Algorithmic Fairness of Statistical Decision Systems

Subhrajyoty Roy
14th August, 2021

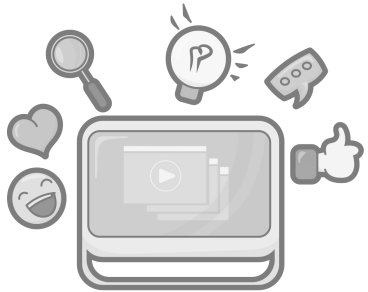
What do Statistical Decision Systems offer?

- Ideal
- Avrio AI Inc.
- Skillate
- Entelo
- Mya Systems

- PredPol
- COMPAS
- many more confidential algorithms.

Credit risk assessment in banks

Recommendations in e-commerce



Personalized news in social media

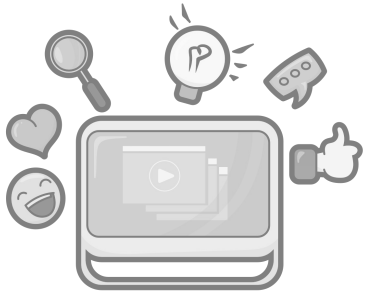
What do Statistical Decision Systems offer?

- Ideal
- Avrio AI Inc.
- Skillate
- Entelo
- Mya Systems

- PredPol
- COMPAS
- many more confidential algorithms.

Credit risk assessment in banks

Recommendations in e-commerce



Personalized news in social media

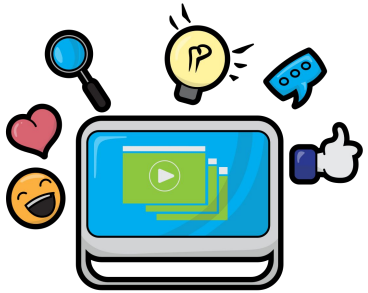
What do Statistical Decision Systems offer?

- Ideal
- Avrio AI Inc.
- Skillate
- Entelo
- Mya Systems

- PredPol
- COMPAS
- many more confidential algorithms.

Credit risk assessment in banks

Recommendations in e-commerce



Personalized news in social media

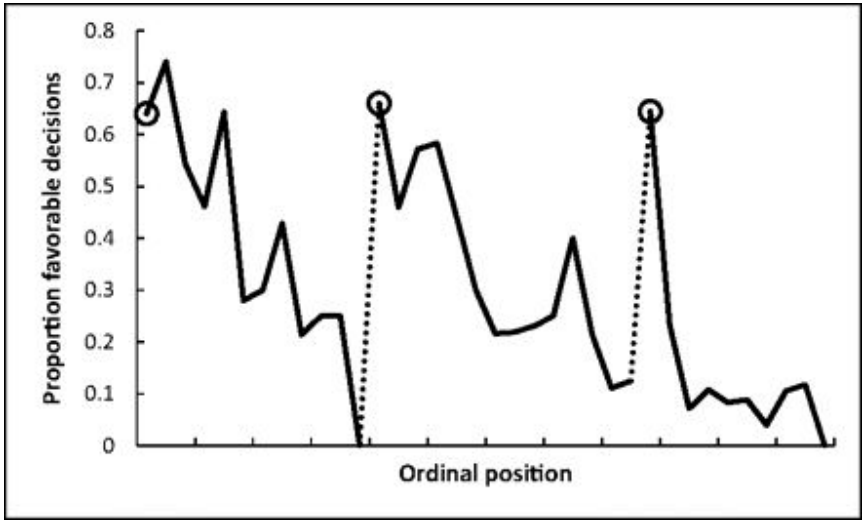
Why need Data-driven Automation?

Automated Hiring and Resume Screening

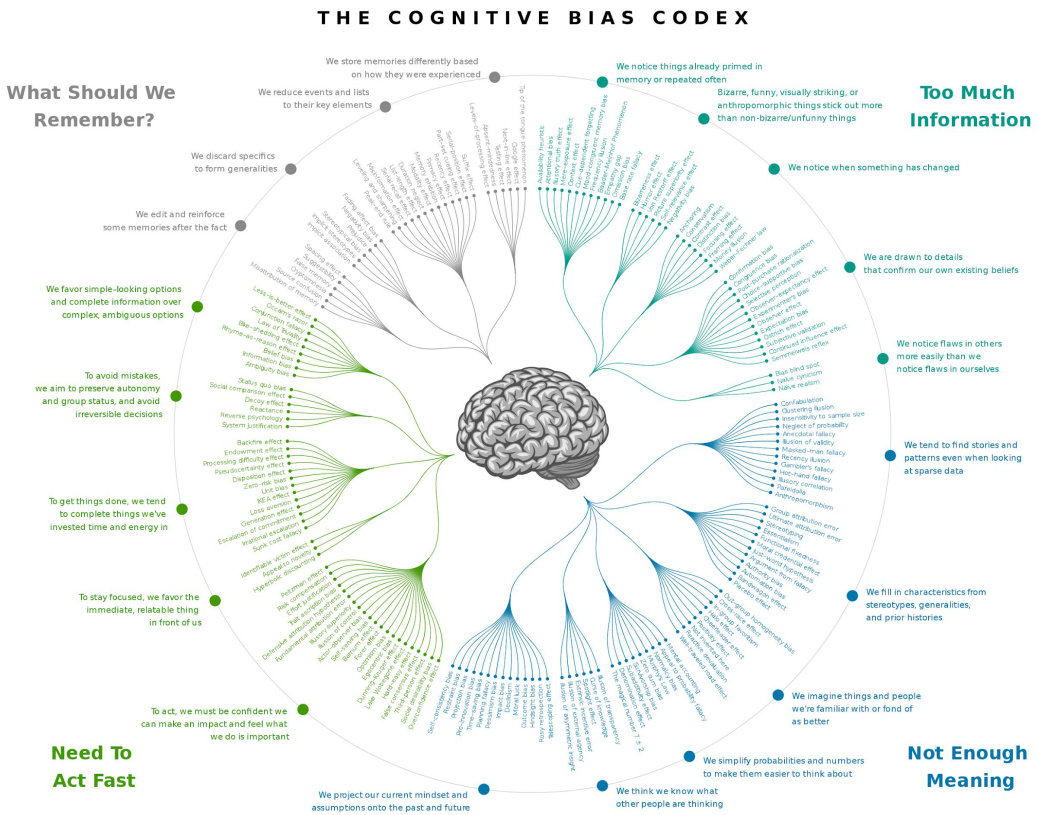
- One million applications per year to Google. That's 2700+/day.
- 40 million active job searchers on LinkedIn.
- 35+ million graduates from bachelor courses every year in India according to MOE report.

Automated Judicial Systems

- Hungry judges problem (2011).



180+ Cognitive Biases of Human

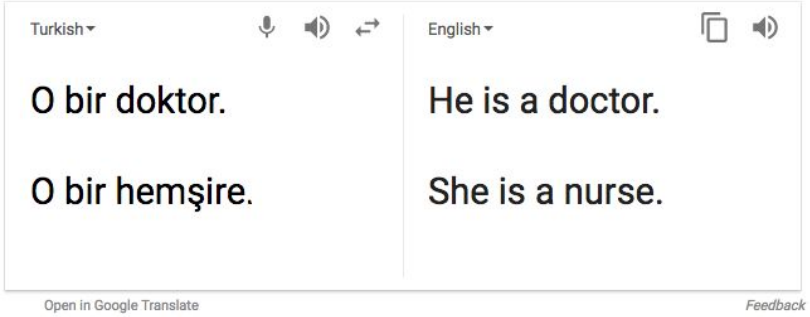


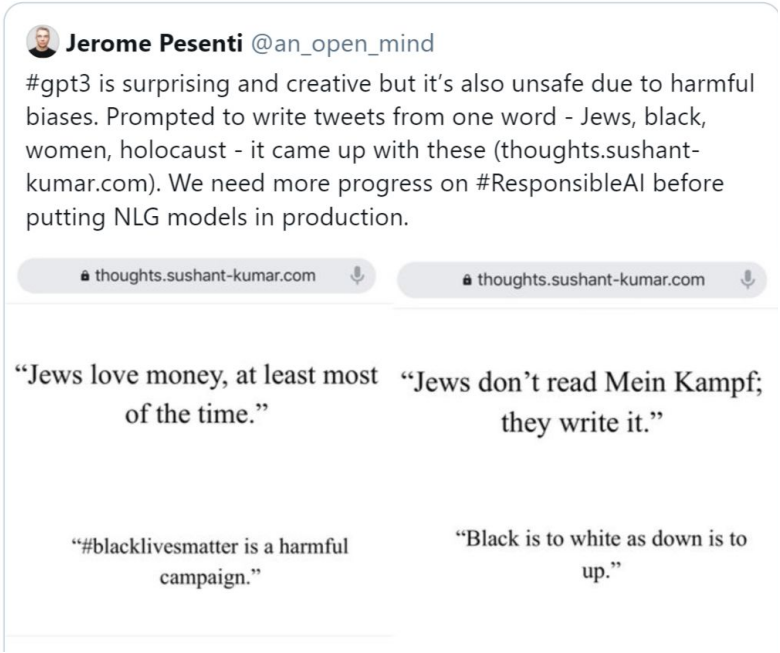
Bias and Fairness Concerns

The New York Times

Even With Affirmative Action, Blacks and Hispanics Are More Underrepresented at Top Colleges Than 35 Years Ago

By JEREMY ASHKENAS, HAEYOUN PARK and ADAM PEARCE AUG. 24, 2017





Bias and Fairness Concerns

	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.)

COMPAS prediction error and follow up analysis show African americans are discriminated against

AI is taking over job hiring, but can it be racist?

by Avi Asher-Schapiro | @AASchapiro | Thomson Reuters Foundation

Monday, 7 June 2021 05:01 GMT

More than 55% of U.S. human resources managers using predictive algorithms to help them make recruitment choices

What is Bias?

- Selection, Sampling and Reporting Bias
- Bias of an estimator
- Bias against the racial and gender minority, discriminations and prejudices in historical context

What is Bias?

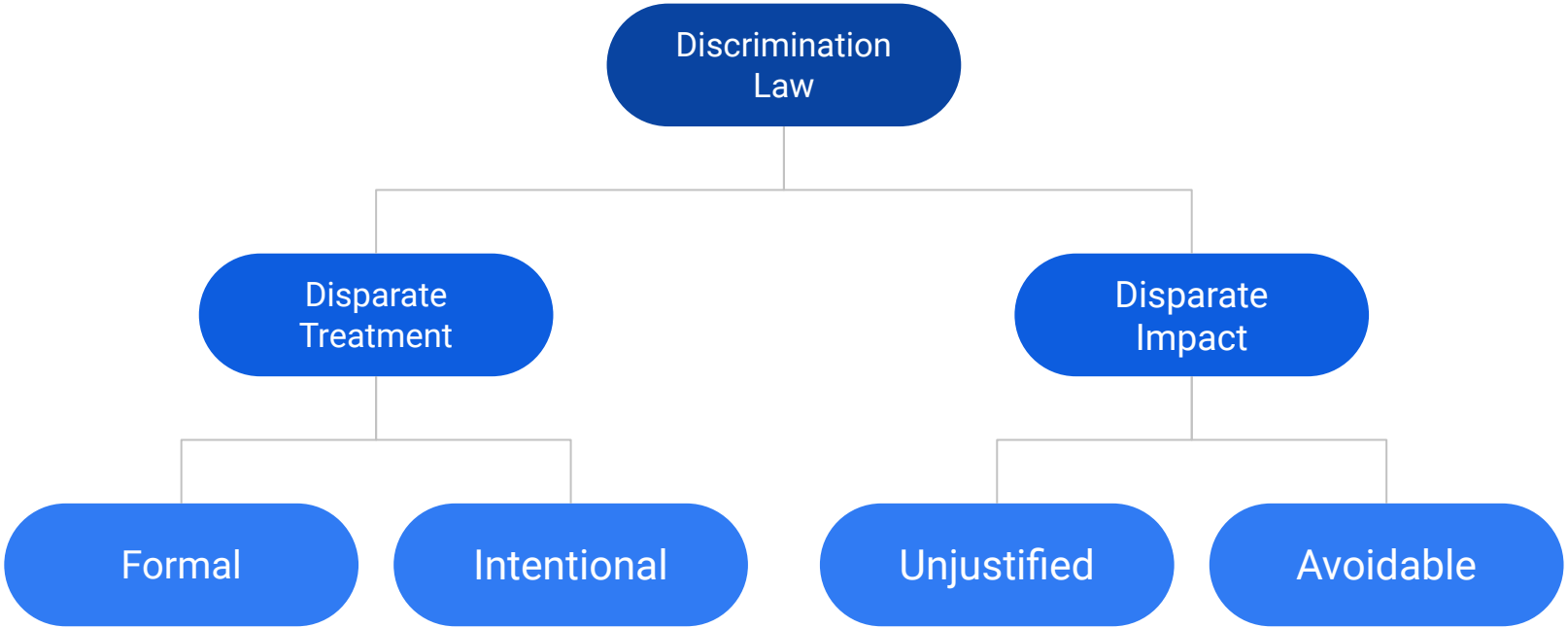
- Selection, Sampling and Reporting Bias
- Bias of an estimator
- Bias against the racial and gender minority, discriminations and prejudices in historical context

Is not AI / ML / DL all about classifying (and discriminate)?

- Unjustified basis of discrimination
- Moral and ethical irrelevance.

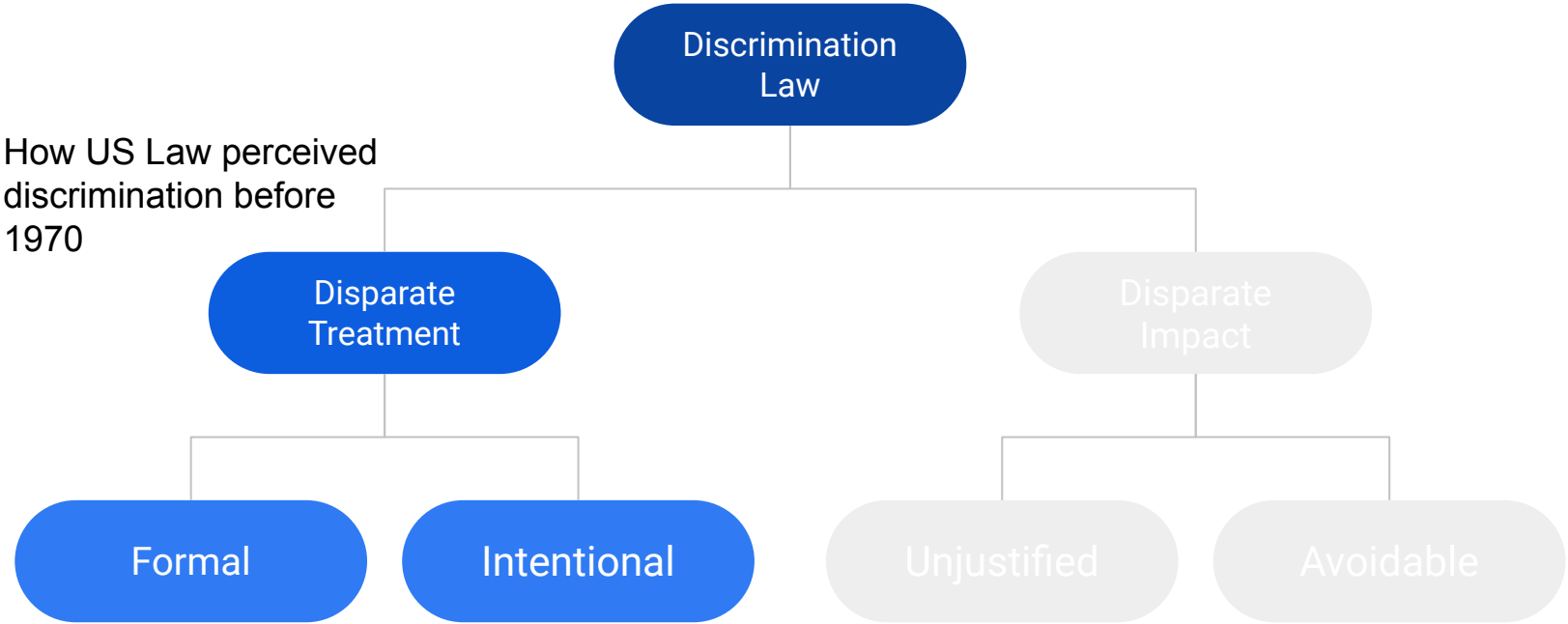
Why should we care? Isn't everything biased?

Two Doctrines

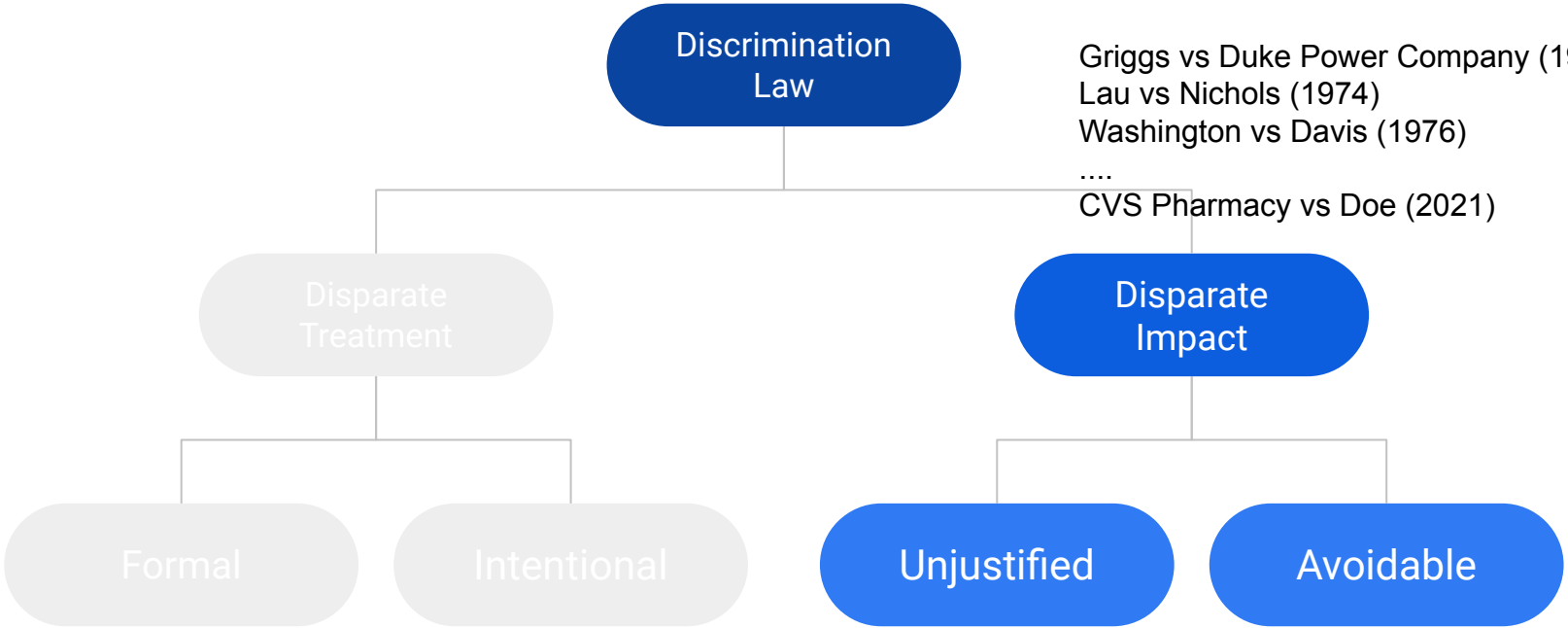


Two Doctrines

How US Law perceived discrimination before 1970



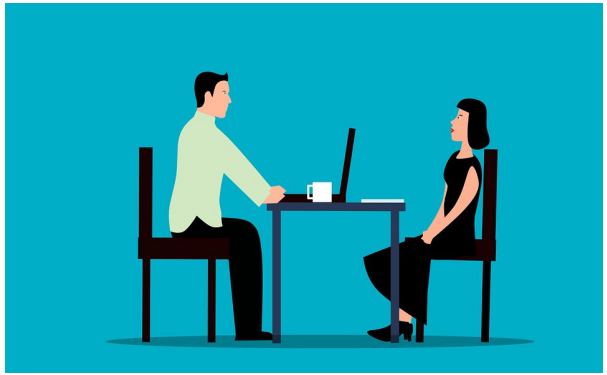
Two Doctrines



Bias from bad selection of predictors

Historical data used by PredPol helps police to deploy more police to regions with more black people.
And so the existing bias reinforces.

Bias from bad selection of targets



Previous hiring decisions?

Year end performance review?

Bias from bad interaction

If one is recommended over other, it would end up overcrowded.

Formal Description

X : The set of predictors

Y : The binary target variable

A : The sensitive attribute

The predicted score $R(X, A) \in [0, 1]$

The predicted binary decision $C(X, A)$

Formal Description

X : The set of predictors

Y : The binary target variable

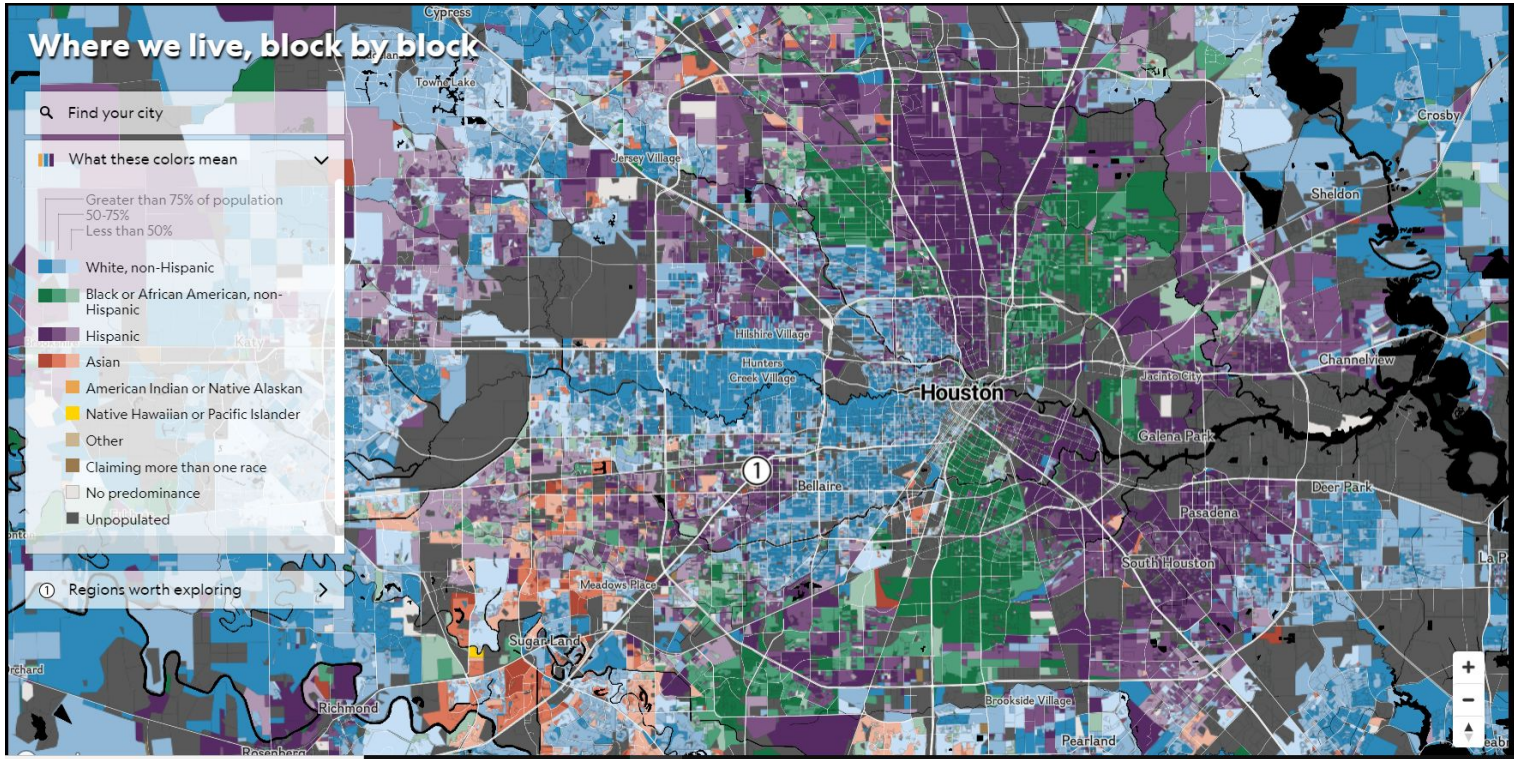
A : The sensitive attribute

Can we be blind to the sensitive attribute and achieve fairness?

The predicted score $R(X, A) \in [0, 1]$

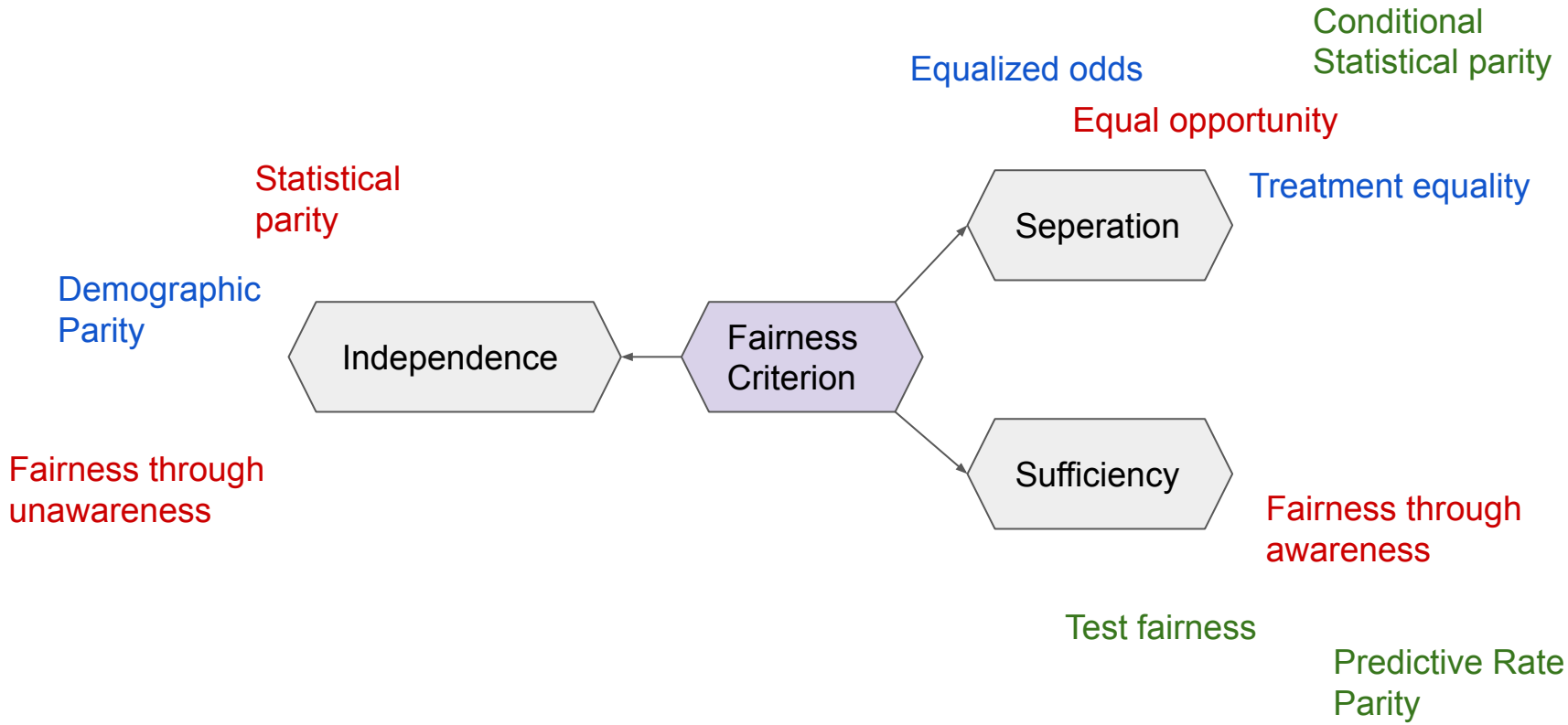
The predicted binary decision $C(X, A)$

Racial Proxies



https://www.nationalgeographic.com/magazine/graphics/diversity-race-ethnicity-united-states-america-interactive-map

Three Fundamental Criterions



Independence

Prediction system is unaware of the objectionable attribute, $C \perp\!\!\!\perp A$

$$P(C = c \mid A = a) = P(C = c \mid A = b)$$

Independence

Prediction system is unaware of the objectionable attribute, $C \perp\!\!\!\perp A$

$$P(C = c \mid A = a) = P(C = c \mid A = b)$$

Approximate versions are more popular, with four-fifth rule for legal proceedings

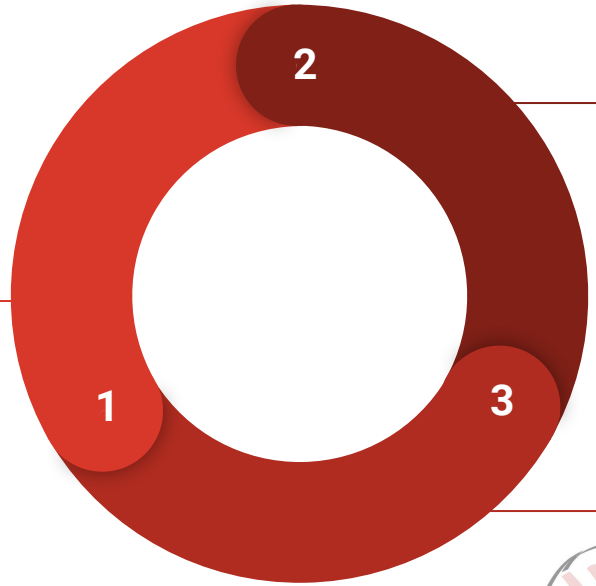
$$|P(C = 1 \mid A = a) - P(C = 1 \mid A = b)| < \epsilon$$

$$(1 - \epsilon) < \frac{P(C = 1 \mid A = a)}{P(C = 1 \mid A = b)} < (1 + \epsilon)$$

How to achieve Independence

Pre Training (Calders et al.)

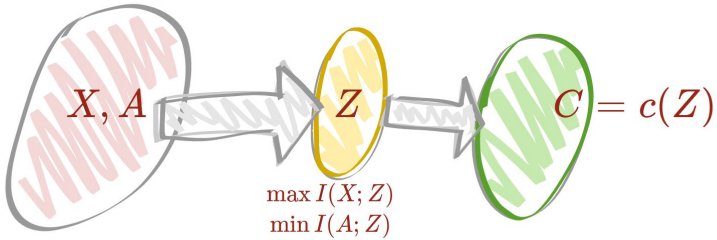
- Change sensitive attribute to create resampled data.
- Change target variables to create resampled data.



Post Training (Feldman et al.)

$$\widetilde{C}(X, A) = \int_a C(X, A = a) dP(a)$$

Representation Learning (Zemel et al.)



How to achieve Independence

Screen qualified good candidates

Screen candidates randomly

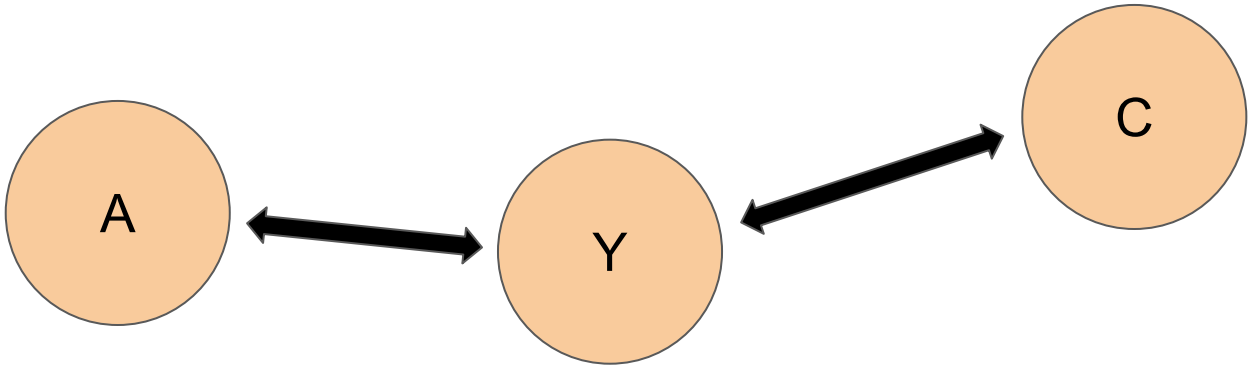
Separation

Prediction does not have more information about objectionable attribute than the original target variable

Separation

Prediction does not have more information about objectionable attribute than the original target variable

$$C \perp\!\!\!\perp A \mid Y$$



Separation

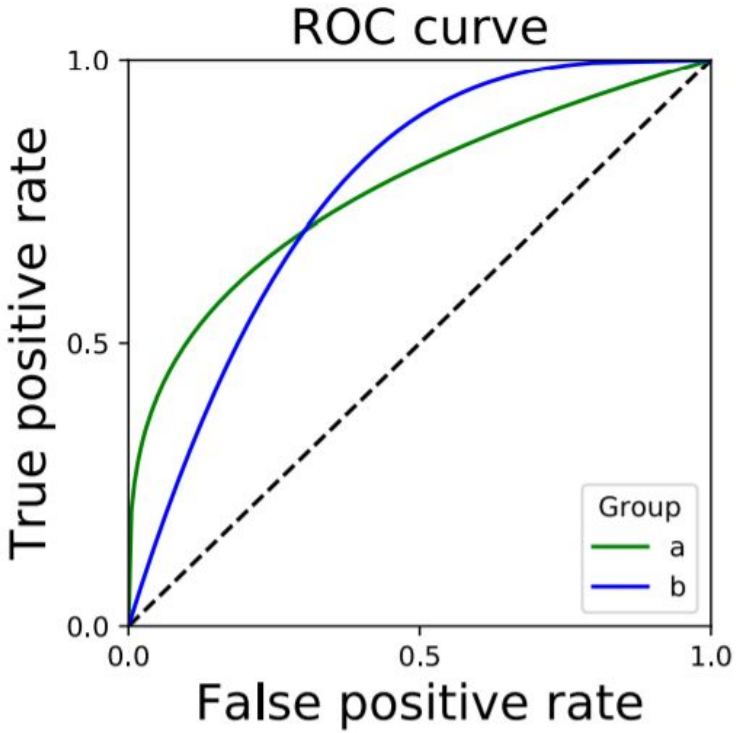
Prediction does not have more information about objectionable attribute than the original target variable

$$C \perp\!\!\!\perp A \mid Y$$

For binary classification setup, false positive (or negative) error rates are equal

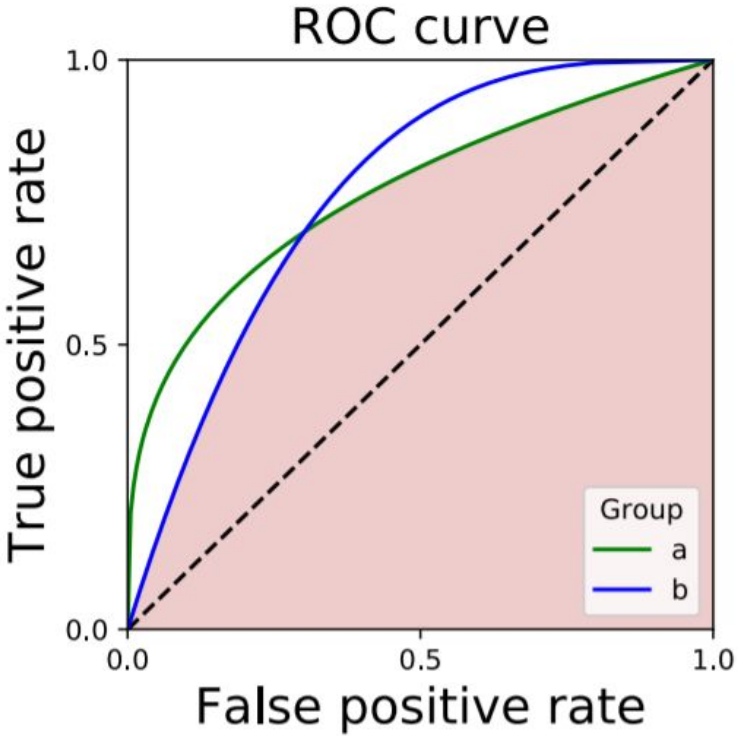
$$P(C = 1 \mid A = a, Y = 1) = P(C = 1 \mid A = b, Y = 1)$$
$$P(C = 1 \mid A = a, Y = 0) = P(C = 1 \mid A = b, Y = 0)$$

How to achieve Separation



Two groups may have different error rates and ROC curve for the same classification score

How to achieve Separation



- Incentive to do better for both groups.
- For two groups, best intersection lies at the intersection point.
- Consequently, perfect prediction is allowed.
- If original score is close to “Bayes optimal”, then post-processing is optimal among all separated scores.

How to achieve Separation

Woodworth et al. (2017) suggested minimizing the loss with a constraint

$$\min_{c \in \mathcal{C}} \ell(c(X, A), Y)$$

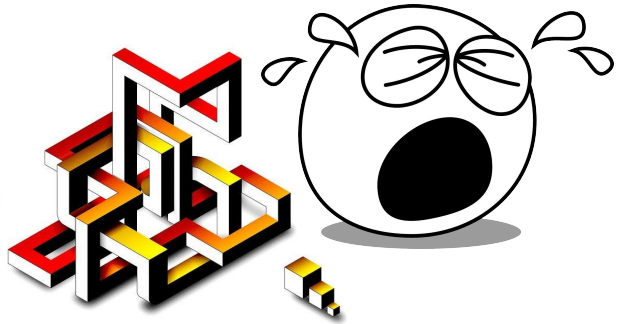
Subject to the constraint $c(X, A) \perp\!\!\!\perp A \mid Y$

How to achieve Separation

Woodworth et al. (2017) suggested minimizing the loss with a constraint

$$\min_{c \in \mathcal{C}} \ell(c(X, A), Y)$$

Subject to the constraint $c(X, A) \perp\!\!\!\perp A \mid Y$



Highly intractable

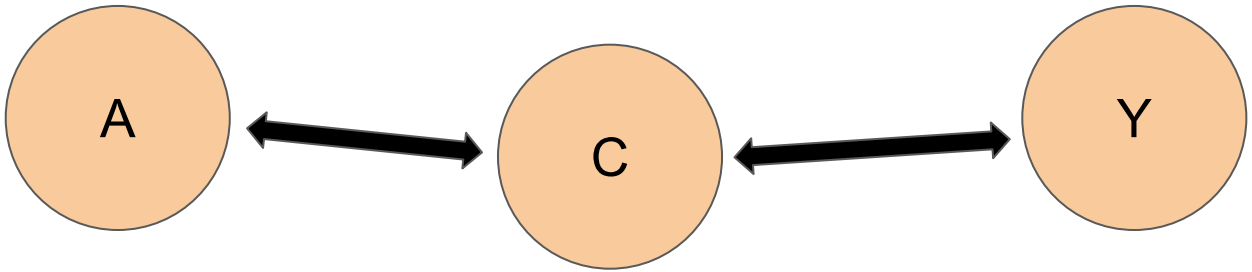
For regression problems with joint normal densities, the constraint can be relaxed to

$$\sigma_{RA}\sigma_Y^2 = \sigma_{RY}\sigma_{YA}$$

Sufficiency

We don’t need to see the sensitive attribute, the prediction is sufficient to explain the responses

$$Y \perp\!\!\!\perp A \mid C$$



Sufficiency

We don’t need to see the sensitive attribute, the prediction is sufficient to explain the responses

$$Y \perp\!\!\!\perp A \mid C$$

Other variants use score values instead of classification decisions

$$Y \perp\!\!\!\perp A \mid R$$

Example: Bayes optimal score, $R(X, A) := E(Y \mid X = x, A = a)$

How to achieve Sufficiency

$$Y \perp\!\!\!\perp A \mid R$$

$$\Rightarrow P(Y = 1 \mid A = a, R = r) = P(Y = 1 \mid A = b, R = r)$$

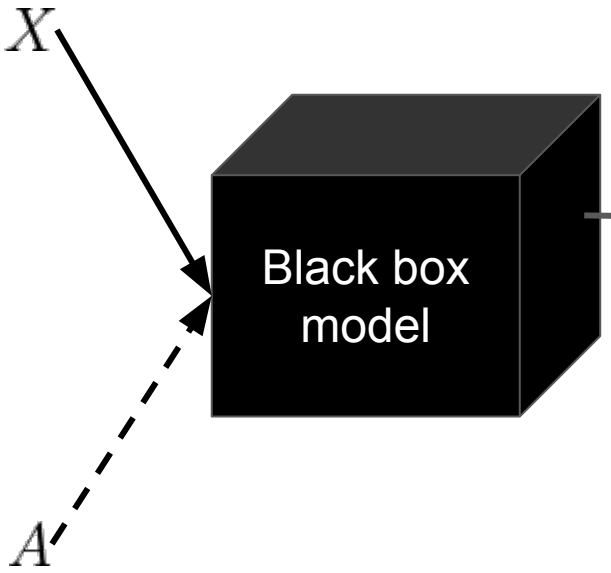
Let, $q(r) = P(Y = 1 \mid A = a, R = r)$

Then, $P(Y = 1 \mid A = a, q(R) = s) = s$

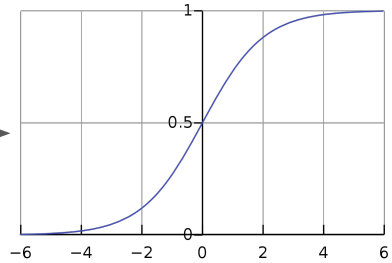
Sufficiency ⇔ Calibration by group

How to find such transformation of score which is calibrated by group

How to achieve Sufficiency (Platt Scaling)



R



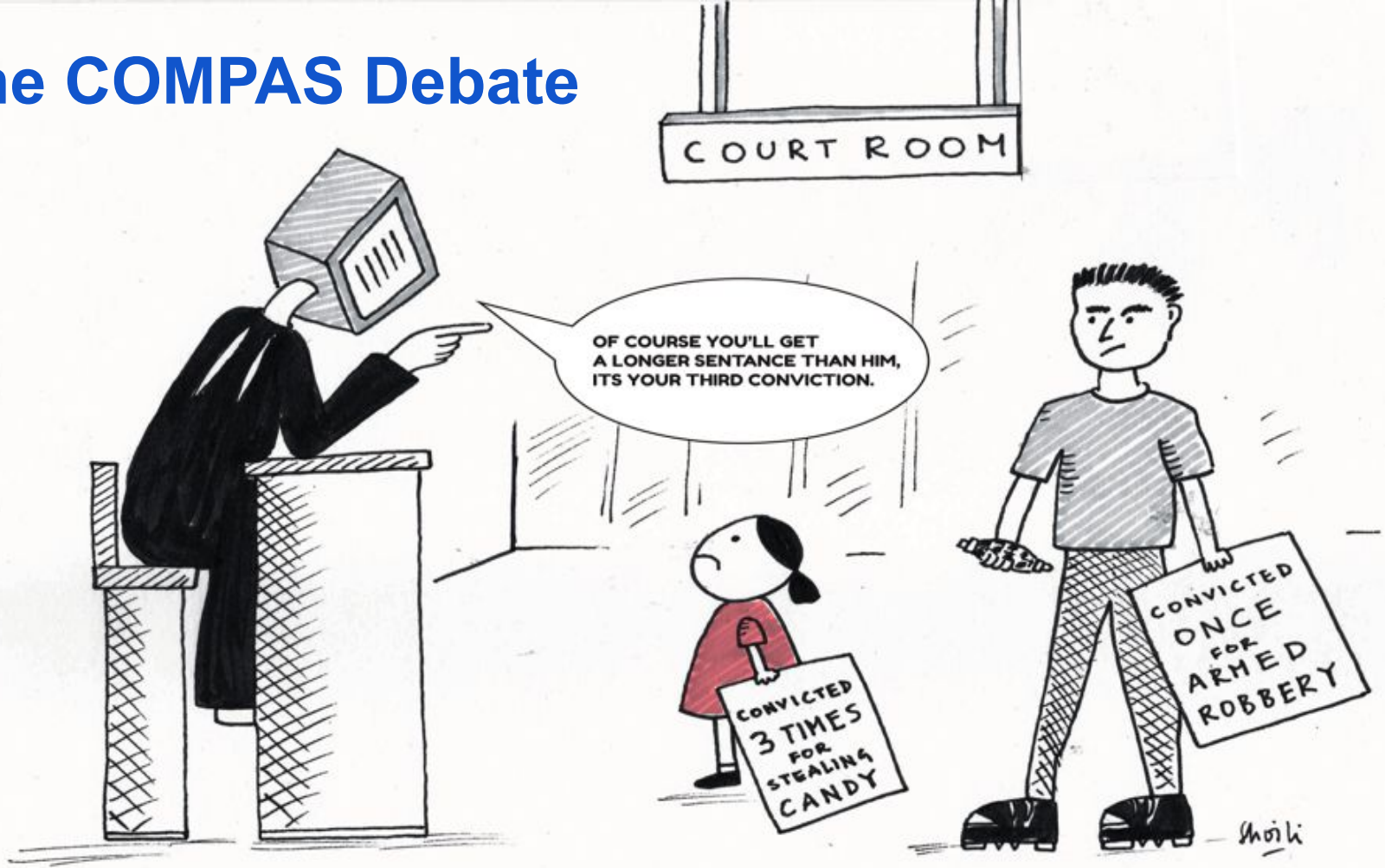
Calibrated score

$q(R)$

$$q(R) = \frac{1}{1 + \exp(aR + b)}$$

Fit a simple logistic regression to match the target variable probabilities from the score function

The COMPAS Debate



The COMPAS Debate

Correctional Offender Management Profiling for Alternative Sanctions is a tool made by Northpointe, Inc. for predicting recidivism.

Till 2016, COMPAS was used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and other jurisdictions for taking the action whether an arrested person should get bail or not.

Data Source: ProPublica obtained two years worth of COMPAS scores from the Broward County Sheriff's Office in Florida in 2013. For more details see https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

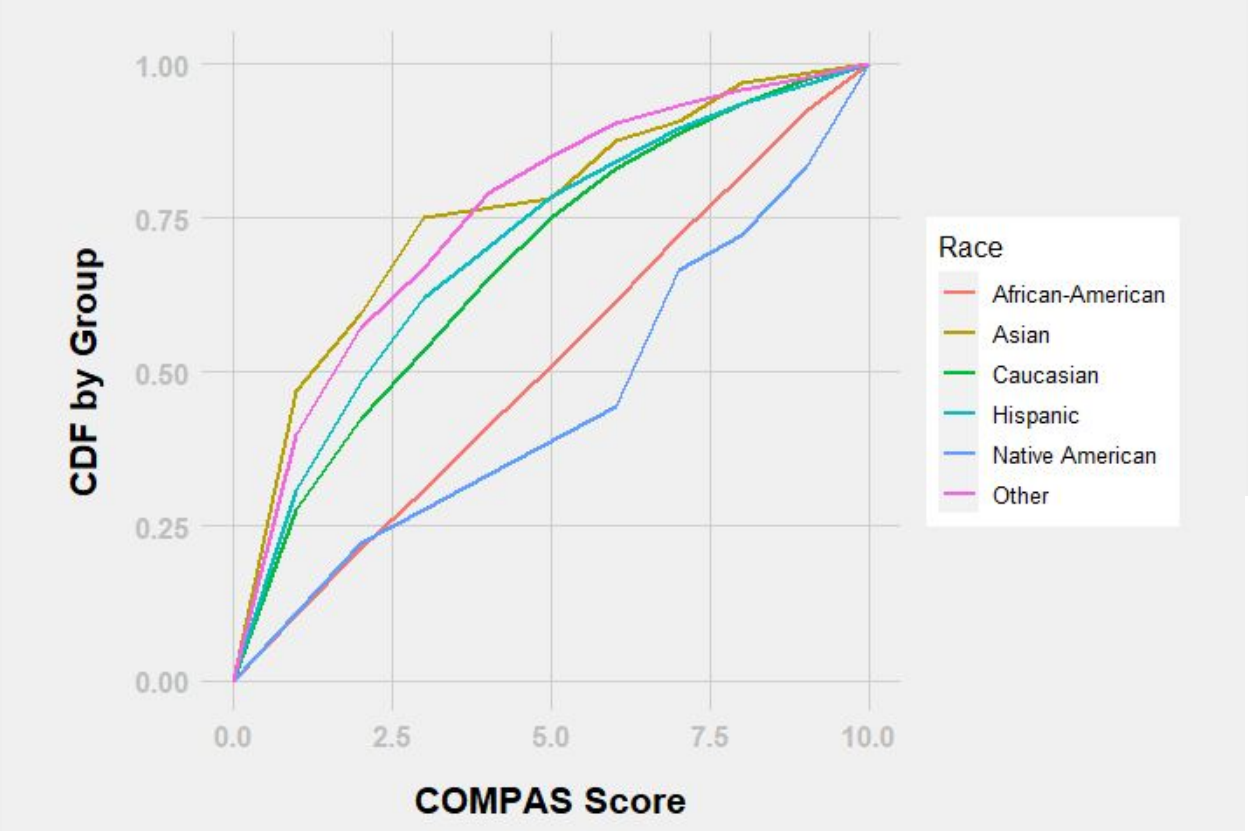
The COMPAS Debate

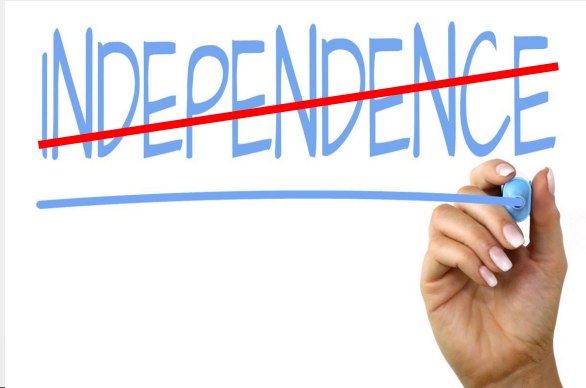
Correctional Offender Management Profiling for Alternative Sanctions is a tool made by Northpointe, Inc. for predicting recidivism.

Till 2016, COMPAS was used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and other jurisdictions for taking the action whether an arrested person should get bail or not.

Data Source: ProPublica obtained two years worth of COMPAS scores from the Broward County Sheriff's Office in Florida in 2013. For more details see https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

The COMPAS Debate

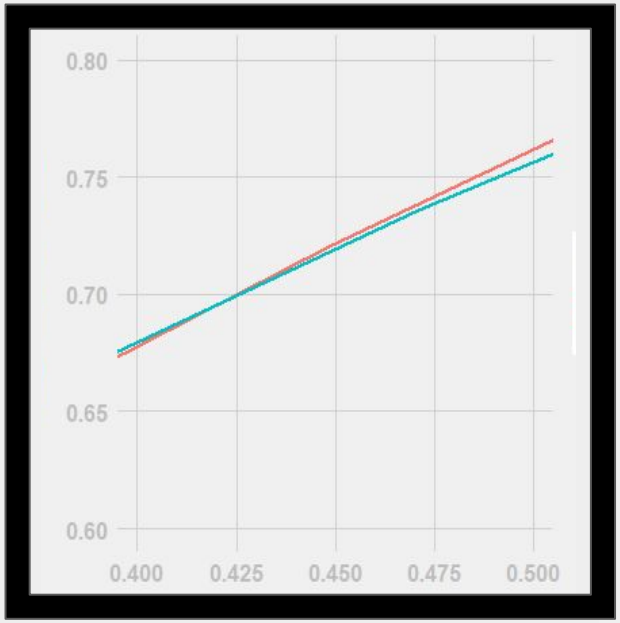




The COMPAS Debate

The score shifts similarly for both groups, implying similar predictive capabilities for both groups

The COMPAS Debate



Achieves separation

The COMPAS Debate

- No independence.
- Blacks are more likely to be assigned high risk score.
- Does not provide individual fairness.

- Trade secret.
- Scores are calibrated by group.
- Accepted to use different thresholds by ethnicity.

Three way battle against discrimination

Assume

$Y \not\perp\!\!\!\perp A$

and

$C \not\perp\!\!\!\perp Y$

and

(Y, C, A)

has positive probabilities (or densities) at all combinations (Price and Sebro)

Space of all classifiers

Three way battle against discrimination

Assume

$Y \not\perp\!\!\!\perp A$

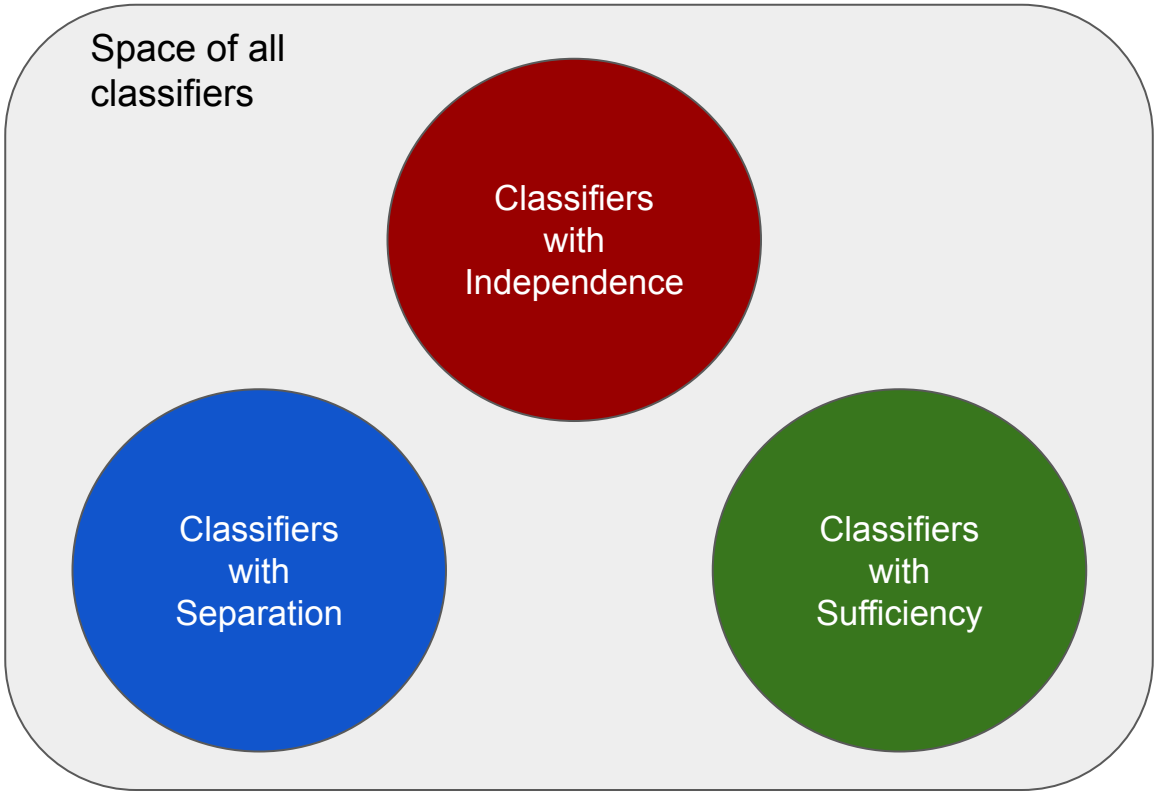
and

$C \not\perp\!\!\!\perp Y$

and

(Y, C, A)

has positive probabilities (or densities) at all combinations (Price and Sebro)



Limitations of Observational Study

A criterion is **observational** if it can be specified using properties of joint distribution of (X, A, C, Y) or (X, A, R, Y)

Price and Sebro (2017) showed that there exists two scenarios where the joint distributions of the variables are exactly same with different interpretations of fairness.

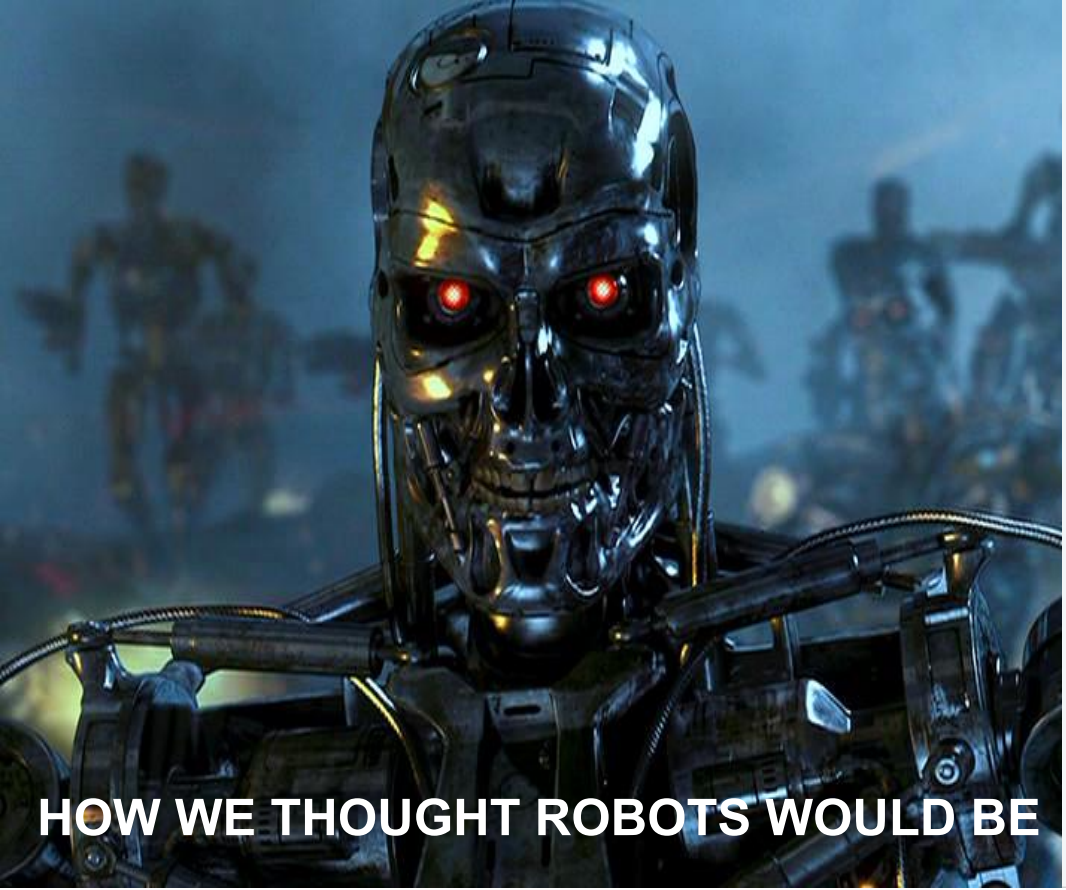
In scenario 1, recruiters hire engineers based on CS degree.
In scenario 2, recruiters hire engineers based on their visit to github.com and pinterest.com

So, the way out?

- Causal Inference.
  - Matching to validate fairness.
  - Difficult to test assumptions in practice.
- Audit Study
  - Similar to randomized control trials.
  - Requires money, time.
- Both needs human expertise to understand the problem domain.
- Google, IBM, Facebook all have developed algorithms to mitigate bias, study long term effects of automated systems.
- We’re all equal (WAE) vs What you see is what you get (WYSIWYG).

Major References

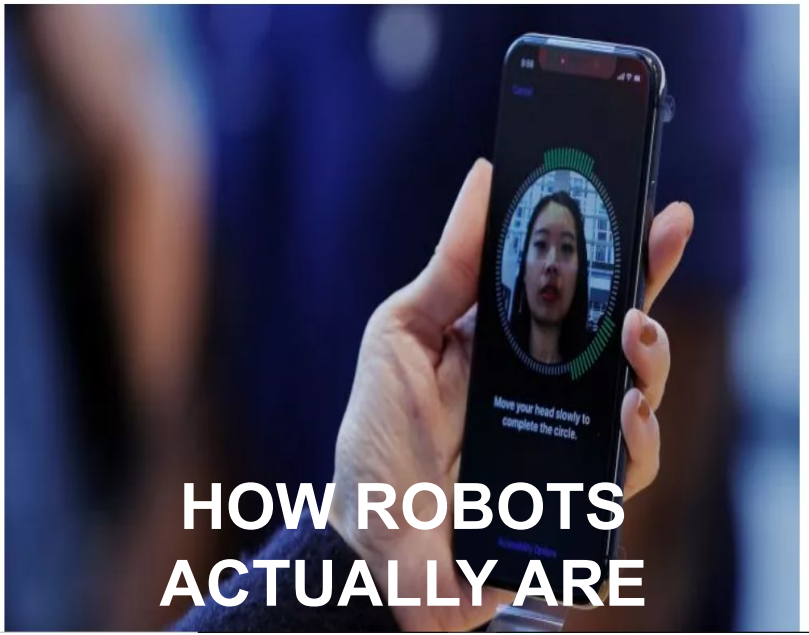
- *Neural Information Processing Systems 2017* Conference talks.
- *FAIRNESS AND MACHINE LEARNING Limitations and Opportunities* by Solon Barocas, Moritz Hardt, Arvind Narayanan
- *21 fairness definitions and their politics* by Arvind Narayanan
- *A Tutorial on Fairness in Machine Learning* by Zihuan Zhong
- *How We Analyzed the COMPAS Recidivism Algorithm* by ProPublica journalists Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin.
- *Fairness in Machine Learning: A Survey* by Simon Caton, Christian Haas.
- *Review of Mathematical frameworks for Fairness in Machine Learning* by Eustasio del Barrio, Paula Gordaliza, Jean-Michel Loubes
- Think with Google ML Fairness for Marketers



WORLD

Is the iPhone X Racist? Apple Refunds Device That Can't Tell Chinese People Apart, Woman Claims

BY CHRISTINA ZHAO ON 12/18/17 AT 12:24 PM EST



THANK YOU