Responsible Data Science

Part 1: intro, algorithmic fairness, diversity

Prof. Julia Stoyanovich

Computer Science and Engineering &
Center for Data Science
New York University
@stoyanoj

https://dataresponsibly.github.io/
https://dataresponsibly.github.io/courses/

NYU

The power of data science

Power

unprecedented data collection capabilities

enormous computational power

ubiquity and broad acceptance

Opportunity

improve people's lives, e.g., recommendation

accelerate scientific discovery, e.g., medicine

boost innovation, e.g., autonomous cars

transform society, e.g., open government

optimize business, e.g., advertisement targeting

goal **- progress**

and now some bad news

Online price discrimination

THE WALL STREET JOURNAL.

WHAT THEY KNOW

Websites Vary Prices, Deals Based on Users' Information

By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and ASHKAN SOLTANI
December 24, 2012

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell's screen, just a few miles away, was $14.29.

A key difference: where Staples seemed to think they were located.

WHAT PRICE WOULD YOU SEE?

lower prices offered to buyers who live in more affluent neighborhoods

https://www.wsj.com/articles/SB10001424127887323777204578189391813881534

Amazon same-day delivery

Bloomberg

Amazon Doesn't Consider the Race of Its Customers. Should It?

"… In six major same-day delivery cities, however, **the service area excludes predominantly black ZIP codes** to varying degrees, according to a Bloomberg analysis that compared Amazon same-day delivery areas with U.S. Census Bureau data."

https://www.bloomberg.com/graphics/2016-amazon-same-day/

Amazon same-day delivery

Bloomberg

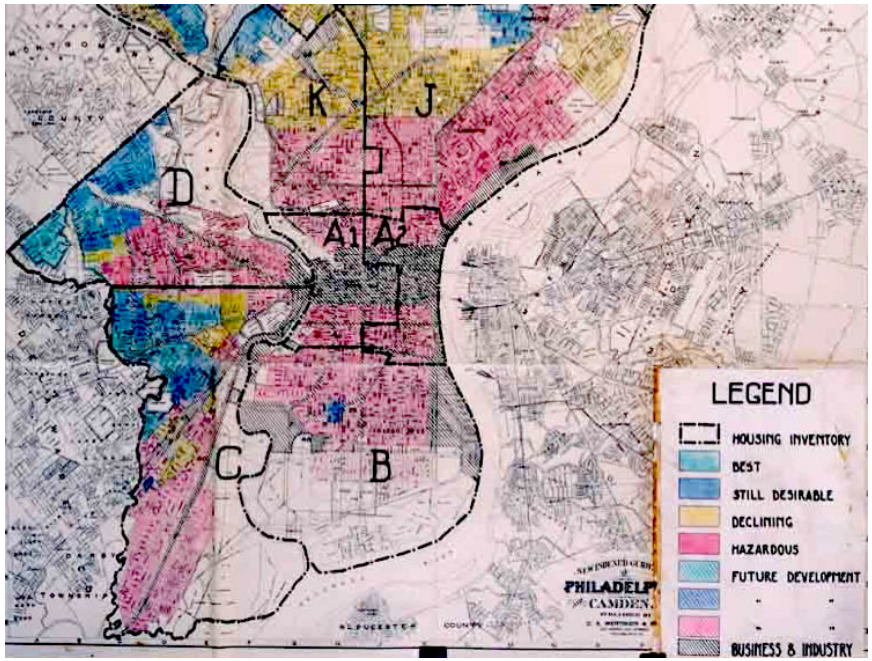
Amazon Doesn't Consider the Race of Its Customers. Should It?

"The most striking gap in Amazon's same-day service is in Boston, where **three ZIP codes encompassing the primarily black neighborhood of Roxbury are excluded** from same-day service, while the neighborhoods that surround it on all sides are eligible."

https://www.bloomberg.com/graphics/2016-amazon-same-day/

Redlining

Redlining is the practice of arbitrarily denying or limiting **financial services** to specific neighborhoods, generally because its residents are people of color or are poor.



A HOLC 1936 security map of Philadelphia showing redlining of lower income neighborhoods

https://en.wikipedia.org/wiki/Redlining

Households and businesses in the **red zones** could not get mortgages or business loans.

Online job ads

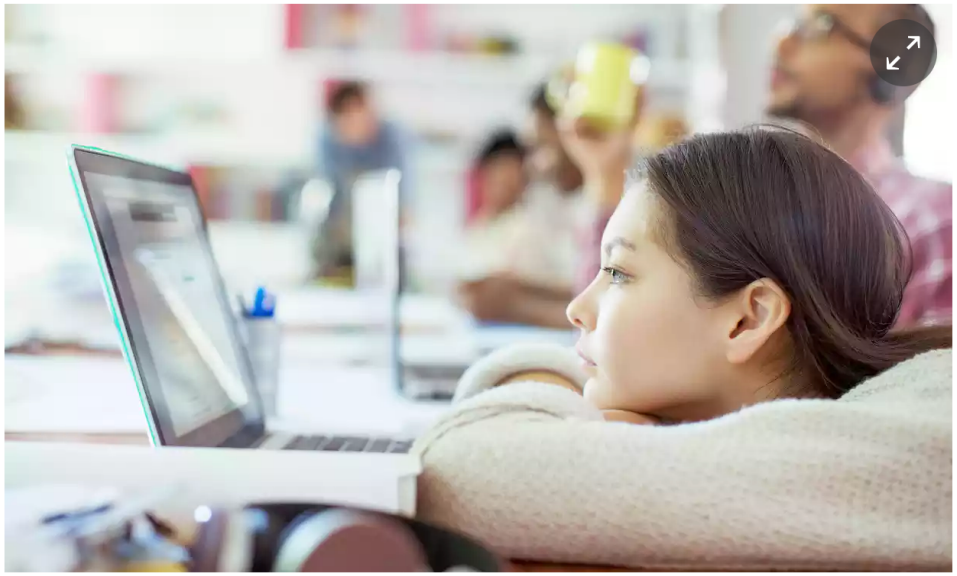
theguardian

Samuel Gibbs

Wednesday 8 July 2015 11.29 BST

Women less likely to be shown ads for high-paid jobs on Google, study shows

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs



One experiment showed that Google displayed adverts for a career coaching service for executive jobs 1,852 times to the male group and only 318 times to the female group. Photograph: Alamy

The AdFisher tool simulated job seekers that did not differ in browsing behavior, preferences or demographic characteristics, except in gender.

One experiment showed that Google displayed ads for a career coaching service for "$200k+" executive jobs **1,852 times to the male group and only 318 times to the female group**. Another experiment, in July 2014, showed a similar trend but was not statistically significant.

https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

Gender bias in recruiting

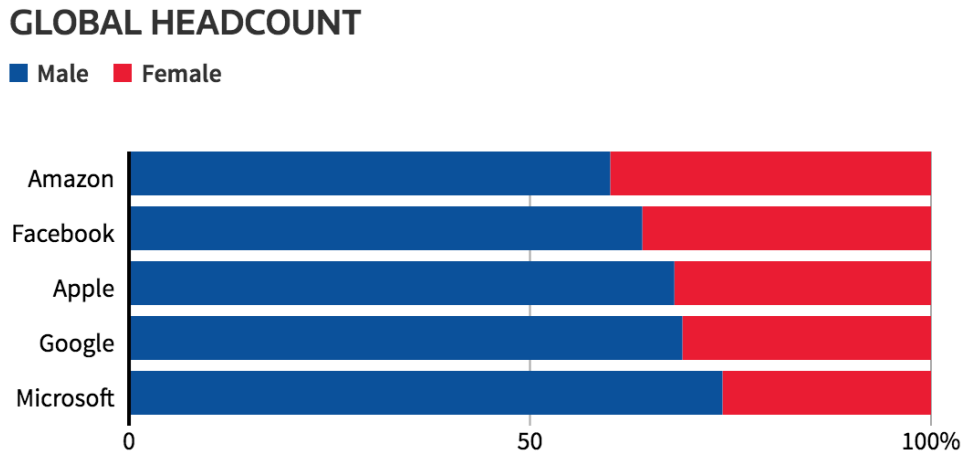
REUTERS

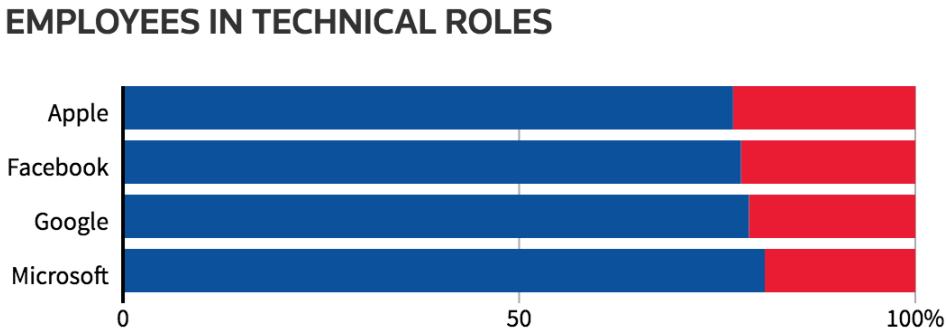
Jeffrey Dastin

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / 6 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

"In effect, **Amazon's system taught itself that male candidates were preferable**. It penalized resumes that included the word "women's," as in "women's chess club captain." And it **downgraded graduates of two all-women's colleges**, according to people familiar with the matter. They did not specify the names of the schools."

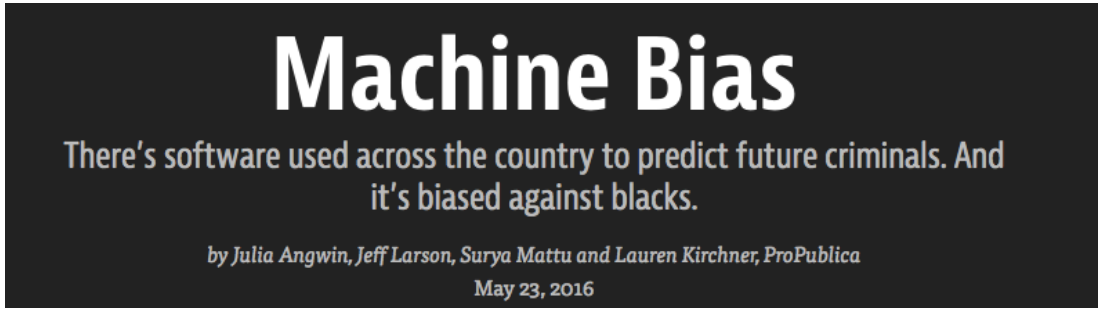




"Note: Amazon does not disclose the gender breakdown of its technical workforce."

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Racial bias in criminal sentencing



A commercial tool **COMPAS** automatically predicts some categories of future crime to assist in bail and sentencing decisions. It is used in courts in the US.

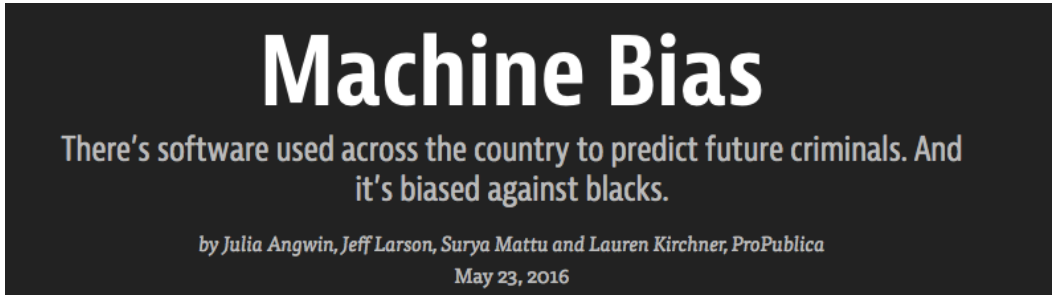
The tool correctly predicts recidivism **61% of the time.**

Blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend.

The tool 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.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Racial bias in criminal sentencing



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

A commercial tool **COMPAS** automatically predicts some categories of future crime to assist in bail and sentencing decisions. It is used in courts in the US.

Prediction Fails Differently for Black Defendants

	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.)*

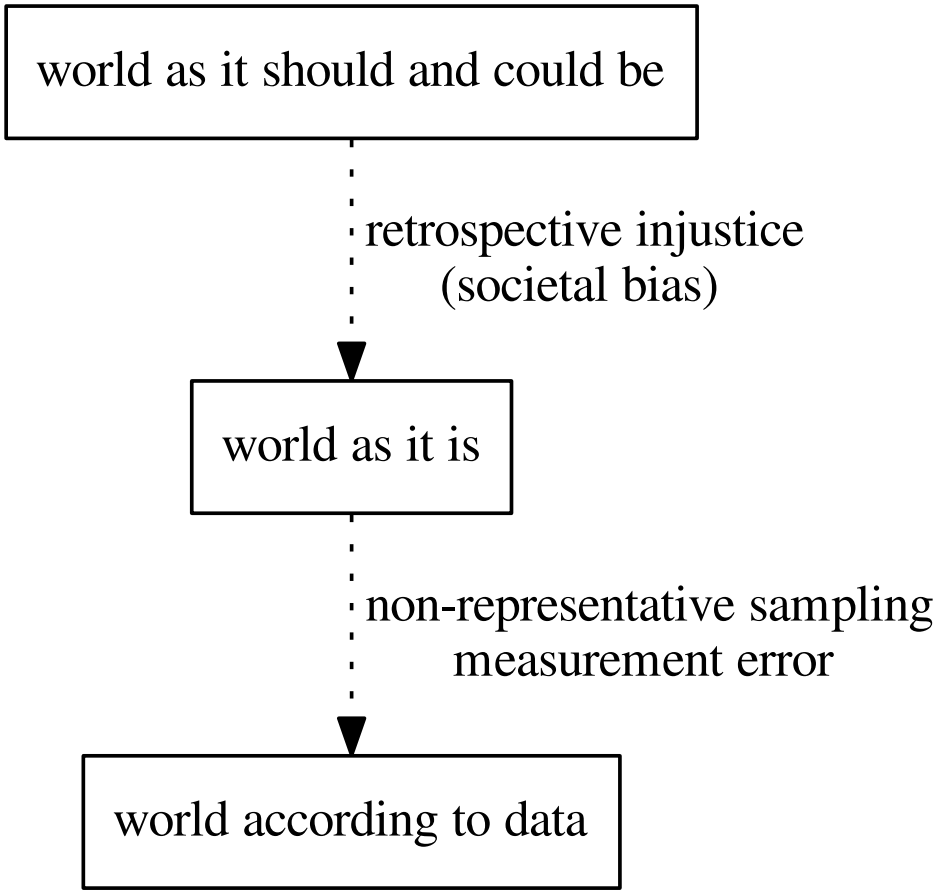
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Fairness is lack of “bias”

- What do we mean by **bias**?
  - **statistical bias**: a model is biased if it doesn’t summarize the data correctly
  - **societal bias**: a dataset or a model is biased if it does not represent the world “correctly”, e.g., data is not representative, there is measurement error, or the **world is “incorrect”**

the world as it is or as it should be?

“Biased data”



from “Prediction-Based Decisions and Fairness” by Mitchell, Potash and Barocas, 2018

when data is about people, bias can lead to discrimination

The evils of discrimination

Disparate treatment is the illegal practice of treating an entity, such as a creditor or employee, differently based on a **protected characteristic** such as race, gender, age, religion, sexual orientation, or national origin.

Disparate impact is the result of systematic disparate treatment, where disproportionate **adverse impact** is observed on members of a **protected class**.

http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

The punchline

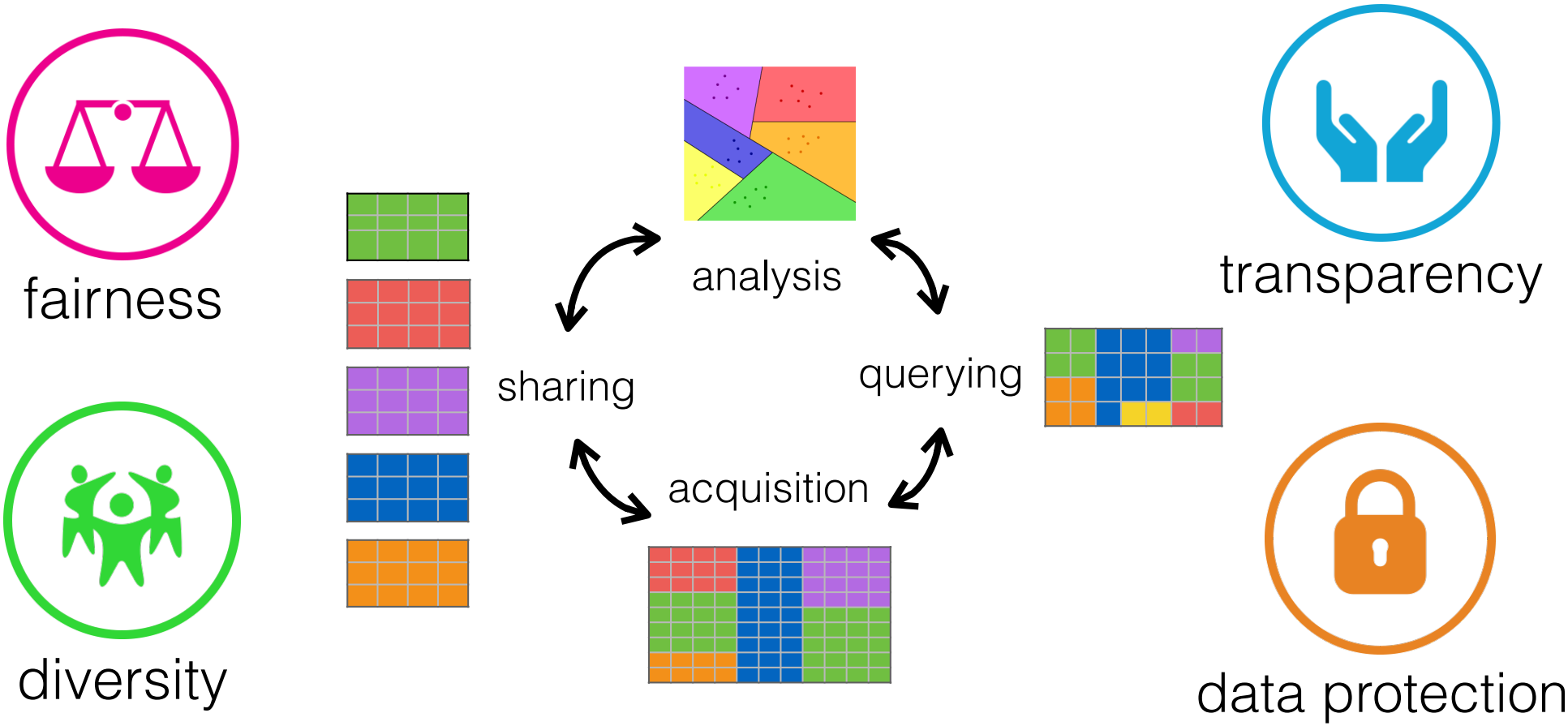
Data science is algorithmic, therefore it cannot be biased! And yet…

- All traditional evils of **discrimination**, and many new ones, exhibit themselves in the data science eco system
- **Bias** that is inherent in the data or in the process, and that is often due to systemic discrimination, is propelled and amplified
- **Transparency** helps prevent discrimination, enable public debate, establish **trust**
- Technology alone won't do: also need **policy**, **user involvement** and **education**

http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

Data, responsibly

Because of its **power**, data science must be used **responsibly**



… with a holistic view of the **lifecycle**

fairness in classification

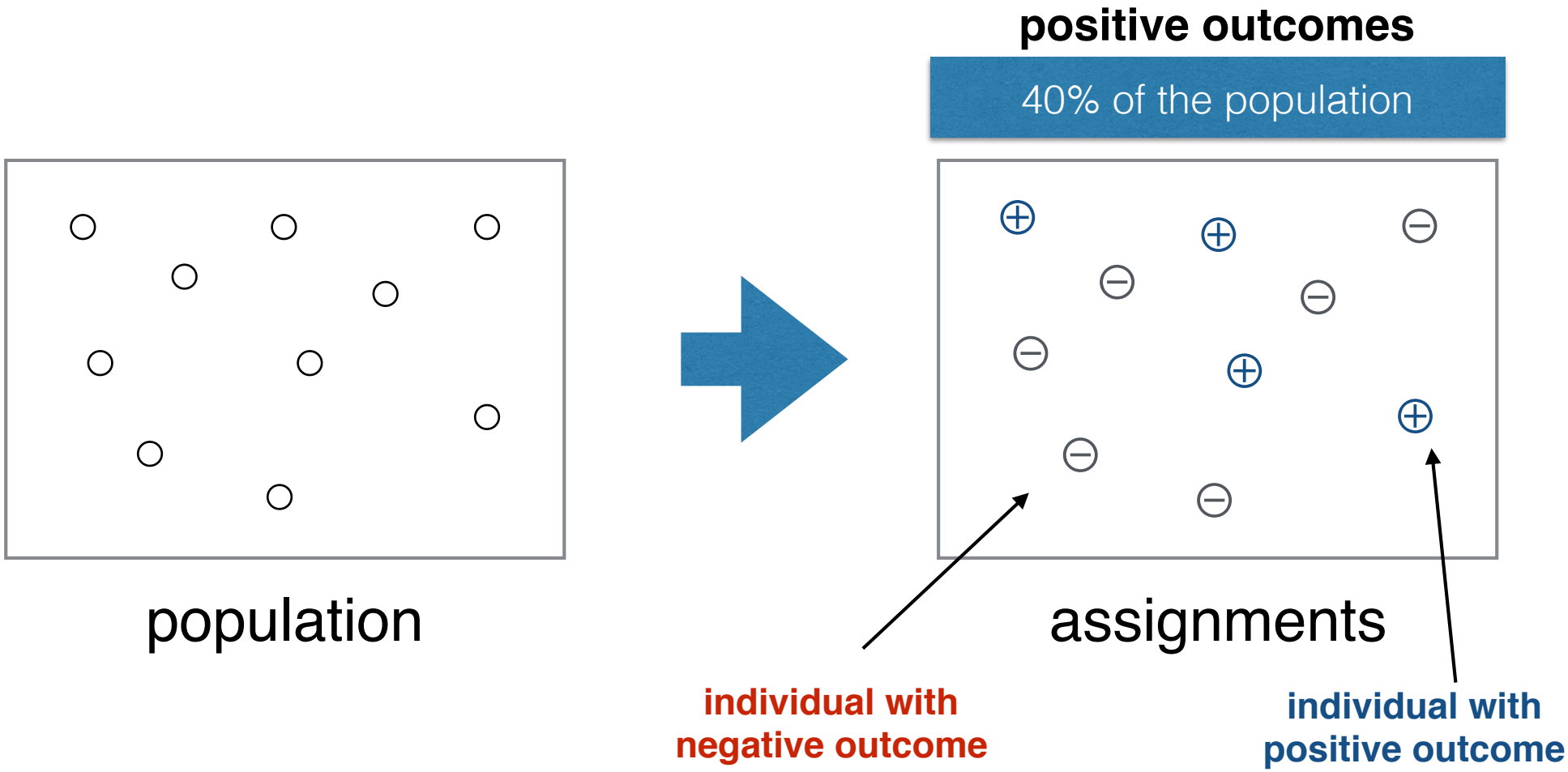
Vendors and outcomes

Consider a **vendor** assigning positive or negative **outcomes** to individuals.

Positive Outcomes	Negative Outcomes
offered employment	not offered employment
accepted to school	not accepted to school
offered a loan	denied a loan
offered a discount	not offered a discount

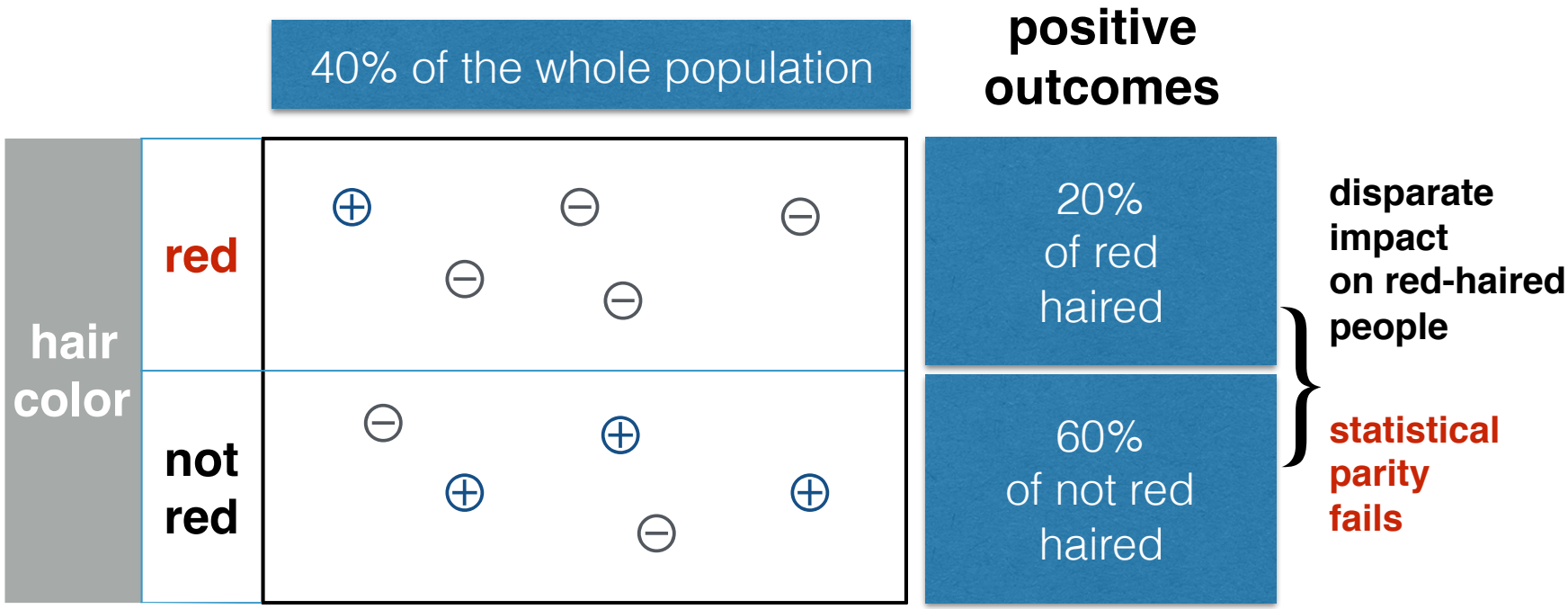
Assigning outcomes to populations

Fairness is concerned with how outcomes are assigned to a population



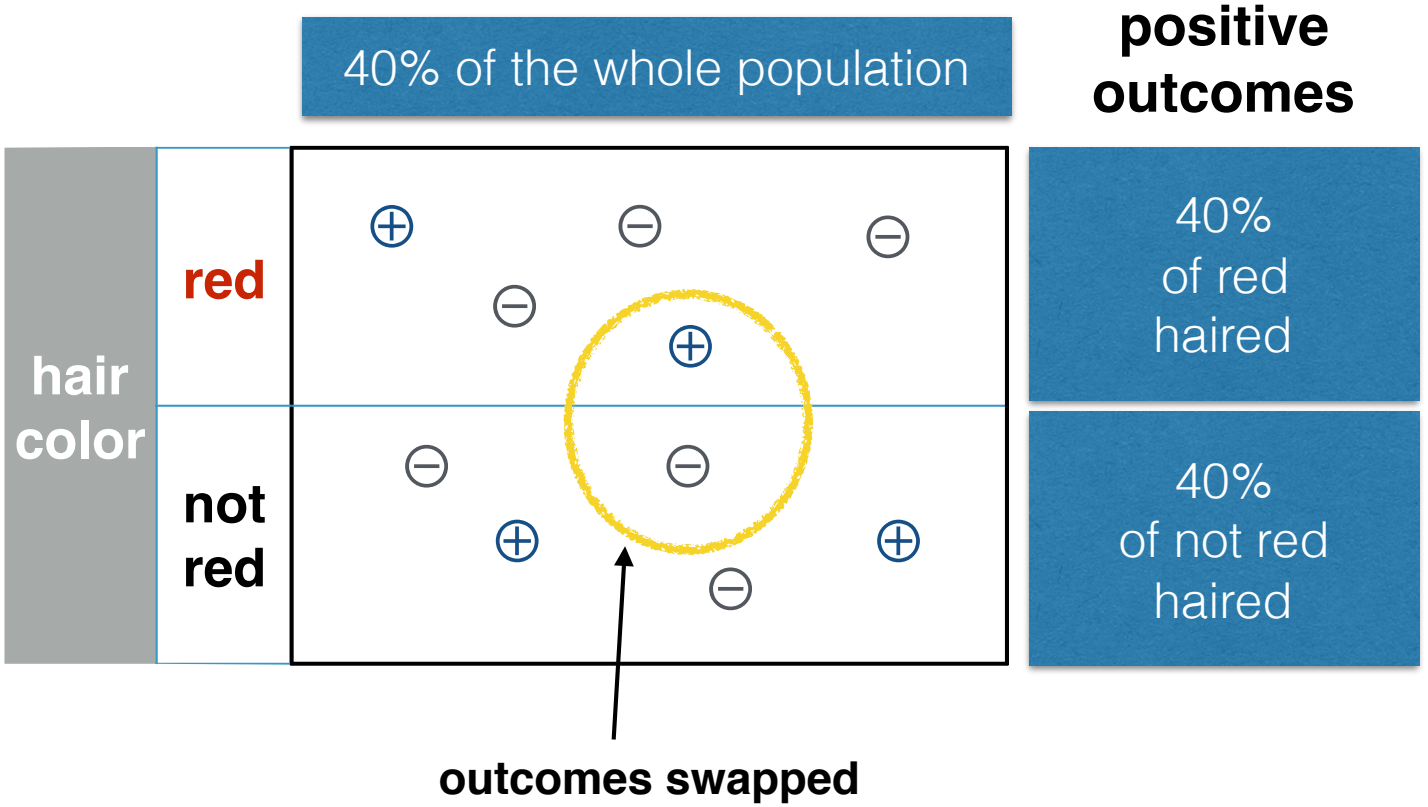
Sub-populations may be treated differently

Sub-population: those with red hair
(under the same assignment of outcomes)



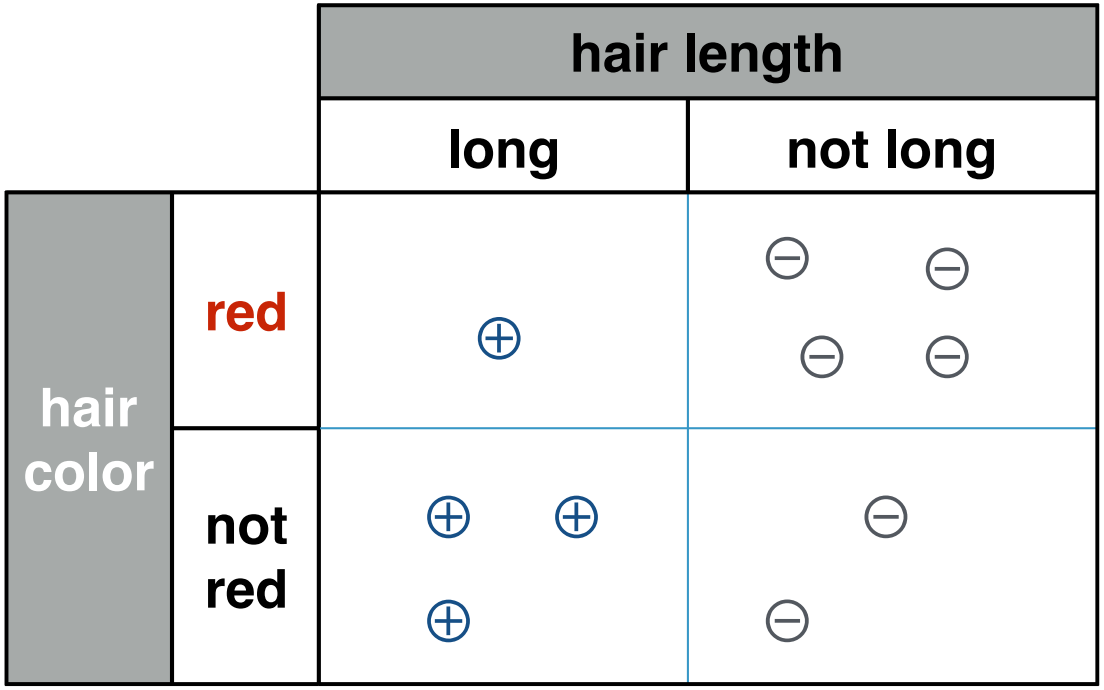
Statistical parity

Statistical parity (a popular **group fairness** measure)
demographics of the individuals receiving any outcome are the same as demographics of the underlying population



Redundant encoding

Now consider the assignments under both **hair color** (protected) and **hair length** (innocuous)



positive outcomes

20% of red haired

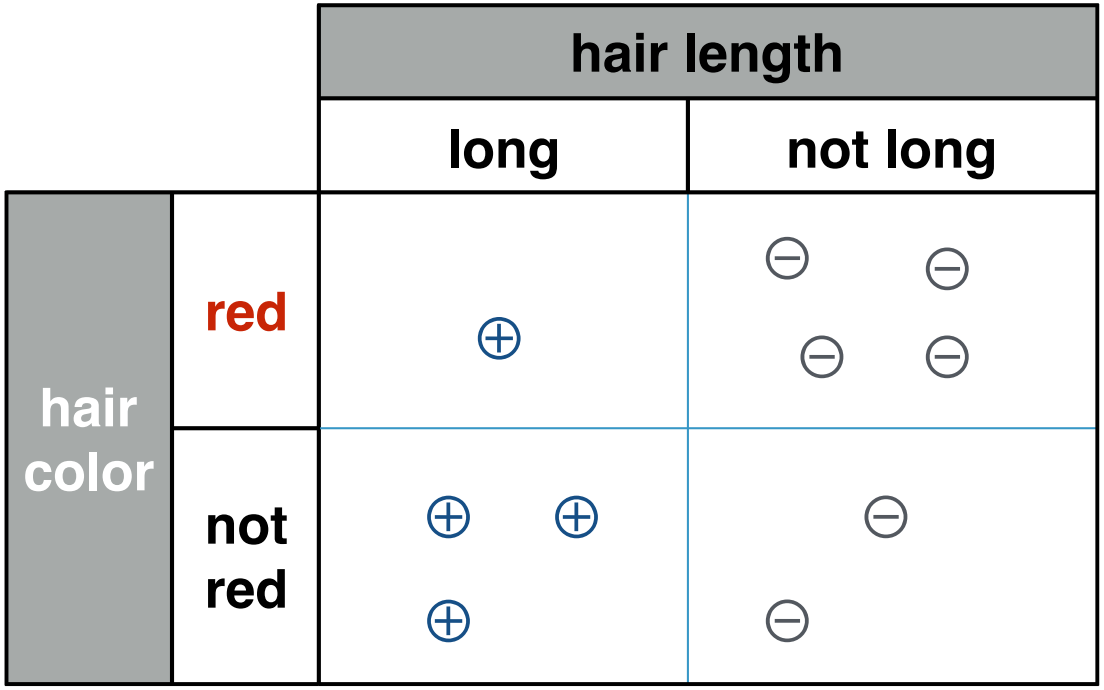
60% of not red haired

Deniability

The vendor has adversely impacted red-haired people, but claims that outcomes are assigned according to hair length.

Blinding is not an excuse

Removing **hair color** from the vendor's assignment process does not prevent discrimination!



positive outcomes

20% of red haired

60% of not red haired

Assessing disparate impact

Discrimination is assessed by the effect on the protected sub-population, not by the input or by the process that lead to the effect.

Redundant encoding

Let's replace hair color with **race** (protected), hair length with **zip code** (innocuous)

		zip code	
		10025	**10027**
race	**black**	⊕	⊖ ⊖ ⊖ ⊖
	white	⊕ ⊕ ⊕	⊖ ⊖

positive outcomes

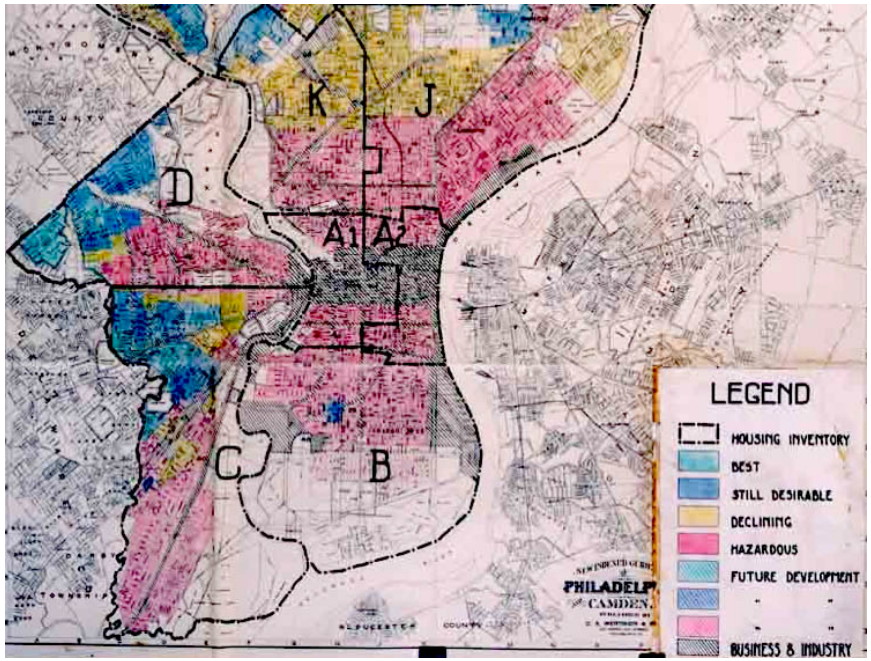
20% of black

60% of white

Redlining

Redlining is the **illegal** practice of arbitrarily denying or limiting financial services to specific neighborhoods, generally because its residents are people of color or are poor.

Philadelphia, 1936



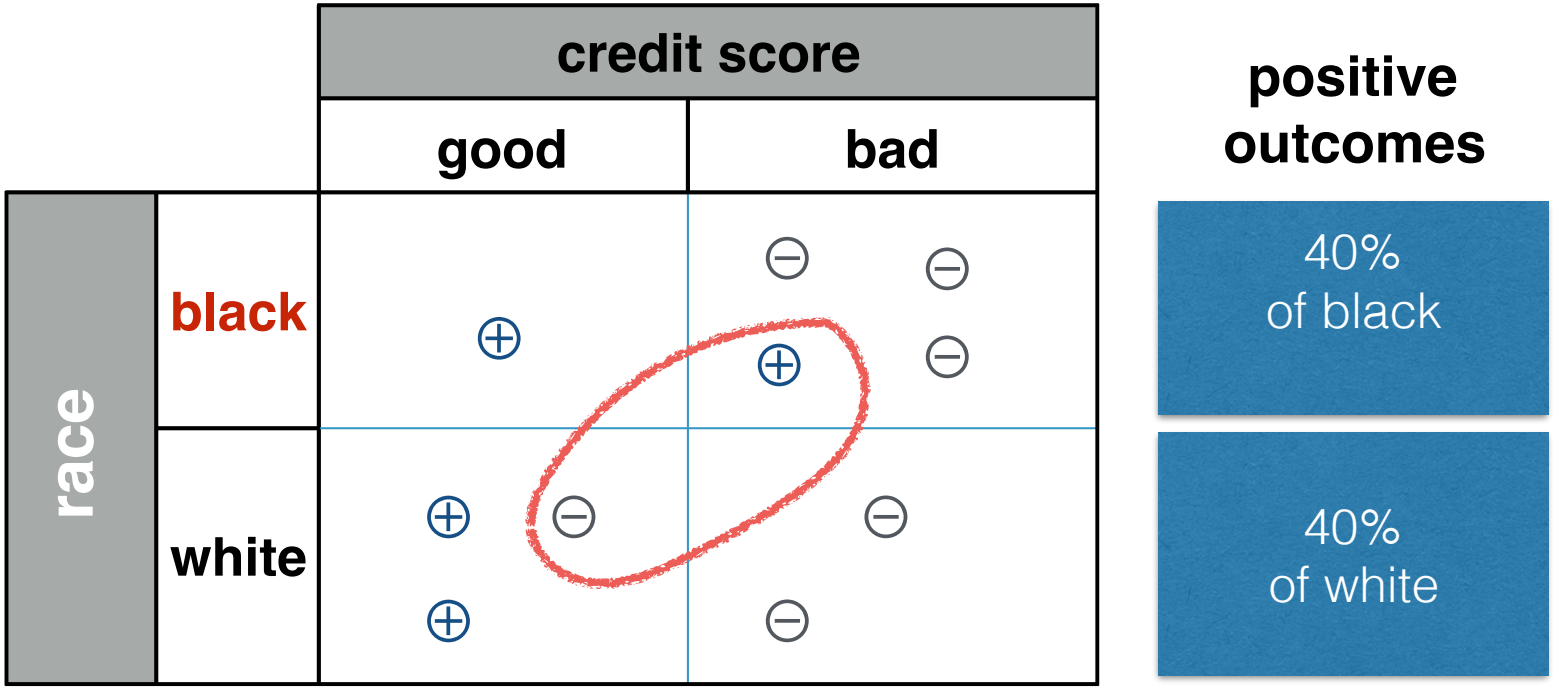
wikipedia

Households and businesses in the red zones could not get mortgages or business loans.

Imposing statistical parity

May be contrary to the goals of the vendor

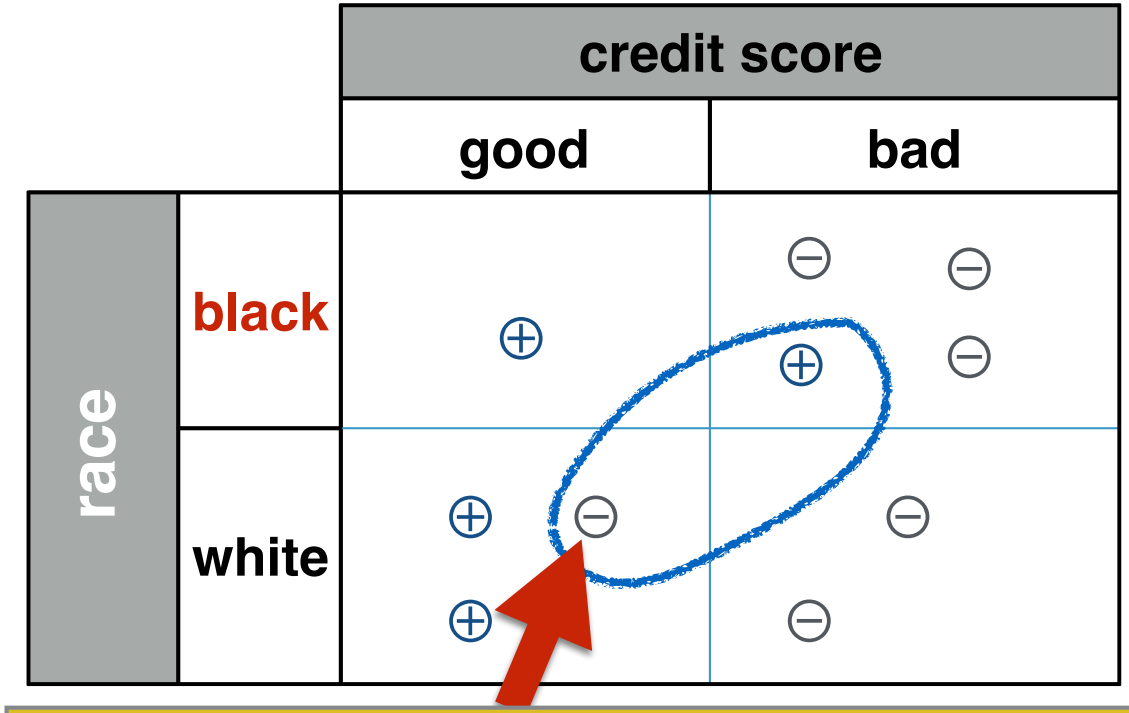
positive outcome: offered a loan



But it's impossible to predict loan payback accurately.
Predictions use past information, which may itself be biased.

Is statistical parity sufficient?

Statistical parity (a popular **group fairness** measure)
demographics of the individuals receiving any outcome are the same as demographics of the underlying population



positive outcomes

40% of black

40% of white

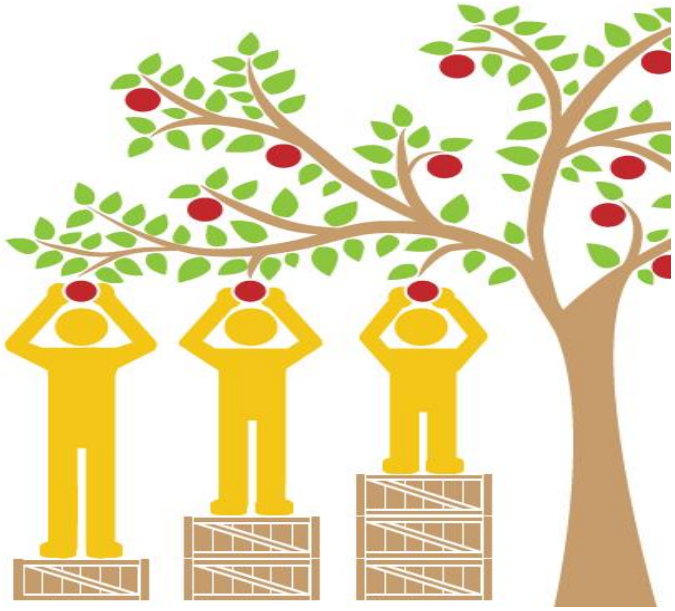
Individual fairness
any two individuals who are similar w.r.t. a particular task should receive similar outcomes

Two notions of fairness

individual fairness

equality

group fairness



equity

two intrinsically different world views

Ricci v. DeStefano (2009)

Supreme Court Finds Bias Against White Firefighters

By **ADAM LIPTAK** JUNE 29, 2009

The New York Times

Karen Lee Torre, left, a lawyer who represented the New Haven firefighters in their lawsuit, with her clients Monday at the federal courthouse in New Haven. Christopher Capozziello for The New York Times

Case opinions	
Majority	Kennedy, joined by Roberts, Scalia, Thomas, Alito
Concurrence	Scalia
Concurrence	Alito, joined by Scalia, Thomas
Dissent	Ginsburg, joined by Stevens, Souter, Breyer

Laws applied

Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e et seq.

On the (im)possibility of fairness

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

Goal: tease out the difference between beliefs about fairness and mechanisms that logically follow from those beliefs.

Construct Space (CS)	Observed Space (OS)	Decision Space (DS)
intelligence	SAT score	performance in college
grit	high-school GPA	
propensity to commit crime	family history	recidivism
risk-averseness	age	

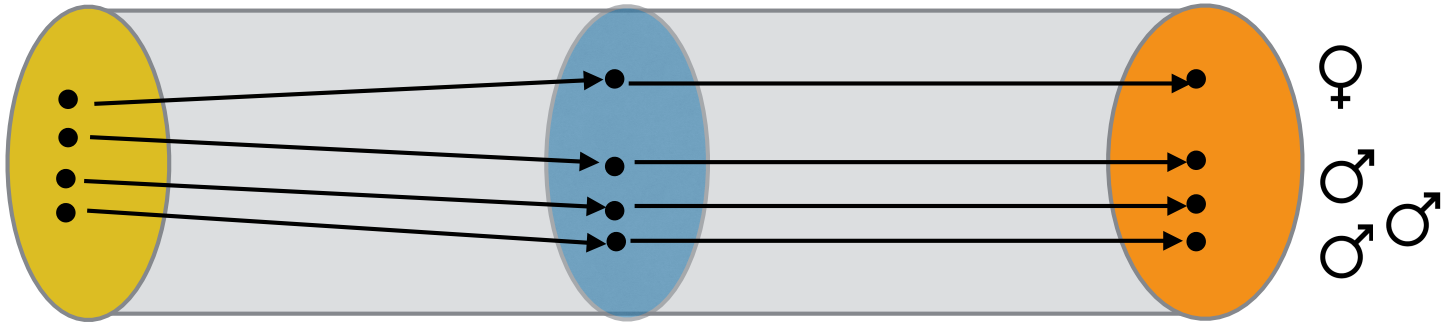
Fairness through mappings

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

Fairness: a mapping from CS to DS is (ε, ε')-fair if two objects that are no further than ε in CS map to objects that are no further than ε' in DS.

$$f : CS \rightarrow DS \qquad d_{CS}(x,y) < \varepsilon \Rightarrow d_{DS}(f(x), f(y)) < \varepsilon'$$

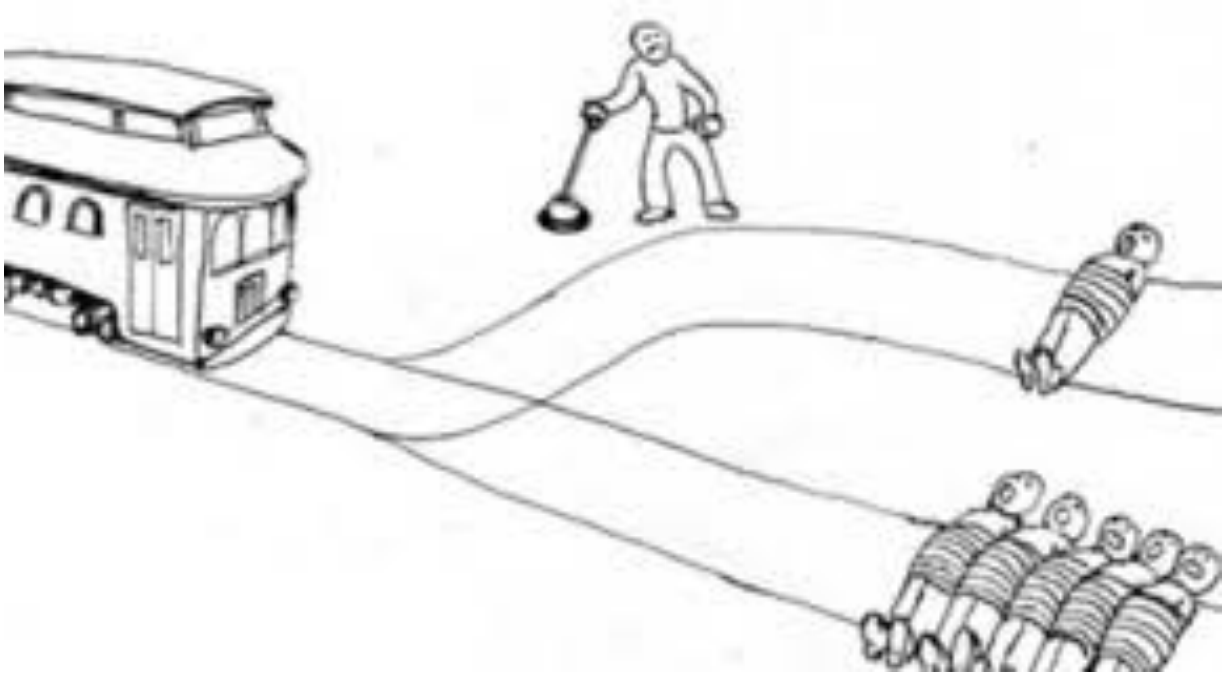
Construct Space (**CS**) Observed Space (**OS**) Decision Space (**DS**)



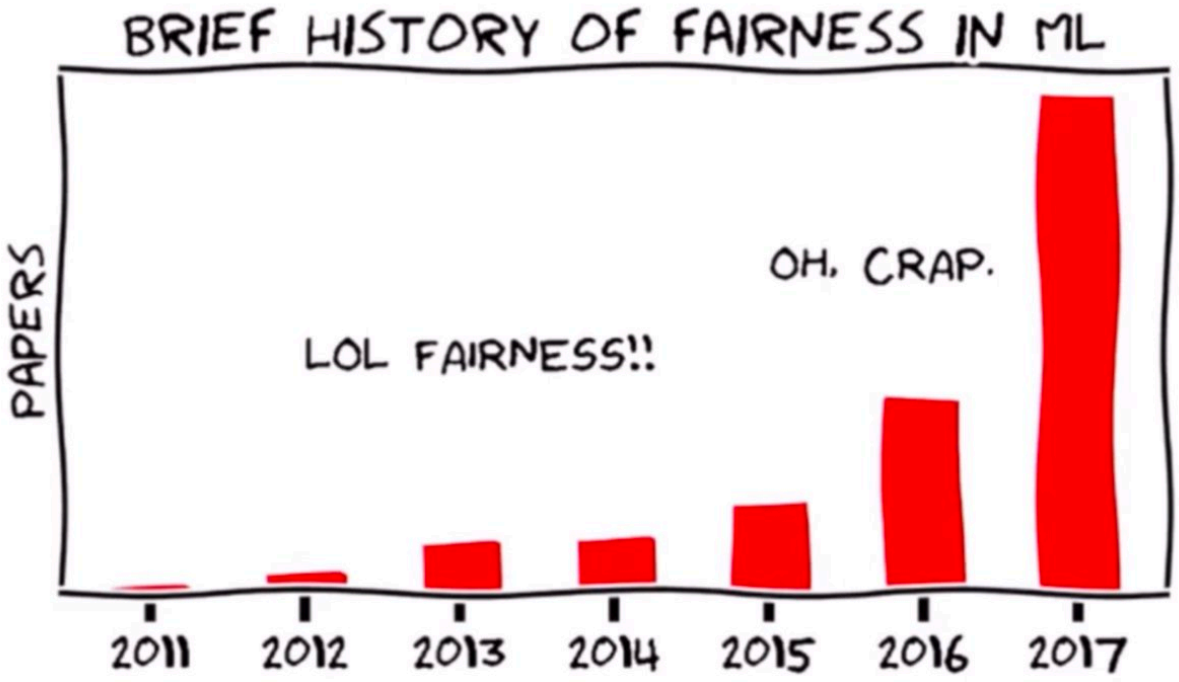
Individual fairness: The mapping from CS to OS has low distortion. That is, OS faithfully represents CS. (**WYSIWYG**)

Group fairness: The mapping from CS to OS has **structural bias,** a distortion that aligns with group structure of CS. (**WAE**)

Fairness definitions as “trolley problems”



Fairness in machine learning

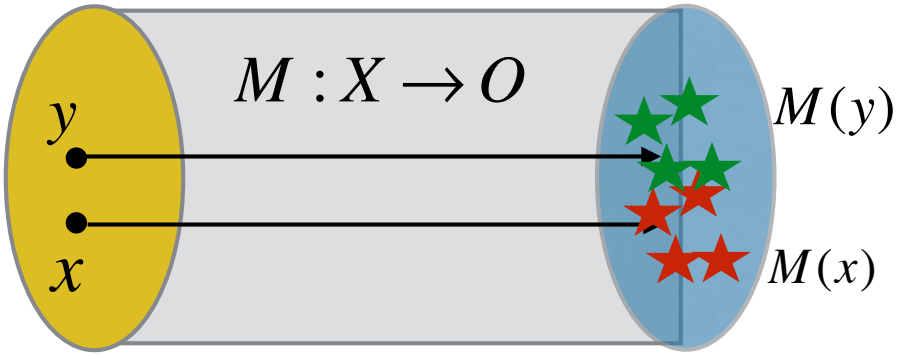


by Moritz Hardt

Fairness through awareness

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

Fairness: Individuals who are **similar** for the purpose of classification task should be **treated similarly**.



X individuals ***O*** outcomes

A task-specific similarity metric is given $d(x,y)$

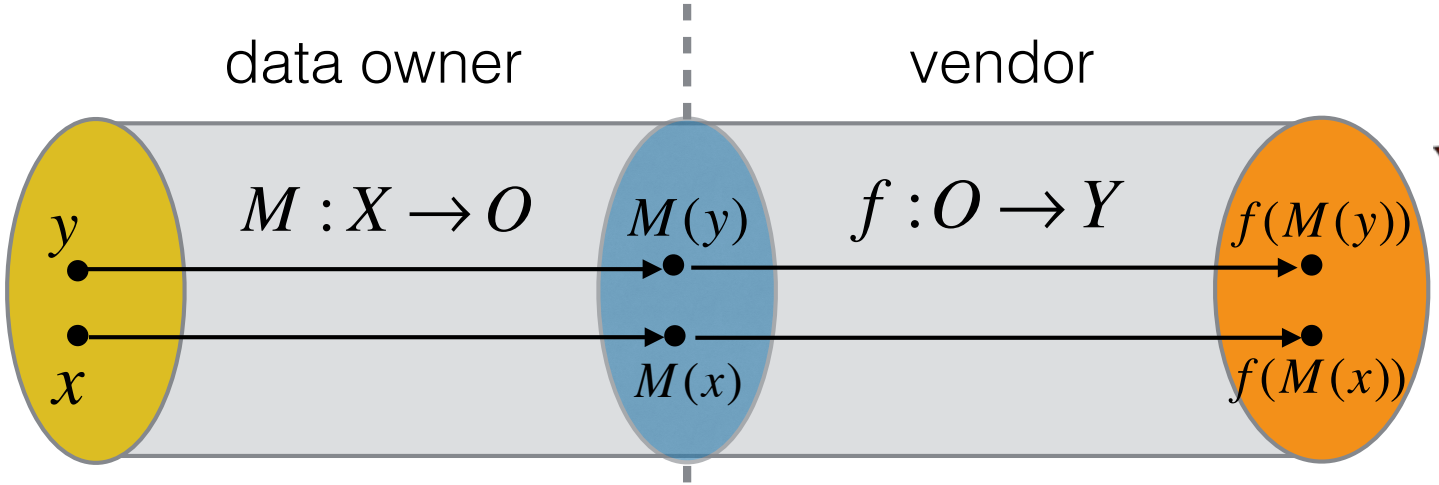
$M : X \to O$ is a **randomized mapping**: an individual is mapped to a distribution over outcomes

M is a Lipschitz mapping if $\forall x, y \in X \quad ||M(x), M(y)|| \leq d(x,y)$

close individuals map to close distributions

Fairness through awareness

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]



simpsons.wikia.com

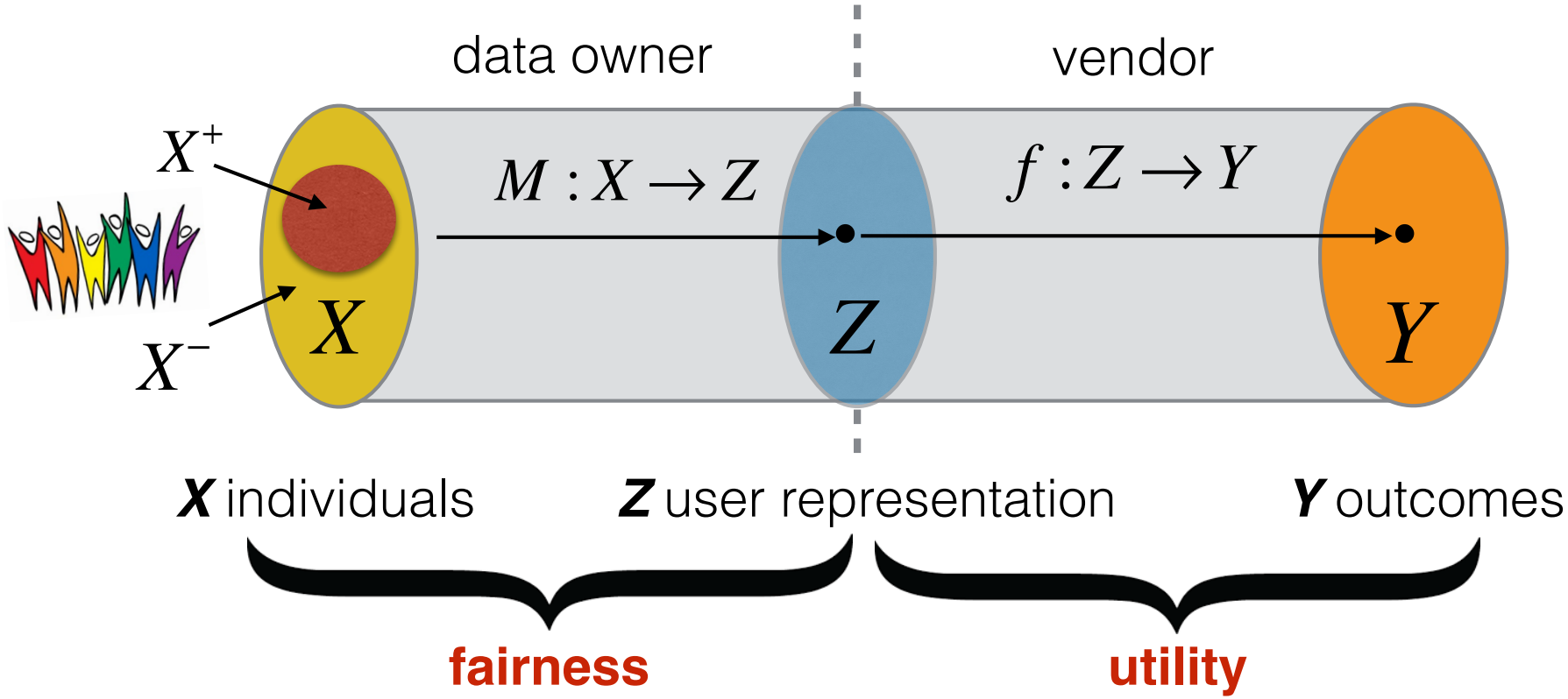
fairness enforced at this step

vendor cannot introduce bias

Vendors can maximize expected utility,
subject to the Lipschitz condition

Learning fair representations

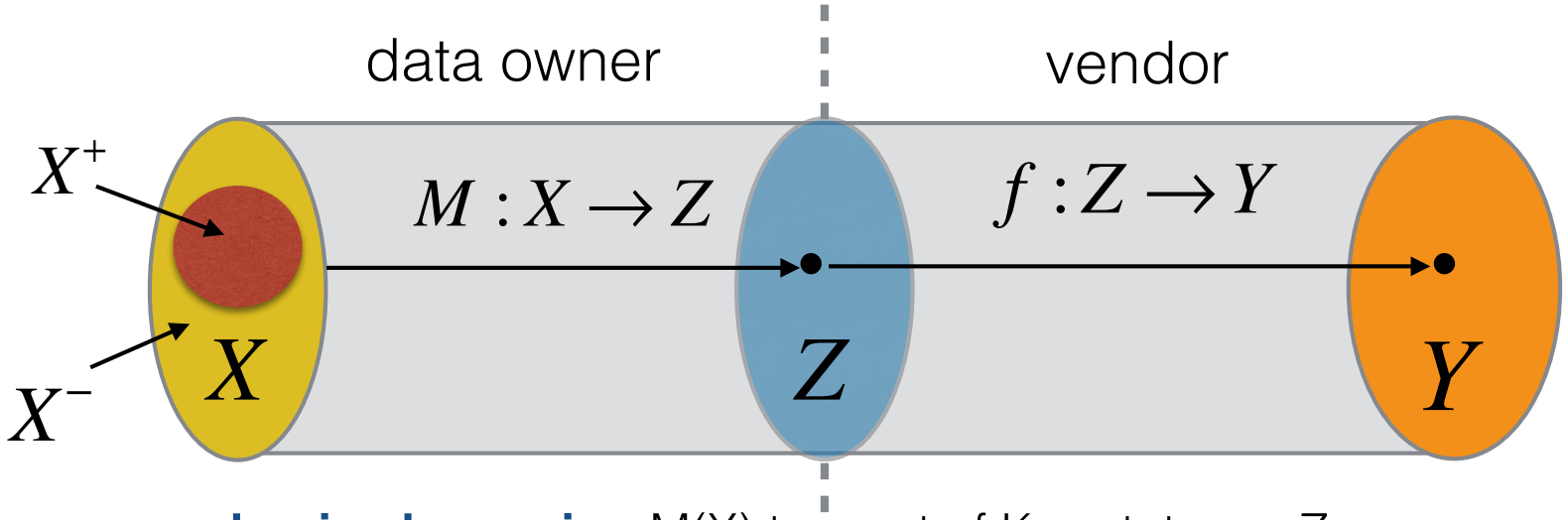
[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]



- **Idea**: remove reliance on a "fair" similarity measure, instead **learn** representations of individuals, distances

Fairness and utility

[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]



Learn a **randomized mapping** M(X) to a set of K prototypes Z

M(X) should lose information about membership in S $P(Z \mid S = 0) = P(Z \mid S = 1)$

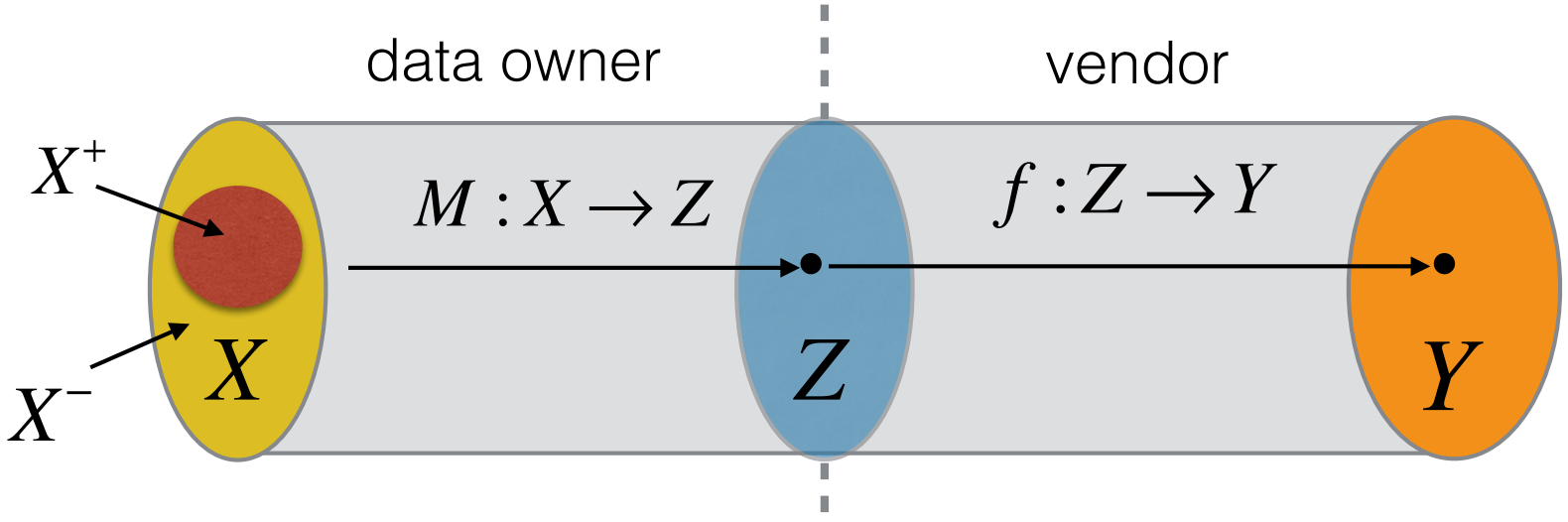
M(X) should preserve other information so that vendor can maximize utility

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

group fairness **individual fairness** **utility**

The objective function

[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]



$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

group fairness — **individual fairness** — **utility**

$$P_k^+ = P(Z = k \mid x \in X^+)$$

$$P_k^- = P(Z = k \mid x \in X^-)$$

$$L_z = \sum \left| P_k^+ - P_k^- \right| \qquad L_x = \sum (x_n - \hat{x}_n)^2$$

$$L_y = \sum_n^k -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n)$$

does this make sense?

Effect on sub-populations

Simpson's paradox
disparate impact at the full population level disappears or reverses when looking at sub-populations!

		grad school admissions	
		admitted	**denied**
gender	**F**	1512	2809
	M	3715	4727

positive outcomes

35% of women

44% of men

UC Berkeley 1973: it appears men were admitted at higher rate.

Effect on sub-populations

Simpson's paradox
disparate impact at the full population level disappears or reverses when looking at sub-populations!

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

favored group	whole population
women	35% of women
women	
men	
women	44% of men
men	
women	

UC Berkeley 1973: women applied to more competitive departments, with low rates of admission among qualified applicants.

Correlation is not causation!

Cannot claim a causal relationship based on observational data alone. Need a story.

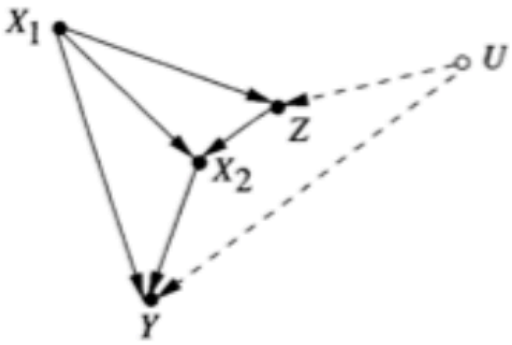


Figure 4.9 Causal relationships relevant to Berkeley's sex discrimination study. Adjusting for department choice (X_2) or career objective (Z) (or both) would be inappropriate in estimating the direct effect of gender on admission. The appropriate adjustment is given in (4.10).

X2 (choice) - "resolving variable",
then the effect of X1 on Y through X2 is "fair"

the direct effect of X1 on Y is unfair

X_1 = applicant's gender;
X_2 = applicant's choice of department;
Z = applicant's (pre-enrollment) career objectives;
Y = admission outcome (accept/reject);
U = applicant's aptitude (unrecorded).

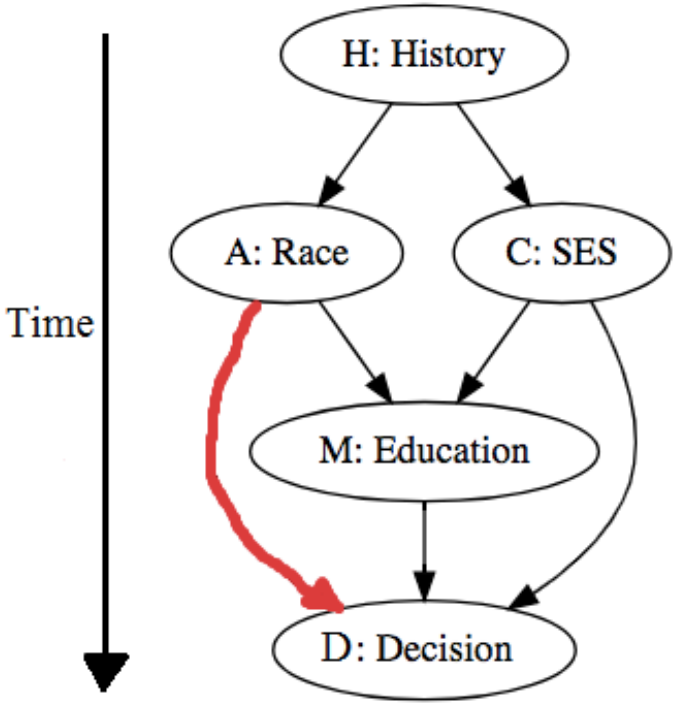
from Pearl's "Causality", page 129

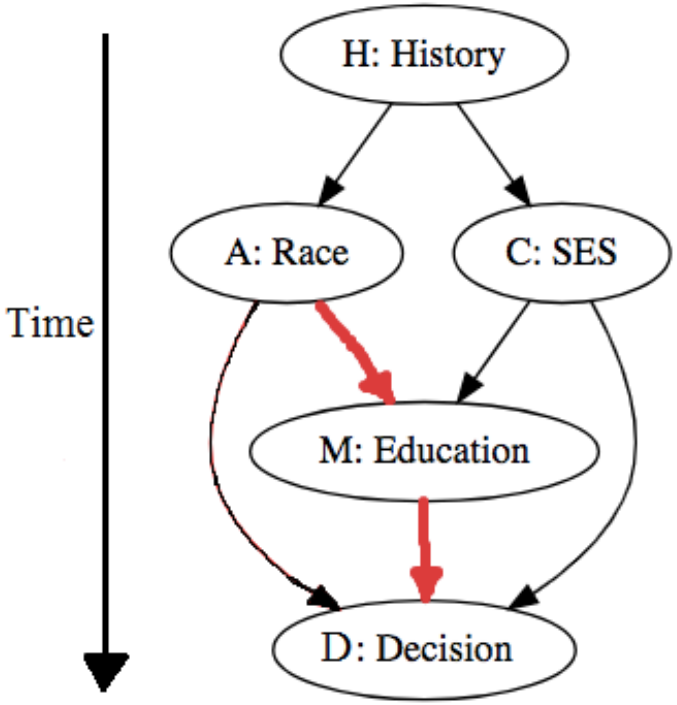
Note that U affects applicant's career objective and also the admission outcome Y (say, through verbal skills (unrecorded)).

Causal interpretations of fairness

[T.J. VanderWeele and W.R. Robinson; Epidemiology *(2014)*]

arrows represent possible causal relationships





we (society) decide which of these are "OK"

Capuchin

[B. Salimi, L. Rodriguez, B. Howe, D. Suciu, ACM SIGMOD (2019)]

CAPUCHIN: CAUSAL DATABASE REPAIR FOR ALGORITHMIC FAIRNESS

A PREPRINT

Babak Salimi
Computer Science and Engineering
University of Washington
Seattle WA
bsalimi@cs.washington.edu

Luke Rodriguez
Information School
University of Washington,
Seattle WA
rodriglr@uw.edu

Bill Howe
Information School
University of Washington,
Seattle WA
rodbillhowe@uw.edu

Dan Suciu
Computer Science and Engineering
University of Washington
Seattle WA
suciu@cs.washington.edu

ACM SIGMOD 2019 best paper award

February 27, 2019

ABSTRACT

Fairness is increasingly recognized as a critical component of machine learning systems. However, it is the underlying data on which these systems are trained that often reflect discrimination, suggesting a database repair problem. Existing treatments of fairness rely on statistical correlations that can be fooled by statistical anomalies, such as Simpson's paradox. Proposals for causality-based definitions of fairness can correctly model some of these situations, but they require specification of the underlying causal models. In this paper, we formalize the situation as a database repair problem, proving sufficient conditions for fair classifiers in terms of admissible variables as opposed to a complete causal model. We show that these conditions correctly capture subtle fairness violations. We then use these conditions as the basis for database repair algorithms that provide provable fairness guarantees about classifiers trained on their training labels. We evaluate our algorithms on real data, demonstrating improvement over the state of the art on multiple fairness metrics proposed in the literature while retaining high utility.

1 Introduction

fairness in risk assessment

Racial bias in criminal sentencing

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

A commercial tool **COMPAS** automatically predicts some categories of future crime to assist in bail and sentencing decisions. It is used in courts in the US.

Prediction Fails Differently for Black Defendants

	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.)*

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Fairness in risk assessment

- A risk assessment tool **gives a probability estimate of a future outcome**
- Used in many domains:
  - insurance, criminal sentencing, medical testing, hiring, banking
  - also in less-obvious set-ups, like online advertising
- **Fairness** is concerned with **how different kinds of errors are distributed among sub-populations**
  - Recall our discussion on fairness in classification - similar?

Desirable properties of risk tools

[J. Kleinberg, S. Mullainathan, M. Raghavan; ITCS *(2017)*]

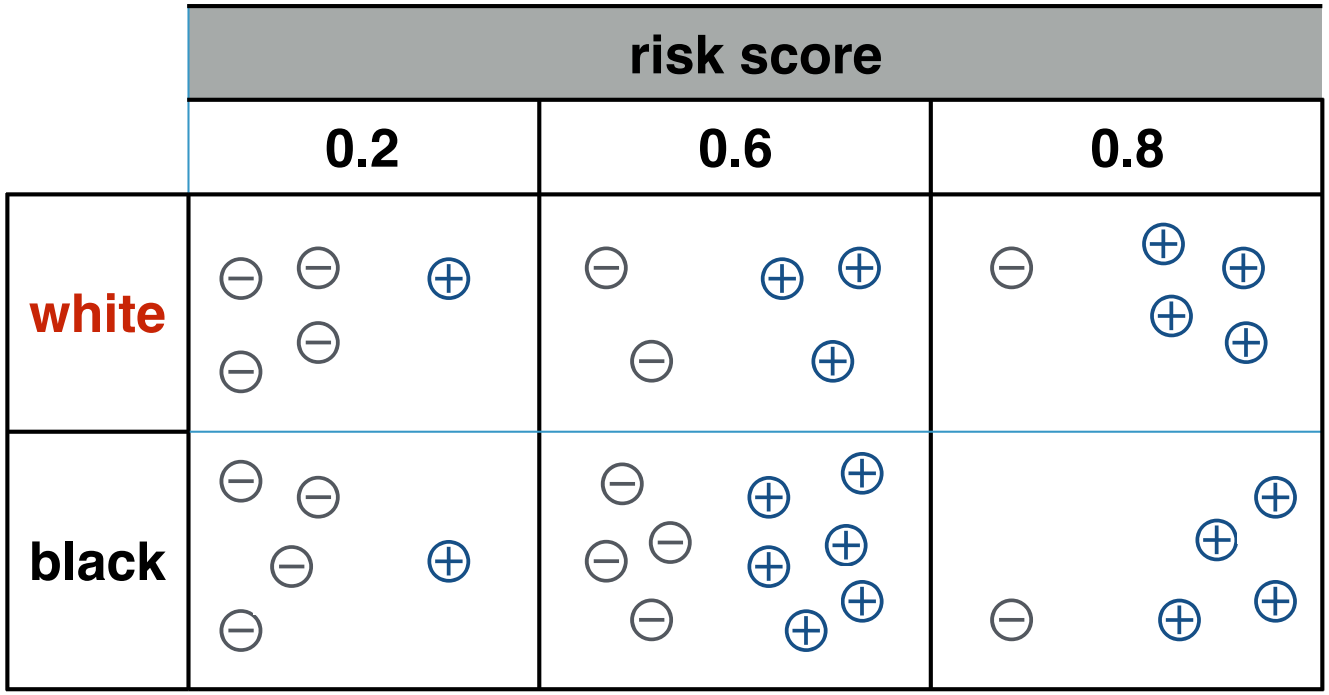
"risk assessment tool / instrument" = "**risk tool / instrument**"
for brevity in the rest of today's slides

- Calibration
- Balance for the positive class
- Balance for the negative class

can we have all these properties?

Calibration

positive outcomes: do recidivate



given the output of a risk tool, likelihood of belonging to the positive class is independent of group membership

0.6 means 0.6 for any defendant - likelihood of recidivism

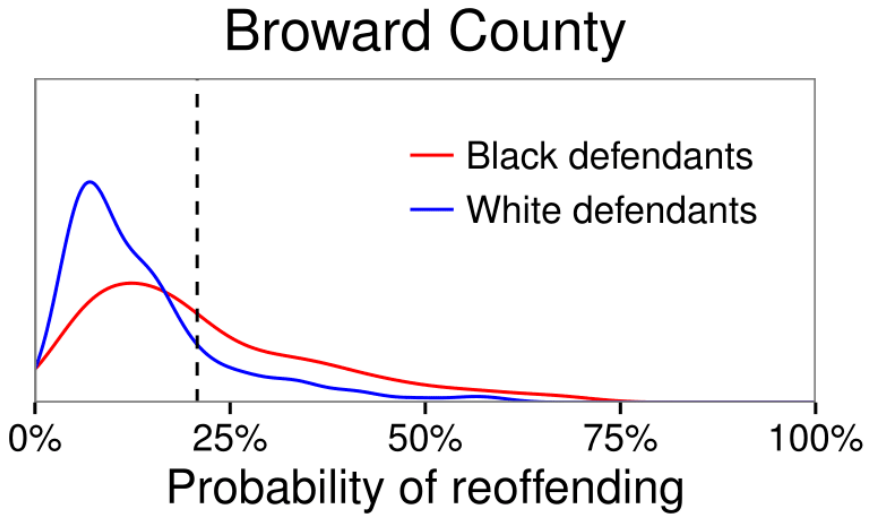
why do we want calibration?

COMPAS as a predictive instrument

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

Predictive parity (also called **calibration**)
an instrument identifies a set of instances as having probability *x* of constituting positive instances, then approximately an *x* fraction of this set are indeed positive instances, over-all and in sub-populations

COMPAS is **well-calibrated**: in the window around 40%, the fraction of defendants who were re-arrested is ~40%, both over-all and per group.



[plot from Corbett-Davies et al.; *KDD 2017*]

Group fairness impossibility result

[A. Chouldechova; arXiv:1610.07524v1 *(2017)*]

If a predictive instrument **satisfies predictive parity**, but the **prevalence** of the phenomenon **differs between groups**, then the instrument **cannot achieve** equal false positive rates and equal false negative rates across these groups

Recidivism rates in the ProPublica dataset are higher for the black group than for the white group

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

What is recidivism?: Northpointe [*the maker of COMPAS*] defined recidivism as "**a finger-printable arrest** involving a charge and a filing for any uniform crime reporting (UCR) code."

A more general desideratum: Balance

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

- **Balance for the positive class**: Positive instances are those who go on to re-offend. The average score of positive instances should be the same across groups.
- **Balance for the negative class**: Negative instances are those who do not go on to re-offend. The average score of negative instances should be the same across groups.
- Generalization of: **Both groups should have equal false positive rates and equal false negative rates.**
- Different from statistical parity!

the chance of making a mistake does not depend on race

Achievable only in trivial cases

[J. Kleinberg, S. Mullainathan, M. Raghavan; ITCS *(2017)*]

- Perfect information: the tool knows who recidivates (score 1) and who does not (score 0)
- Equal base rates: the fraction of positive-class people is the same for both groups

cannot even find a good approximate solution

a negative result, need tradeoffs

proof sketch (starts about 12 minutes in)

https://www.youtube.com/watch?v=UUC8tMNxwV8

Fairness for whom?

Decision-maker: of those I've labeled high-risk, how many will recidivate?

Defendant: how likely am I to be incorrectly classified high-risk?

Society: (think positive interventions) is the selected set demographically balanced?

based on a slide by Arvind Narayanan

	labeled low-risk	labeled high-risk
did not recidivate	TN	**FP**
recidivated	**FN**	TP

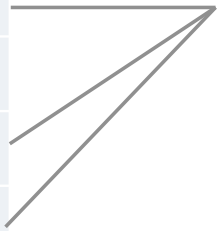
different metrics matter to different stakeholders

https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story

Impossibility theorem

based on a slide by Arvind Narayanan

Metric	Equalized under
Selection probability	Demographic parity
Pos. predictive value	Predictive parity
Neg. predictive value	
False positive rate	Error rate balance
False negative rate	Error rate balance
Accuracy	Accuracy equity



Chouldechova paper

All these metrics can be expressed in terms of FP, FN, TP, TN

If these metrics are equal for 2 groups, some trivial algebra shows that the prevalence (in the COMPAS example, of recidivism, as measured by re-arrest) is also the same for 2 groups

Nothing special about these metrics, can pick any 3!

Ways to evaluate binary classifiers

based on a slide by Arvind Narayanan

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

364 impossibility theorems :)

What's the right answer?

there is no single answer!

need transparency and public debate

- Consider harms and benefits to different stakeholders
- Be transparent about which fairness criteria we use, how we trade them off
- Recall "Learning Fair Representations": a typical ML approach

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

group fairness **individual fairness** **utility**

apples + oranges + fairness = ?

diversity

Selection in presence of bias

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)

We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. 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. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U.S. labor market. (JEL J71, J64).

Selection in presence of bias

17 MAY 2017 RESEARCH & IDEAS

Minorities Who 'Whiten' Job Resumes Get More Interviews

by Dina Gerdeman

African American and Asian job applicants who mask their race on resumes seem to have better success getting job interviews, according to research by **Katherine DeCelles** and colleagues.

Selection in presence of bias

Blacks get more job interview callbacks when they "whiten" their resumes. Graphic by Blair Storie-Johnson (*Source: "Whitened Resumes: Race and Self-Presentation in the Labor Market"*)

AI's White Guy Problem

The New York Times

Artificial Intelligence's White Guy Problem

By KATE CRAWFORD JUNE 25, 2016

Like all technologies before it, artificial intelligence will reflect the values of its creators. So **inclusivity matters** — from who designs it to who sits on the company boards and which ethical perspectives are included.

Otherwise, **we risk constructing machine intelligence that mirrors a narrow and privileged vision of society**, with its old, familiar biases and stereotypes.

problems are beyond AI, whatever your definition of AI

A technical review paper

REVIEW

Diversity in Big Data: A Review

Marina Drosou[1], H.V. Jagadish[2], Evaggelia Pitoura[1], and Julia Stoyanovich[3,*]

Big Data
Volume 5 Number 2, 2017

DOI: 10.1089/big.2016.0054

Abstract

Big data technology offers unprecedented opportunities to society as a whole and also to its individual members. At the same time, this technology poses significant risks to those it overlooks. In this article, we give an overview of recent technical work on diversity, particularly in selection tasks, discuss connections between diversity and fairness, and identify promising directions for future work that will position diversity as an important component of a data-responsible society. We argue that diversity should come to the forefront of our discourse, for reasons that are both ethical—to mitigate the risks of exclusion—and utilitarian, to enable more powerful, accurate, and engaging data analysis and use.

Keywords: data; diversity; empirical studies; models and algorithms; responsibly

Step 1: The Rooney Rule

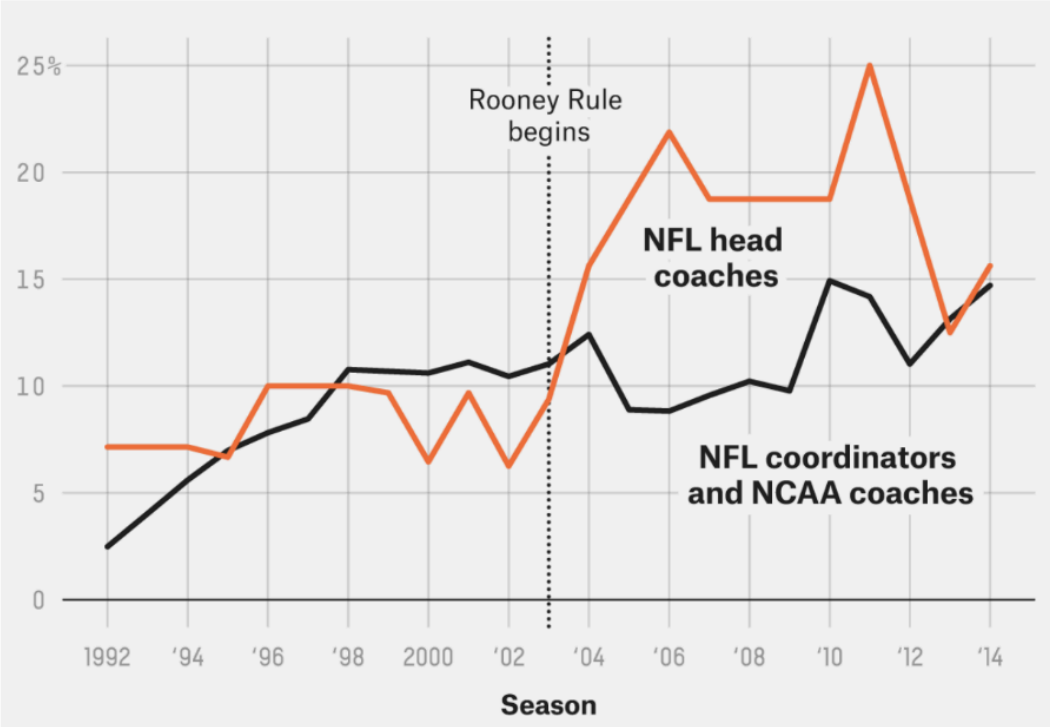
Named for a protocol adopted by the National Football League (NFL) in 2002, to increase the number of African-American head coaches

Requires that **at least one minority candidate** be interviewed for a position

Currently also used by the tech giants, to increase hiring of women and members of under-represented minorities (URM)

Push-back based on a **utility argument**: does the quality of the hired candidate / candidates decrease if the Rooney Rule is implemented?

Step 1: The Rooney Rule



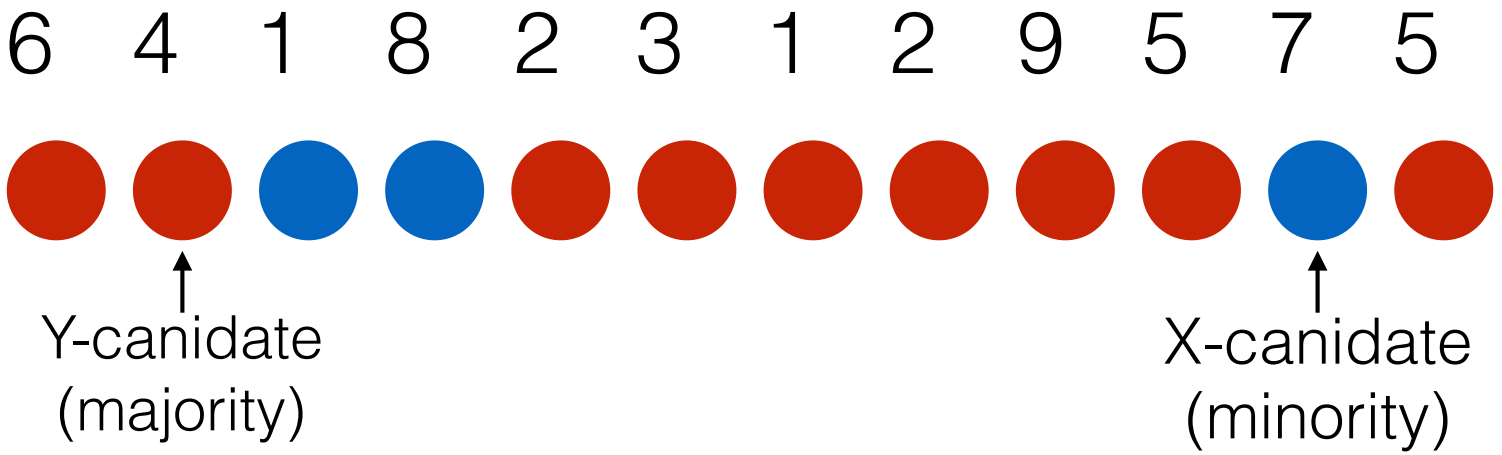
DuBois 2016

Selection with implicit bias

[J. Kleinberg, M. Raghavan, ITCS (*2018*)]

Goal: Given some estimates of the **extent of bias**, and the **prevalence** of available minority candidates, develop a mathematical model to quantify the **expected quality** of the candidates interviewed by a hiring committee.

Potential: of each candidate drawn from Z, the **same power law distribution for X and Y!**



α proportion of X-candidates

δ exponent of the power law

Bias in scoring and ranking

MENU

nature
International journal of science

Commentary | Published: 22 May 1997

Nepotism and sexism in peer-review

Christine Wennerås & Agnes Wold

Nature **387**, 341–343 (1997) | Download Citation

In the first-ever analysis of peer-review scores for postdoctoral fellowship applications, the system is revealed as being riddled with prejudice. The policy of secrecy in evaluation must be abandoned.

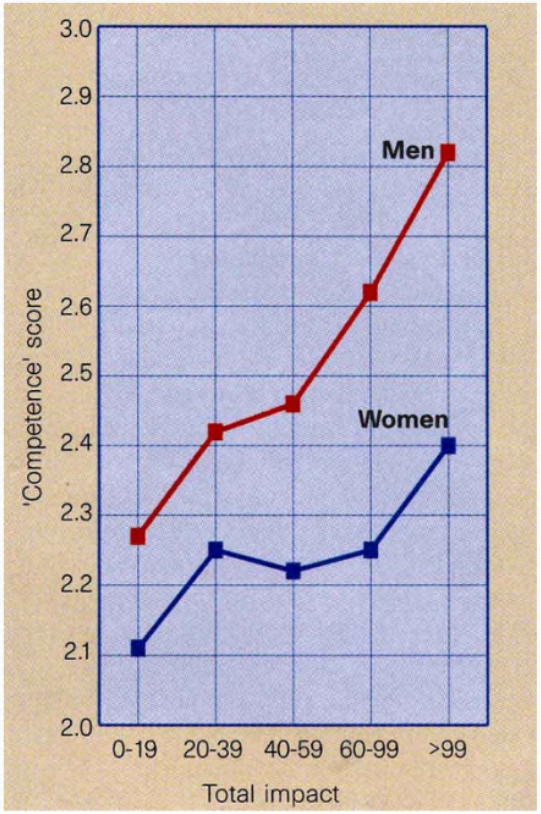


Figure 1 **The mean competence score given to male (red squares) and female (blue squares) applicants by the MRC reviewers as a function of their scientific productivity, measured as total impact. One impact point equals one paper published in a journal with an impact factor of 1. (See text for further explanation.)**

Multiplicative bias

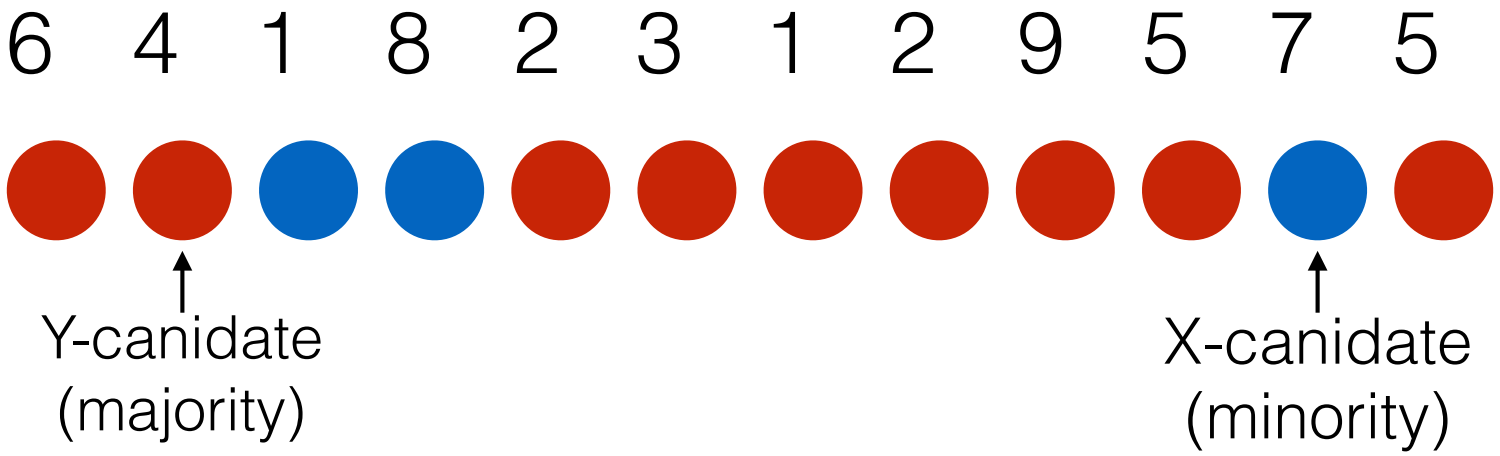
[J. Kleinberg, M. Raghavan, ITCS (*2018*)]

Goal: Pick **k** finalists to interview, maximizing expected utility: sum of potentials of the chosen candidates.

Bias: Committee correctly estimates the potential of Y-candidates and they under-estimate the potential of X-candidates.

$$\tilde{X}_i = X_i / \beta$$

$$\beta > 1$$



Process: Estimate potentials of all candidates, rank, pick the best **k**.

Main result

[J. Kleinberg, M. Raghavan, ITCS (*2018*)]

▶ **Theorem 1.** *For $k = 2$ and sufficiently large n, the Rooney Rule produces a positive expected change if and only if $\phi_2(\alpha, \beta, \delta) > 1$ where*

$$\phi_2(\alpha, \beta, \delta) = \frac{\alpha^{1/(1+\delta)}\left[1 - (1 + c^{-1})^{-\delta/(1+\delta)}\left[1 + \frac{\delta}{1+\delta}(1 + c)^{-1}\right]\right]}{\frac{\delta}{1+\delta}(1 + c)^{-1-\delta/(1+\delta)}} \quad (1)$$

and $c = \alpha\beta^{-(1+\delta)}$. Moreover, $\phi_2(\alpha, \beta, \delta)$ is increasing in β, so for fixed α and δ there exists β^* such that $\phi_2(\alpha, \beta, \delta) > 1$ if and only if $\beta > \beta^*$.

Selecting a seemingly sub-optimal candidate can improve utility!

Illustration: Infinite bias (all Y ranked higher than all X), pick k=2 candidates. Rooney rule improves utility if and only if

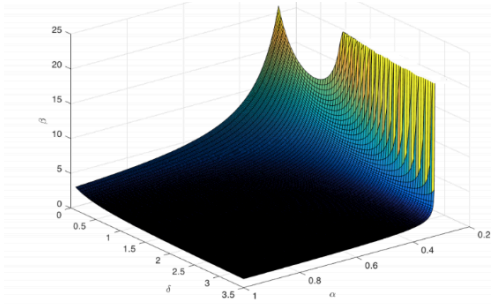
$$\alpha > \left(\frac{\delta}{1+\delta}\right)^{1+\delta}$$

Intuition

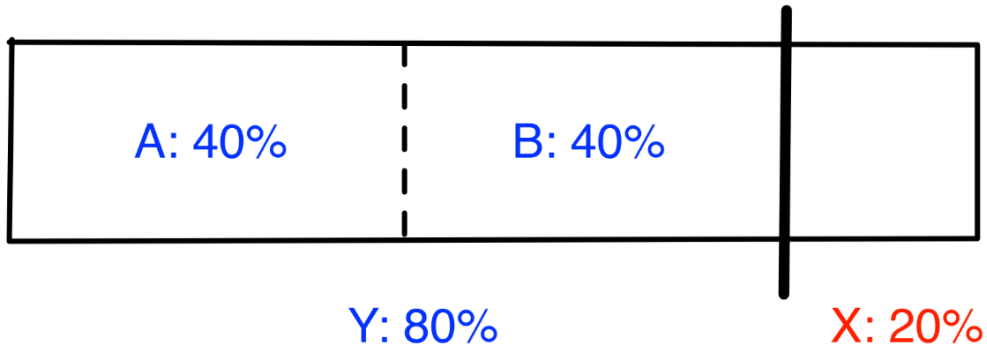
[J. Kleinberg, M. Raghavan, ITCS (*2018*)]

For which (α, δ) pairs does the Rooney Rule improve utility as $\beta \to \infty$?

- When should we reserve a slot for an X-candidate in the case of infinite bias? (Let's focus on $k = 2$.)

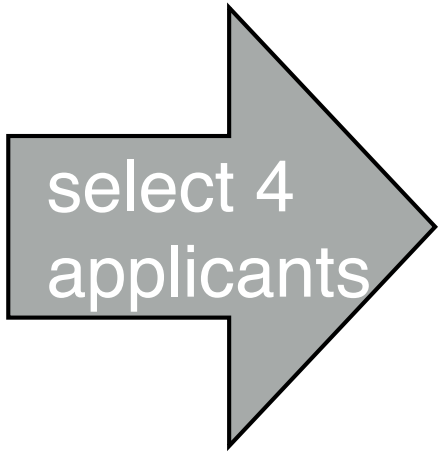


Surprising fact: No matter how small the fraction of X-candidates ($\alpha > 0$), there is a small enough power-law exponent ($\delta > 0$) so that the Rooney Rule improves utility.



slide from Jon Kleinberg's FAT* 2019 keynote

Another take: Online job applicant selection



ranked

proportional

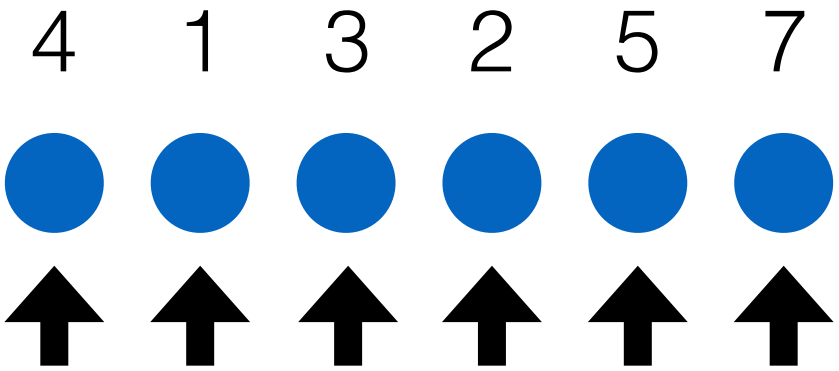
equal

Can state all these as constraints:

for each category ***i***, pick K_i elements, with $floor_i \leq K_i \leq ceil_i$

Hiring a job candidate

Goal: Hire a candidate with a high score



Candidates arrive one-by-one

A candidate's score is revealed when the candidate arrives

Decision to accept or reject a candidate made on the spot

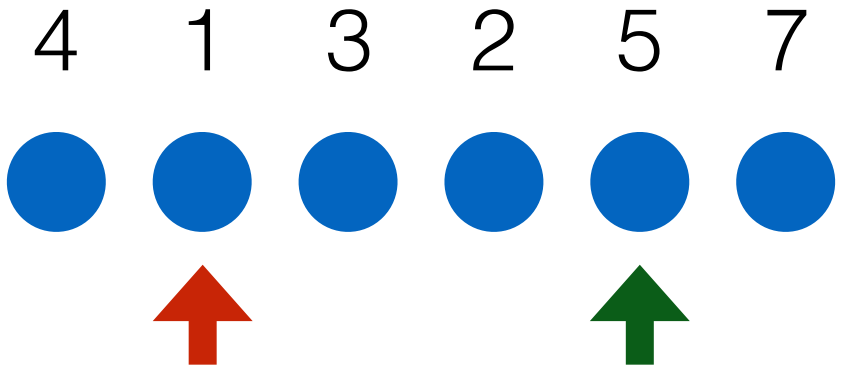
The Secretary Problem

Goal: Design an algorithm for picking **one** element of a **randomly ordered** sequence, to maximize the probability of picking the **maximum element** of the entire sequence.

$N = 6$

$S = \left\lfloor N/e \right\rfloor = 2$

$T = 4$



Competitive ratio

$1/e$

the best possible!

Consider, and reject, the first ***S*** candidates

Record ***T***, the best seen score among the first ***S*** candidates

Accept the next candidate with score better than ***T***

K-choice Secretary

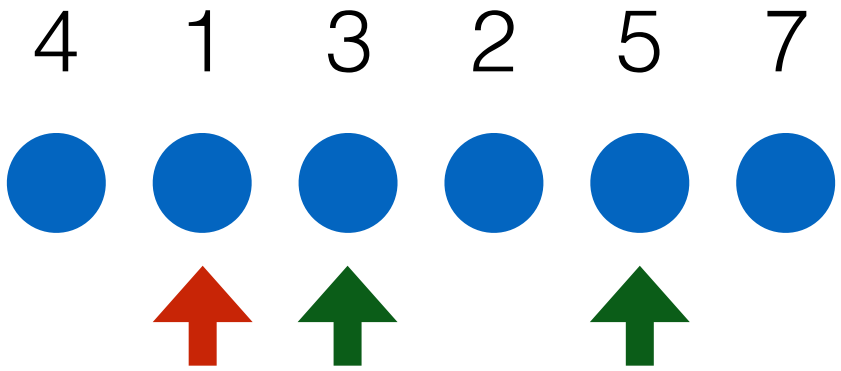
[Babaioff et al., 2007]

Goal: Design an algorithm for picking **K** elements of a **randomly ordered** sequence, to maximize their **expected sum**.

$N = 6 \quad K = 2$

$S = \left\lfloor N/e \right\rfloor = 2$

$T = \{1, 4\}$



Competitive ratio

$1/e$

far from optimal

Consider, and reject, the first *S* candidates

Record *K* best scores among the first *S* candidates, call this *T*

Whenever a candidate arrives whose score is higher than the minimum in *T*, accept the candidate and delete the minimum from *T*

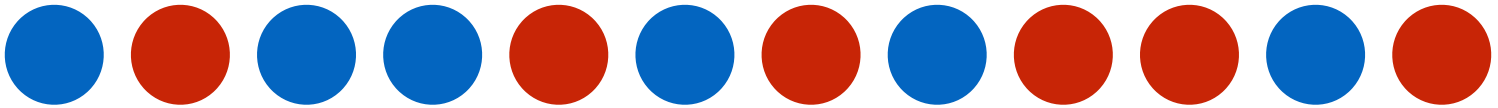
K-choice Secretary

[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

Goal: Design an algorithm for picking **K** elements of a **randomly ordered** sequence, to maximize their **expected sum**.

For each **category** *i*, pick K_i elements, with $floor_i \leq K_i \leq ceil_i$

6 4 1 8 2 3 1 2 9 5 7 5



$N_{red} = N_{blue} = 6$

$K = 3$

$1 \leq K_{red}, K_{blue} \leq 2$

Accept *floor* items for each category from per-category streams

$$slack = K - (floor_{red} + floor_{blue})$$

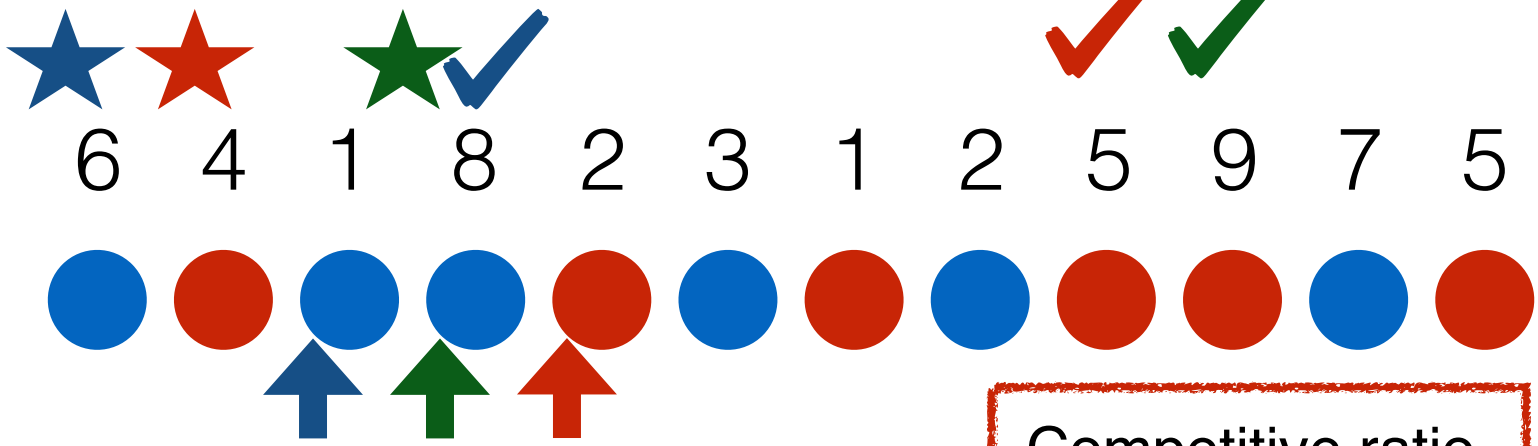
Accept the remaining *slack* items irrespective of category membership, but subject to *ceil*

Diverse K-choice Secretary

[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

$N_{red} = N_{blue} = 6$ $\quad slack = 1$

$K = 3 \quad 1 \le K_{red}, K_{blue} \le 2$ $\quad S_{red} = S_{blue} = 2 \quad S = 4$



Competitive ratio

$1/e$

far from optimal

Adding a deferred list

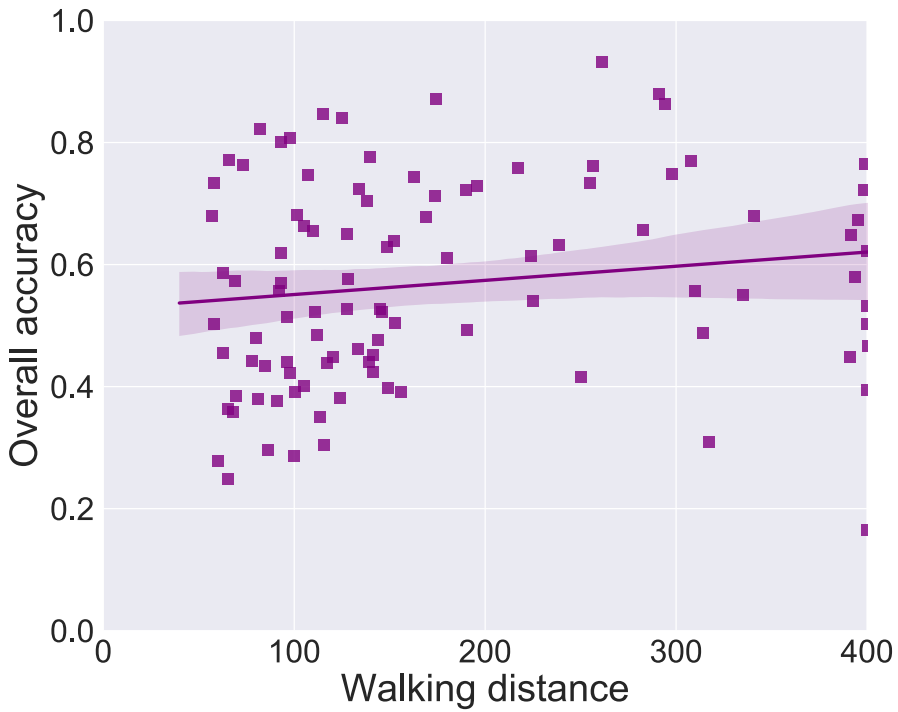
[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

- An improvement on Diverse K-choice Secretary
- Do not immediately reject or accept items: keep a deferred list D_i per category i of size up to $ceil_i$
- Stop reading the input, post-warm up, once all $floor_i$ constraints are met, and once there are K items in the union of the deferred lists
- Main advantage: often avoids reading items from the end of the stream

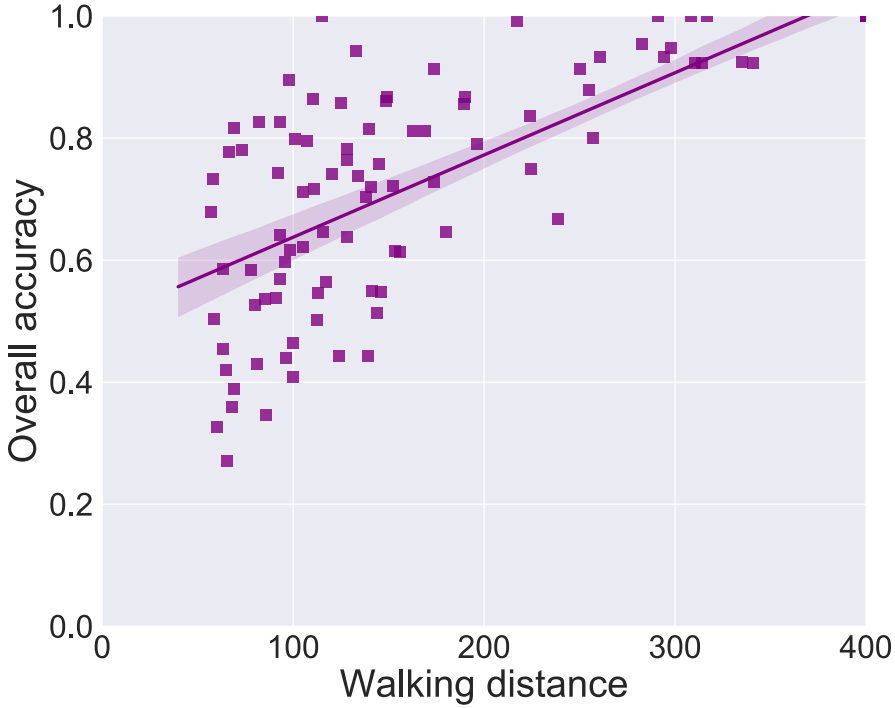
Diversity is achievable

[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

deferred list



with deferred list

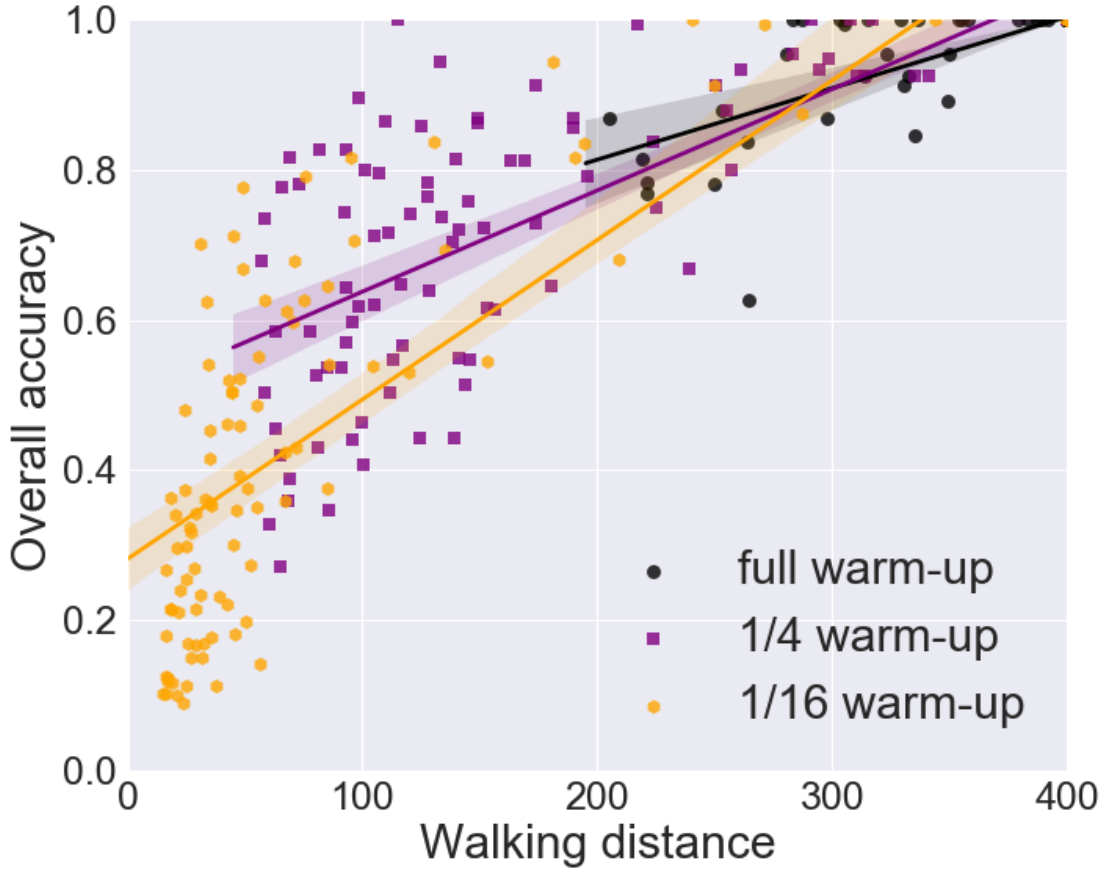


Forbes US Richest: N=400, K=4 (27 female, 373 male)

diversity on gender: select 2 per gender

Warm-up can be shorter

[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

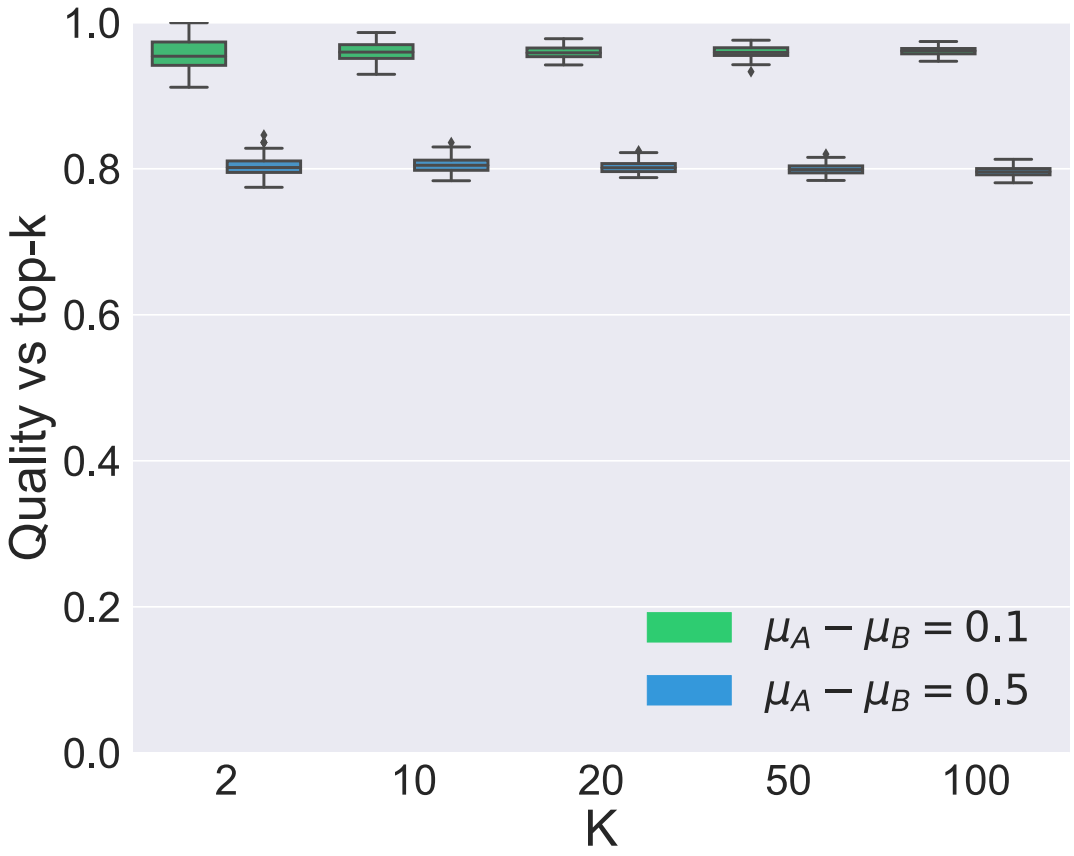


Forbes US Richest: N=400, K=4 (27 female, 373 male)

deferred list variant, diversity on gender: select 2 per gender

The cost of diversity

[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

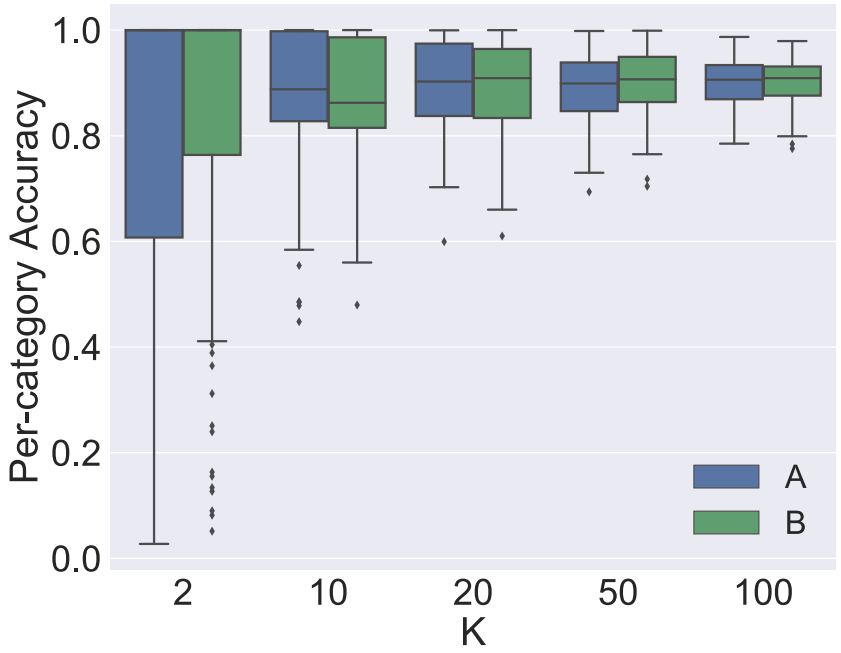


static variant (see paper), synthetic data in categories A and B, score lower for B

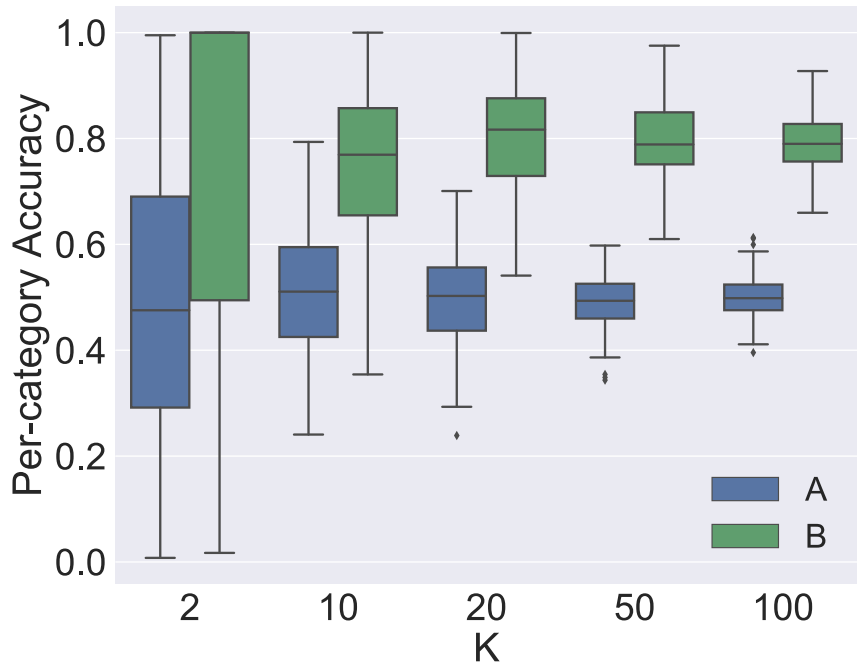
Per-category warm-up is crucial

[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

Per-category warm-up period



Common warm-up period



synthetic data with categories A and B, score depends on category, lower for A

diversity by design

Balanced ranking with diversity

[K. Yang, V. Gkatzelis, J. Stoyanovich, IJCAI (2019)]

Balanced Ranking with Diversity Constraints

Ke Yang[1*], **Vasilis Gkatzelis**[2], **Julia Stoyanovich**[1]
[1]New York University, Department of Computer Science and Engineering
[2]Drexel University, Department of Computer Science
ky630@nyu.edu, gkatz@drexel.edu, stoyanovich@nyu.edu

Abstract

Many set selection and ranking algorithms have recently been enhanced with *diversity constraints* that aim to explicitly increase representation of historically disadvantaged populations, or to improve the overall representativeness of the selected set. An unintended consequence of these constraints, however, is reduced *in-group fairness*: the selected candidates from a given group may not be the best ones, and this unfairness may not be well-balanced across groups. In this paper we study this phenomenon using datasets that comprise multiple sensitive attributes. We then introduce additional constraints, aimed at balancing the in-group fairness across groups, and formalize the induced optimization problems as integer linear programs. Using these programs, we conduct an experimental evaluation with real datasets, and quantify the feasible trade-offs between balance and overall performance in the presence of diversity constraints.

are increasingly recognized by sociologists and political scientists [Page, 2008; Surowiecki, 2005]. Last but not least, diversity constraints can be used to ensure dataset representativeness, for example when selecting a group of patients to study the effectiveness of a medical treatment, or to understand the patterns of use of medical services [Cohen *et al.*, 2009], an example we will revisit in this paper.

Our goal in this paper is to evaluate and mitigate an unintended consequence that such diversity constraints may have on the outcomes of set selection and ranking algorithms. Namely, we want to ensure that these algorithms do not systematically select lower-quality items in particular groups. In what follows, we make our set-up more precise.

Given a set of items, each associated with multiple sensitive attribute labels and with a quality score (or utility), a set selection algorithm needs to select k of these items aiming to maximize the overall utility, computed as the sum of *utility scores* of selected items. The score of an item is a single scalar that may be pre-computed and stored as a physical attribute, or it may be computed on the fly. The output of traditional set selection algorithms, however, may lead to

Balanced ranking with diversity

[K. Yang, V. Gkatzelis, J. Stoyanovich, IJCAI (2019)]

Goal: pick k=4 candidates, including 2 of each gender, and at least one candidate per ethnicity, maximizing the total score of the selected candidates.

	Male		Female	
White	A (99)	B (98)	C (96)	D (95)
Black	E (91)	F (91)	G (90)	H (89)
Asian	I (87)	J (87)	K (86)	L (83)

score=373

Table 1: A set of 12 individuals with sensitive attributes `race` and `gender`. Each cell lists an individual's ID, and score in parentheses.

Problem: **In-group fairness fails** for Female (C and D not picked, which G and K are), Black (E and F are not picked, while G is), and Asian (I and J are not picked, while K is). **In-group fairness holds** for White and Male groups though (those with higher scores)!

Insight: while in-group fairness will inevitably fail to some extent because of diversity constraints, this loss should be **balanced** across groups.

Balanced ranking with diversity

[K. Yang, V. Gkatzelis, J. Stoyanovich, IJCAI (2019)]

Goal: pick k=4 candidates, including 2 of each gender, and at least one candidate per ethnicity, maximizing the total score of the selected candidates.

	Male		Female	
White	A (99)	B (98)	C (96)	D (95)
Black	E (91)	F (91)	G (90)	H (89)
Asian	I (87)	J (87)	K (86)	L (83)

score=372

Table 1: A set of 12 individuals with sensitive attributes `race` and `gender`. Each cell lists an individual's ID, and score in parentheses.

Problem: **In-group fairness fails** for Female (C and D not picked, which G and K are), Black (E and F are not picked, while G is), and Asian (I and J are not picked, while K is). **In-group fairness holds** for White and Male groups though (those with higher scores)!

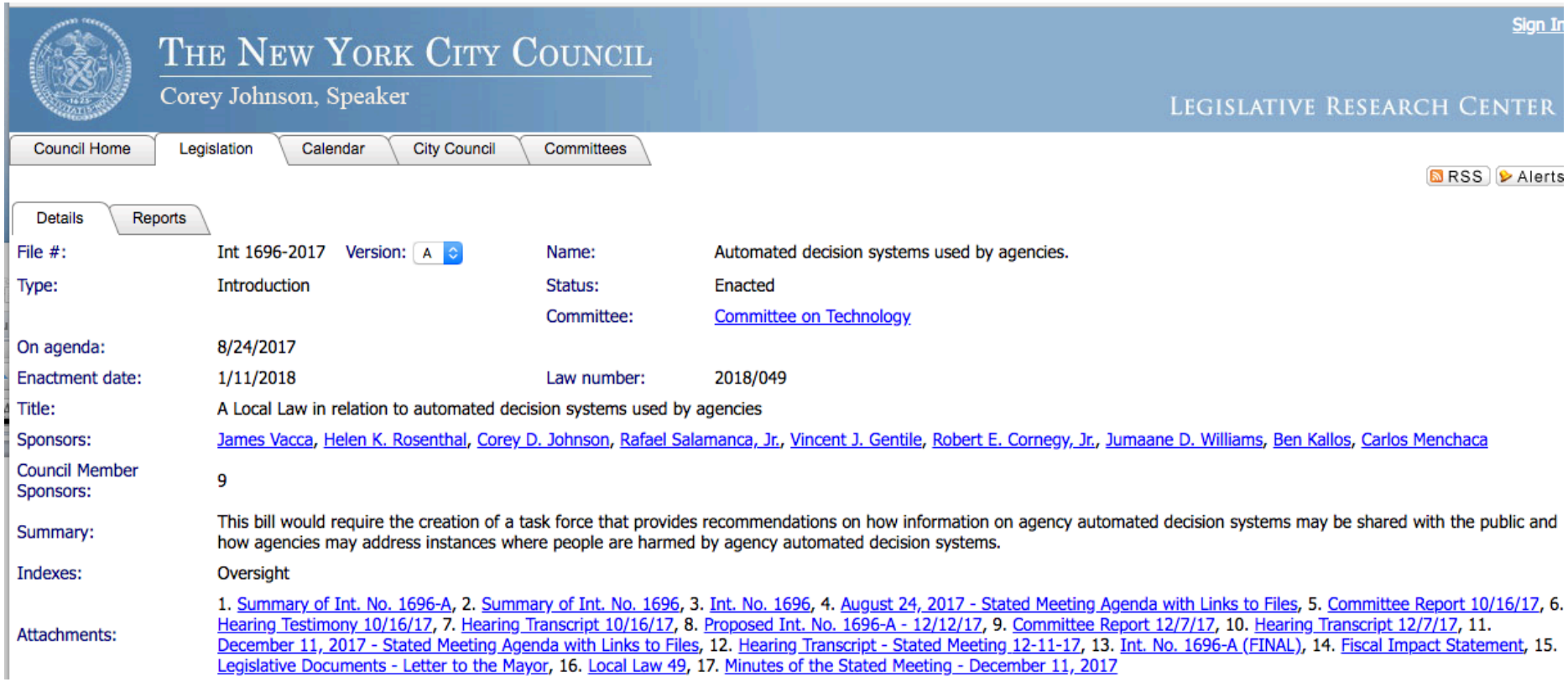
Insight: while in-group fairness will inevitably fail to some extent because of diversity constraints, this loss should be **balanced** across groups.

wrapping up

NYC algorithmic transparency law

1/11/2018

A Local Law 49 of 2018, in relation to automated decision systems used by agencies



THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker

Sign In

LEGISLATIVE RESEARCH CENTER

Council Home | Legislation | Calendar | City Council | Committees

RSS Alerts

Details | Reports

File #:	Int 1696-2017 Version: A	Name:	Automated decision systems used by agencies.
Type:	Introduction	Status:	Enacted
		Committee:	Committee on Technology
On agenda:	8/24/2017		
Enactment date:	1/11/2018	Law number:	2018/049
Title:	A Local Law in relation to automated decision systems used by agencies		
Sponsors:	James Vacca, Helen K. Rosenthal, Corey D. Johnson, Rafael Salamanca, Jr., Vincent J. Gentile, Robert E. Cornegy, Jr., Jumaane D. Williams, Ben Kallos, Carlos Menchaca		
Council Member Sponsors:	9		
Summary:	This bill would require the creation of a task force that provides recommendations on how information on agency automated decision systems may be shared with the public and how agencies may address instances where people are harmed by agency automated decision systems.		
Indexes:	Oversight		
Attachments:	1. Summary of Int. No. 1696-A, 2. Summary of Int. No. 1696, 3. Int. No. 1696, 4. August 24, 2017 - Stated Meeting Agenda with Links to Files, 5. Committee Report 10/16/17, 6. Hearing Testimony 10/16/17, 7. Hearing Transcript 10/16/17, 8. Proposed Int. No. 1696-A - 12/12/17, 9. Committee Report 12/7/17, 10. Hearing Transcript 12/7/17, 11. December 11, 2017 - Stated Meeting Agenda with Links to Files, 12. Hearing Transcript - Stated Meeting 12-11-17, 13. Int. No. 1696-A (FINAL), 14. Fiscal Impact Statement, 15. Legislative Documents - Letter to the Mayor, 16. Local Law 49, 17. Minutes of the Stated Meeting - December 11, 2017		

NYC algorithmic transparency law

10/16/2017

THE NEW YORKER

By Julia Powles December 20, 2017

ELEMENTS

NEW YORK CITY'S BOLD, FLAWED ATTEMPT TO MAKE ALGORITHMS ACCOUNTABLE

Automated systems guide the allocation of everything from firehouses to food stamps. So why don't we know more about them?

Photograph by Mario Tama / Getty

The original draft

8/16/2017

Int. No. 1696

By Council Member Vacca

A Local Law to amend the administrative code of the city of New York, in relation to automated processing of data for the purposes of targeting services, penalties, or policing to persons

Be it enacted by the Council as follows:

Section 1. Section 23-502 of the administrative code of the city of New York is amended to add a new subdivision g to read as follows:

g. Each agency that uses, for the purposes of targeting services to persons, imposing penalties upon persons or policing, an algorithm or any other method of automated processing system of data shall:

1. Publish on such agency's website, the source code of such system; and

2. Permit a user to (i) submit data into such system for self-testing and (ii) receive the results of having such data processed by such system.

§ 2. This local law takes effect 120 days after it becomes law.

MAJ
LS# 10948
8/16/17 2:13 PM

this is **NOT** what was adopted

Summary of Local Law 49 of 2018

1/11/2018

Form an automated decision systems (**ADS**) task force that surveys current use of algorithms and data in City agencies and develops procedures for:

- requesting and receiving an **explanation** of an algorithmic decision affecting an individual (3(b))
- interrogating ADS for **bias** and **discrimination** against members of legally-protected groups (3(c) and 3(d))
- allowing the **public** to **assess** how ADS function and are used (3(e)), and archiving ADS together with the data they use (3(f))

we've come a long way from the original draft!

The ADS task force

5/16/2018

Visit **alpha.nyc.gov** to help us test out new ideas for NYC's website.

The Official Website of the City of New York

NYC

简体中文 ▸ Translate ▼ Text Size

NYC Resources | NYC311 | Office of the Mayor | Events | Connect | Jobs | Search

Mayor | First Lady | News | Officials

SHARE

Email

Print

Mayor de Blasio Announces First-In-Nation Task Force To Examine Automated Decision Systems Used By The City

May 16, 2018

NEW YORK— Today, Mayor de Blasio announced the creation of the Automated Decision Systems Task Force which will explore how New York City uses algorithms. The task force, the first of its kind in the U.S., will work to develop a process for reviewing "automated decision systems," commonly known as algorithms, through the lens of equity, fairness and accountability.

"As data and technology become more central to the work of city government, the algorithms we use to aid decision making must be aligned with our goals and values," said **Mayor de Blasio**. "The establishment of the Automated Decision Systems Task Force is an important first step towards greater transparency and equity in our use of technology."

The ADS task force

THE VERGE

5/15/2019

POLICY | REPORT | US & WORLD

New York City's algorithm task force is fracturing

Some members say the city isn't being transparent

By Colin Lecher | @colinlecher | Apr 15, 2019, 8:43am EDT

give us examples!

DS-GA 3001.009: Special Topics in Data Science: Responsible Data Science

New York University, Center for Data Science, Spring 2019

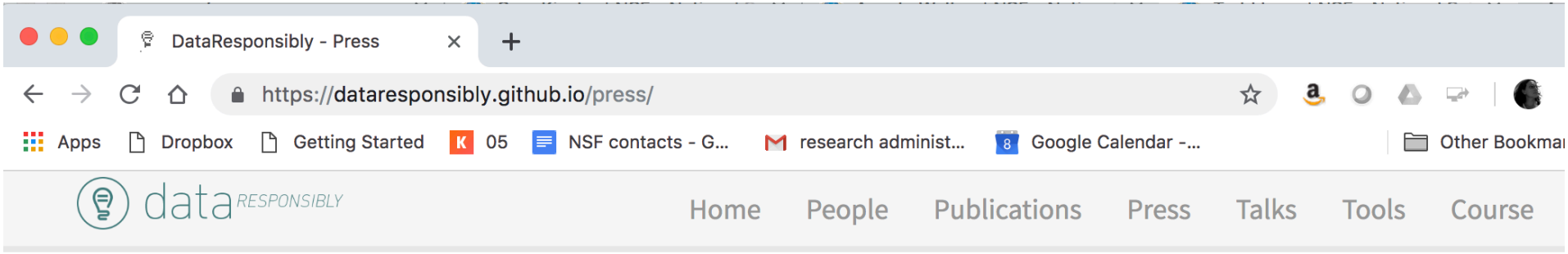
Lecture: Mondays from 11am-12:40pm; Lab: Thursdays from 5:20pm-6:10pm

Instructor: Julia Stoyanovich, Assistant Professor of Data Science, Computer Science and Engineering

The first wave of data science focused on accuracy and efficiency -- on what we *can* do with data. The second wave focuses on responsibility -- on what we *should* and *shouldn't* do. Irresponsible use of data science can cause harm on an unprecedented scale. Algorithmic changes in search engines can sway elections and incite violence; irreproducible results can influence global economic policy; models based on biased data can legitimize and amplify racist policies in the criminal justice system; algorithmic hiring practices can silently and scalably violate equal opportunity laws, exposing companies to lawsuits and reinforcing the feedback loops that lead to lack of diversity. Therefore, as we develop and deploy data science methods, we are compelled to think about the effects these methods have on individuals, population groups, and on society at large.

Responsible Data Science is a technical course that tackles the issues of ethics, legal compliance, data quality, algorithmic fairness and diversity, transparency of data and algorithms, privacy, and data protection. The course is developed and taught by Julia Stoyanovich, Assistant Professor at the Center for Data Science and at the Tandon School of Engineering, and member of the NYC Automated Decision Systems Task Force.

Prerequisites: Introduction to Data Science, Introduction to Computer Science, or relevant courses



Press

Follow the Data! Algorithmic Transparency Starts with Data Transparency

The data revolution that is transforming every sector of science and industry has been slow to reach the local and municipal governments and NGOs that deliver vital human services in health, housing, and mobility. Urbanization has made the issue acute in 2016, more than half of North Americans lived in cities with at least 500,000 inhabitants.

Julia Stoyanovich and Bill Howe

The Ethical Machine, November 27, 2018

An Algorithmic Approach to Correct Bias in Urban Transportation Datasets

While a significant amount of attention and research has addressed individual privacy concerns in private companies' datasets, data owners and publishers also want to avoid revealing certain patterns—even in anonymized datasets—that might compromise a competitive advantage or perpetuate discrimination against any group of people. Data published by urban transportation companies is highly valuable for research, policy, and public accountability.

NYU Center for Data Science, October 30, 2018

Responsible Data Science

Part 2: transparency and interpretability

Prof. Julia Stoyanovich

Computer Science and Engineering &
Center for Data Science
New York University
@stoyanoj

https://dataresponsibly.github.io/
https://dataresponsibly.github.io/courses/

NYU

Thank you!

dataresponsibly.github.io

@stoyanoj