
Crime and Justice Institute at C
Christopher T. Lowenkar
Administrative Office of the United States Cou
Probation and Pretrial Services Off

COMPAS Risk Scales:
Demonstrating
Accuracy Equity and Predictive Parity

PERFORMANCE
OF THE COMPAS RISK SCALES
IN BROWARD COUNTY

NORTHPOINTE INC.
RESEARCH DEPARTMENT

WILLIAM DIETERICH, PH.D.
CHRISTINA MENDOZA, M.S.
TIM BRENNAN, PH.D.

JULY 8, 2016

Data*

Black

	Low Risk	High Risk
Did not recidivate	1000	800
Recidivate	550	1400

White

	Low Risk	High Risk
Did not recidivate	1150	350
Recidivate	450	500

Defendant POV

$\Pr[HR \mid No\ rec.]$ $\quad \frac{800}{1800} \approx 44\% \quad > \quad \frac{350}{1450} \approx 24\%$

Predictor POV

$\Pr[No\ Rec. \mid HR]$ $\quad \frac{800}{2200} \approx 36\% \quad < \quad \frac{350}{850} \approx 41\%$

Fairness and causaility

Berkeley graduate admissions, 1973

44% of male applicants admitted

35% of female applicants admitted

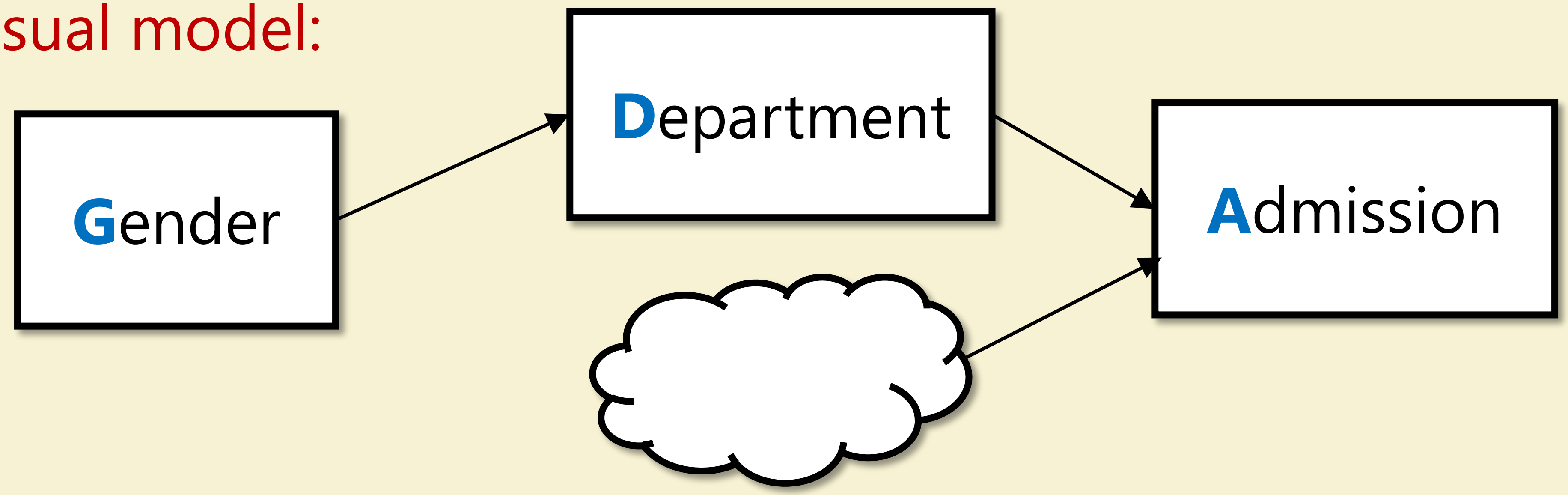
Department level:

Female acceptance rate *higher*

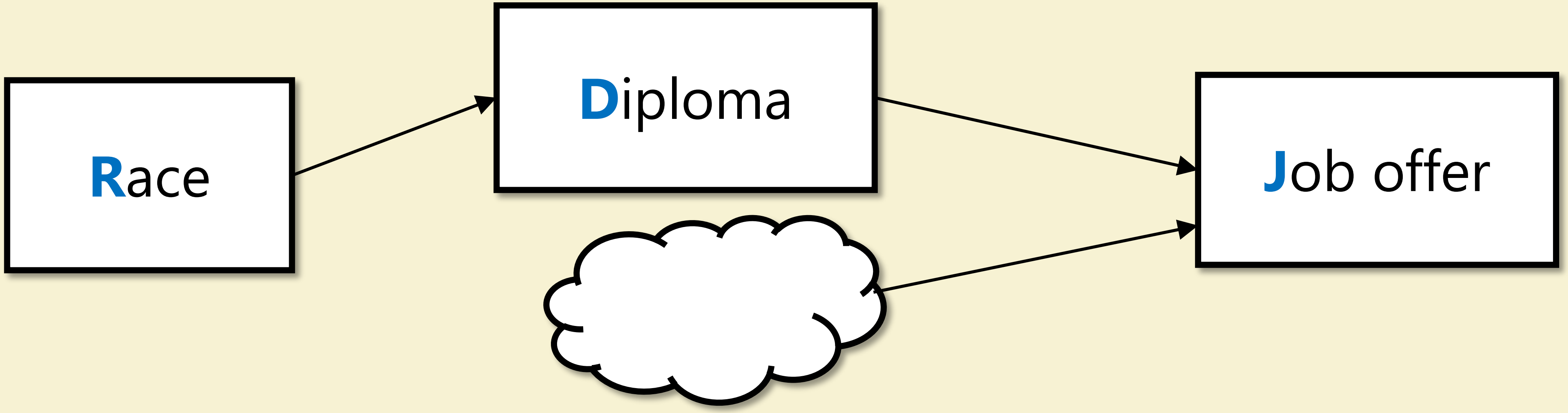
UC Berkeley admissions data from 1973.

	Men		Women	
Department	Applied	Admitted (%)	Applied	Admitted (%)
A	825	62	108	**82**
B	520	60	25	**68**
C	325	**37**	593	34
D	417	33	375	**35**
E	191	**28**	393	24
F	373	6	341	**7**

“Fair” casual model:



Content of boxes matter (e.g. Griggs v. Duke Power Co., 1971)

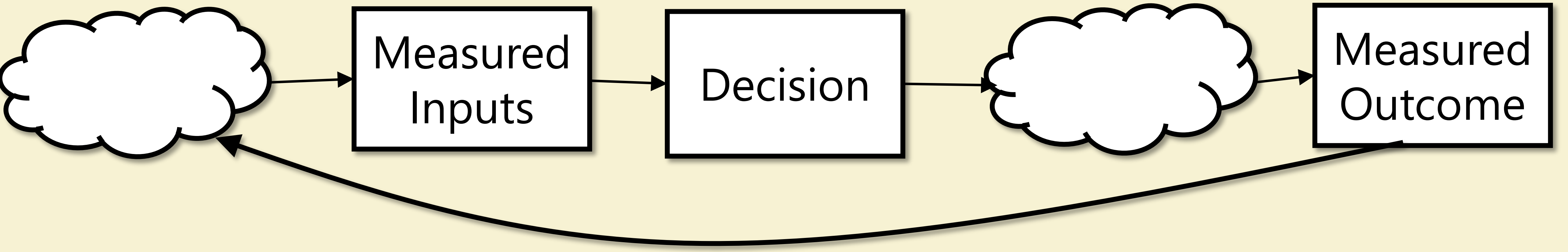


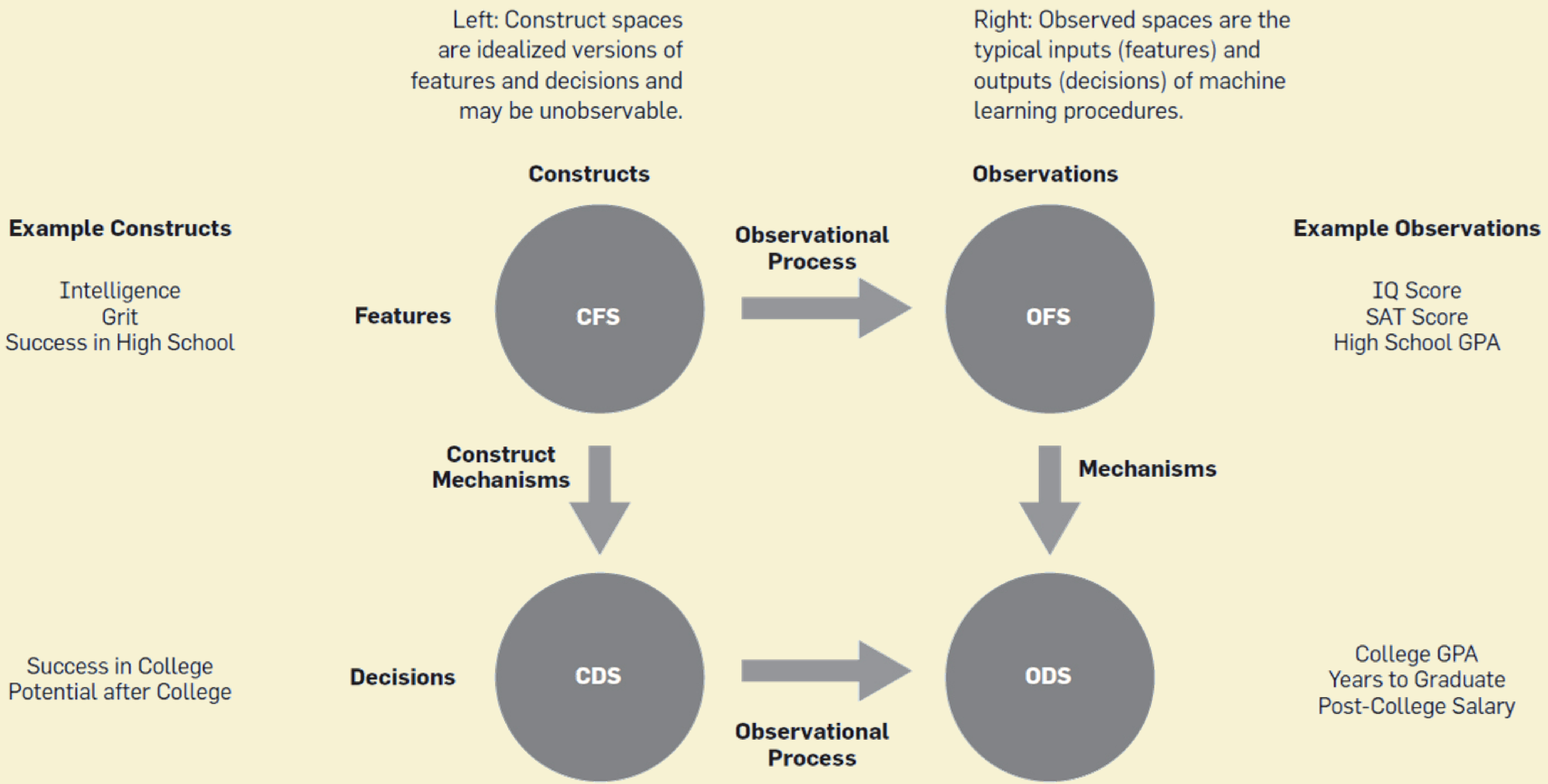
Bottom line

Can't come up with universal observational fairness criteria

Fairness is based on assumptions on:

- Representation of data
- Relation to unmeasured inputs and outcomes
- Causal relation of inputs, predictions, outcomes





Friedler, Scheidegger, Venkatasubramanian 2021