DS-GA 3001.009: Responsible Data Science

Introduction and Overview.
Algorithmic Fairness

Prof. Julia Stoyanovich
Center for Data Science
Computer Science and Engineering at Tandon

@stoyanoj

http://stoyanovich.org/
https://dataresponsibly.github.io/
https://dataresponsibly.github.io/courses/spring20/

Instructor

Julia Stoyanovich @stoyanoj

- Assistant Professor of Data Science at CSE and CDS
- PhD in Computer Science from Columbia University
- BS in Computer Science & Math from UMass Amherst
- worked in start-ups between college and graduate school

Research: data and knowledge management (aka “databases”)

- **Responsible Data Science**: ethics, legal compliance through the DS lifecycle
- **Portal**: querying and analysis of evolving graphs
- **DB4Pref**: preference data management, computational social choice

Office hours: Mondays 1:30-3pm and by appointment
at CDS (60 5th Avenue), room 703

Course staff

Brina Seidel, section leader

Office hours: Thursdays 3:30-4:30pm and by appointment, at CDS (60 5th Avenue), room TBD

Prasanthi Gurumurthi, grader

Office hours: Wednesdays 10:30-11:30am and by appointment, at CDS (60 5th Avenue), room TBD

Course logistics

Website: https://dataresponsibly.github.io/courses/spring20/

Assigned readings are from a variety of sources. Best to read before class. Let me know how you find them!

We are working to improve the course, will be doing assessment. Please give your feedback informally as well!

Course logistics

Website: https://dataresponsibly.github.io/course

Grading: labs - 10% (attend 10 labs in person for full credit)
homeworks - 10% x 3 = 30%
project - 30%
final - 30%

No credit for late homeworks. 2 late days over the term, no questions asked. If a homework is submitted late — a day is used in full.

What's "Responsible Data Science"?

As advertised: ethics, legal compliance, personal responsibility.
But also: **data quality**!

A technical course, with content drawn from:

1. data engineering - **yep, I'll teach you some useful database stuff!**
2. security and privacy
3. fairness, accountability and transparency

We will learn **algorithmic techniques** for data analysis.
We will also learn about recent **laws** / regulatory frameworks.

Bottom line: we will learn that many of the problems are **socio-technical**, and so cannot be "solved" with technology alone.

My perspective: a pragmatic engineer, **not** a technology skeptic.

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

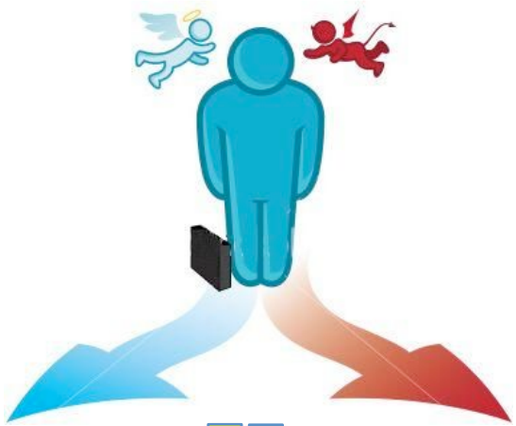
transform society, e.g., open government

optimize business, e.g., advertisement targeting

goal **- progress**

Example: personalized medicine

Analysis of a person's medical data, genome, social data



personalized medicine

personalized care and predictive measures

personalized insurance

expensive, or unaffordable, for those at risk

the same technology makes both possible!

Online price discrimination

THE WALL STREET JOURNAL.

WHAT THEY KNOW

Websites Vary Prices, Deals Based on Users' Information

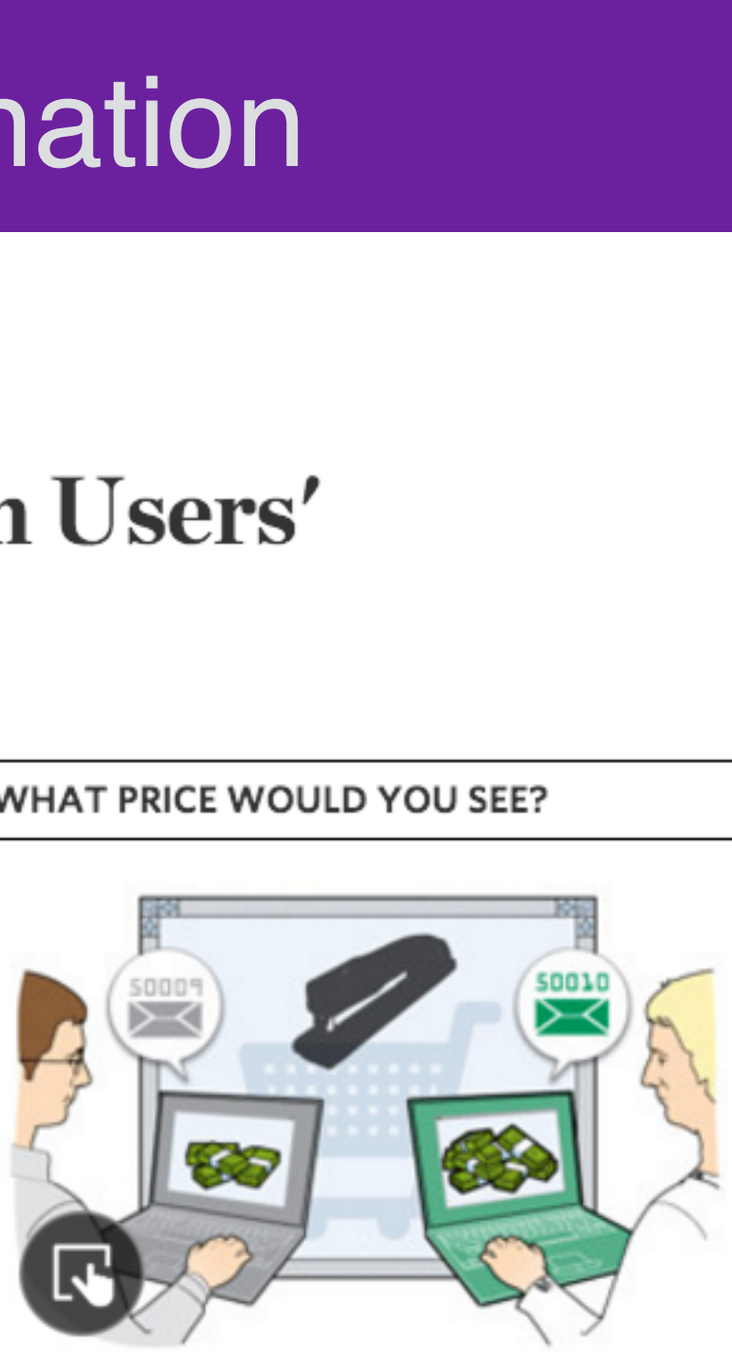
By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and ASHKAN SOLTANI

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

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."

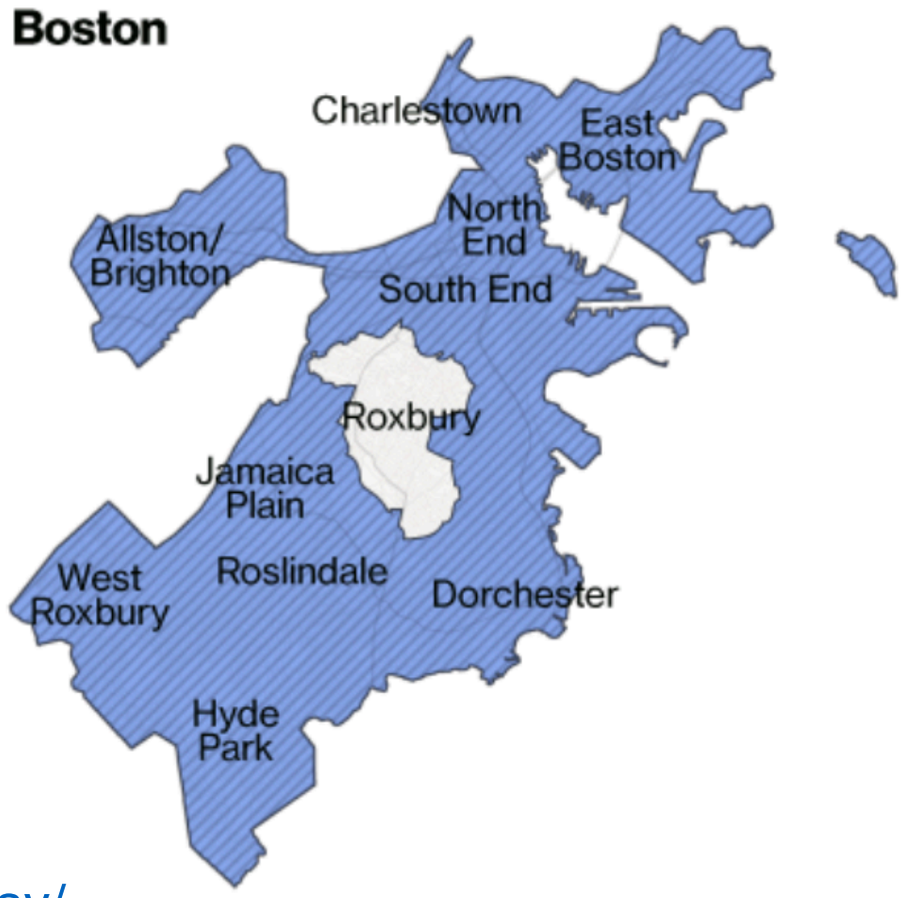
https://www.bloomberg.com/graphics/2016-amazon-same-day/

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/

Online job ads

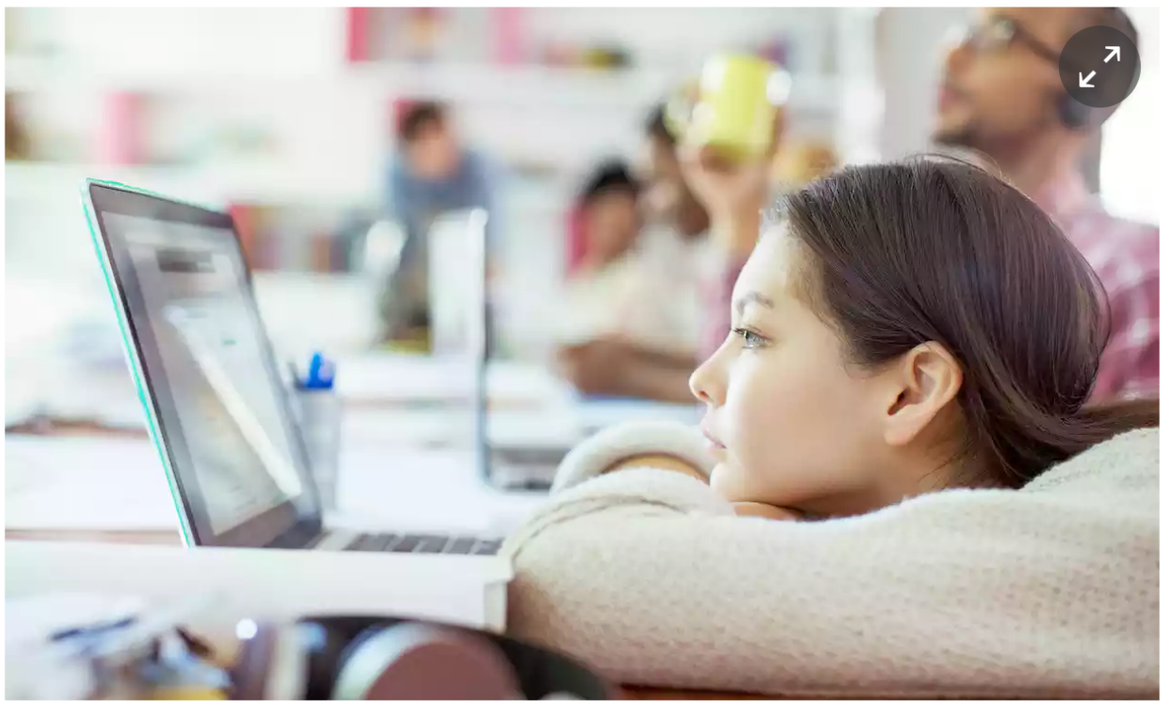
theguardian

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

Job-screening personality tests

THE WALL STREET JOURNAL.

Are Workplace Personality Tests Fair?

Growing Use of Tests Sparks Scrutiny Amid Questions of Effectiveness and Workplace Discrimination

By LAUREN WEBER and ELIZABETH DWOSKIN

Sept. 29, 2014 10:30 p.m. ET

Kyle Behm accused Kroger and six other companies of discrimination against the mentally ill through their use of personality tests. *TROY STAINS FOR THE WALL STREET JOURNAL*

The Equal Employment Opportunity commission is **investigating whether personality tests discriminate against people with disabilities**.

As part of the investigation, officials are trying to determine if the tests **shut out people suffering from mental illnesses** such as depression or bipolar disorder, even if they have the right skills for the job.

http://www.wsj.com/articles/are-workplace-personality-tests-fair-1412044257

Gender bias in recruiting

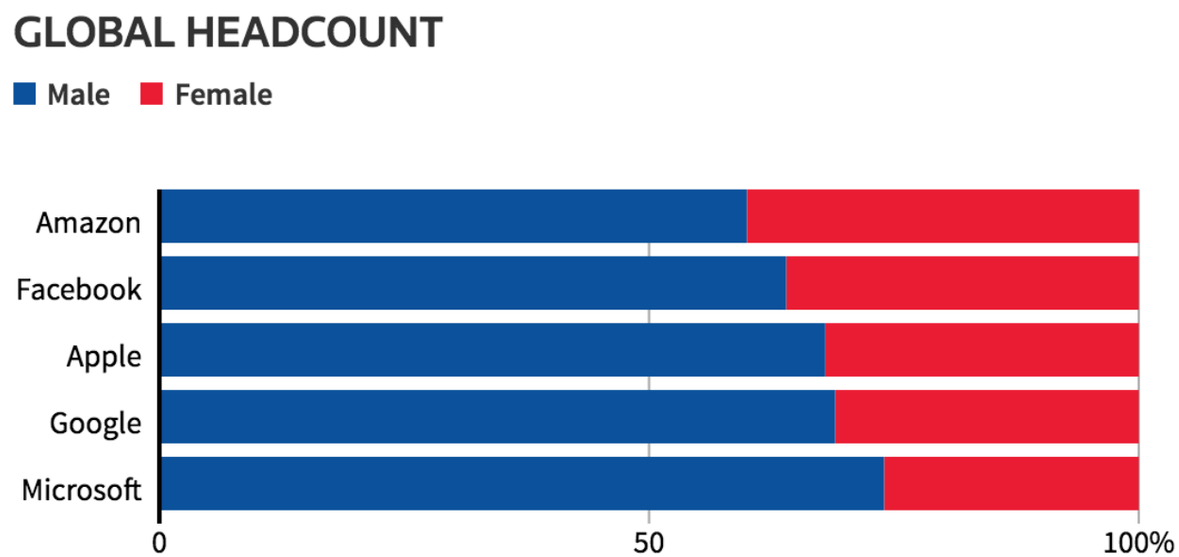
REUTERS

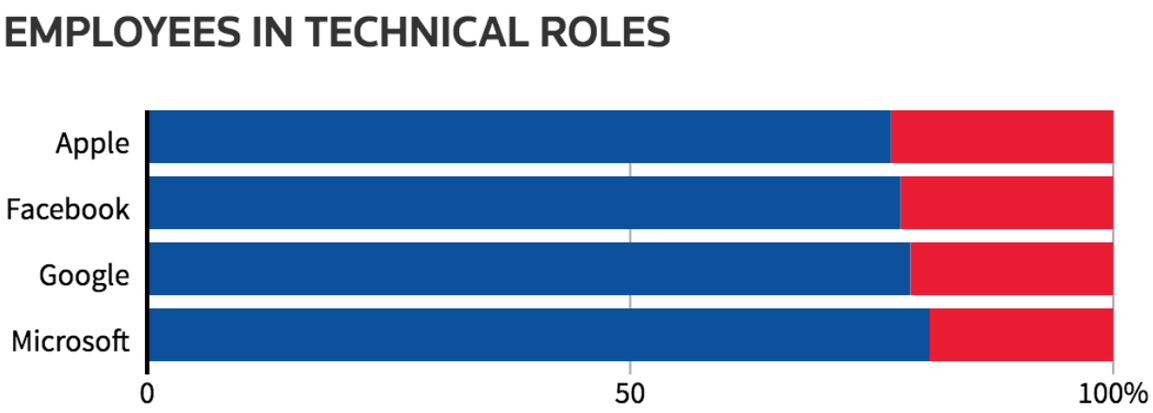
Jeffrey Dastin

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

Job screening: before AI

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).

Job screening: AI?

HARVARD | BUSINESS | SCHOOL

WORKING KNOWLEDGE

Business Research for Business Leaders

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.

Job screening: AI?

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"*)

Racially identifying names

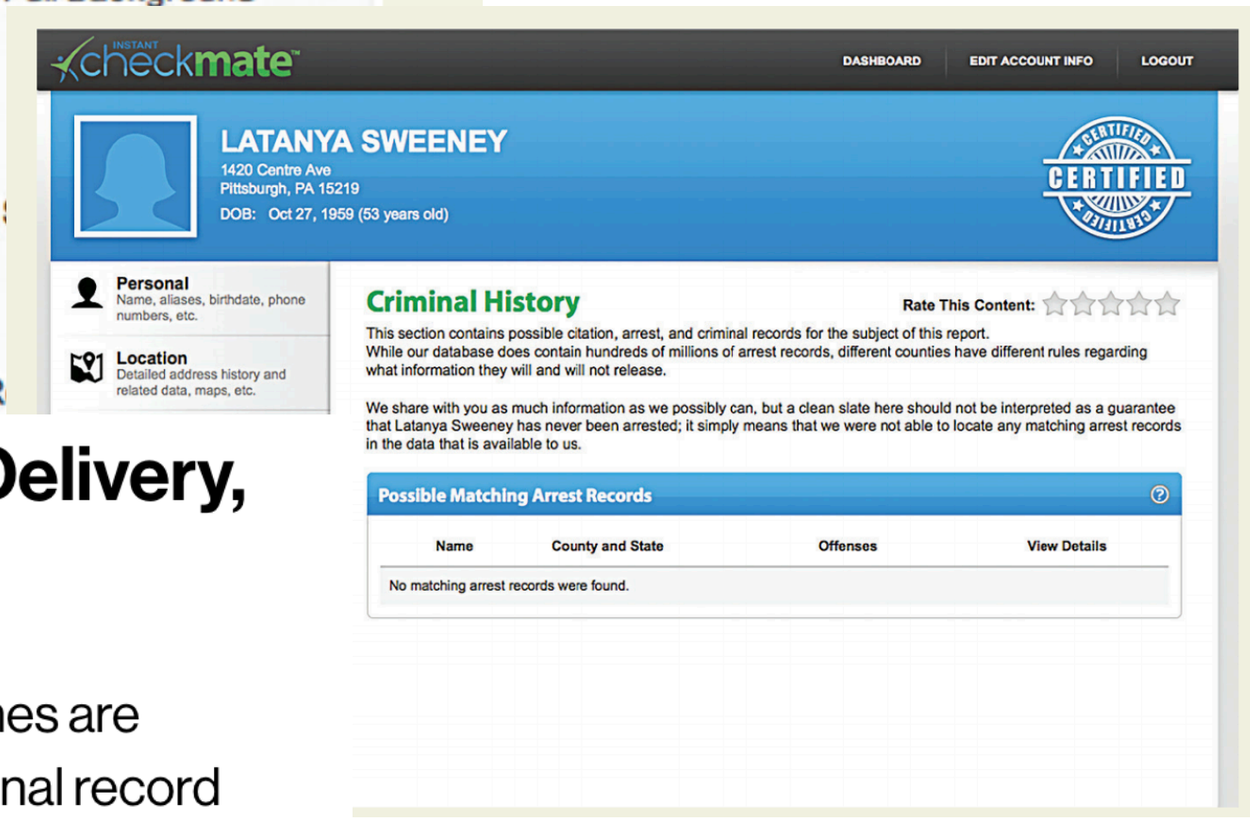
Google AdSense

Ads by Google

Latanya Sweeney, Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

Latanya Sweeney
Public Records Found For: **Latanya**
www.publicrecords.com/

La Tanya
Search for La Tanya Look Up Fast R



Racism is Poisoning Online Ad Delivery, Says Harvard Professor

Google searches involving black-sounding names are more likely to serve up ads suggestive of a criminal record than white-sounding names, says computer scientist

racially identifying names trigger ads suggestive of a criminal record

https://www.technologyreview.com/s/510646/racism-is-poisoning-online-ad-delivery-says-harvard-professor/

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

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

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

Racial bias in health-care algorithms

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2,*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5,*,†]
+ See all authors and affiliations

Science

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and **affecting millions of patients**, exhibits significant **racial bias**: **At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses**. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm **predicts health care costs rather than illness**, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, **despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise**. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

Fixing bias in algorithms?

The New York Times

By Sendhil Mullainathan

ECONOMIC VIEW

Dec. 6, 2019

Biased Algorithms Are Easier to Fix Than Biased People

Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.

Tim Cook

In one study published 15 years ago, **two people applied for a job**. Their résumés were about as similar as two résumés can be. One person was named Jamal, the other Brendan.

In a study published this year, **two patients sought medical care**. Both were grappling with diabetes and high blood pressure. One patient was black, the other was white.

Both studies documented **racial injustice**: In the first, the applicant with a black-sounding name got fewer job interviews. In the second, the black patient received worse care.

But they differed in one crucial respect. In the first, hiring managers made biased decisions. In the second, the culprit was a computer program.

https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

Is data science impartial?

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

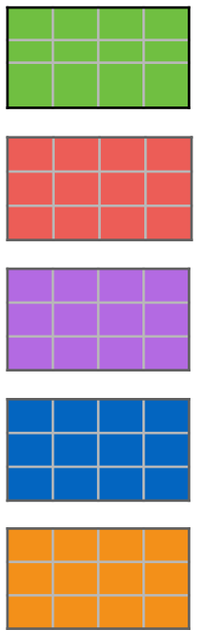
- All traditional evils of **discrimination**, and many new ones, exhibit themselves in the data science eco system
- **Bias** that is inherent in the data or in the process, and that is often due to systemic discrimination, is propelled and amplified
- **Transparency** helps prevent discrimination,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

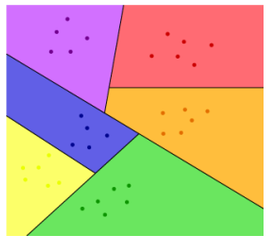
http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

Data, responsibly

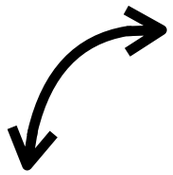
Because of its **power**, data science must be used **responsibly**

fairness

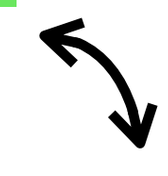




analysis



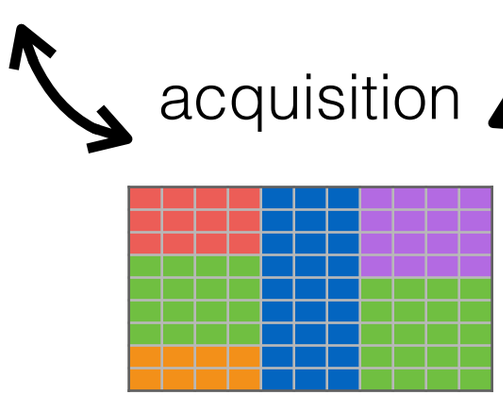
sharing



querying

transparency

diversity



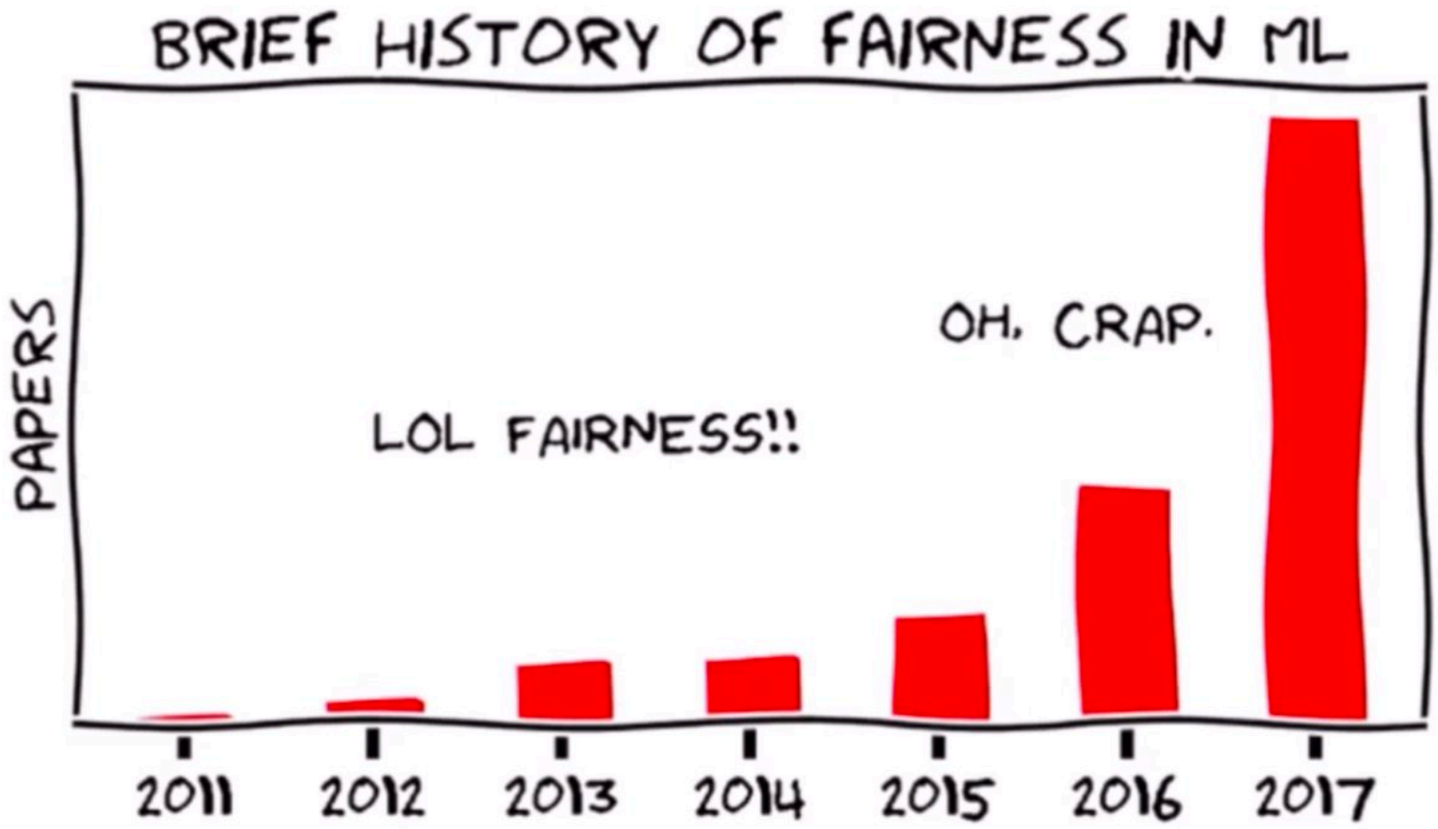
acquisition

data protection

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

Fairness

Fairness in ML



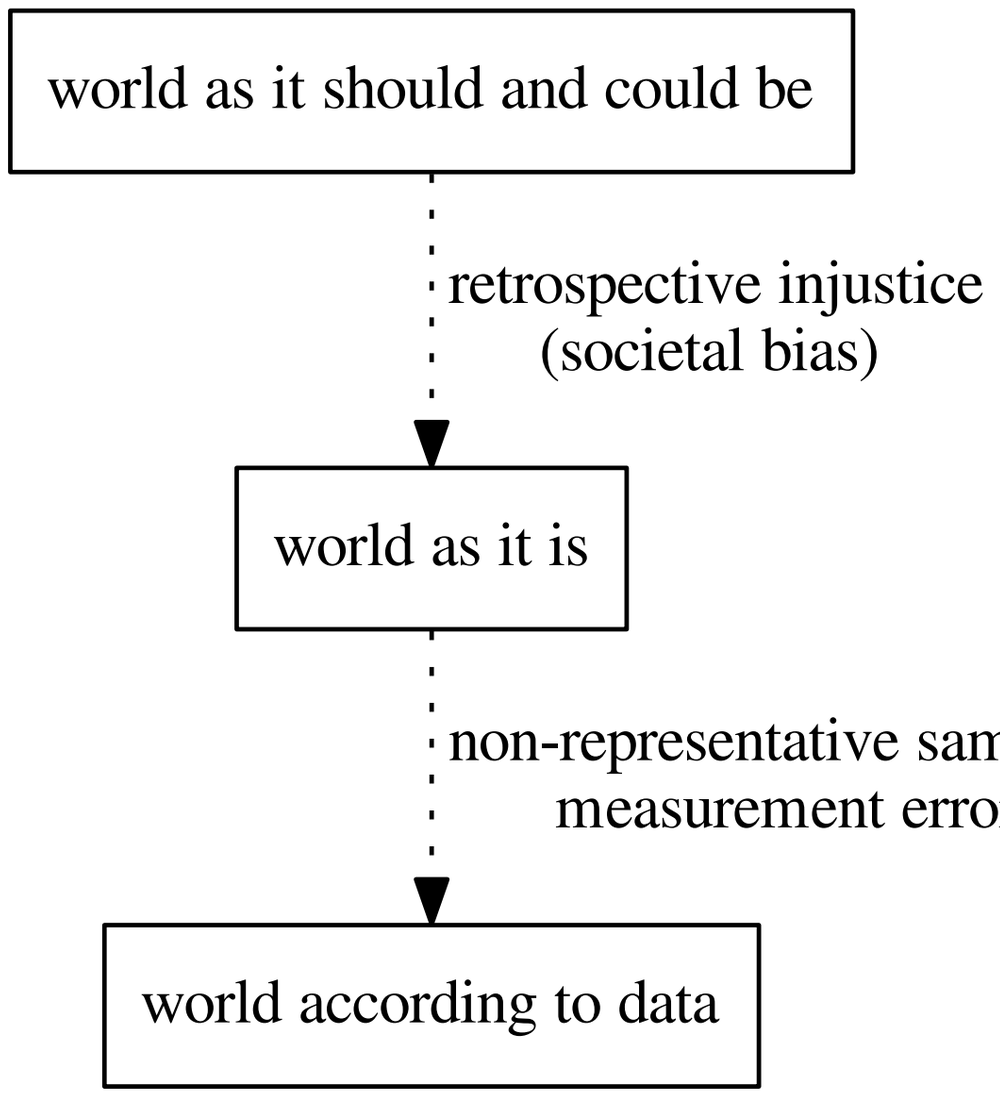
by Moritz Hardt

Fairness is lack of “bias”

- What are the tasks we are interested in?
  - for now, let’s say: predictive analytics
- 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

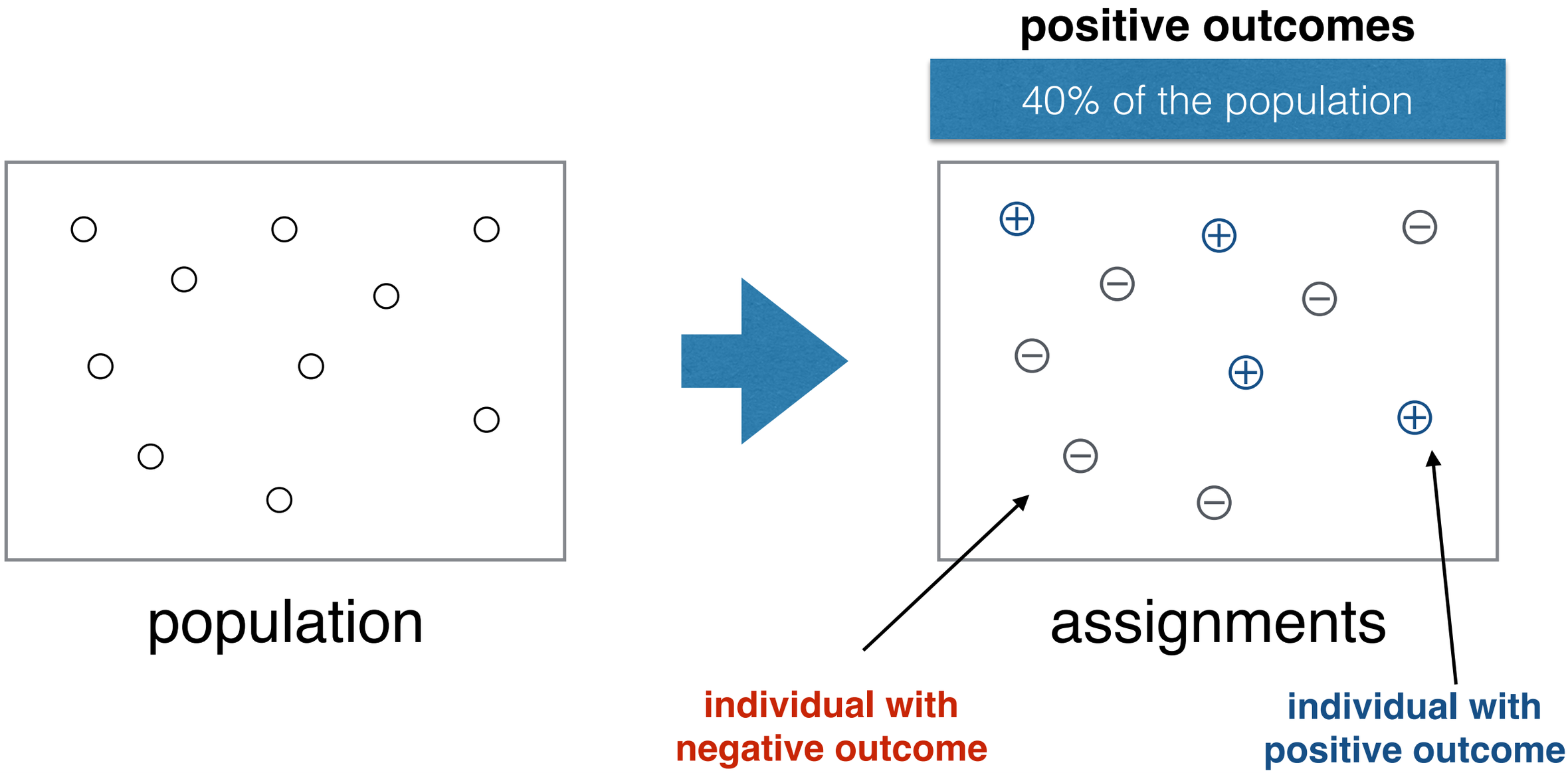
Vendors and outcomes

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

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

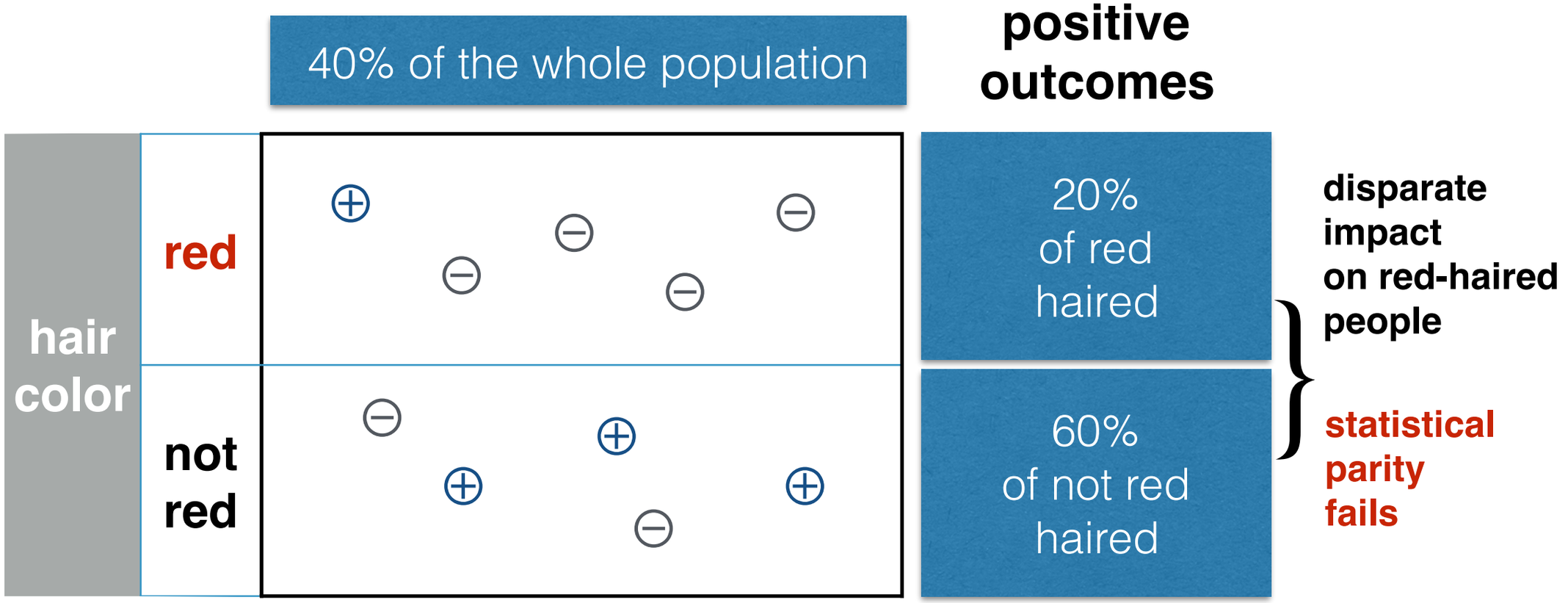
Assigning outcomes to populations

Fairness is concerned with how outcomes are assigned to a population



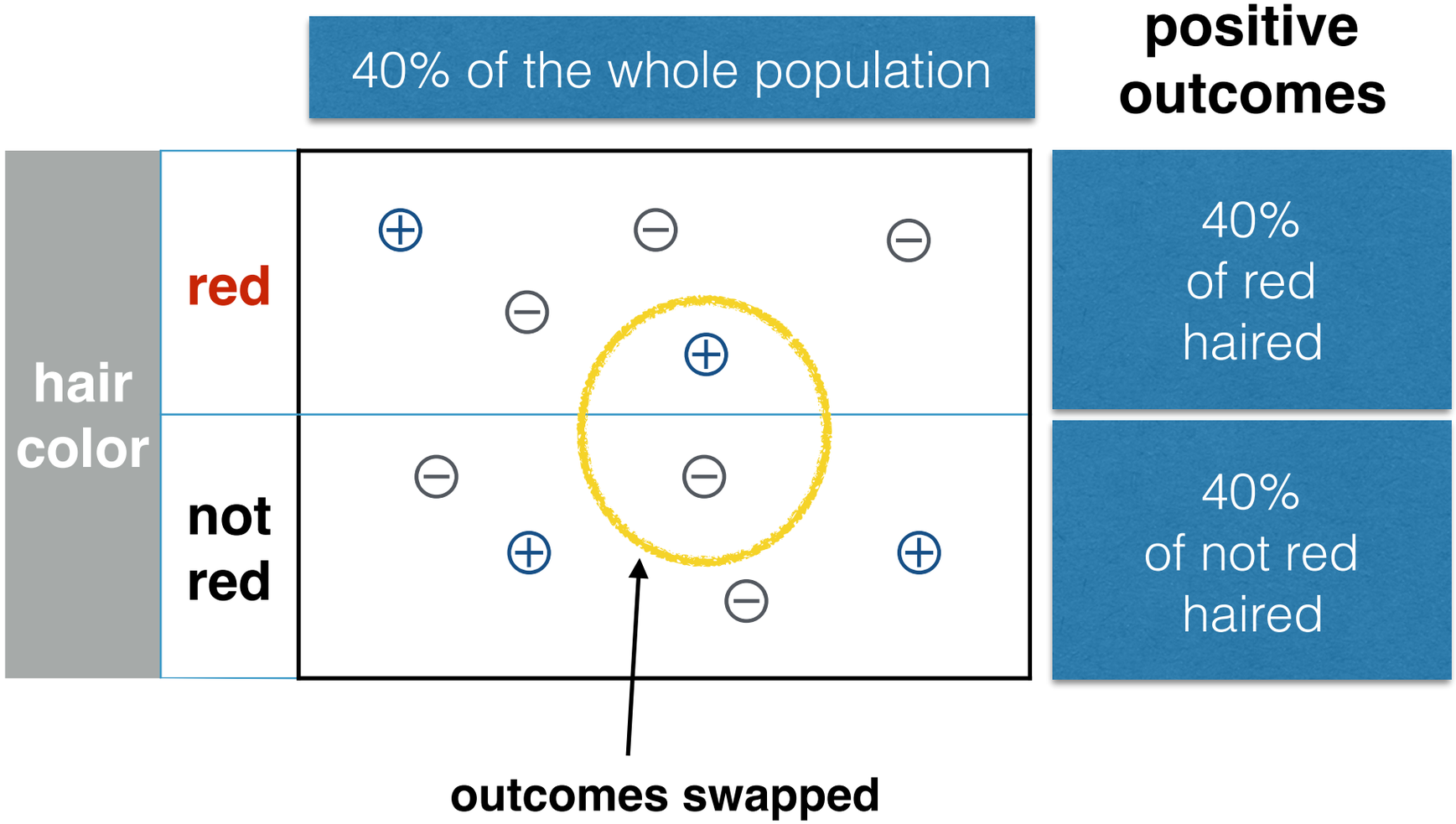
Sub-populations may be treated differently

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)

		hair length: long	hair length: not long
hair color	red	⊕	⊖ ⊖ ⊖
hair color	not red	⊕ ⊕ ⊕	⊖ ⊖

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!

		hair length: long	hair length: not long
hair color	red	⊕	⊖ ⊖ ⊖
	not red	⊕ ⊕ ⊕	⊖ ⊖

positive outcomes

20% of red haired

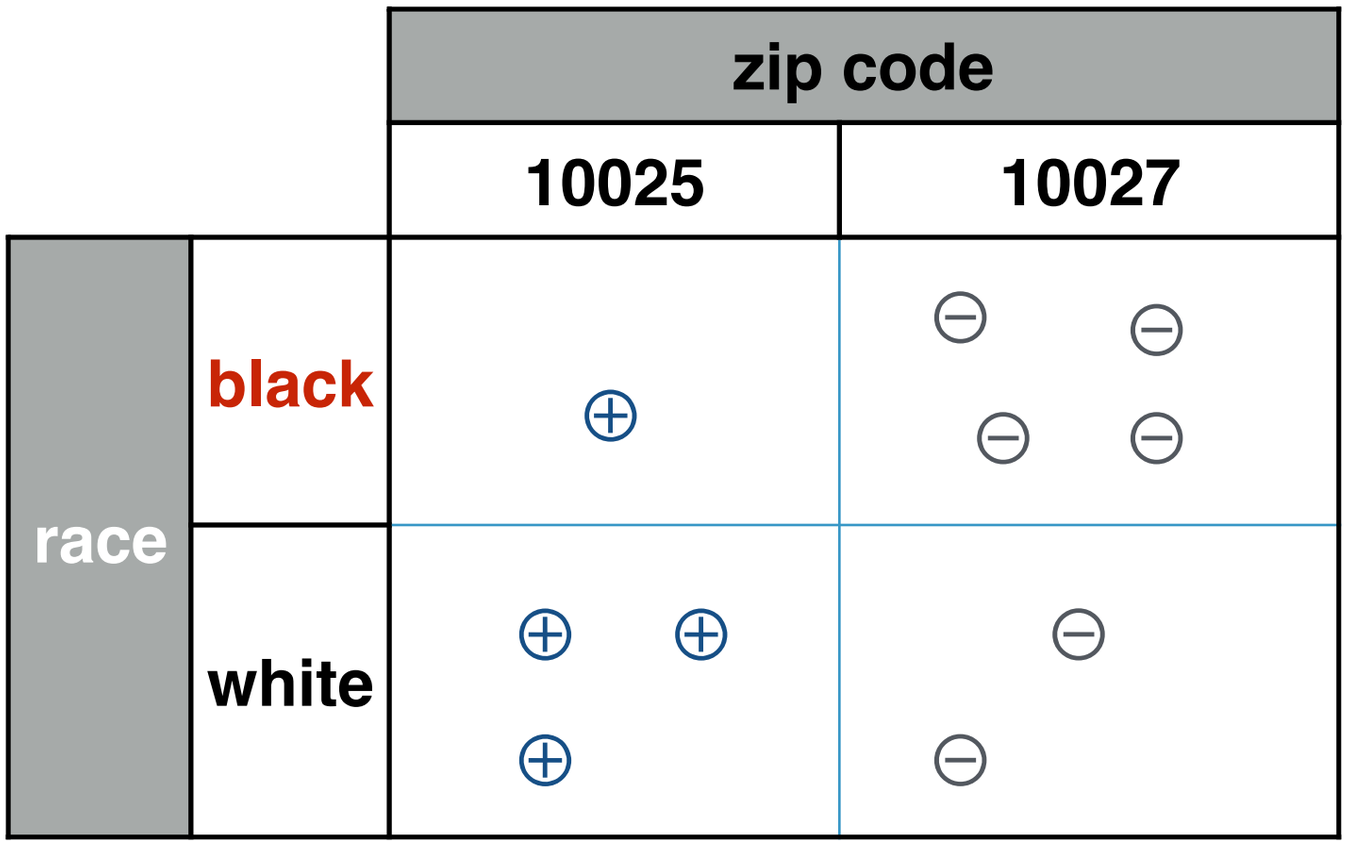
60% of not red haired

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)



positive outcomes

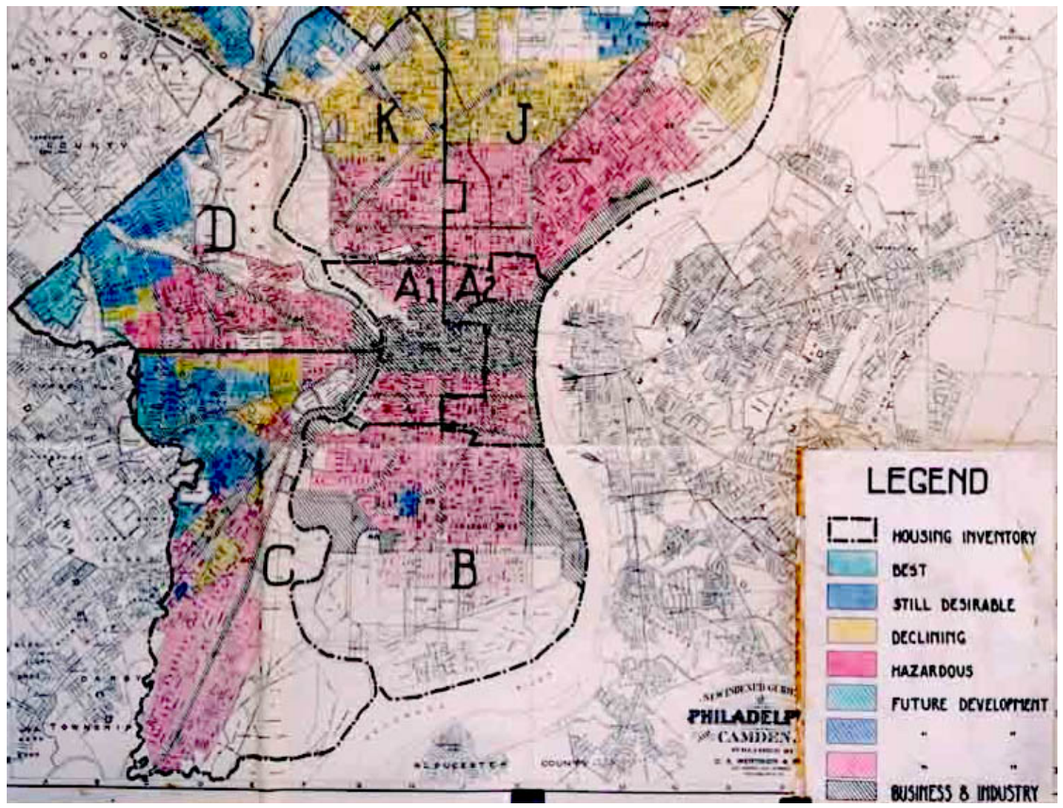
20% of black

60% of white

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.

Philadelphia, 1936



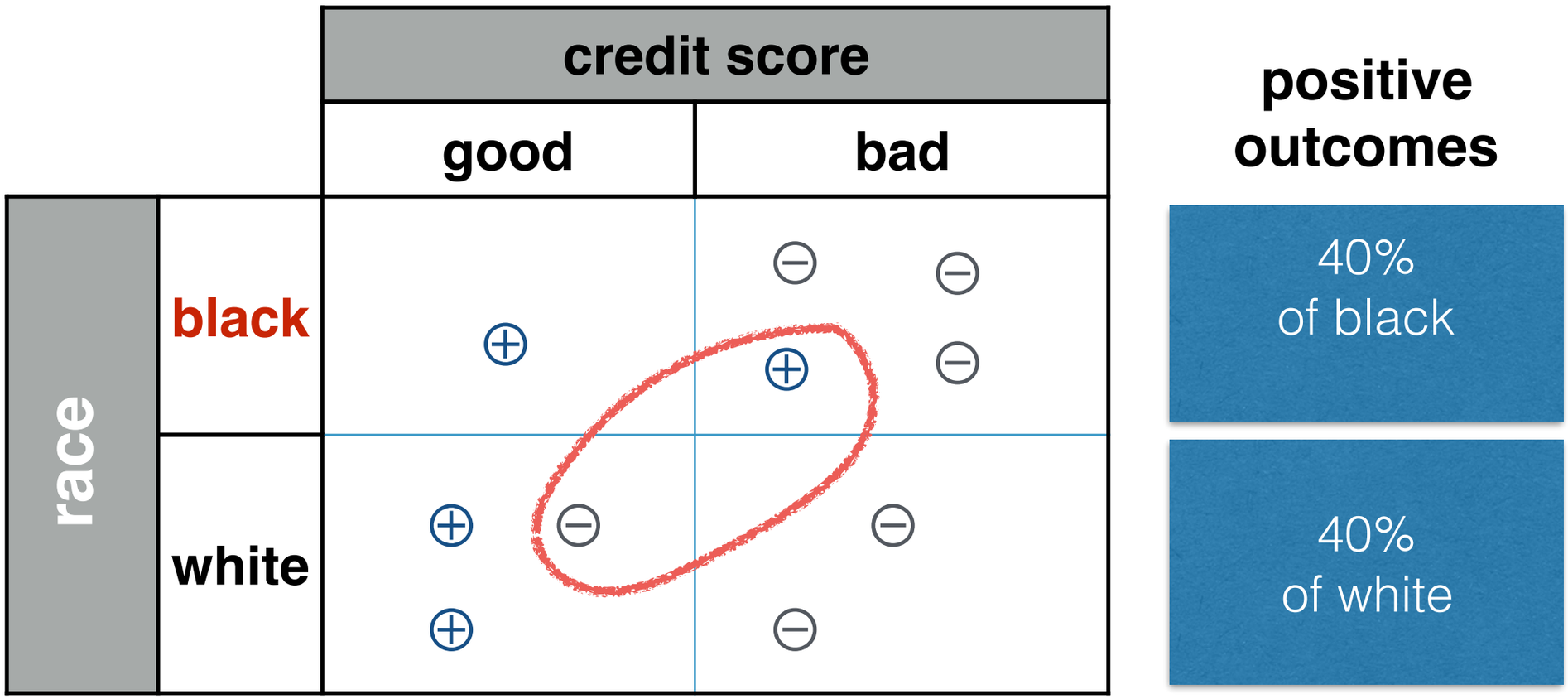
wikipedia

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



positive outcomes

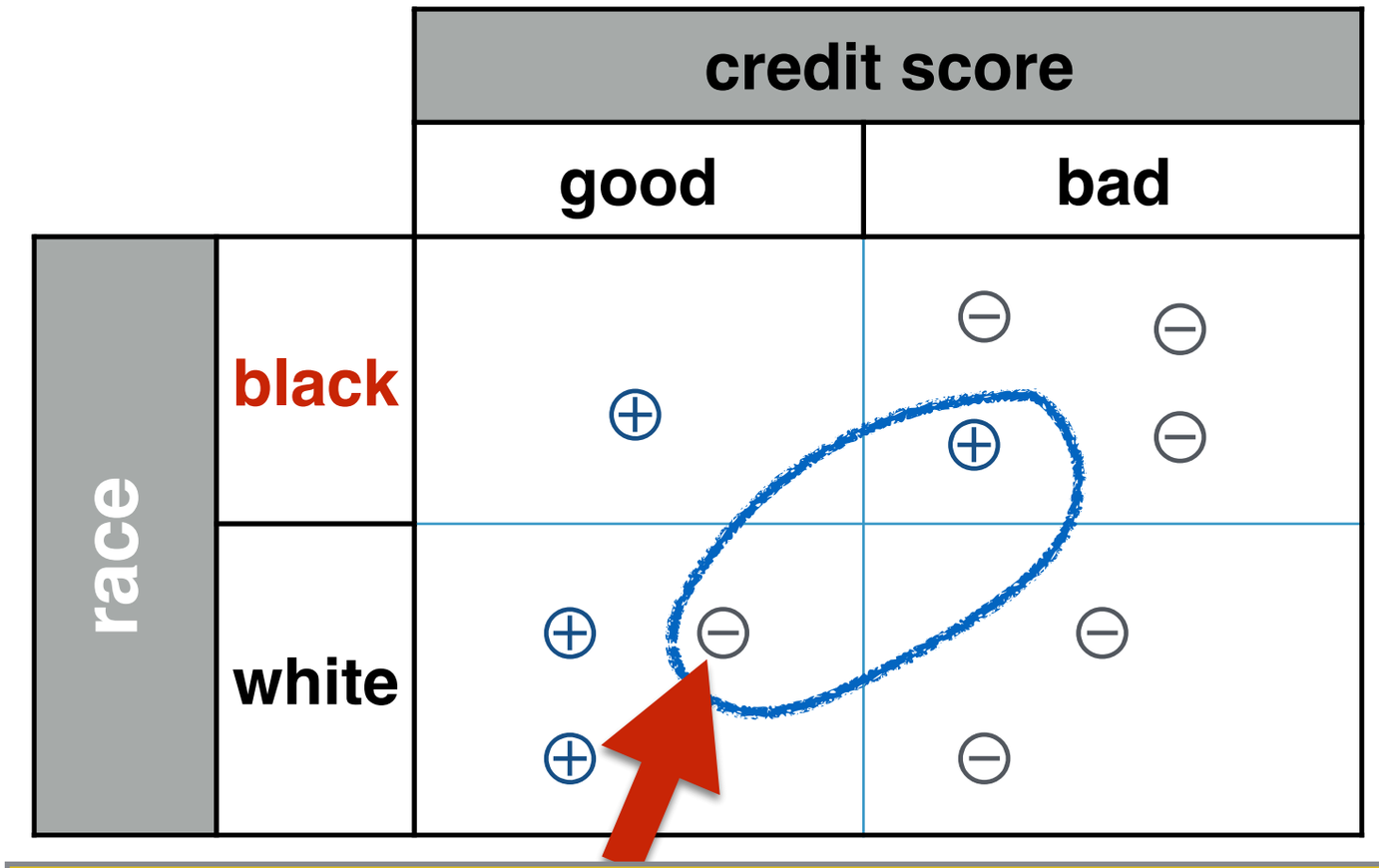
40% of black

40% of white

Impossible to predict loan payback accurately.
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

Ricci v. DeStefano (2009)

Supreme Court Finds Bias Against White Firefighters

By **ADAM LIPTAK** JUNE 29, 2009

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.	

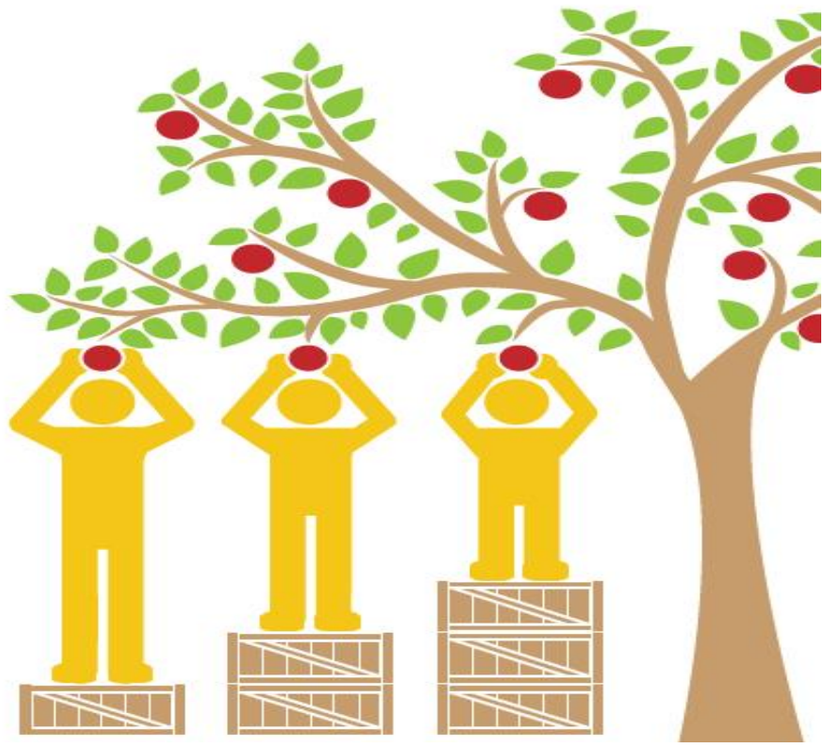
Karen Lee Torre, left, a lawyer who represented the New Haven firefighters in their lawsuit, with her clients Monday at the federal courthouse in New Haven. Christopher Capozziello for The New York Times

Two notions of fairness

individual fairness

equality

group fairness



equity

two intrinsically different world views

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.

A word of caution: Observational data

Correlation is not causation!

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

Discrimination-aware data analysis

- Detecting discrimination
  - mining for discriminatory patterns in (input) data
  - verifying data-driven applications
- Preventing discrimination
  - data pre-processing
  - model post-processing
  - model regularization
  - data post-processing

[Ruggieri *et al.*; 2010]
[Luong *et al.*; 2011]
[Pedresci *et al.*; 2012]
[Romei *et al.*; 2012]
[Hajian & Domingo-Ferrer; 2013]
[Mancuhan & Clifton; 2014]
[Kamiran & Calders; 2009]
[Kamishima *et al.*; 2011]
[Mancuhan & Clifton; 2014]
[Feldman *et al.*; 2015]
[Dwork *et al.*; 2012]
[Zemel *et al.*; 2013]
many more….

both rely on discrimination criteria

Discrimination criteria

[I. Zliobaite, Data Mining & Knowledge Discovery (2017)]

- Statistical tests check how likely the difference between groups is due to chance - is there discrimination?
- Absolute measures express the absolute difference between groups, quantifying the magnitude of discrimination
- Conditional measures express how much of the difference between groups cannot be explained by other attributes, while also quantifying the magnitude of discrimination
- Structural measures how wide-spread is discrimination? Measures the number of individuals impacted by direct discrimination.

Discrimination measures

[I. Zliobaite, Data Mining & Knowledge Discovery (2017)]

a proliferation of task-specific measures

Table III. Summary of absolute measures. Checkmark (✓) indicates that it is directly applicable in a given machine learning setting. Tilde (∼) indicates that a straightforward extension exists (for instance, measuring pairwise).

	Protected variable			Target variable		
Measure	Binary	Categoric	Numeric	Binary	Ordinal	Numeric
Mean difference	✓	∼		✓		✓
Normalized difference	✓	∼		✓		
Area under curve	✓	∼		✓	✓	✓
Impact ratio	✓	∼		✓		
Elift ratio	✓	∼		✓		
Odds ratio	✓	∼		✓		
Mutual information	✓	✓	✓	✓	✓	✓
Balanced residuals	✓	∼		∼	✓	✓
Correlation	✓		✓	✓		✓

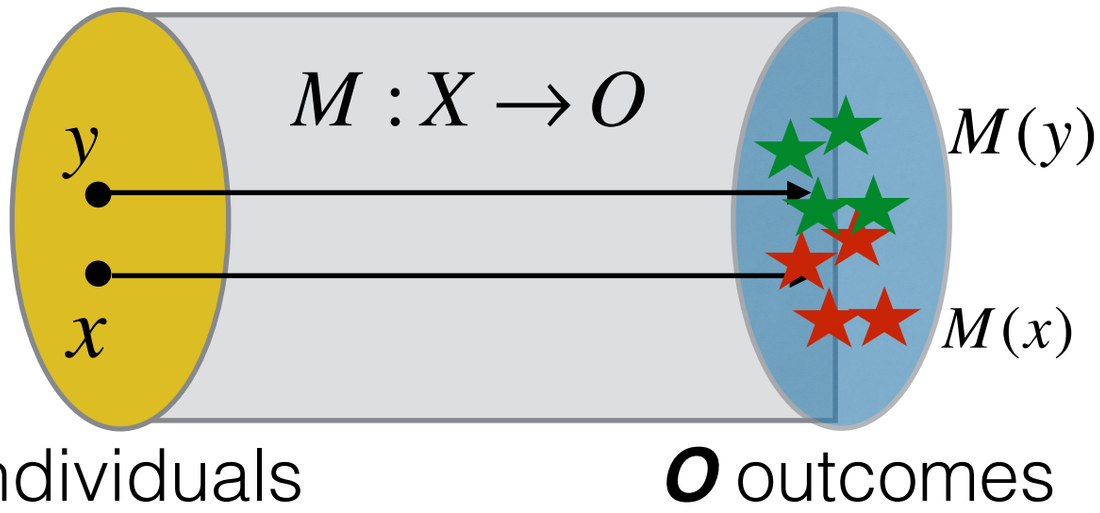
used for statistical parity:

$$\frac{\%\text{ of + for protected class}}{\%\text{ of + for population}}$$

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**.



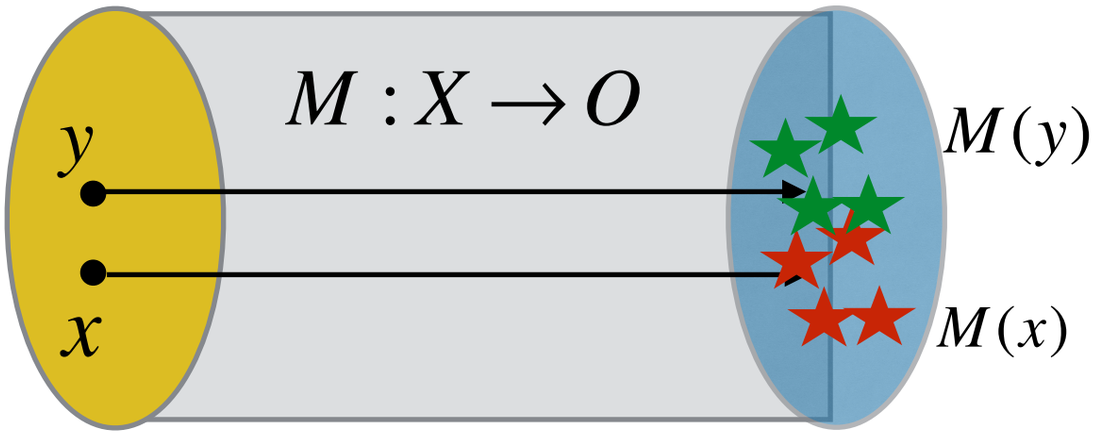
A task-specific similarity metric is given $d(x,y)$

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

Fairness through a Lipschitz mapping

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

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



X individuals ***O*** intermediate mapping

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

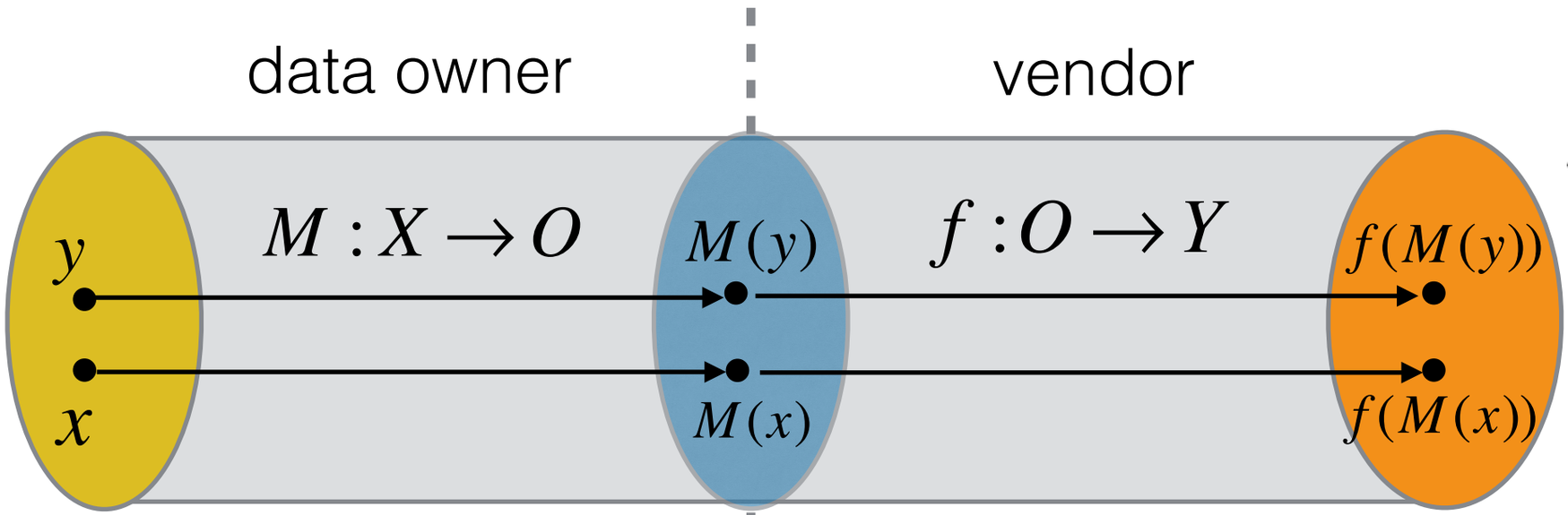
M is a Lipschitz mapping if $\forall x,y \in X \quad ||M(x),M(y)|| \leq d(x,y)$

close individuals map to close distributions

there always exists a Lipschitz mapping - which?

Fairness through awareness

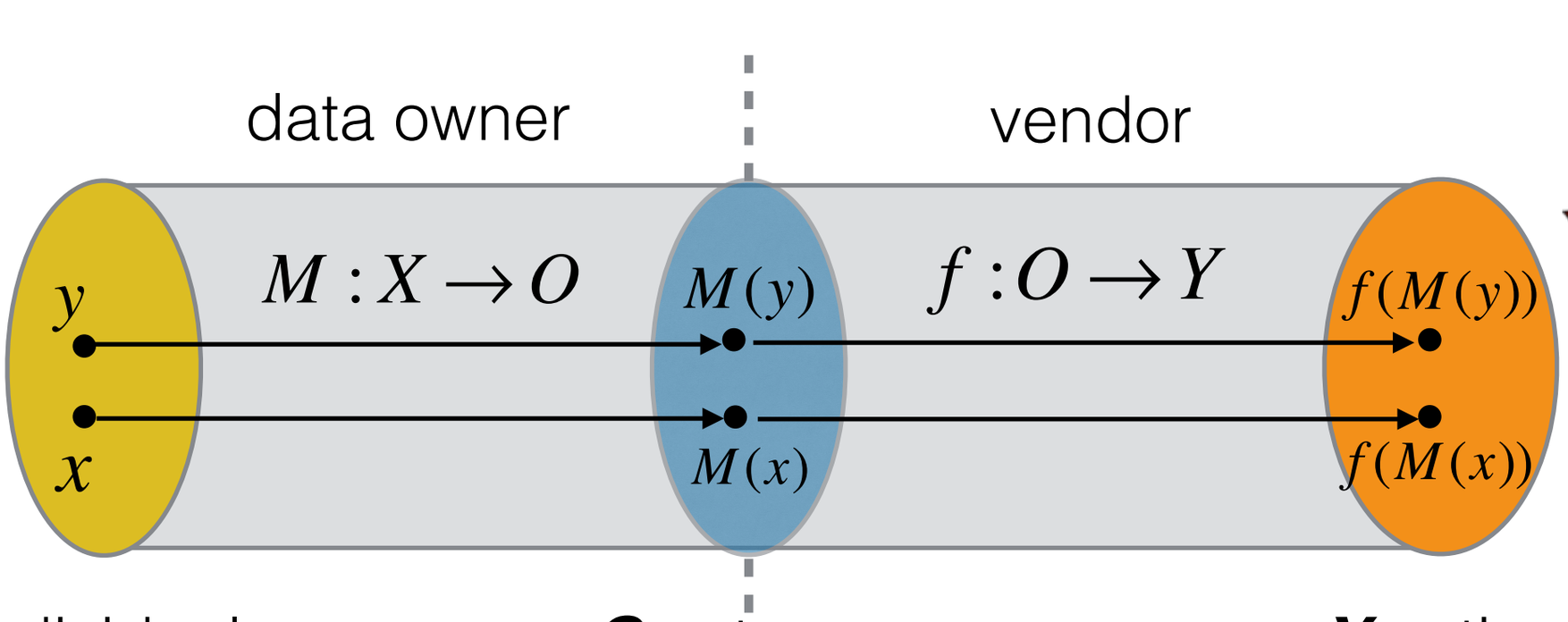
[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]



simpsons.wikia.com

Objective of a data owner

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



simpsons.wikia.com

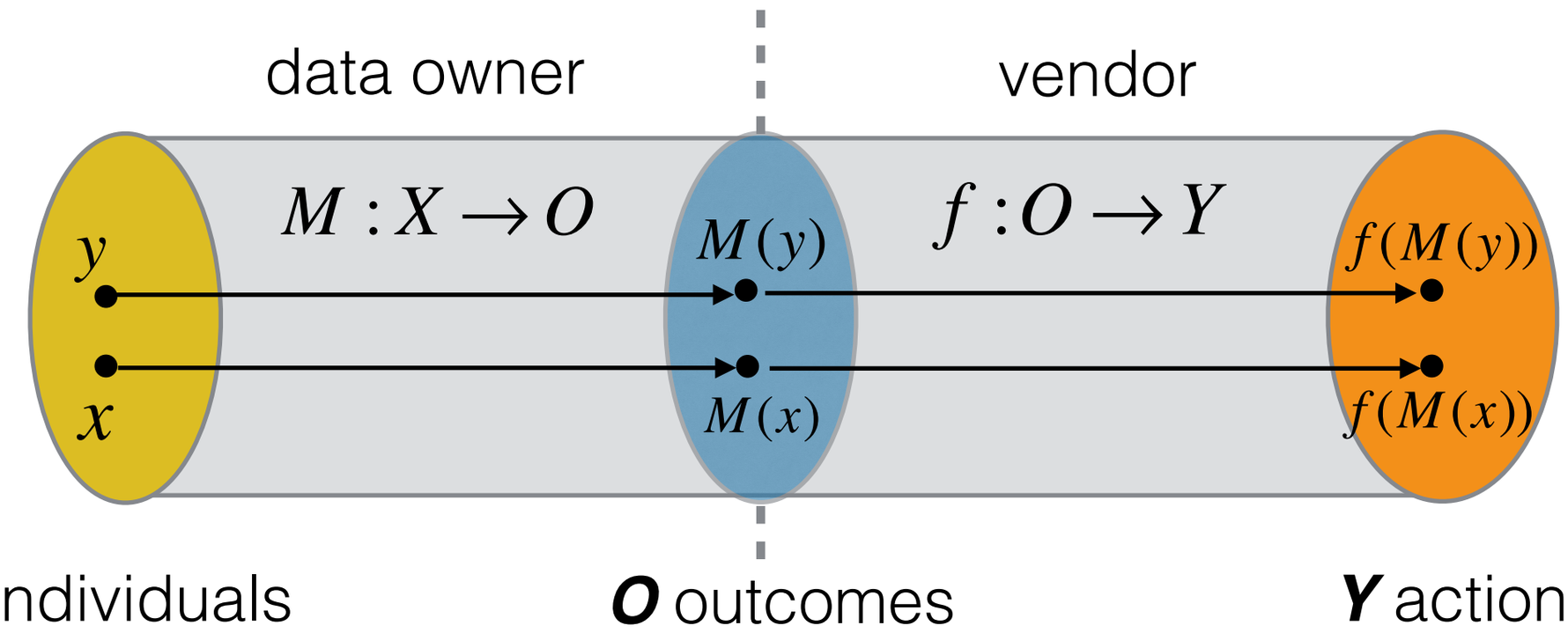
X individuals ***O*** outcomes ***Y*** actions

Find a mapping from individuals to distributions over outcomes that minimizes expected loss, **subject to the Lipschitz condition**. Optimization problem: minimize an arbitrary loss function.

What about the vendor?

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

Vendors can efficiently maximize expected utility,
subject to the Lipschitz condition



Computed with a linear program of size $poly(|X|,|Y|)$

the same mapping can be used by multiple vendors

Some philosophical background

[C. Calsamiglia; *PhD thesis 2005*]

"**Equality of opportunity** defines an important welfare criterion in political philosophy and policy analysis.
Philosophers define equality of opportunity as the requirement that an individual's well being be independent of his or her irrelevant characteristics. **The difference among philosophers is mainly about which characteristics should be considered irrelevant.**

Policymakers, however, are often called upon to address more specific questions: How should admissions policies be designed so as to provide equal opportunities for college? Or how should tax schemes be designed so as to equalize opportunities for income? These are called local distributive justice problems, because each policymaker is in charge of achieving equality of opportunity to a specific issue."

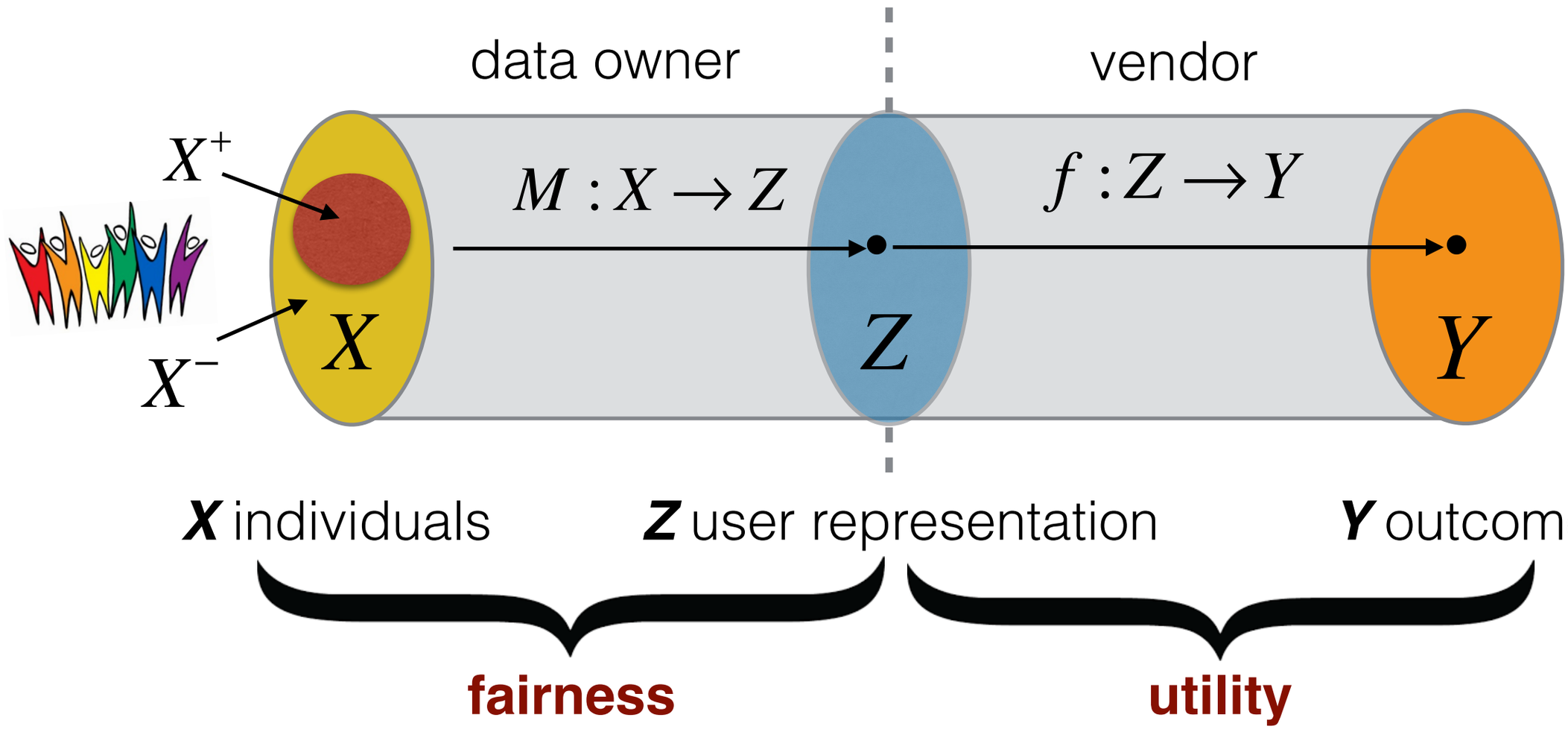
Fairness through awareness: summary

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

- An early work in this space, proposes a principled data pre-processing approach
- Stated as an individual fairness condition but also sometimes leads to group fairness
- Relies on an externally-supplied task-specific similarity metric - magic!
- Is not formulated as a learning problem, does not generalize to unseen data

Learning fair representations

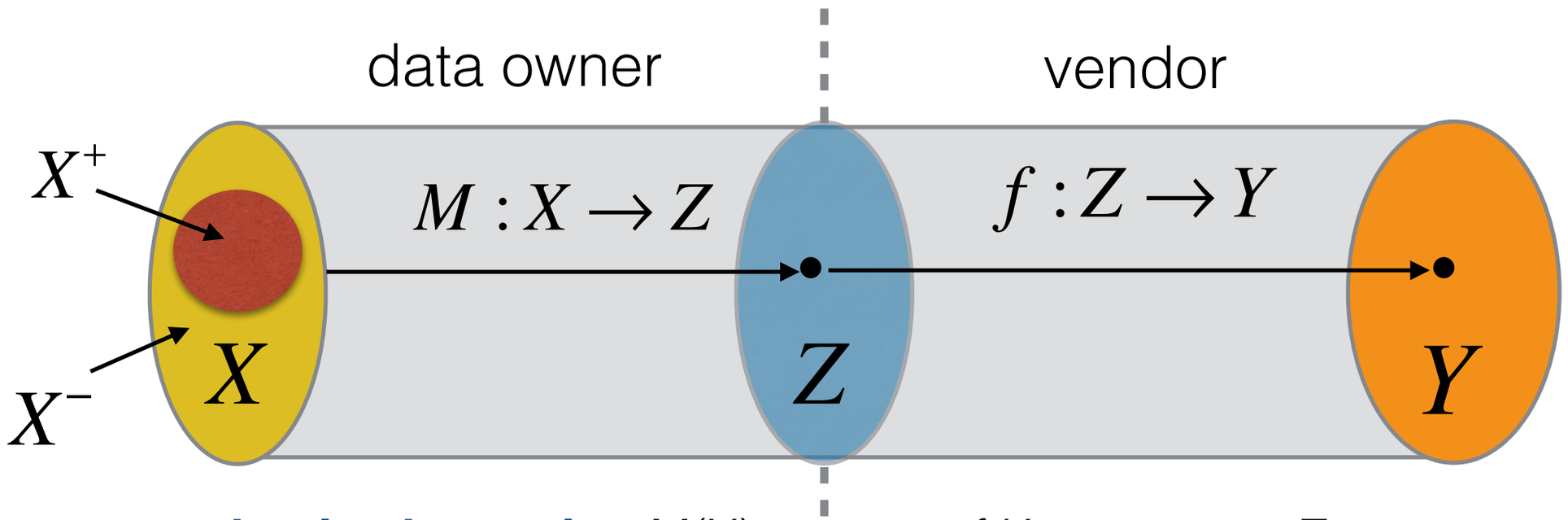
[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]



- **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 $P(Z \mid S = 0) = P(Z \mid S = 1)$

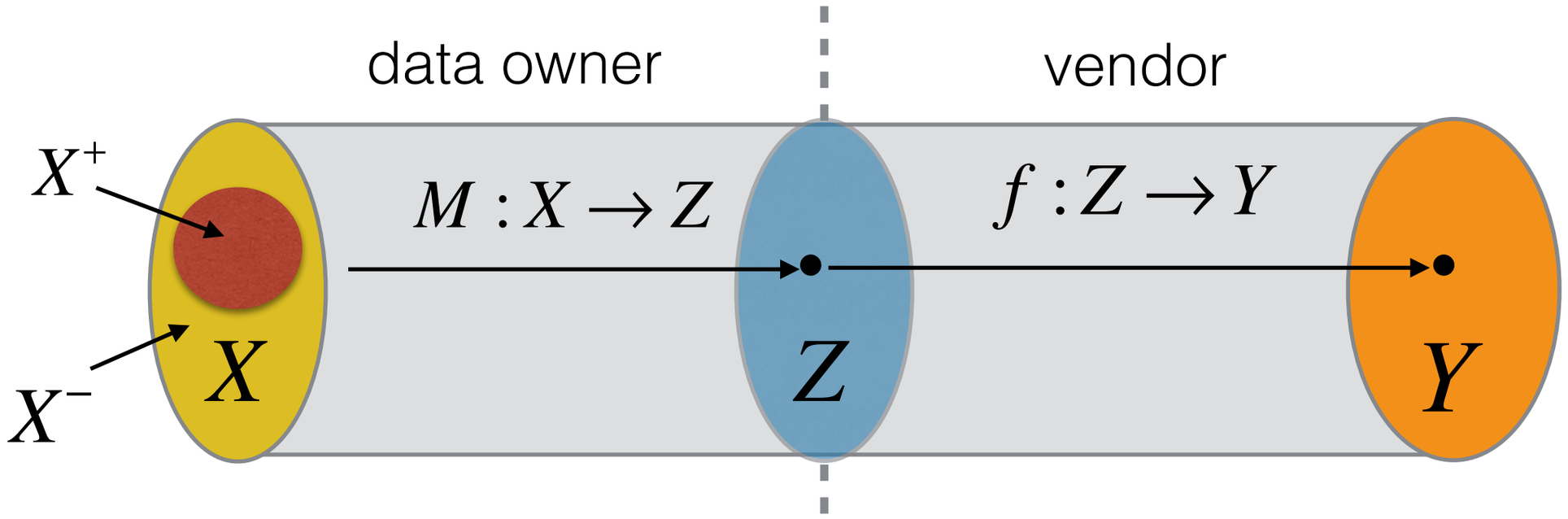
M(X) should preserve other information so that vendor can maximize utility

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

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**

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

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

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

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

does this make sense?

Learning fair representations: summary

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

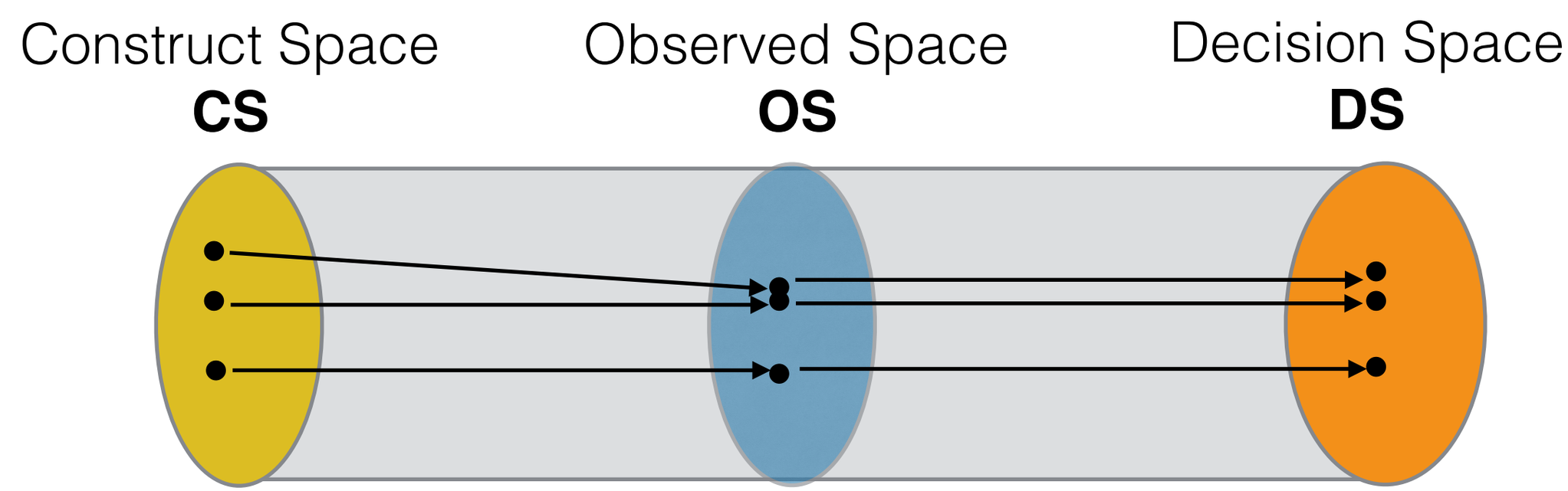
- A principled learning framework in the data pre-processing / classifier regularization category
- **Evaluation** of accuracy, discrimination (group fairness) and consistency (individual fairness), promising results on real datasets
- Not clear how to set *K*, so as to trade off accuracy / fairness
- The mapping is **task-specific**

On the (im)possibility of fairness

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

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

Main insight: To study algorithmic fairness is to study the interactions between different spaces that make up the decision pipeline for a task



Examples of features and outcomes

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

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

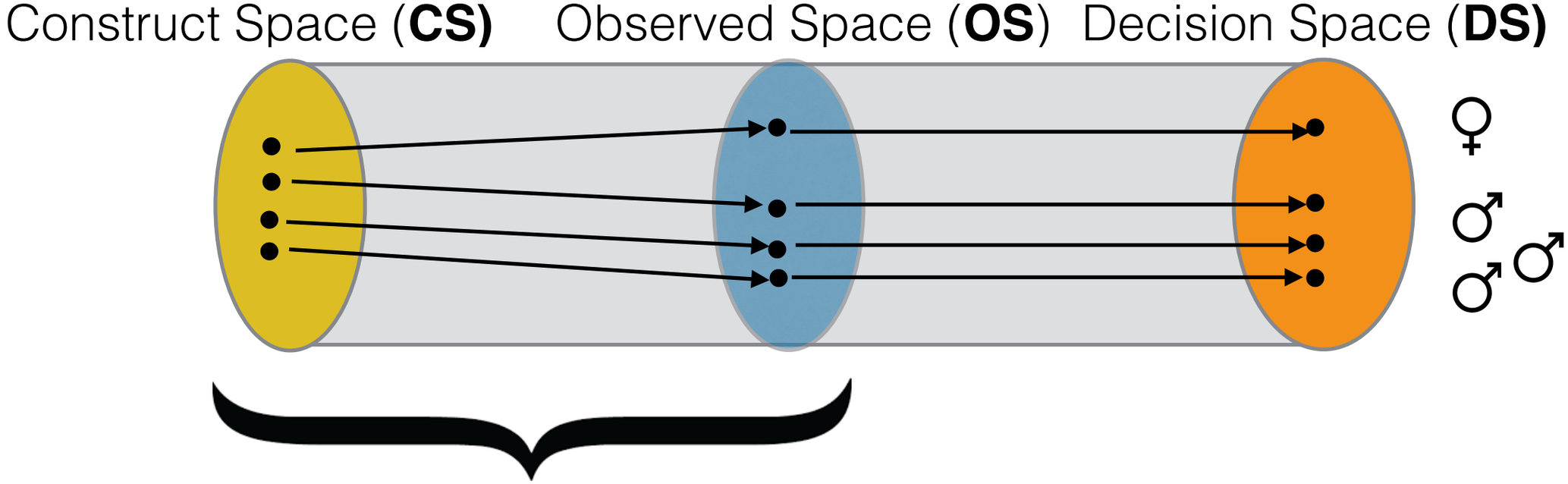
define fairness through properties of mappings between CS, OS and DS

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**.

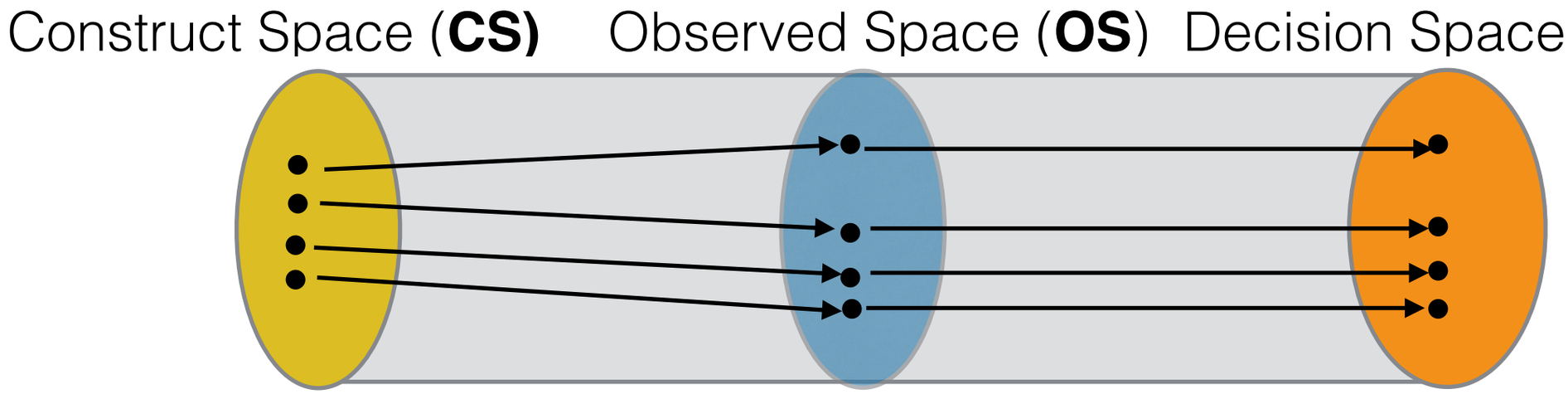
$$f : CS \to DS \qquad d_{CS}(x,y) < \varepsilon \Rightarrow d_{DS}(f(x), f(y)) < \varepsilon'$$



let's focus on this portion

A world view: What you see is what you get

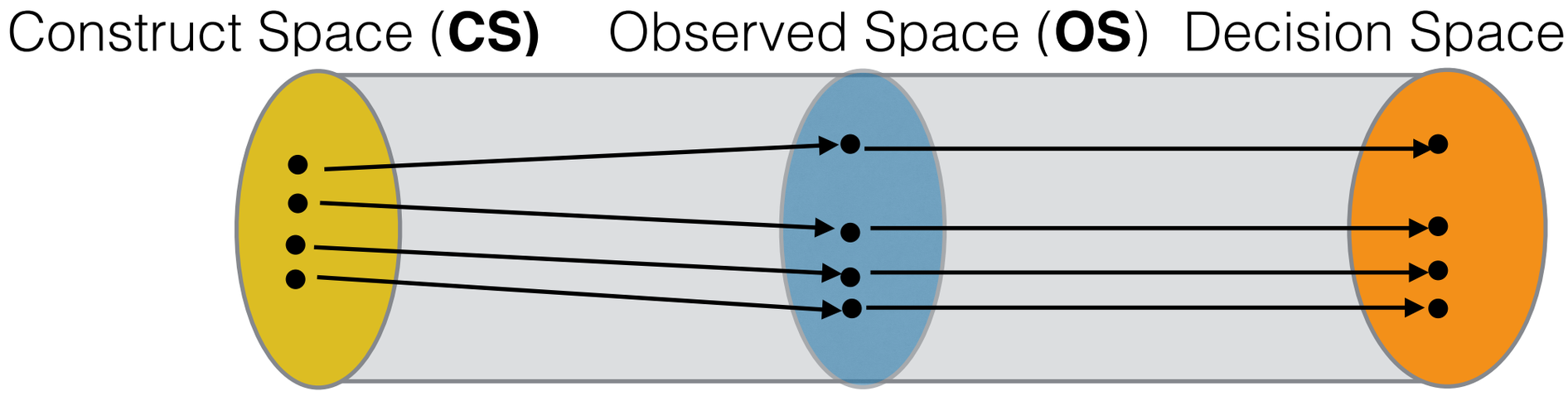
[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]



What you see is what you get (**WYSIWYG**): there exists a mapping from **CS** to **OS** that has low distortion. That is, we believe that OS faithfully represents CS. **This is the individual fairness world view.**

A world view: Structural bias

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



We are all equal (**WAE**): the mapping from CS to OS introduces **structural bias** - there is a distortion that aligns with the group structure of CS. **This is the group fairness world view.**

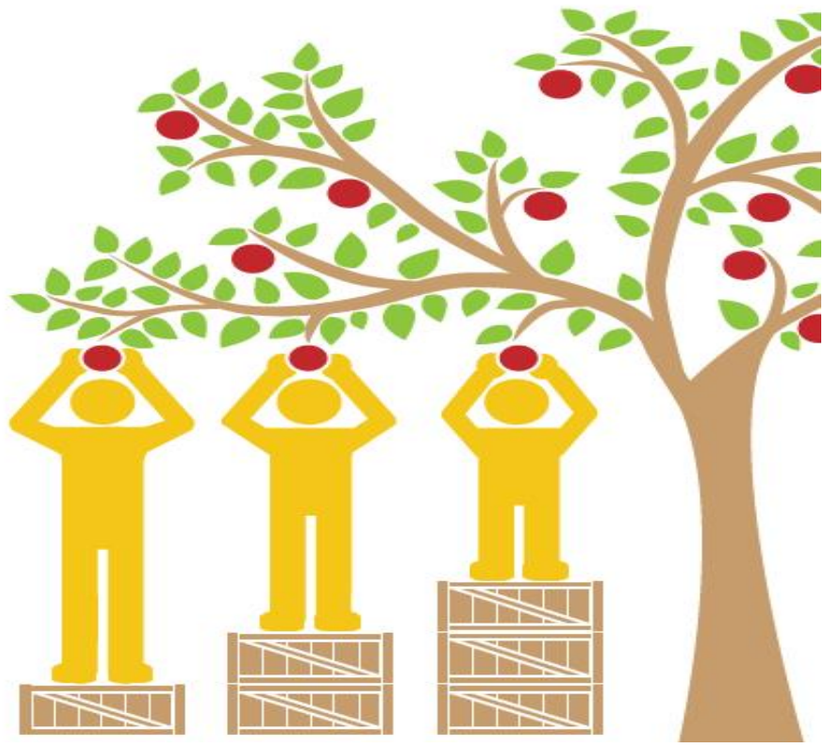
Structural bias examples: SAT verbal questions function differently in the African-American and in the Caucasian subgroups in the US. Other examples?

Two notions of fairness

individual fairness

equality

group fairness



equity

two intrinsically different world views

What's the right answer?

There is no single answer!

Need transparency and public debate

- Consider harms and benefits to different stakeholders
- Being 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 = ?

Fairness definitions as “trolley problems”