Fair Causal Reasoning

May 22, 2020

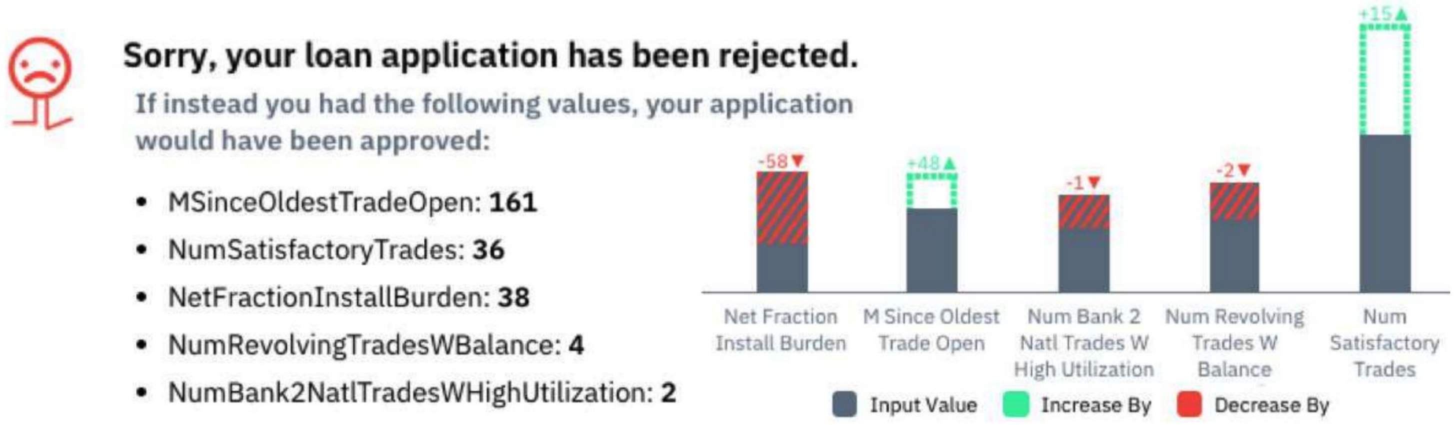
Dr. Wei Wei, Prof. James Landay

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning

Stanford University

Recap

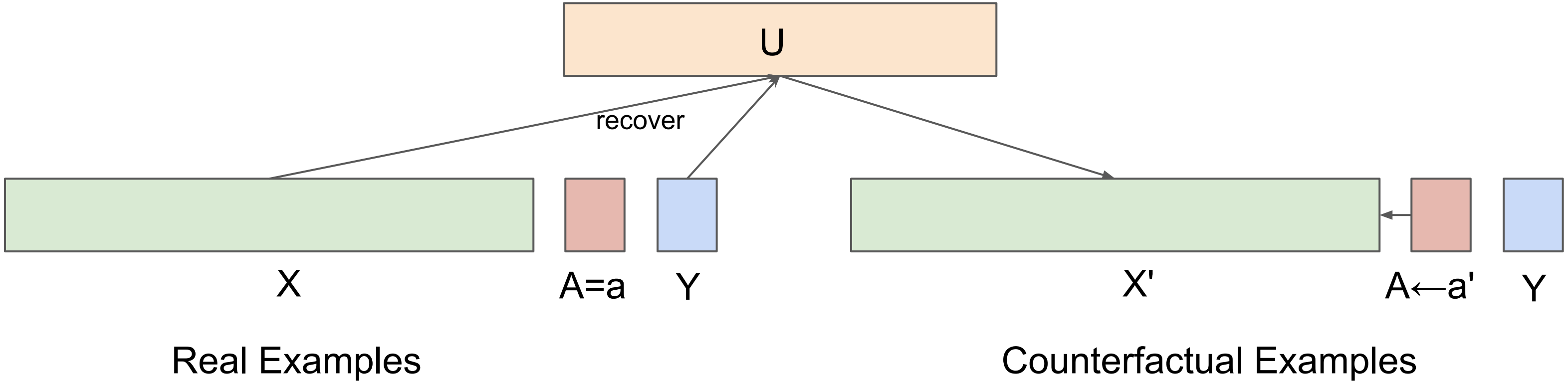
- Counterfactual Explanations



[Grath et al, 2018]()

Recap

- Counterfactual Fairness

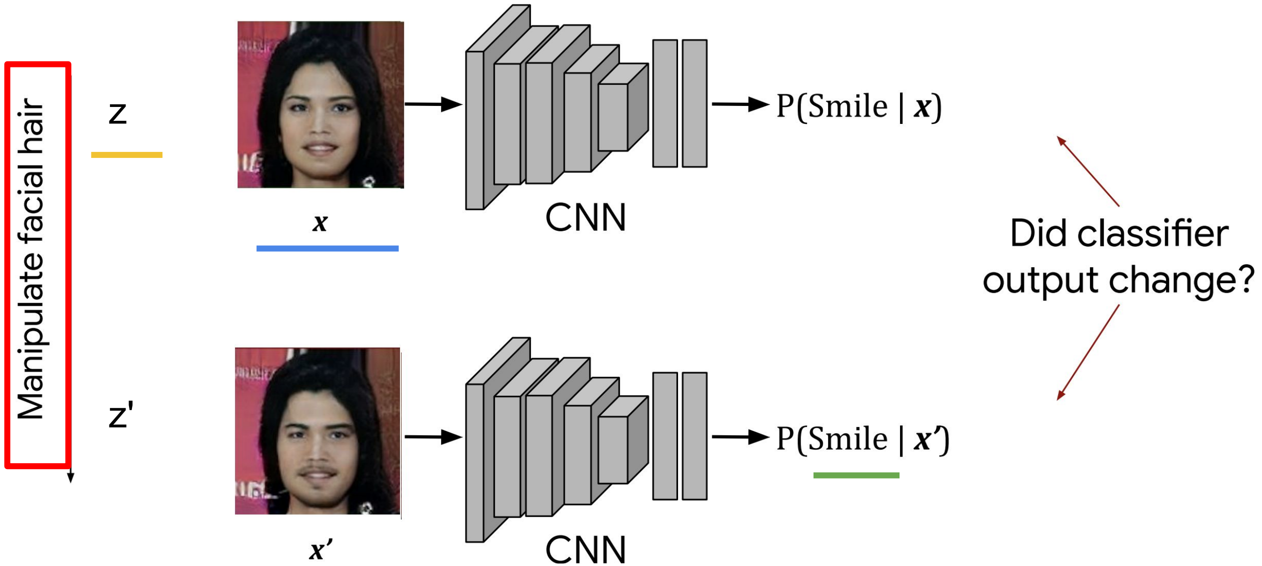


$$\underbrace{P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a)}_{\text{Real Examples}} = \underbrace{P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)}_{\text{Counterfactual Examples}}$$

[Kusner et al, 2017](#)

Recap

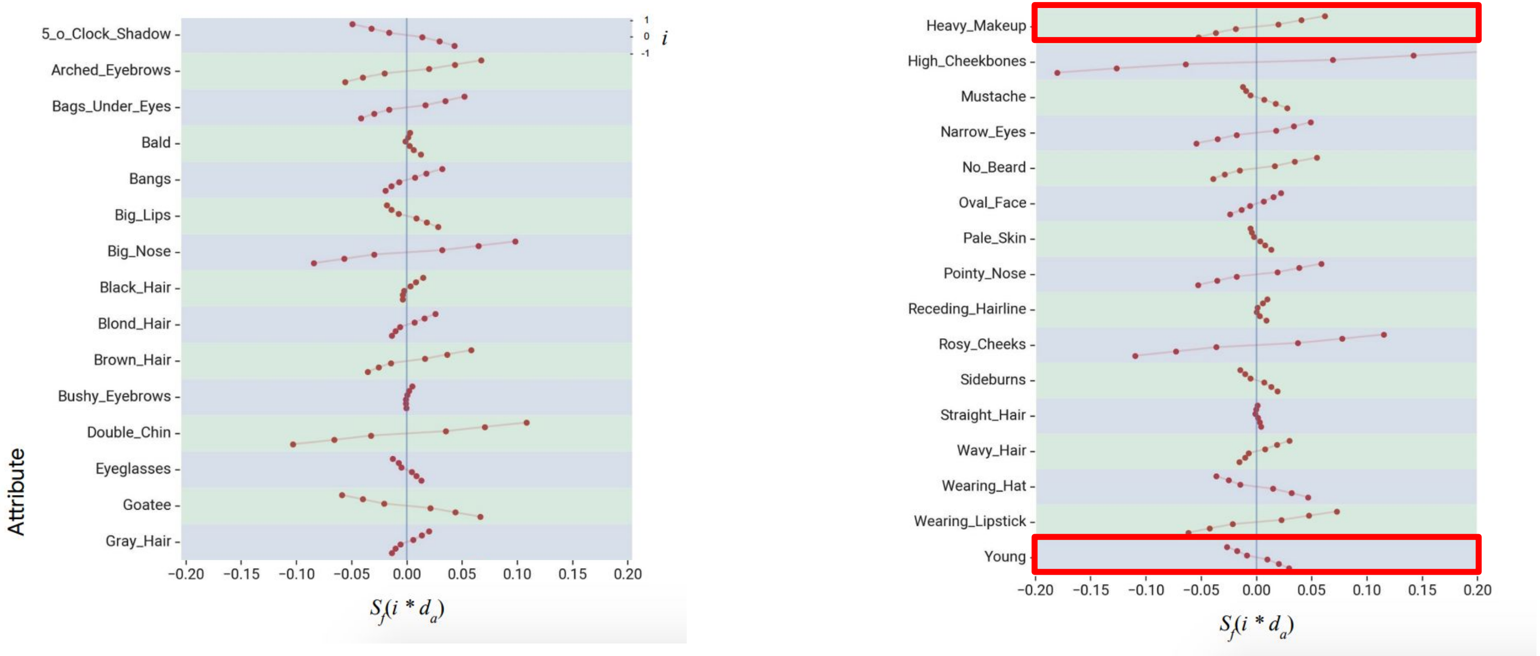
- Counterfactual Face Attribution



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

Recap

$$S_f(d) = \mathbb{E}_{z \sim p(z)}[f(G(z+d)) - f(G(z))]$$



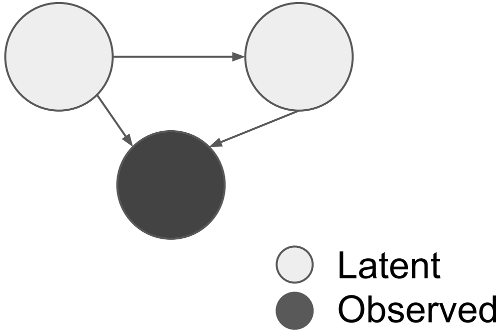
Outline

- Fair Causal Reasoning
- Counterfactual Fairness
  - Formal Methods
  - Law School
  - Crime Rates in NYC
- Equalized Counterfactual Odds
- Multiple Causal Worlds

Fair Causal Reasoning

Causal Graph

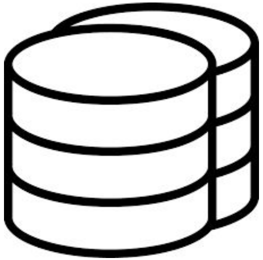
- Observed Data
- Latent Data
- Relations



Causal Fairness Criteria

- Counterfactual Fairness

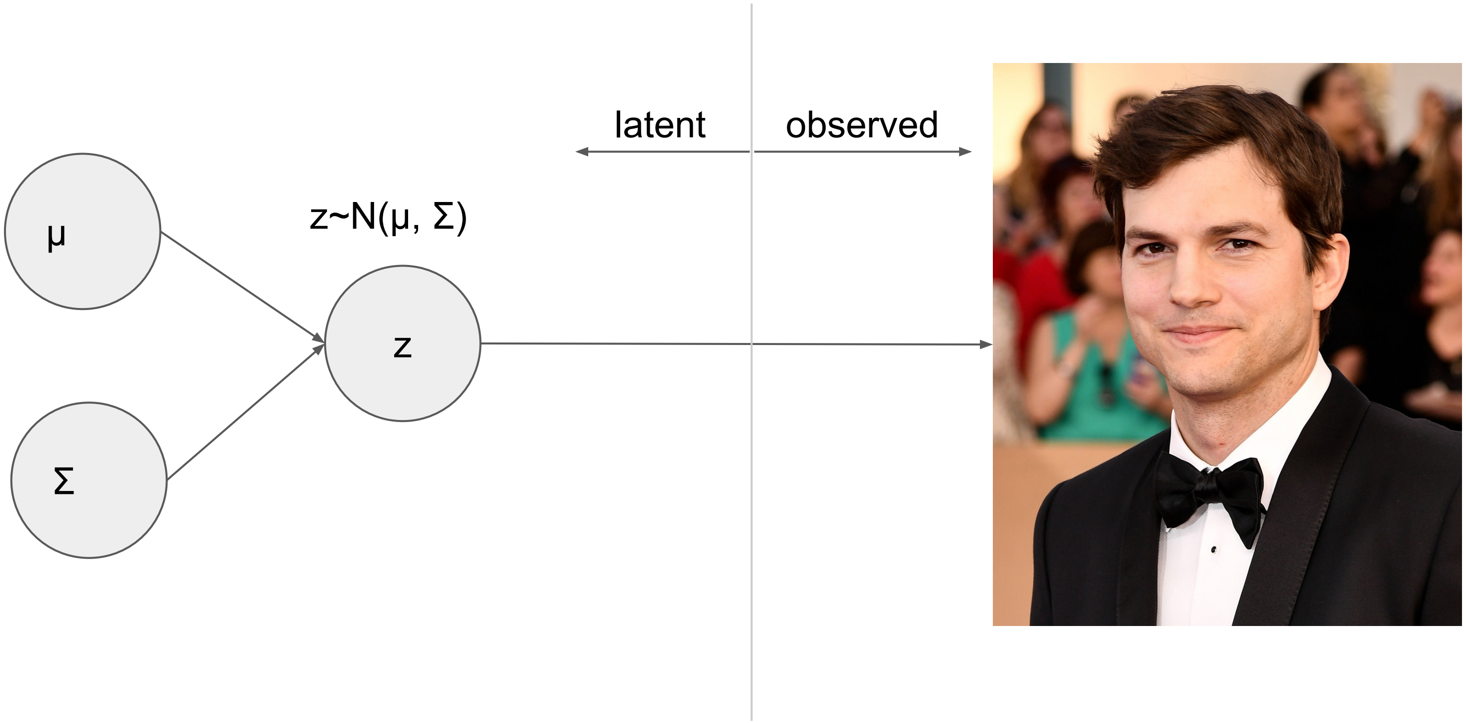
Observed Data



Observational Fairness Criteria

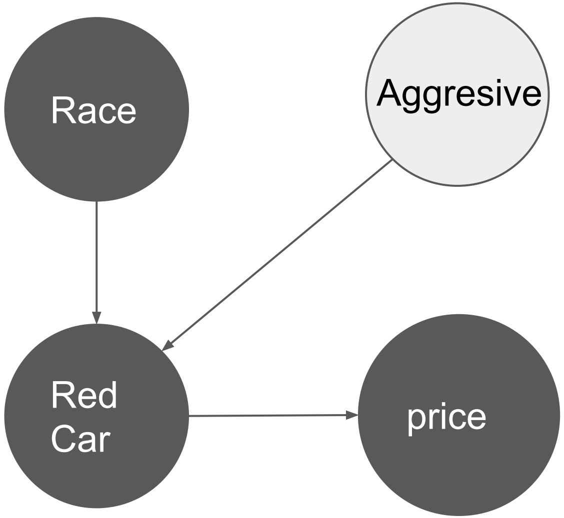
- Fairness Through Unawareness
- Demographic Parity
- Equalized Odds/Opp

Causal Graph



Why Do We Need Causal Fairness?

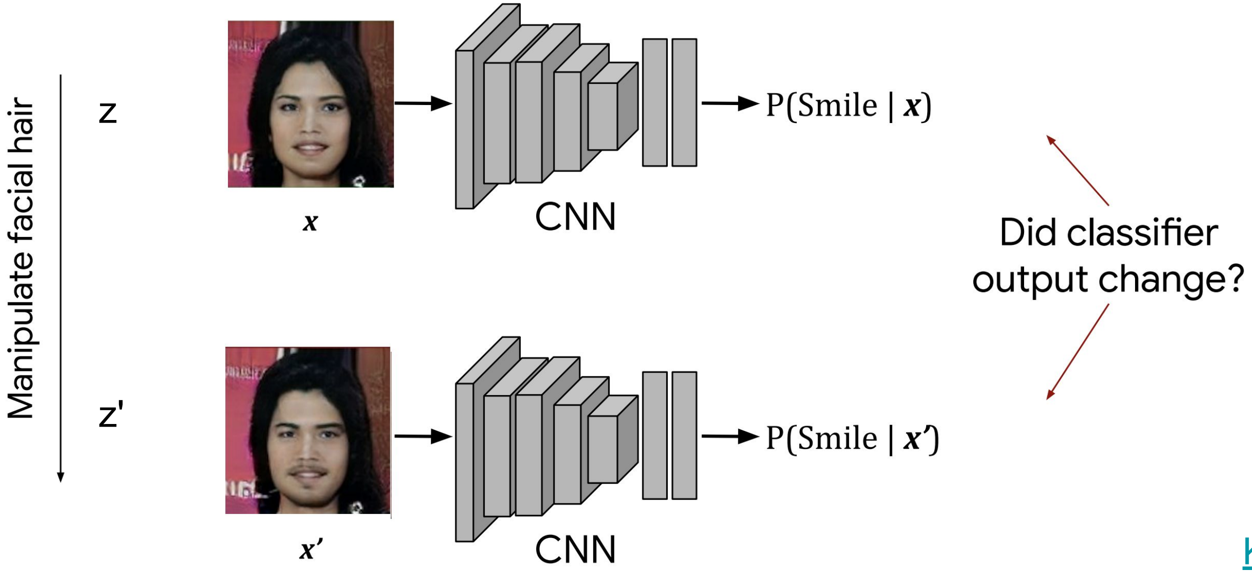
- Recover Latent Variables



Kusner et al, 2018

Why Do We Need Causal Fairness?

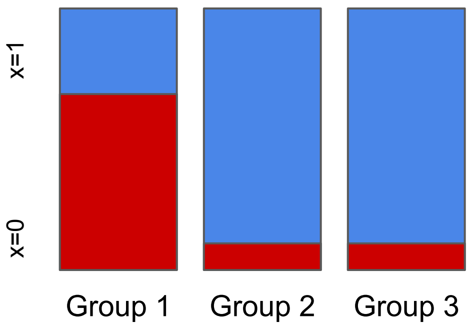
- Recover Latent Variables



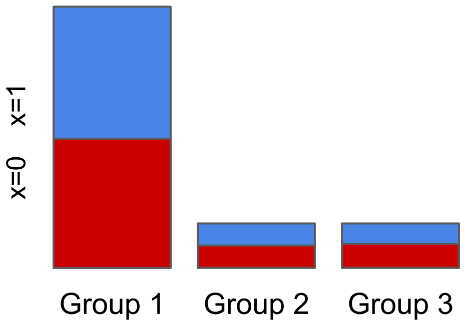
Kusner et al, 2018

Why Do We Need Causal Fairness?

- Dealing with Inherent bias



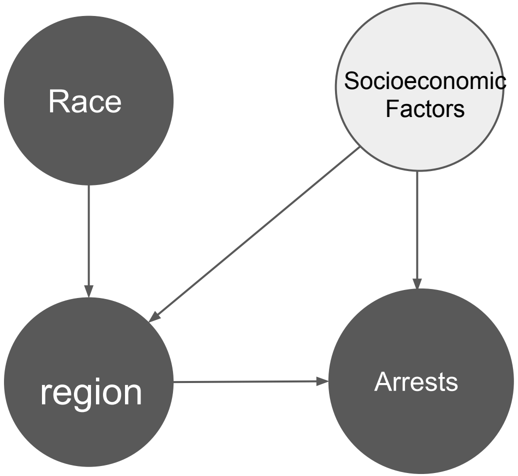
Inherent Biases



Sampling Biases

Inherent bias

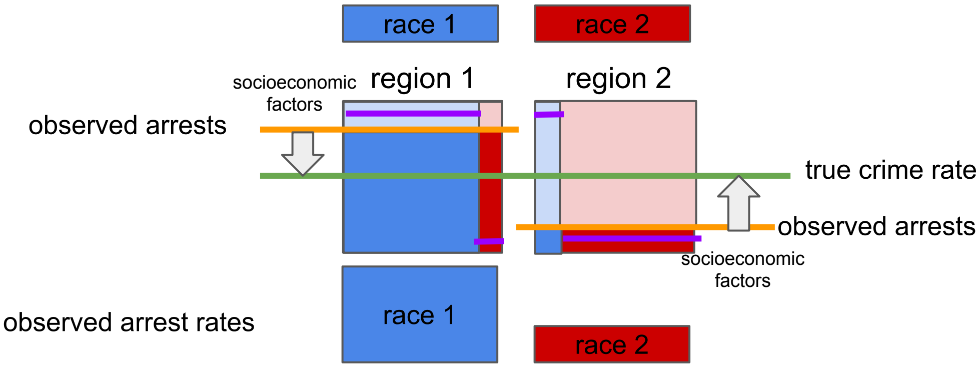
- Race groups live in certain regions due to socioeconomic status
- Latent Socioeconomic factors
  - More police resources in regions with low economic status
  - Results in more arrests



Kusner et al, 2018

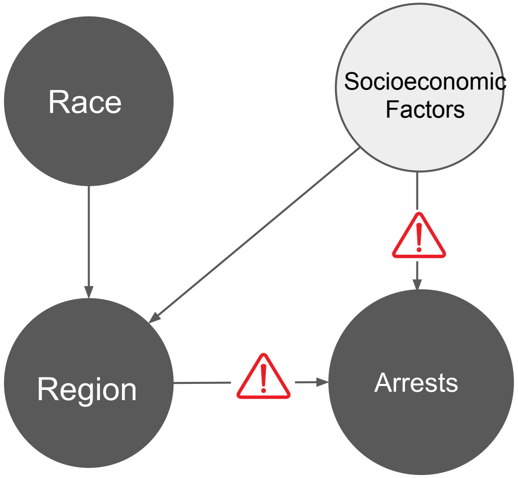
Inherent bias

- Observational Fairness Criteria Won't work
  - Dataset (observed variables) contains inherent selection biases
  - Concentration of police resources resulted in high arrests
  - Attributing regions (and eventual race) unfairly to arrests in the dataset



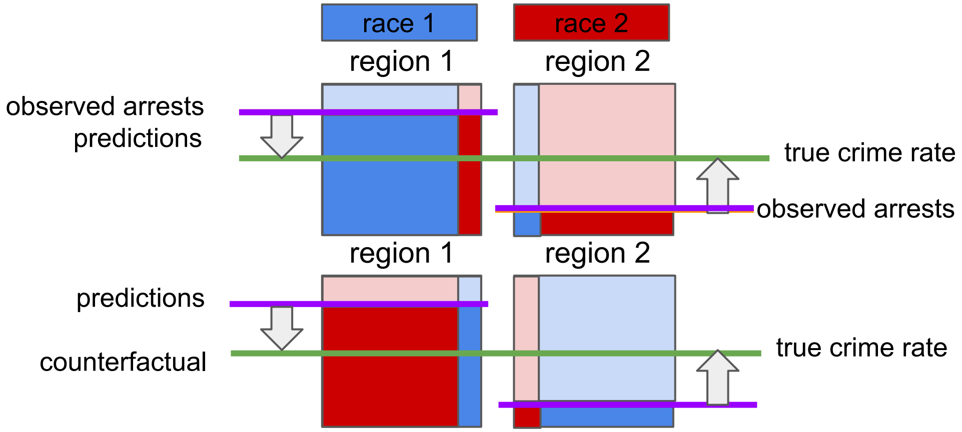
Observational Criteria

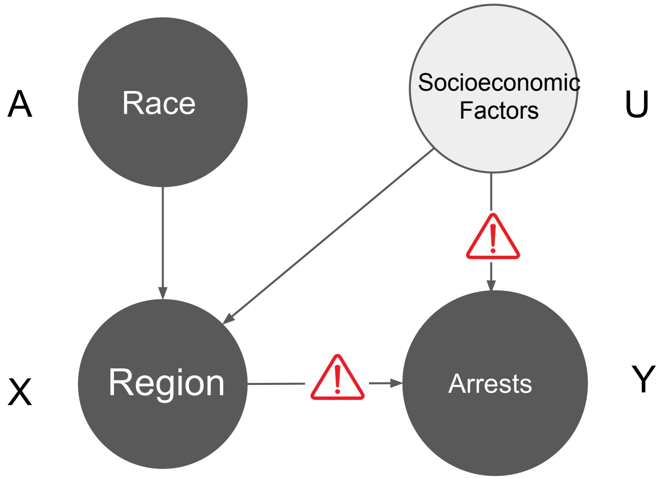
$$P(\hat{Y} = 1 | A = 1) = P(\hat{Y} = 1 | A = 0)$$



Inherent bias

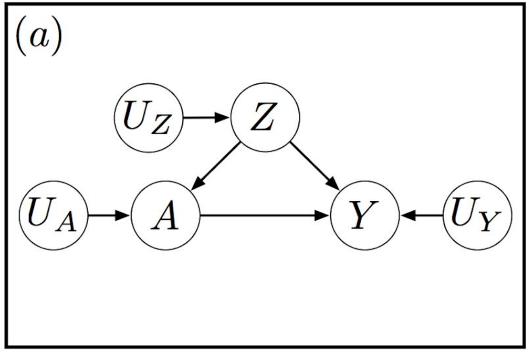
- Causal Fairness
  - Intervene variables in a causal graph
  - Generating samples with races that live in neighborhood that have high police resources



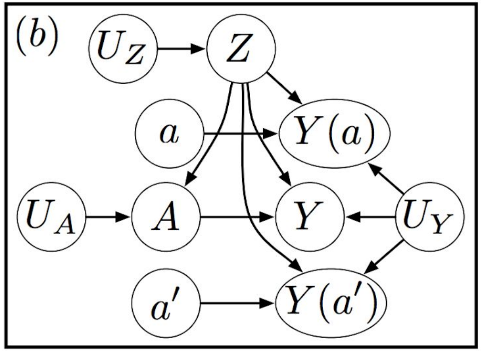


$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a)$$

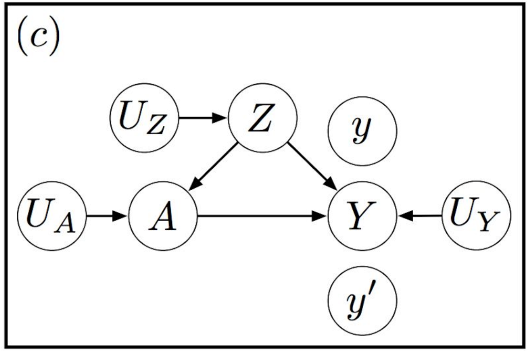
Intervention on Causal Graphs



Causal Graph with A, Z, Y



Intervene on A



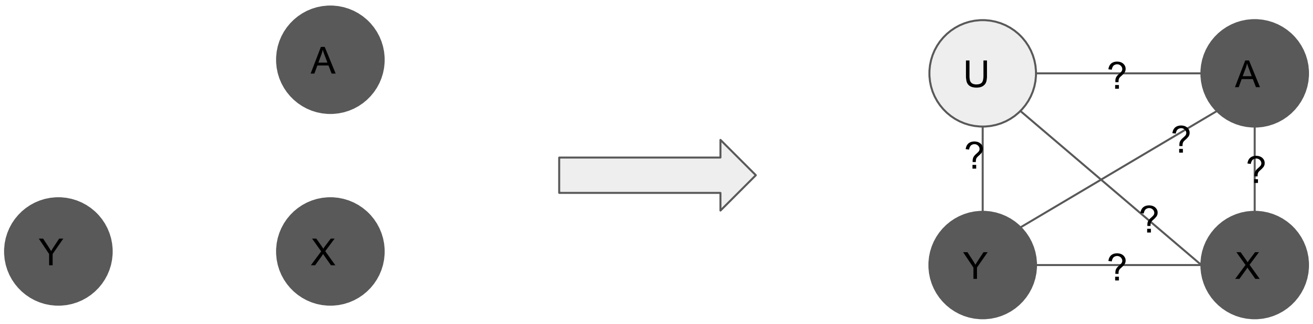
Intervene on Y

[Loftus et al, 2018]

Outline

- Fair Causal Reasoning
- Counterfactual Fairness
  - Formal Methods
  - Law School
  - Crime Rates in NYC
- Equalized Counterfactual Odds
- Multiple Causal Worlds

Counterfactual Fairness Revisited

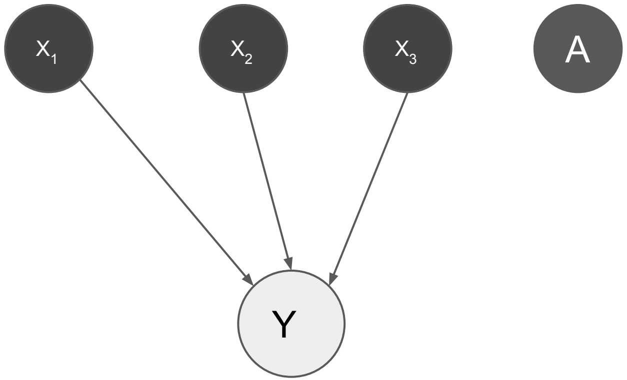


$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a)$$

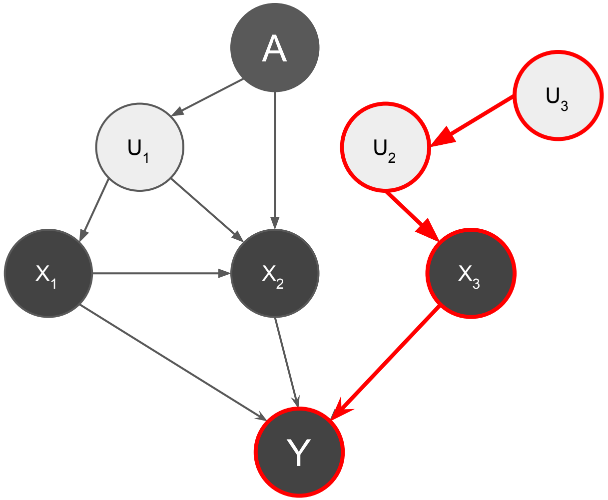
Real Examples	Counterfactual Examples
Intervention on A← a	Intervention on A←a'

Counterfactual Fairness

- Level 1
  - Build predictors using only the observable non-descendants of A

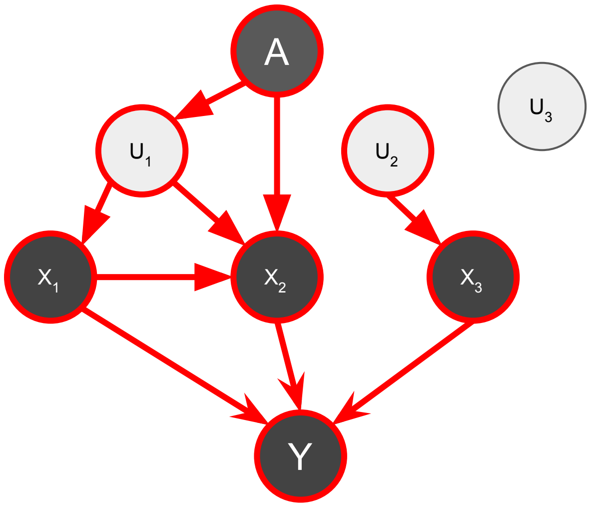


Fairness Through Unawareness



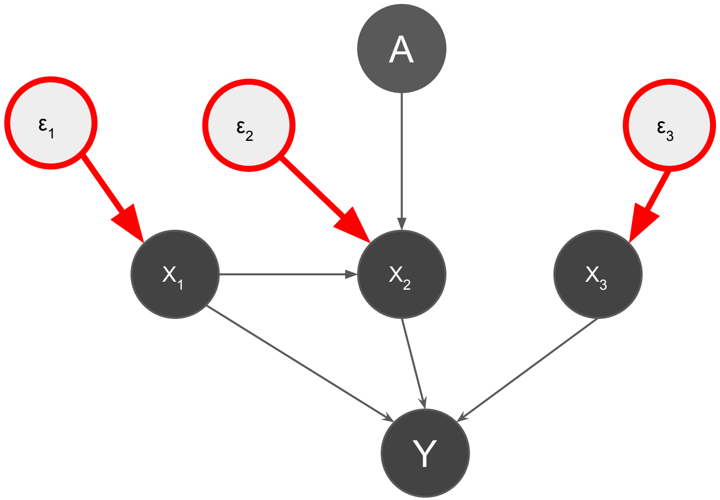
Counterfactual Fairness

- Level 2
  - Build Predictors using the parents of the observable variables



Counterfactual Fairness

- Level 3
  - Build Predictors by adding independent error terms



Outline

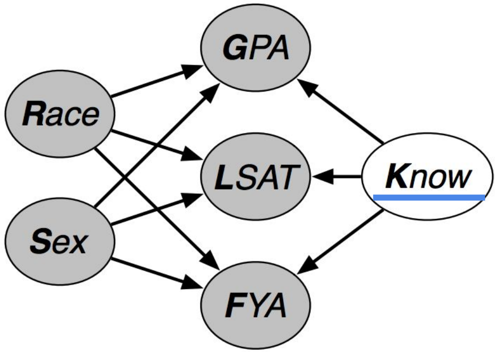
- Fair Causal Reasoning
- Counterfactual Fairness
  - Formal Methods
  - Law School
  - Crime Rates in NYC
- Equalized Counterfactual Odds
- Multiple Causal Worlds

Law School Success Dataset

- Conducted by Law School Admission Council in US
  - 21,790 law students
  - Entrance exam scores (LSAT)
  - Grade-point average (GPA) collected prior to law school
  - Prediction Y = first year average grade (FYA)
  - Protected features = {Gender, Race}

Level 2 Counterfactual Fairness

- Build Predictors using the parents of the observable variables



$$\mathrm{K} \sim \mathcal{N}(0,1)$$

$$\mathrm{FYA} \sim \mathcal{N}(w_F^K K + w_F^R R + w_F^S S, 1) \qquad \mathrm{GPA} \sim \mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G)$$

$$\mathrm{LSAT} \sim \mathrm{Poisson}(\exp(b_L + w_L^K K + w_L^R R + w_L^S S))$$

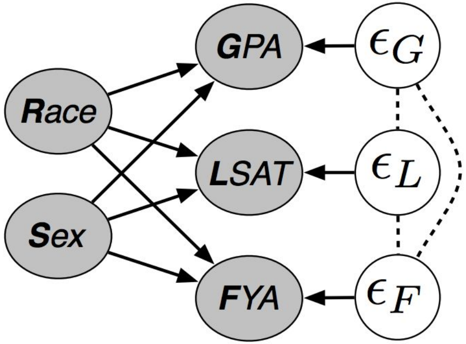
Gaussian Dist.

Parameters

[Kusner et al, 2018](#)

Level 3 Counterfactual Fairness

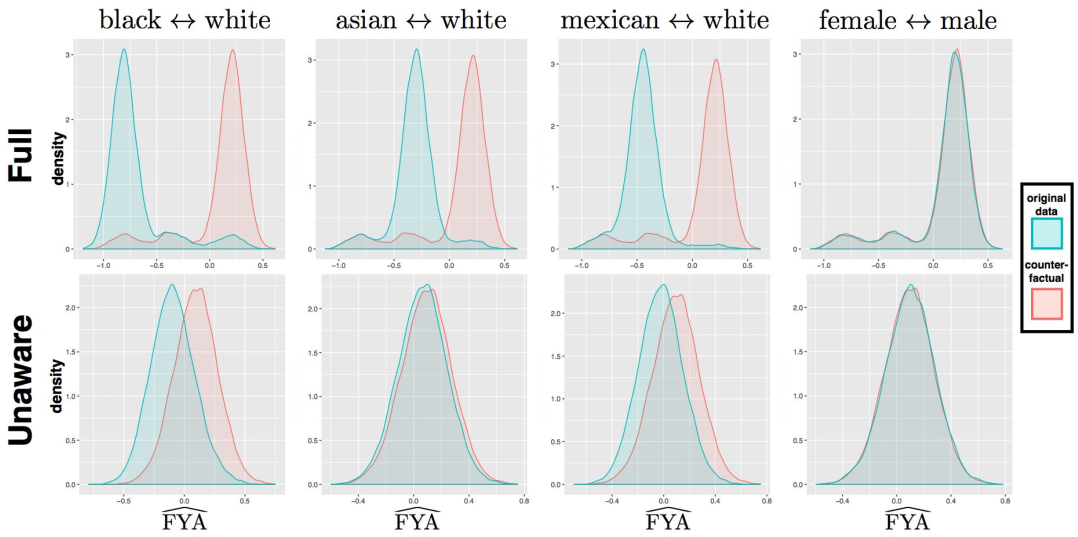
- Build Predictors by adding independent error terms



$$\text{GPA} = b_G + w_G^R R + w_G^S S + \epsilon_G, \quad \epsilon_G \sim p(\epsilon_G)$$
$$\text{LSAT} = b_L + w_L^R R + w_L^S S + \epsilon_L, \quad \epsilon_L \sim p(\epsilon_L)$$
$$\text{FYA} = b_F + w_F^R R + w_F^S S + \epsilon_F, \quad \epsilon_F \sim p(\epsilon_F)$$

[Kusner et al, 2018](#)

Baselines



full - using all features
unaware - fairness through unawareness

[Kusner et al, 2018]

Results

	Baseline	Baseline	Level 2	Level 3
	Full	**Unaware**	**Fair K**	**Fair Add**
RMSE	0.873	0.894	0.929	0.918

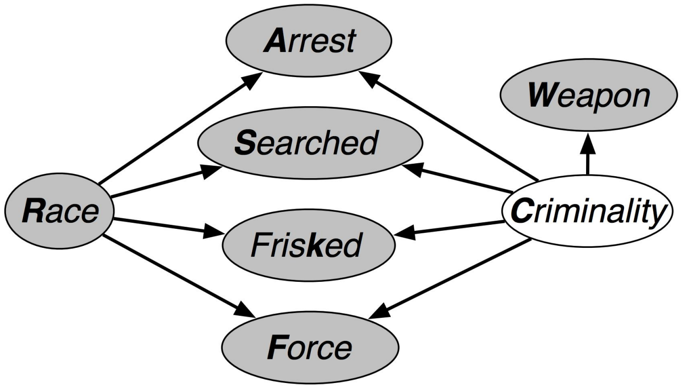
Kusner et al, 2018

Outline

- Fair Causal Reasoning
- Counterfactual Fairness
  - Formal Methods
  - Law School
  - Crime Rates in NYC
- Equalized Counterfactual Odds
- Multiple Causal Worlds

Causal Graph

- Assess the fairness of the NYC arrest dataset
  - 38,609 records
  - White individuals (4492)
  - Black Hispanic individuals (2414)



Kusner et al, 2018

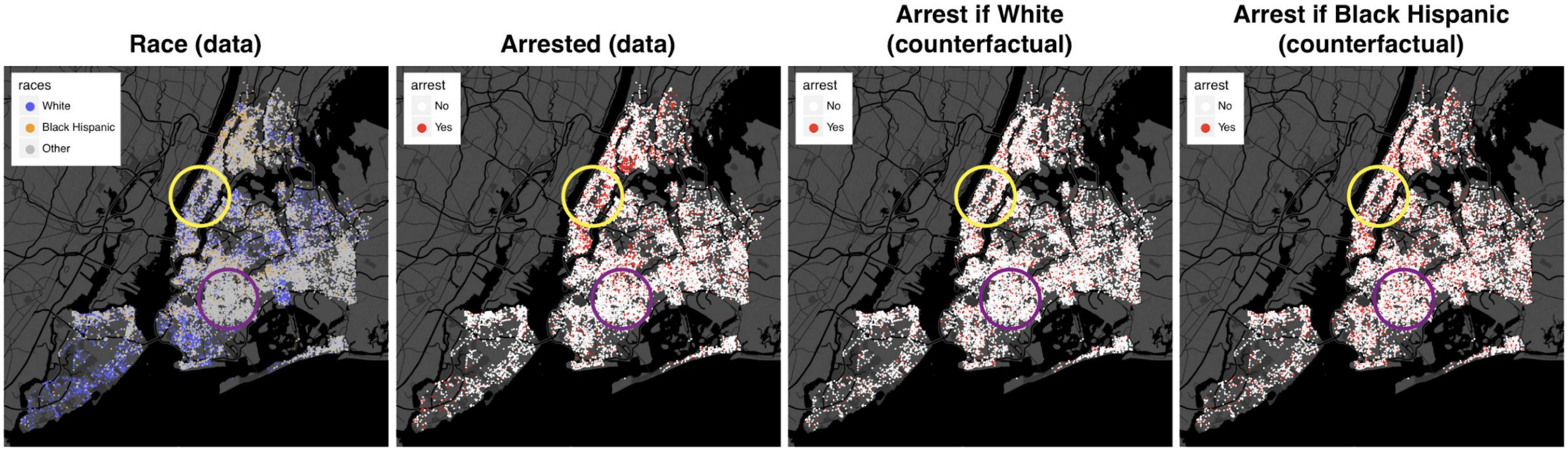
Assessment Results

White (4492)
Black Hispanic (2414)

White (12.1%)
Black Hispanic (19.8%)

Arrests decreases from 5659 to 3722

Arrests increases from 5659 to 6439



Kusner et al, 2018

Outline

- Fair Causal Reasoning
- Counterfactual Fairness
  - Formal Methods
  - Law School
  - Crime Rates in NYC
- Equalized Counterfactual Odds
- Multiple Causal Worlds

Equalized Counterfactual Odds

Equality of Odds

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

Counterfactual Fairness

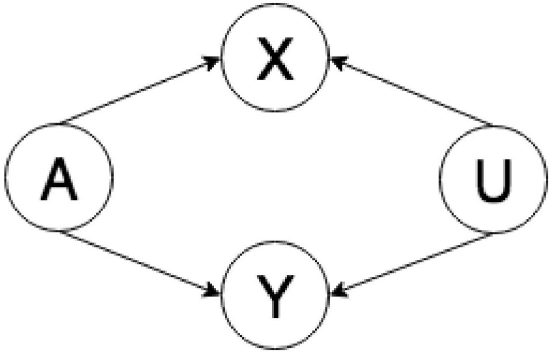
$$P(\underline{\hat{Y}_{A \leftarrow a}(U)} = y \mid X = x, A = a) = P(\underline{\hat{Y}_{A \leftarrow a'}(U)} = y \mid X = x, A = a)$$

Equalized Counterfactual Odds

$$p(\underline{\hat{Y}_{A \leftarrow a}(U)} \mid X = x, \underline{Y_{A \leftarrow a}} = y, A = a) = p(\underline{\hat{Y}_{A \leftarrow a'}(U)} \mid X = x, \underline{Y_{A \leftarrow a'}} = y, A = a)$$

Healthcare Equality

- Protected Features A = {Gender}
- Features X, vector representation of coded diagnoses, procedures, medication orders, lab results, and clinical notes
- Prediction Y, a binary indicator of the occurrence of a clinically relevant outcome



$$u \sim p(U) = \mathrm{Normal}(0, I)$$
$$a \sim p(A) = \mathrm{Categorical}(A \mid \pi)$$
$$x, y \sim p(X, Y \mid U, A) = p(X \mid U, A)p(Y \mid U, A)$$

Training Objective

- σ - sigmoid function
- h - predictor
- J - cross entropy loss

$$\mathcal{L} = J(h_\theta(x, a), y) + \lambda_{\mathrm{CF}} \sum_{a_k \in \mathcal{A}} \mathbb{1}[a \neq a_k] J(h_\theta(\underline{x_{A \leftarrow a_k}}, a_k), \underline{y_{A \leftarrow a_k}}) +$$

$$\lambda_{\mathrm{CLP}} \sum_{a_k \in \mathcal{A}} \mathbb{1}[a \neq a_k] \mathbb{1}[y = \underline{y_{A \leftarrow a_k}}] \Big(\underline{\sigma^{-1}(h_\theta(\underline{x_{A \leftarrow a_k}}, a_k))} - \underline{\sigma^{-1}(h_\theta(x, a))} \Big)^2$$

logits

Dataset Overview

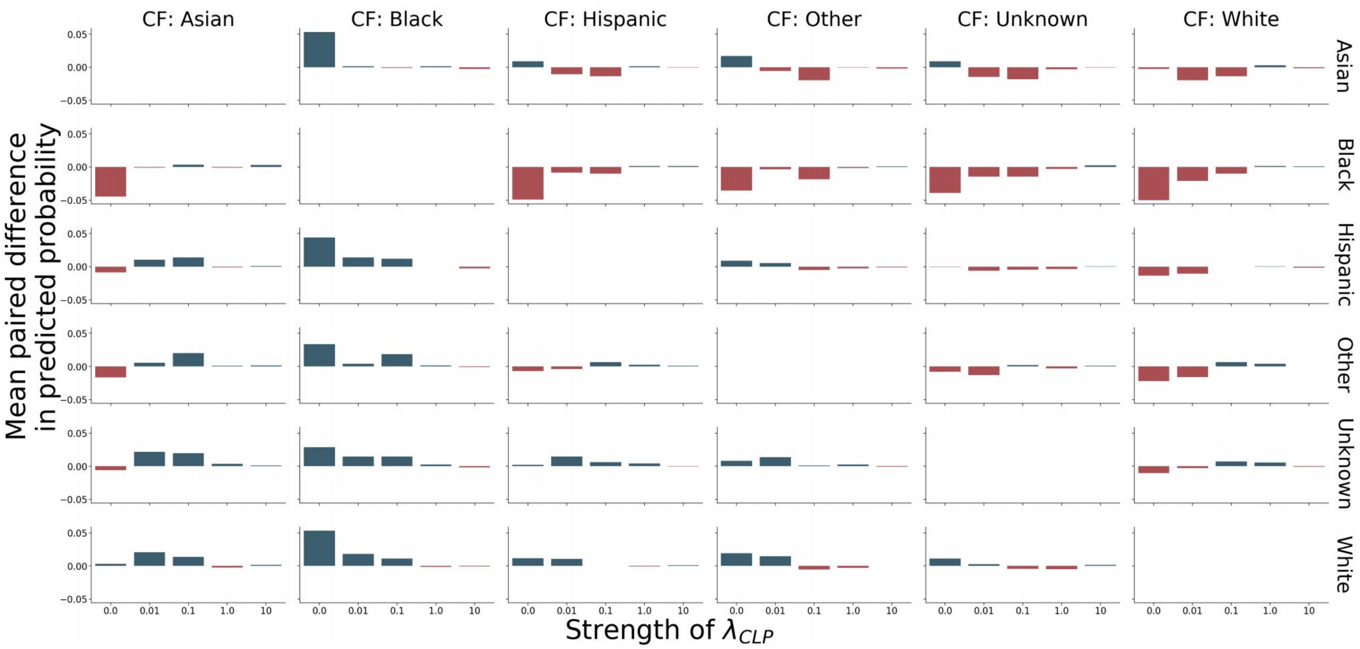
Group	Count	Length of Stay $\geq$ 7 Days	Inpatient Mortality
Asian	17,465	0.187	0.025
Black	5,202	0.239	0.020
Hispanic	21,978	0.196	0.019
Other	11,004	0.200	0.022
Unknown	3,593	0.201	0.072
White	70,391	0.204	0.021
Female	72,556	0.167	0.018
Male	57,076	0.245	0.029
[18, 30)	15,291	0.180	0.007
[30, 45)	27,155	0.140	0.007
[45, 65)	43,529	0.222	0.025
[65, 89)	43,658	0.226	0.036
All	129,633	0.201	0.023

Results

		λ_{CLP}					
Group	Metric	N/A	0.0	0.01	0.1	1.0	10.0
Asian	AUC-PRC	0.605	0.563	0.555	0.561	0.56	0.562
	AUC-ROC	0.86	0.853	0.853	0.854	0.849	0.851
	Brier	0.106	0.11	0.109	0.109	0.11	0.112
Black	AUC-PRC	0.579	0.548	0.55	0.545	0.563	0.573
	AUC-ROC	0.838	0.825	0.82	0.825	0.823	0.823
	Brier	0.124	0.135	0.129	0.128	0.127	0.129
Hispanic	AUC-PRC	0.592	0.558	0.565	0.57	0.564	0.56
	AUC-ROC	0.862	0.855	0.856	0.861	0.853	0.854
	Brier	0.113	0.117	0.115	0.114	0.117	0.118
Other	AUC-PRC	0.549	0.557	0.557	0.563	0.553	0.561
	AUC-ROC	0.824	0.827	0.819	0.824	0.819	0.827
	Brier	0.122	0.124	0.121	0.121	0.122	0.124
Unknown	AUC-PRC	0.675	0.616	0.616	0.606	0.614	0.633
	AUC-ROC	0.9	0.891	0.888	0.893	0.891	0.887
	Brier	0.104	0.106	0.103	0.103	0.105	0.111
White	AUC-PRC	0.575	0.568	0.564	0.559	0.562	0.563
	AUC-ROC	0.847	0.84	0.839	0.838	0.838	0.837
	Brier	0.118	0.12	0.118	0.12	0.12	0.121

Results

- Difference in the counterfactual versus factual predicted probability

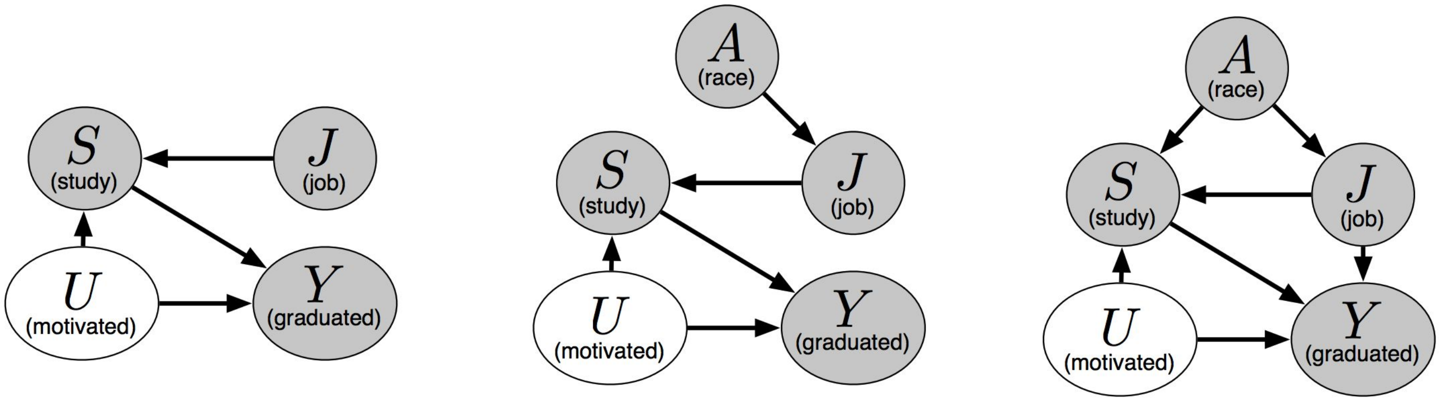


Outline

- Fair Causal Reasoning
- Counterfactual Fairness
    - Formal Methods
    - Law School
    - Crime Rates in NYC
- Equalized Counterfactual Odds
- Multiple Causal Worlds

Multiple Causal Graphs

- Whether a student can graduate on time



Russell et al, 2017

Alternative Definitions of Counterfactual Fairness

Exact Formulation

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

ε - Approximate Formulation

$$\left| f(\mathbf{x}_{A \leftarrow a}, a) - f(\mathbf{x}_{A \leftarrow a'}, a') \right| \leq \epsilon$$

(δ, ε) - Approximate Formulation

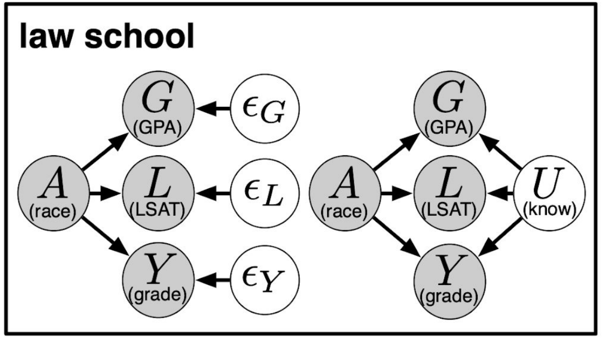
$$\mathbb{P}(\left| f(\mathcal{X}_{A \leftarrow a}, a) - f(\mathcal{X}_{A \leftarrow a'}, a') \right| \leq \epsilon \mid \mathcal{X} = \mathbf{x}, A = a) \geq 1 - \delta$$

Multi-world Counterfactual Fairness

$$\min_{f} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\ell(f(\mathbf{x}_i, a_i), y_i)}_{\text{loss of the data}} + \lambda \underbrace{\sum_{j=1}^{m}}_{\text{world j}} \frac{1}{n} \sum_{i=1}^{n} \sum_{a' \neq a_i} \mu_j(f, \underbrace{\mathbf{x}_i,}_{\text{counterfactual examples}} a_i, a')$$

$$\mu_j(f, \mathbf{x}_i, a_i, a') := \underbrace{\frac{1}{S} \sum_{s=1}^{S}}_{\text{Monte-carlo Samples}} \max\{0, \underbrace{\left| f(\mathbf{x}^{s}_{i, A^j \leftarrow a_i}, a_i) - f(\mathbf{x}^{s}_{i, A^j \leftarrow a'}, a') \right| - \epsilon\}}_{\varepsilon\text{ - Approximate Counterfactual Fairness}}$$

Law Graduate School



$$G = b_G + w_G^A A + \epsilon_G$$
$$L = b_L + w_L^A A + \epsilon_L$$
$$Y = b_Y + w_Y^A A + \epsilon_Y$$
$$\epsilon_G, \epsilon_L, \epsilon_Y \sim \mathcal{N}(0,1)$$

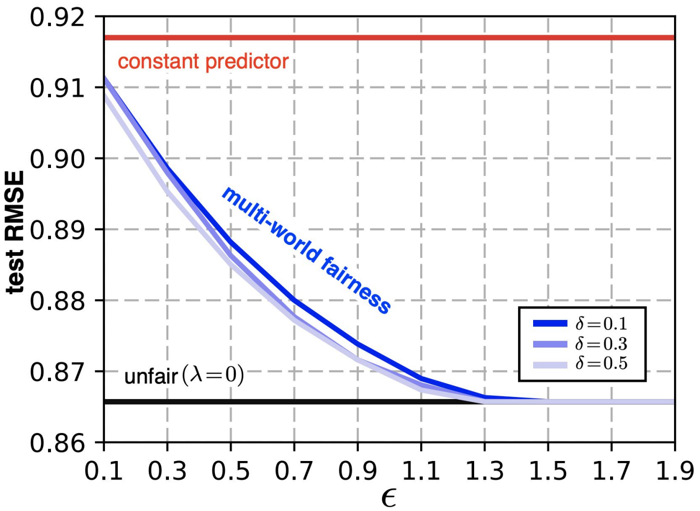
L3 Method

$$G \sim \mathcal{N}(b_G + w_G^A A + w_G^U U, \sigma_G)$$
$$L \sim \text{Poisson}(\exp(b_L + w_L^A A + w_L^U U))$$
$$Y \sim \mathcal{N}(w_Y^A A + w_Y^U U, 1)$$
$$U \sim \mathcal{N}(0,1)$$

L2 Method

[Russell et al, 2017](#)

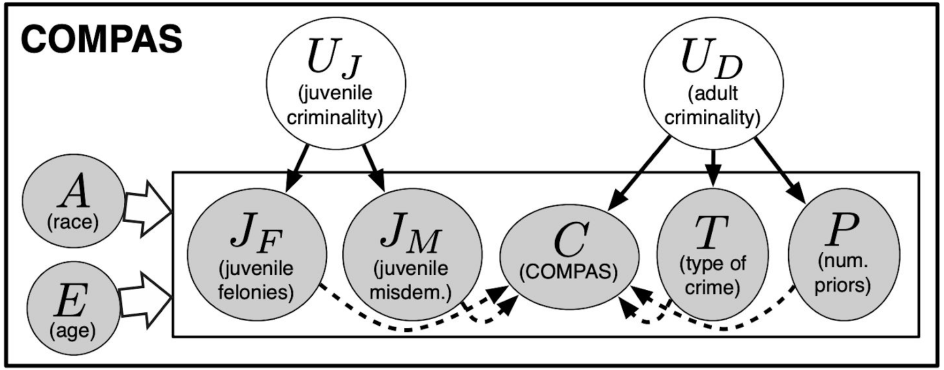
Results



$$\left| f(\mathbf{x}_{A \leftarrow a}, a) - f(\mathbf{x}_{A \leftarrow a'}, a') \right| \leq \epsilon$$

[Russell et al, 2017](#)

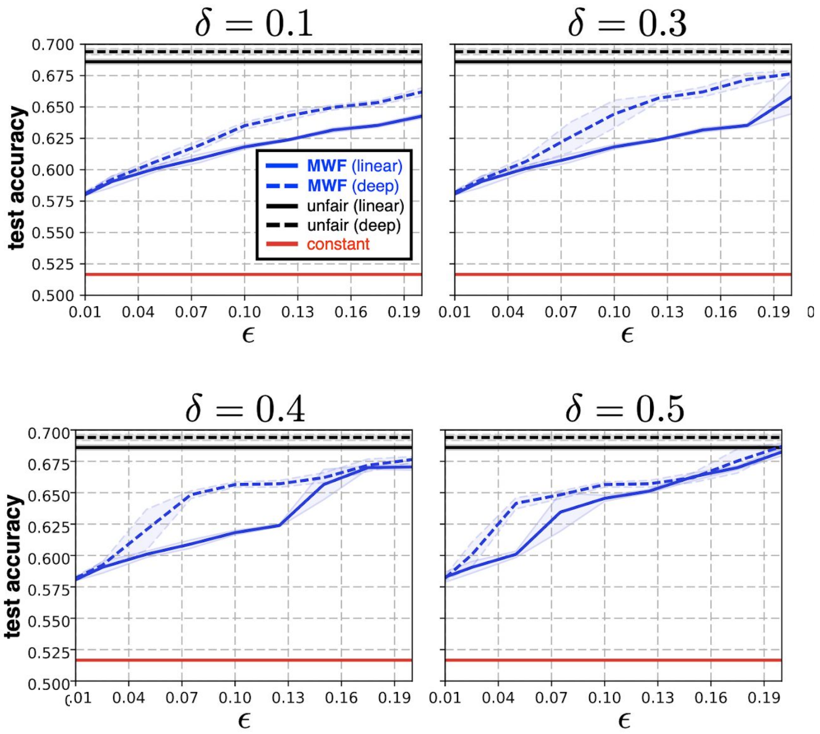
COMPAS



$$
\begin{aligned}
T &\sim \text{Bernoulli}(\phi(b_T + w_C^{U_D} U_D + w_C^E E + w_C^A A) \\
C &\sim \mathcal{N}(b_C + w_C^{U_D} U_D + w_C^E E + w_C^A A + w_C^T T + w_C^P P + w_C^{J_F} J_F + w_C^{J_M} J_M, \sigma_C) \\
P &\sim \text{Poisson}(\exp(b_P + w_P^{U_D} U_D + w_P^E E + w_P^A A)) \\
J_F &\sim \text{Poisson}(\exp(b_{J_F} + w_{J_F}^{U_J} + w_{J_F}^E E + w_{J_F}^A A)) \\
J_M &\sim \text{Poisson}(\exp(b_{J_M} + w_{J_M}^{U_J} + w_{J_M}^E E + w_{J_M}^A A)) \\
[U_J, U_D] &\sim \mathcal{N}(0, \Sigma)
\end{aligned}
$$

[Russell et al, 2017](#)

Results



Russell et al, 2017

Reading Assignments

- Wu, Yongkai, Lu Zhang, Xintao Wu, and Hanghang Tong. PC-Fairness: A Unified Framework for Measuring Causality-based Fairness, NeurIPS 2019
- Chiappa, Silvia. Path-specific counterfactual fairness, AAAI 2019
- Balcan, Maria-Florina F., Travis Dick, Ritesh Noothigattu, and Ariel D. Procaccia. Envy-free classification, NeurIPS 2019
- Qureshi, Bilal, Faisal Kamiran, Asim Karim, Salvatore Ruggieri, and Dino Pedreschi. Causal inference for social discrimination reasoning, Journal of Intelligent Information Systems 2019
- Zhang, Junzhe, and Elias Bareinboim. Equality of opportunity in classification: A causal approach, NeurIPS 2018