Responsible Data Science

Algorithmic Fairness

Prof. Julia Stoyanovich

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

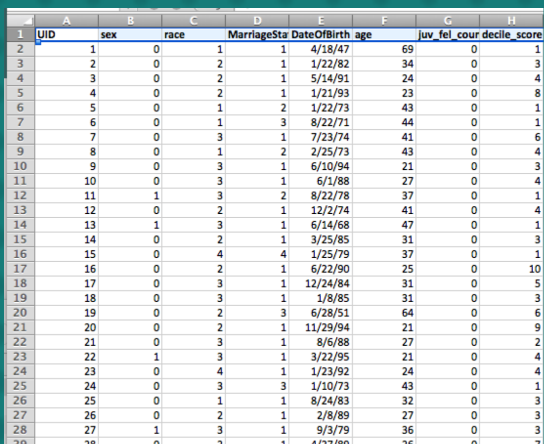
data RESPONSIBLY

NYU

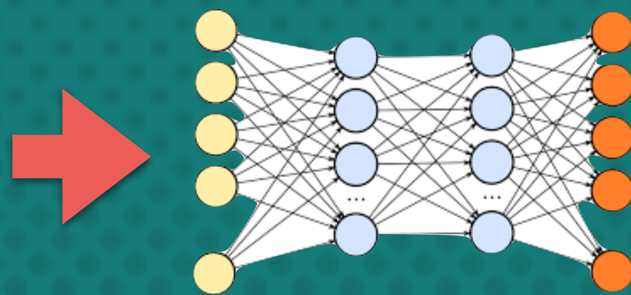
recall:

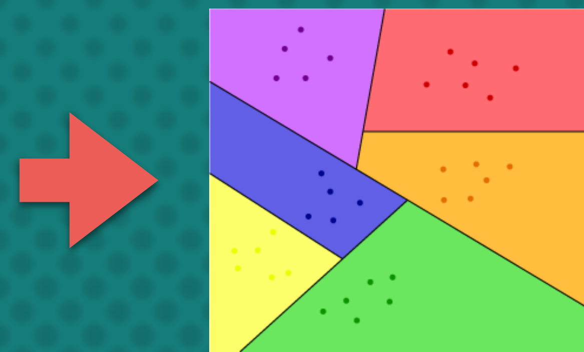
pre-existing bias

“Bias” in predictive analytics



	A	B	C	D	E	F	G	H
1	UID	sex	race	MarriageSta	DateOfBirth	age	juv_fel_cour	decile_score
2	1	0	1	1	4/18/47	69	0	1
3	2	0	2	1	1/22/82	34	0	3
4	3	0	2	1	5/14/91	24	0	4
5	4	0	2	1	1/21/93	23	0	8
6	5	0	1	2	1/22/73	43	0	1
7	6	0	1	3	8/22/71	44	0	1
8	7	0	3	1	7/23/74	41	0	6
9	8	0	1	2	2/25/73	43	0	4
10	9	0	3	1	6/10/94	21	0	3
11	10	0	3	1	6/1/88	27	0	4
12	11	1	3	2	8/22/78	37	0	1
13	12	0	2	1	12/2/74	41	0	4
14	13	1	3	1	6/14/68	47	0	1
15	14	0	2	1	3/25/85	31	0	3
16	15	0	4	4	1/25/79	37	0	1
17	16	0	2	1	6/22/90	25	0	10
18	17	0	3	1	12/24/84	31	0	5
19	18	0	3	1	1/8/85	31	0	3
20	19	0	2	3	6/28/51	64	0	6
21	20	0	2	1	11/29/94	21	0	9
22	21	0	3	1	8/6/88	27	0	2
23	22	1	3	1	3/22/95	21	0	4
24	23	0	4	1	1/23/92	24	0	4
25	24	0	3	3	1/10/73	43	0	1
26	25	0	1	1	8/24/83	32	0	3
27	26	0	2	1	2/8/89	27	0	3
28	27	1	3	1	9/3/79	36	0	3





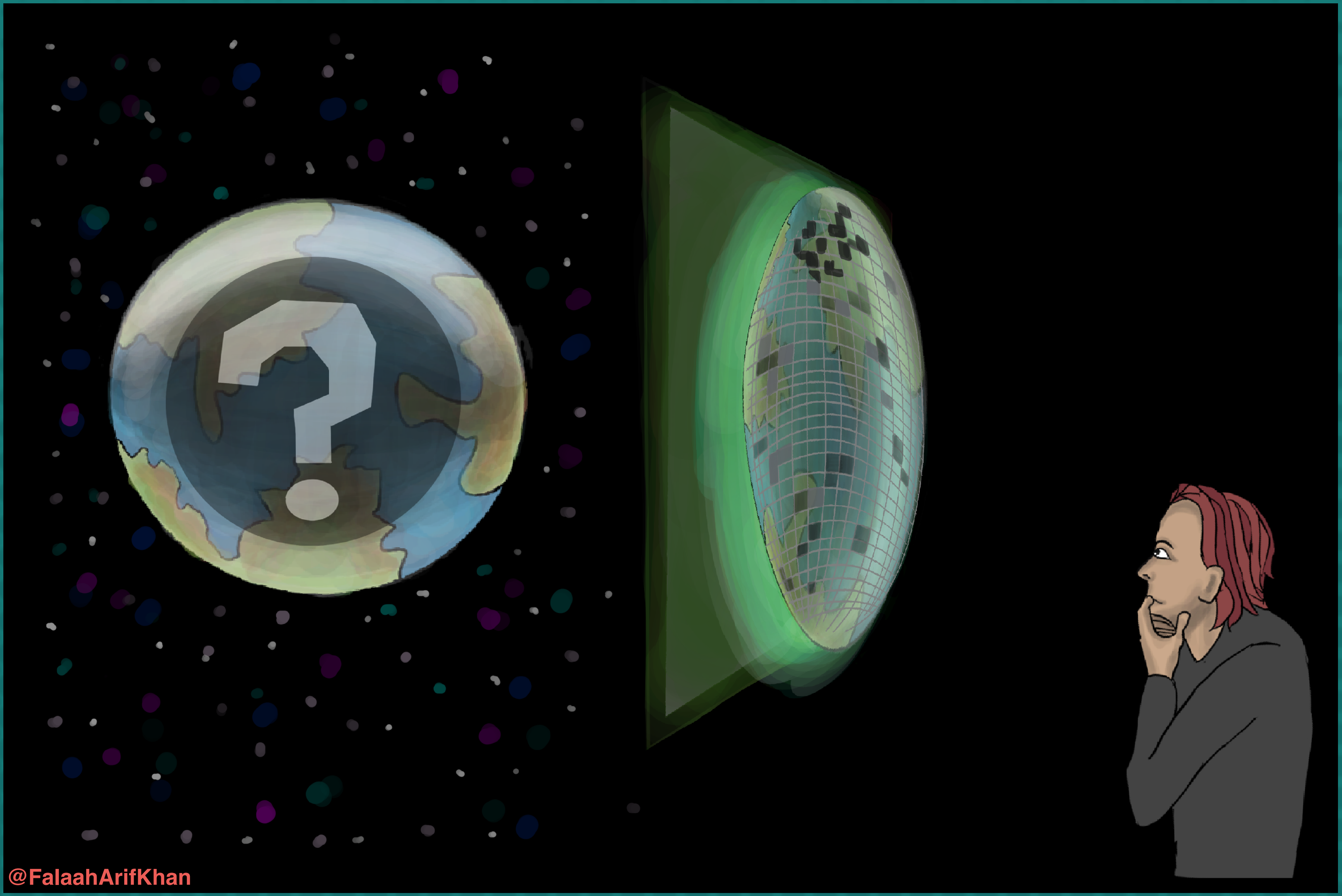
Statistical

model does not summarize the data correctly

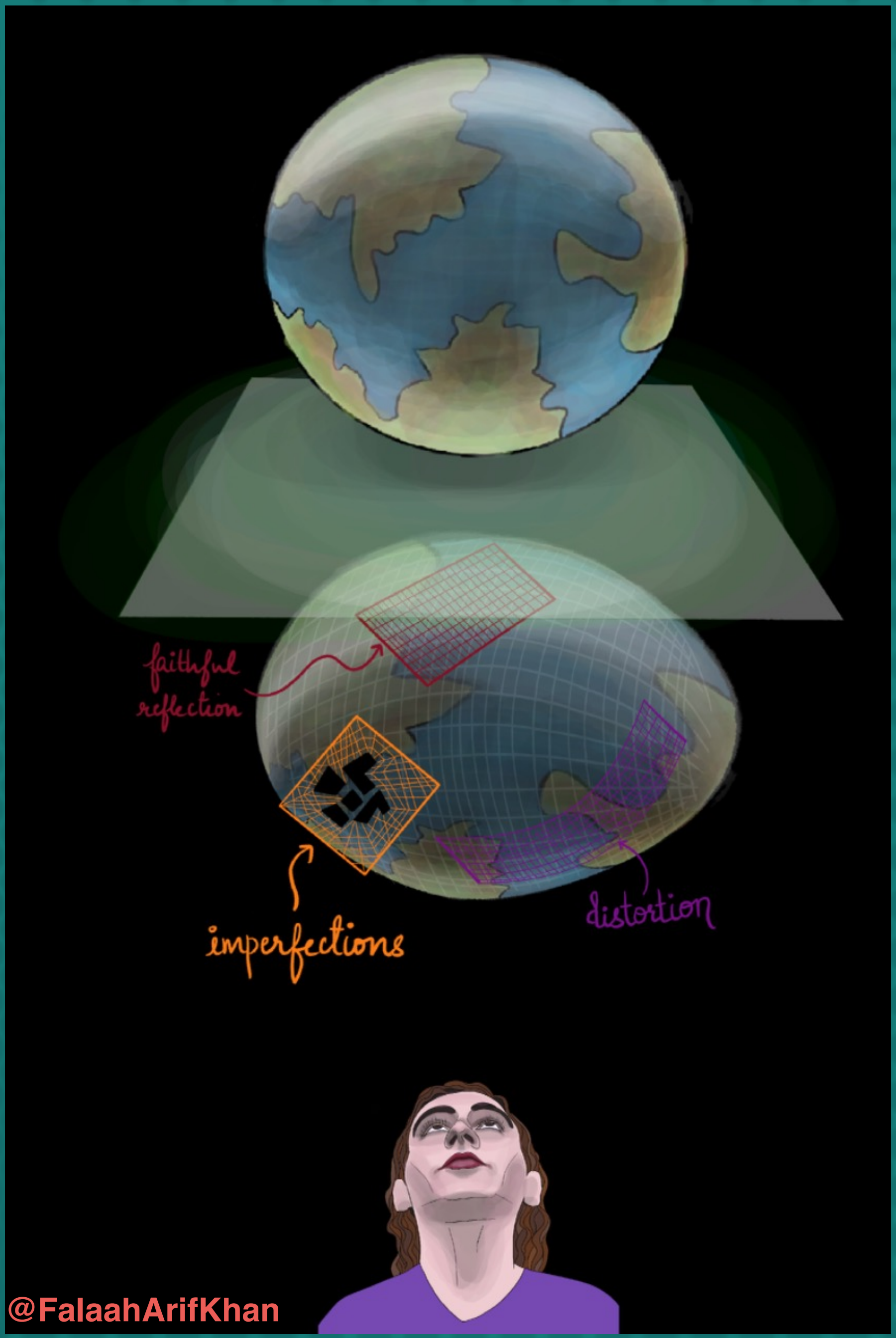
Societal

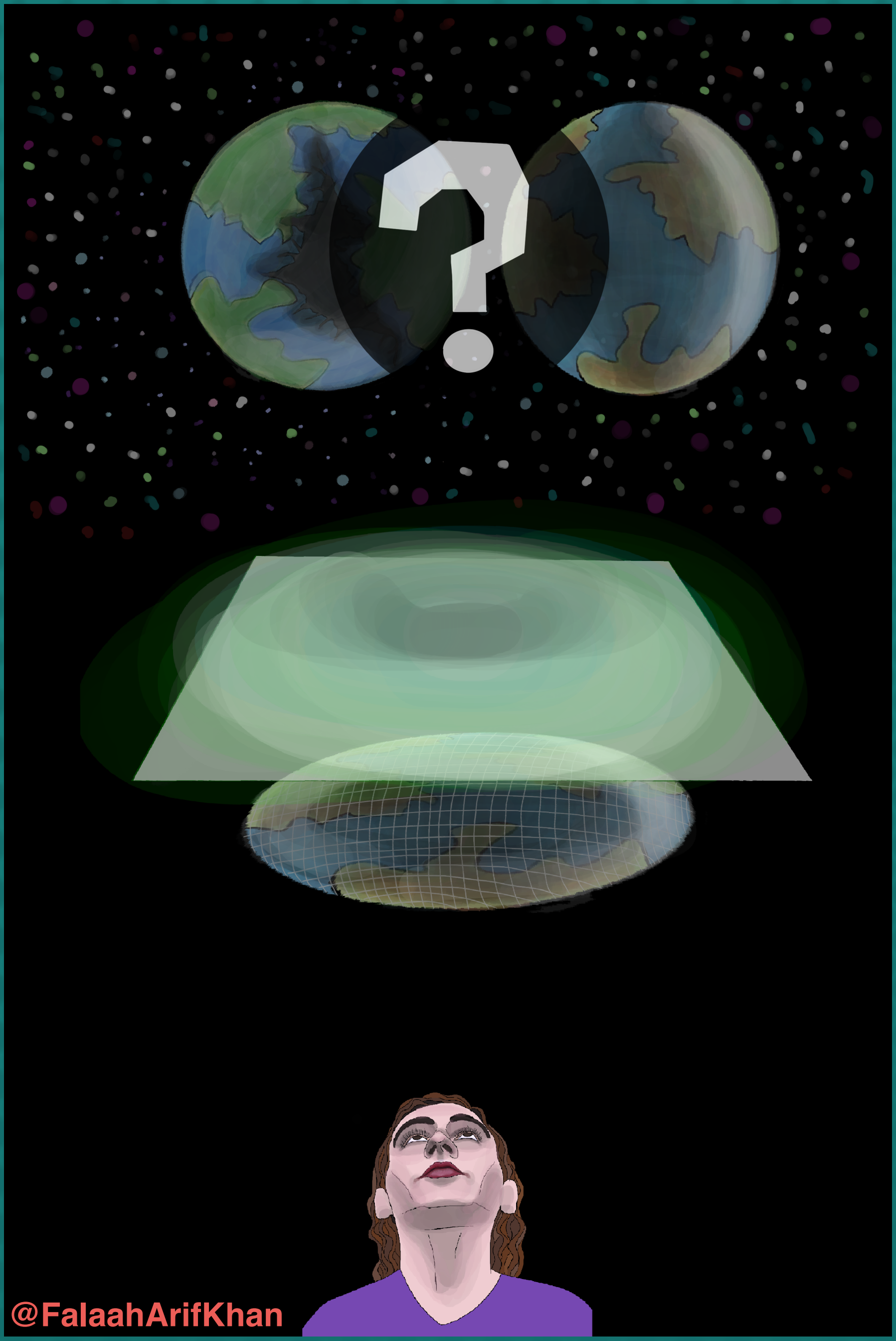
data does not represent the world correctly

Data, a reflection of the world

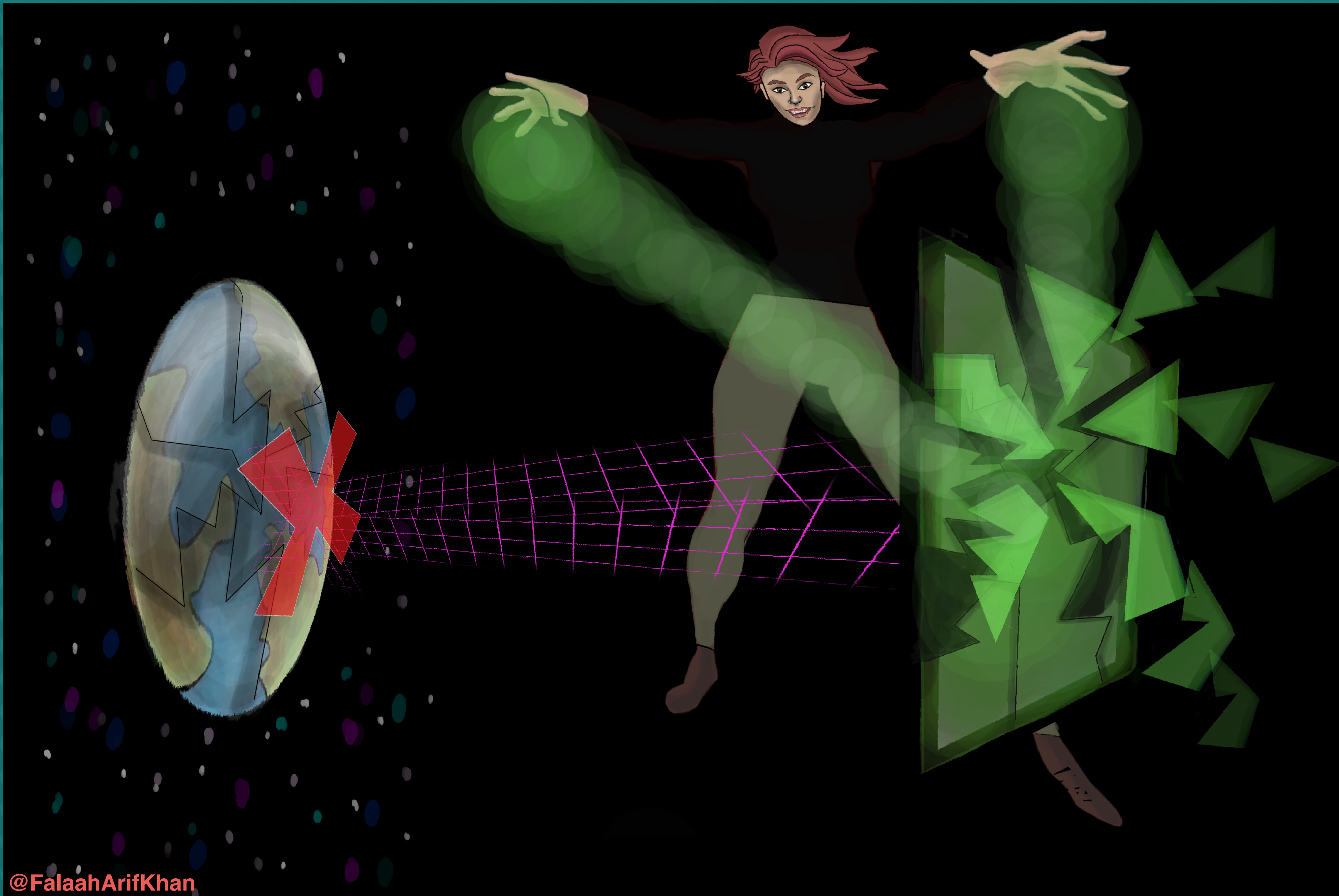


Data, a reflection of the world





Changing the reflection won’t change the world



bias can lead to discrimination

The evils of discrimination

Disparate treatment

is the illegal practice of treating an entity, such as a job applicant or an 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**.

fairness in classification

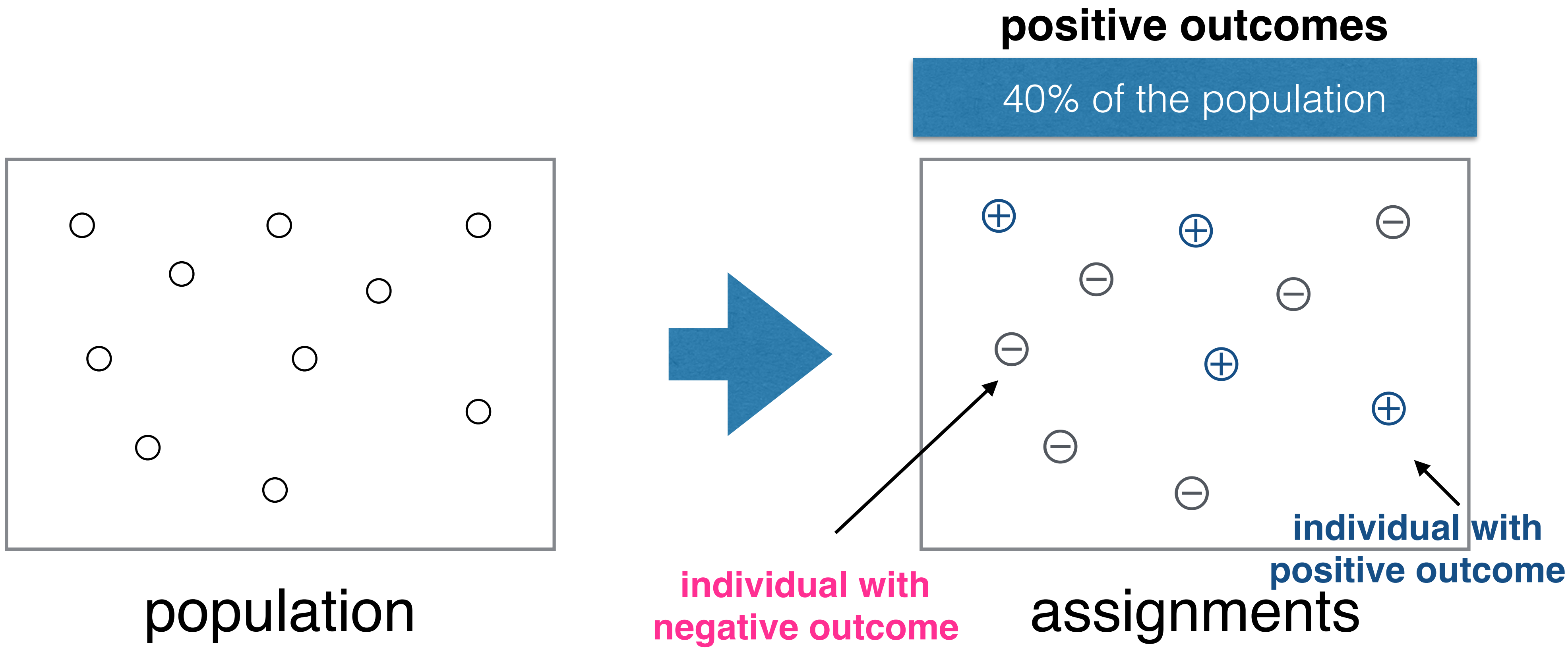
Vendors and outcomes

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

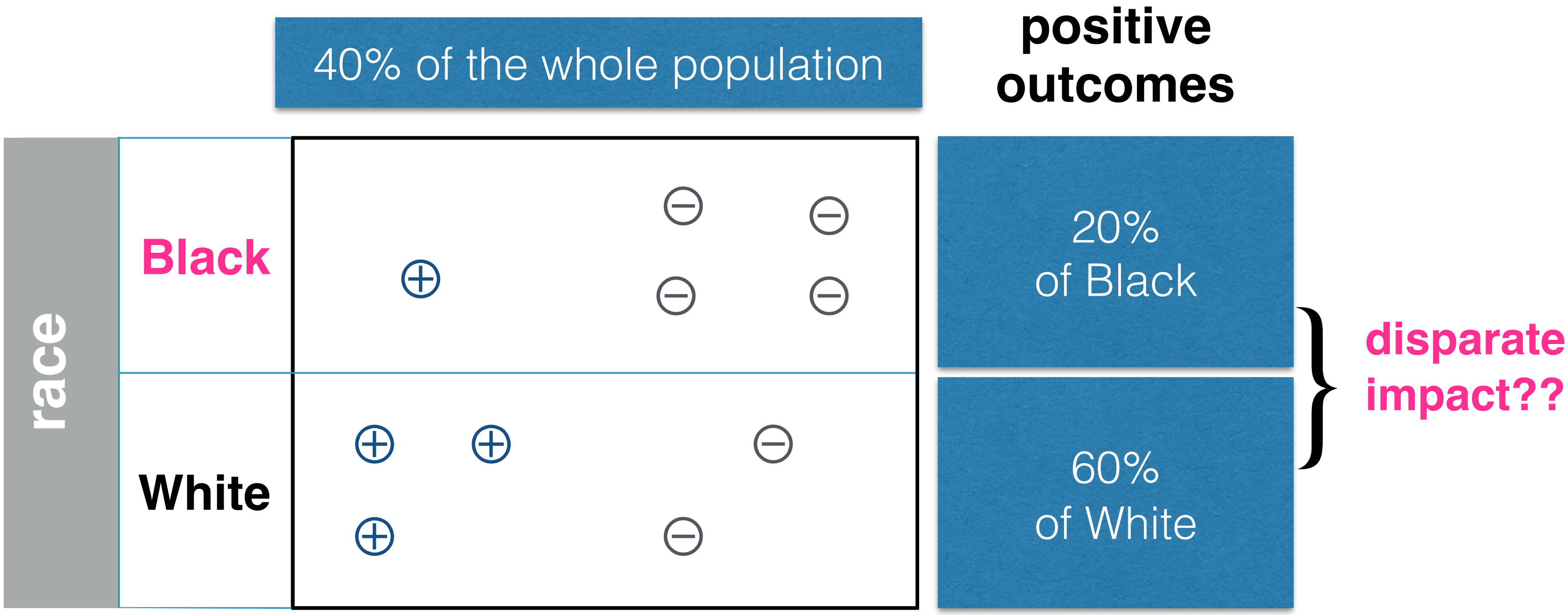
Fairness in classification

Fairness in classification is concerned with how outcomes are assigned to a population



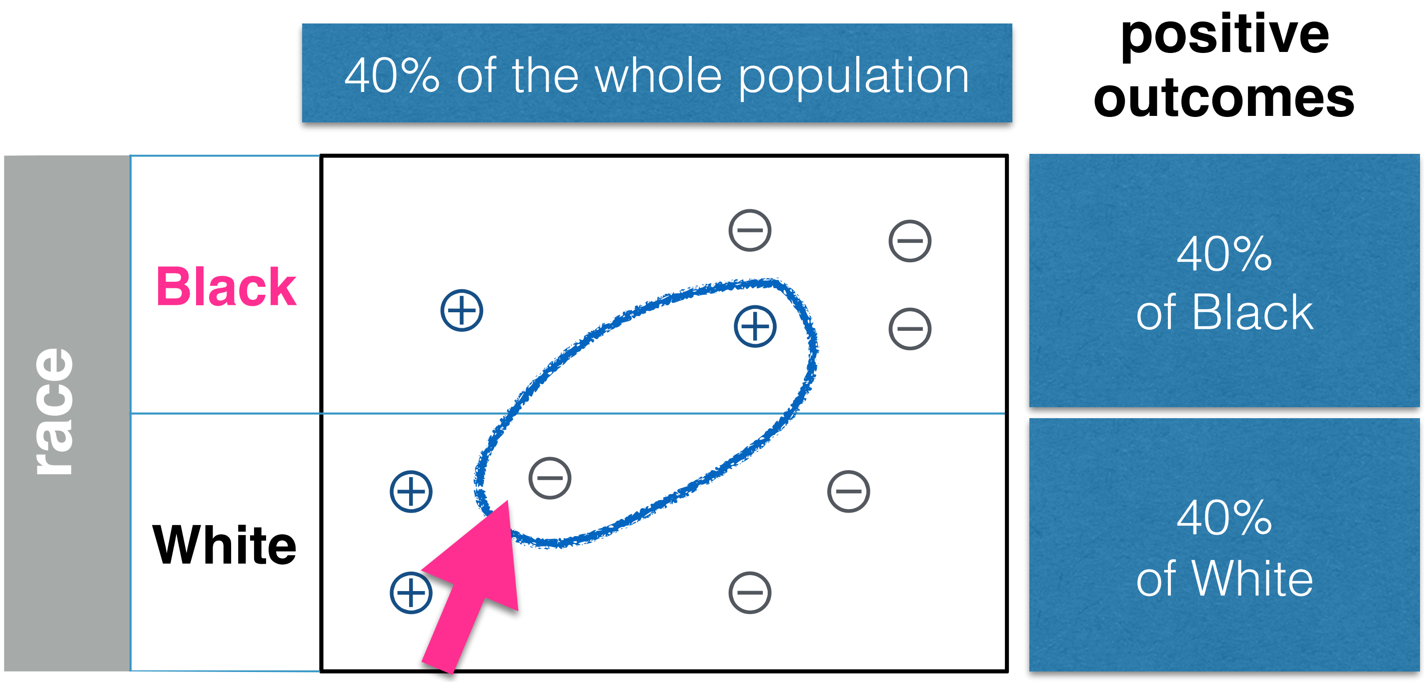
Fairness in classification

Sub-populations may be treated differently



Fairness in classification

Sub-populations may be treated differently



Fairness in classification

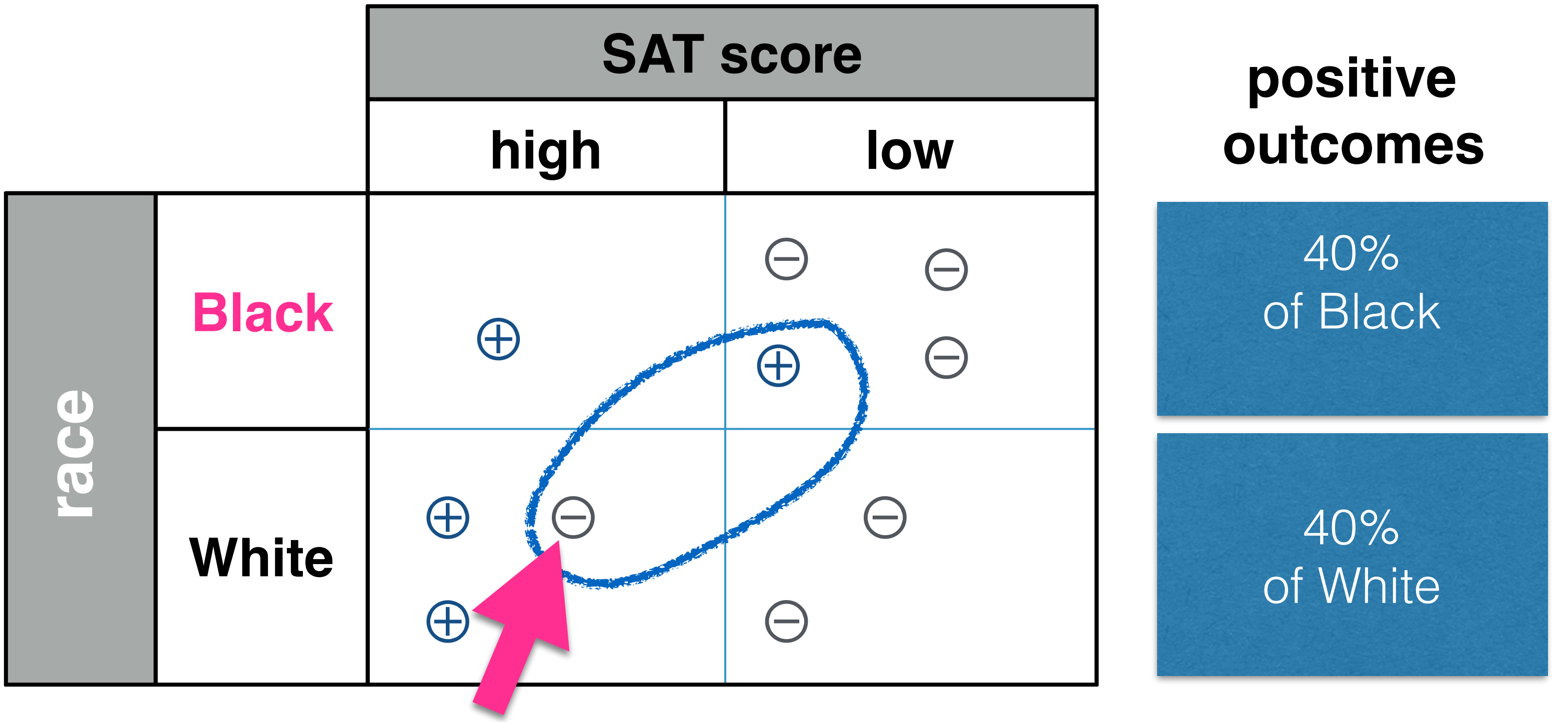
race		SAT score: high	SAT score: low
race	**Black**	⊕	⊖ ⊖ ⊖ ⊖
race	**White**	⊕ ⊕ ⊕	⊖ ⊖

positive outcomes

20%
of Black

60%
of White

Swapping outcomes



Two families of fairness measures

Group fairness (*here* statistical parity)

demographics of the individuals receiving any outcome - positive or negative - should be the same as demographics of the underlying population

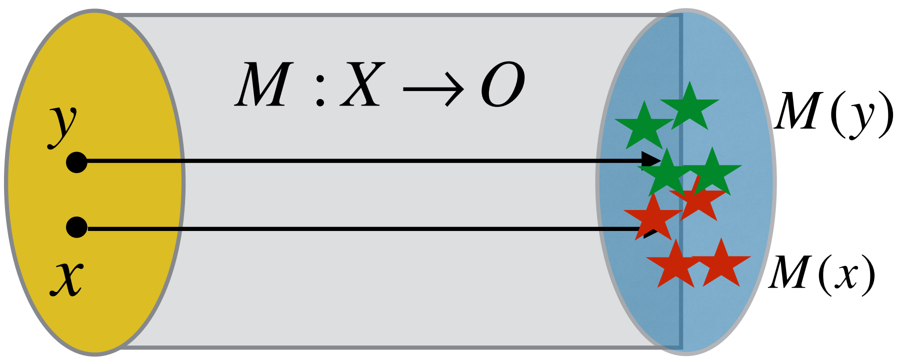
Individual fairness

any two individuals who are similar **with respect to a task** should receive similar outcomes fairness through awareness

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

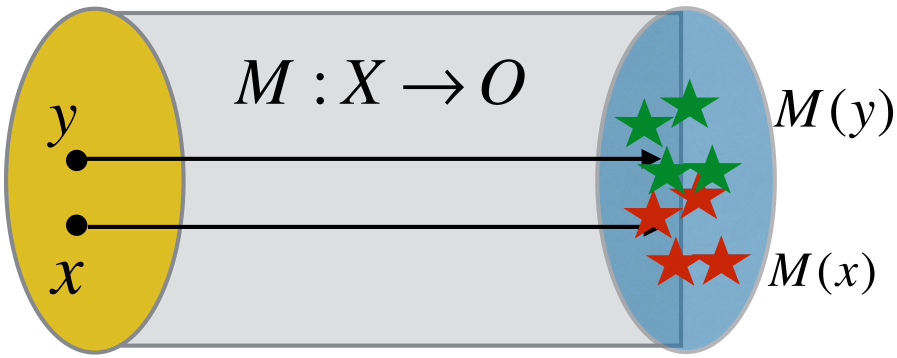
A task-specific similarity metric is given $d(x,y)$

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

Fairness through a Lipschitz mapping

[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 $d(x,y)$

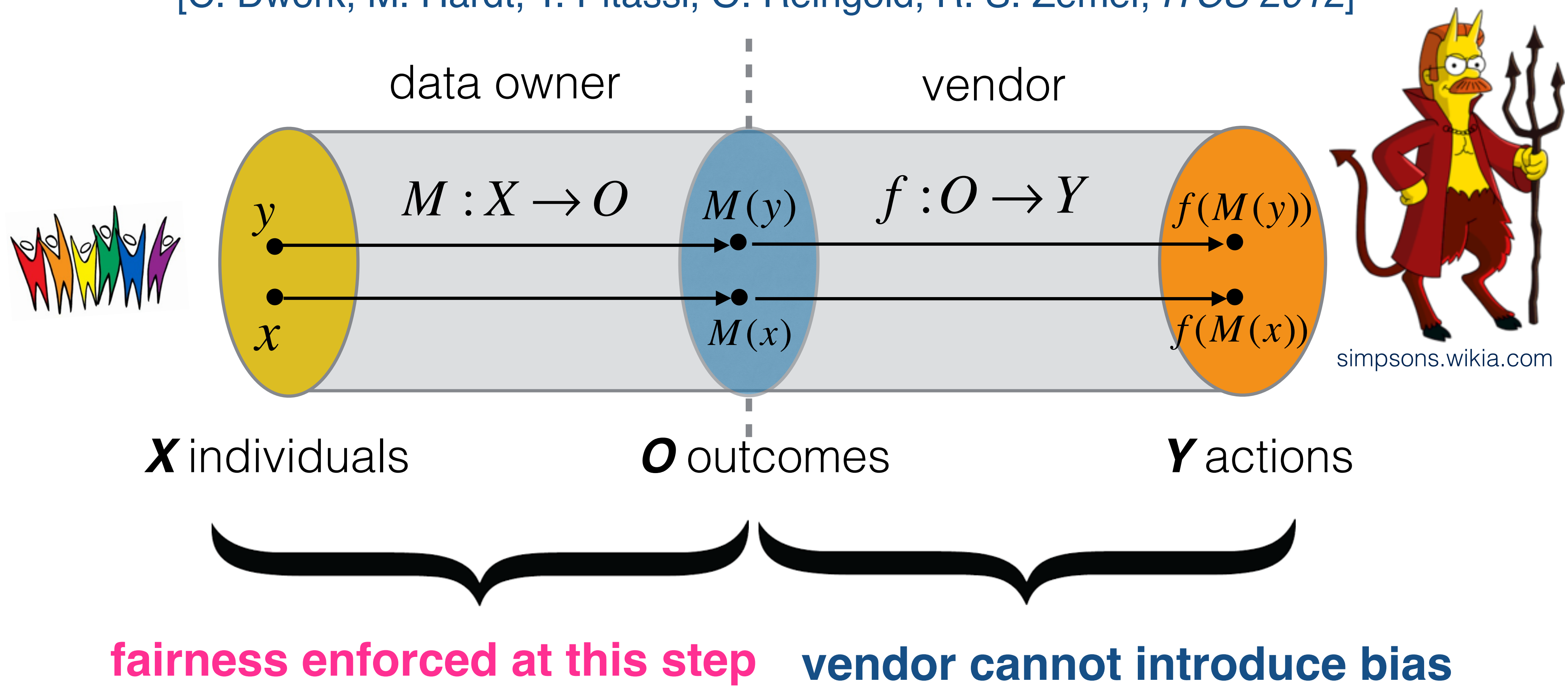
M is a Lipschitz mapping if $\forall x,y \in X \quad \| M(x), M(y) \| \leq d(x,y)$

close individuals map to close distributions

there always exists a Lipschitz mapping - which?

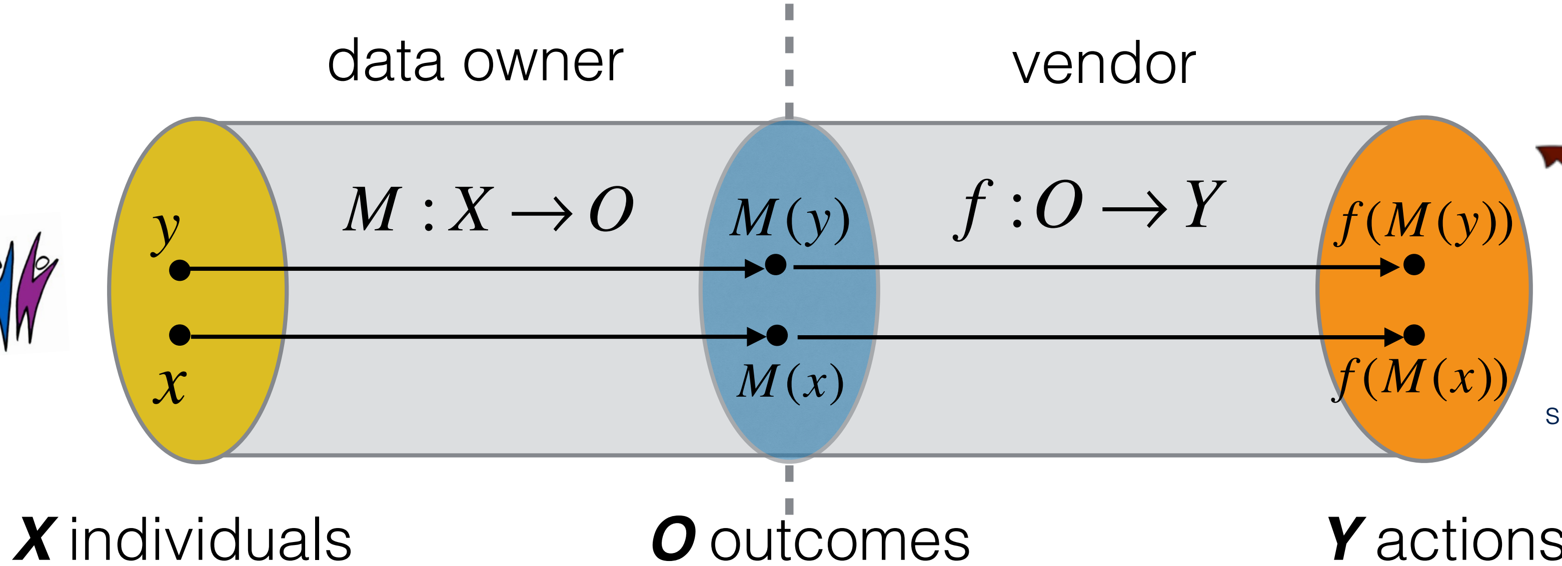
Fairness through a Lipschitz mapping

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



Fairness through a Lipschitz mapping

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

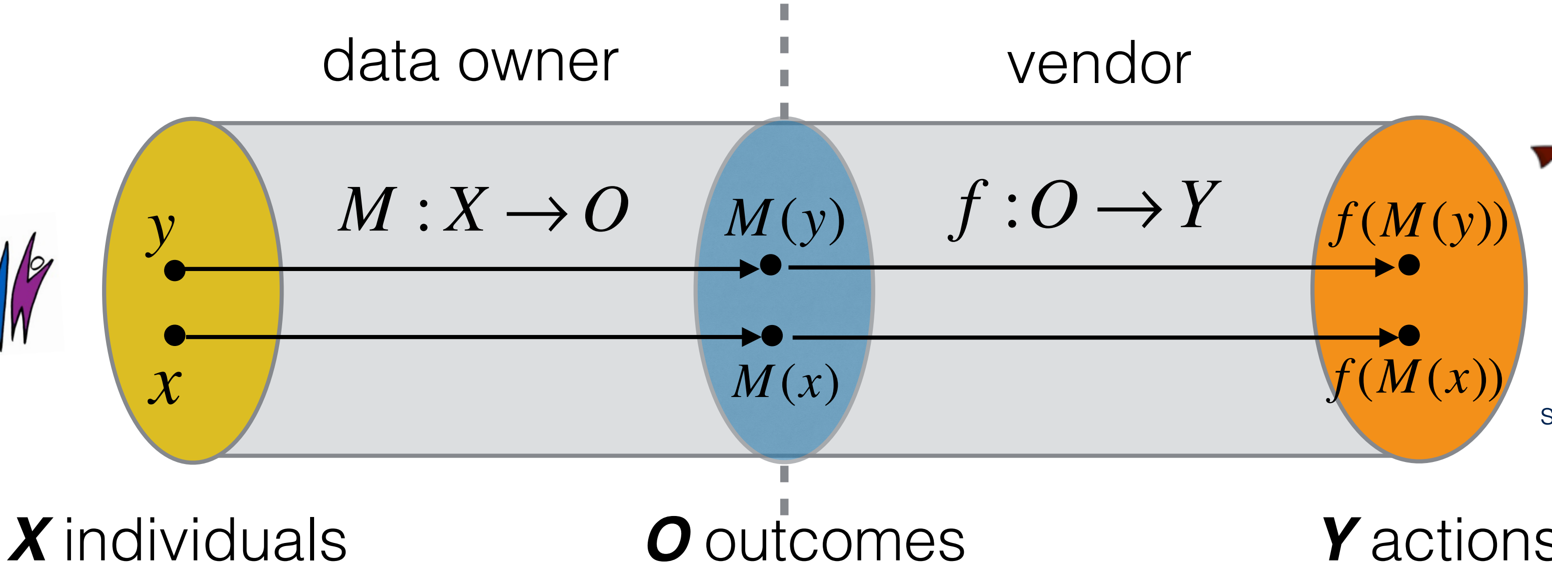


simpsons.wikia.com

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.

Fairness through a Lipschitz mapping

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



simpsons.wikia.com

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*]

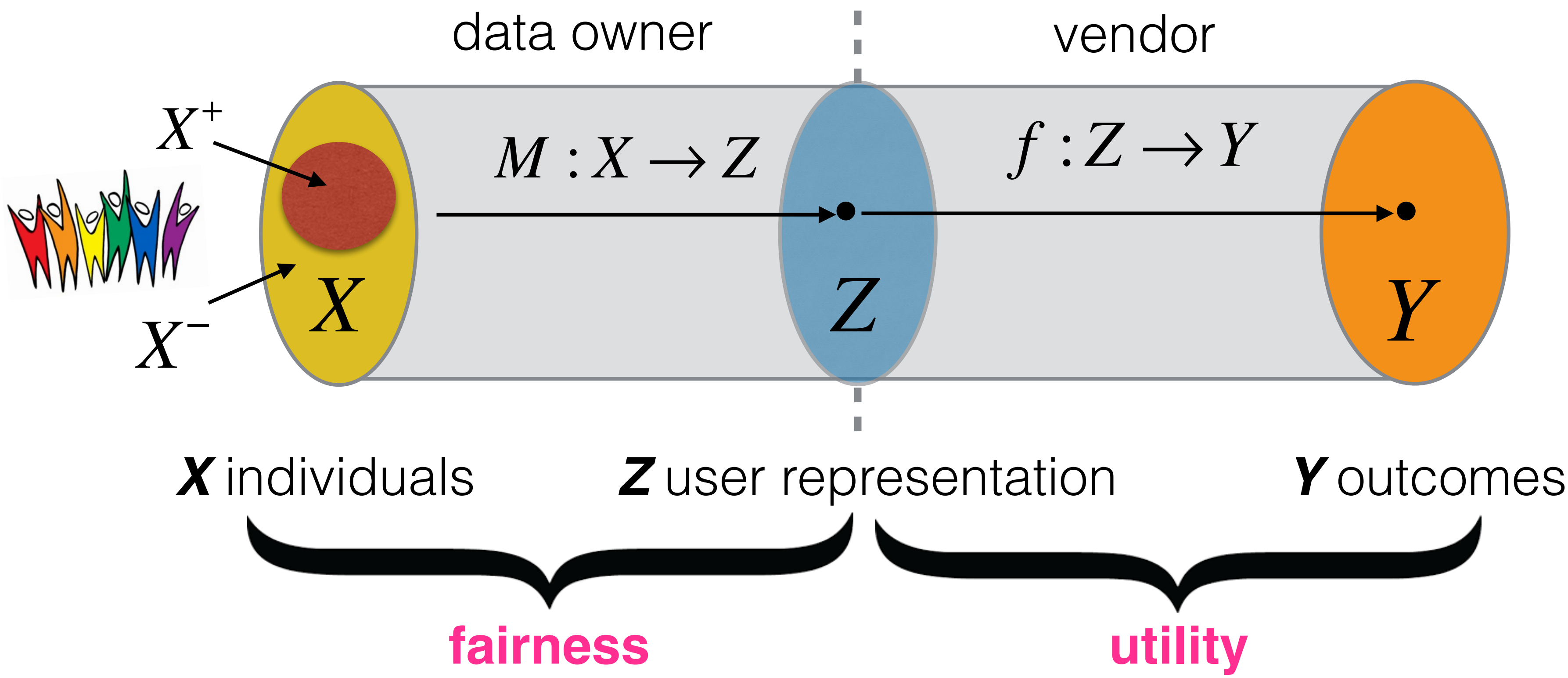
"**Equality of opportunity** defines an important welfare criterion in political philosophy and policy analysis. **Philosophers** define equality of opportunity as the requirement that an individual's well being be independent of his or her irrelevant characteristics. **The difference among philosophers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

learning fair representations

Learning fair representations

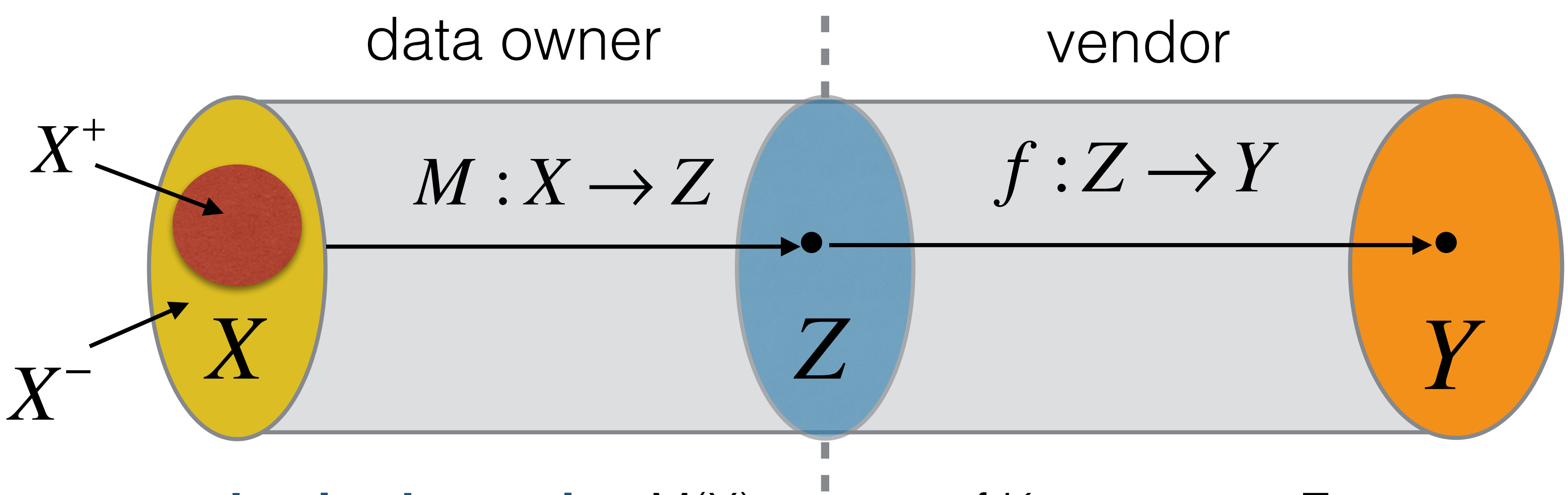
[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]



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 $\quad P(Z \mid S=0) = P(Z \mid S=1)$

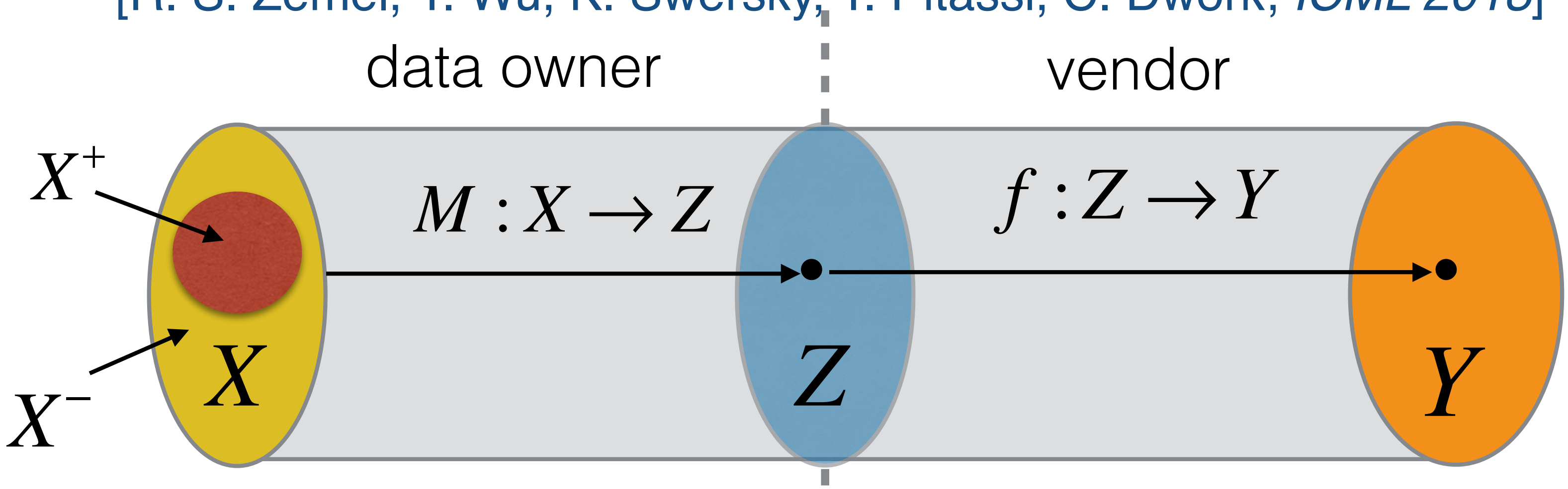
M(X) should preserve other information so that vendor can maximize utility

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

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

Fairness and utility

[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**

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

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

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

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

does this make sense?

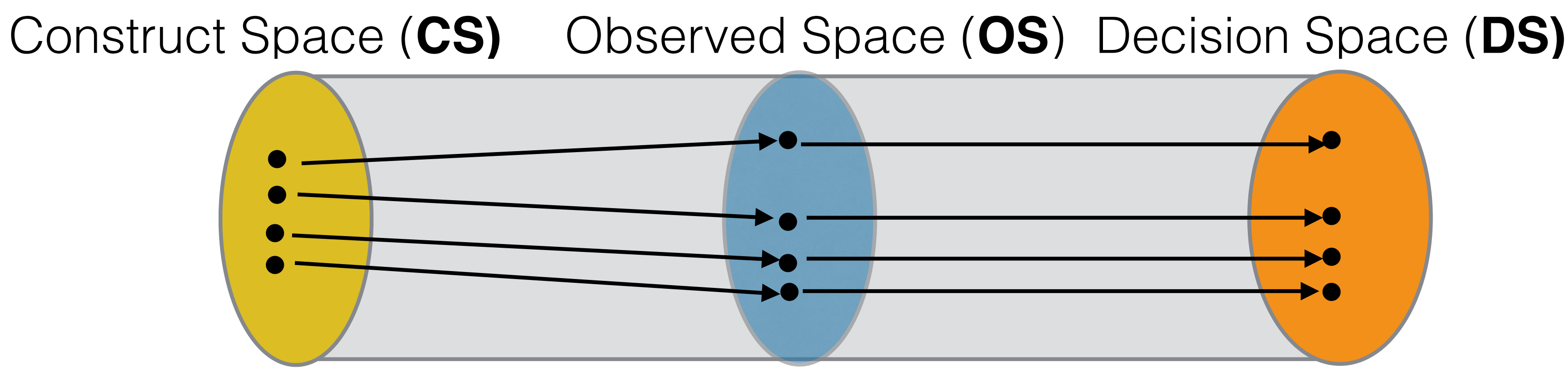
on the (im)possibility of fairness

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



On the (im)possibility of fairness

[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	

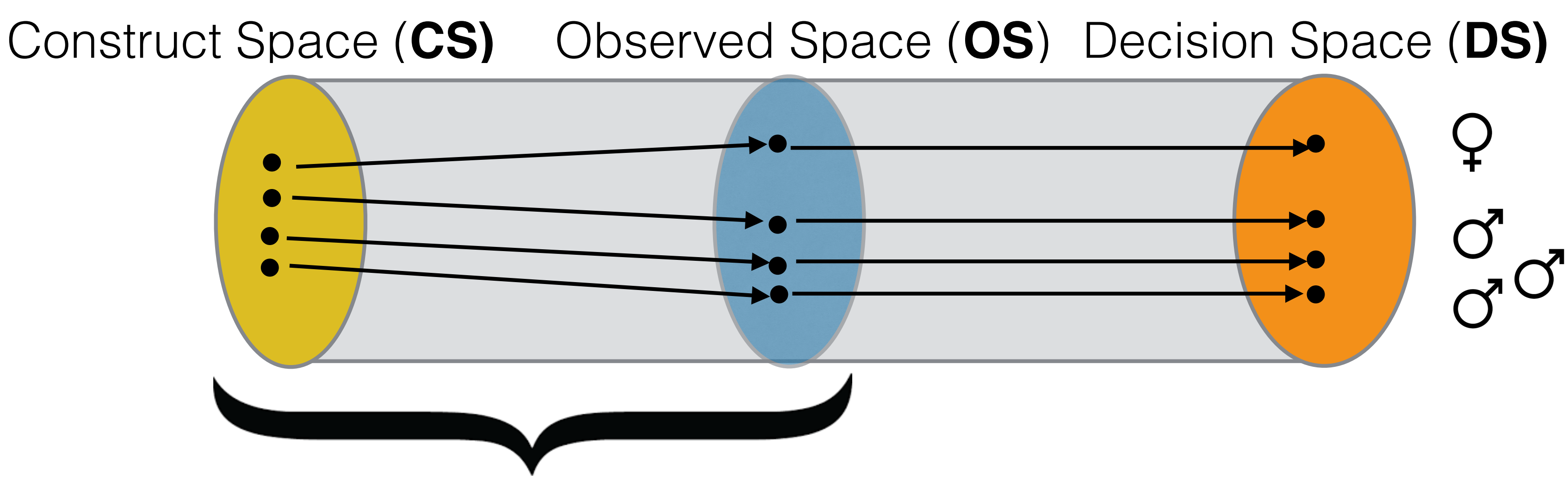
define fairness through properties of mappings

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**.

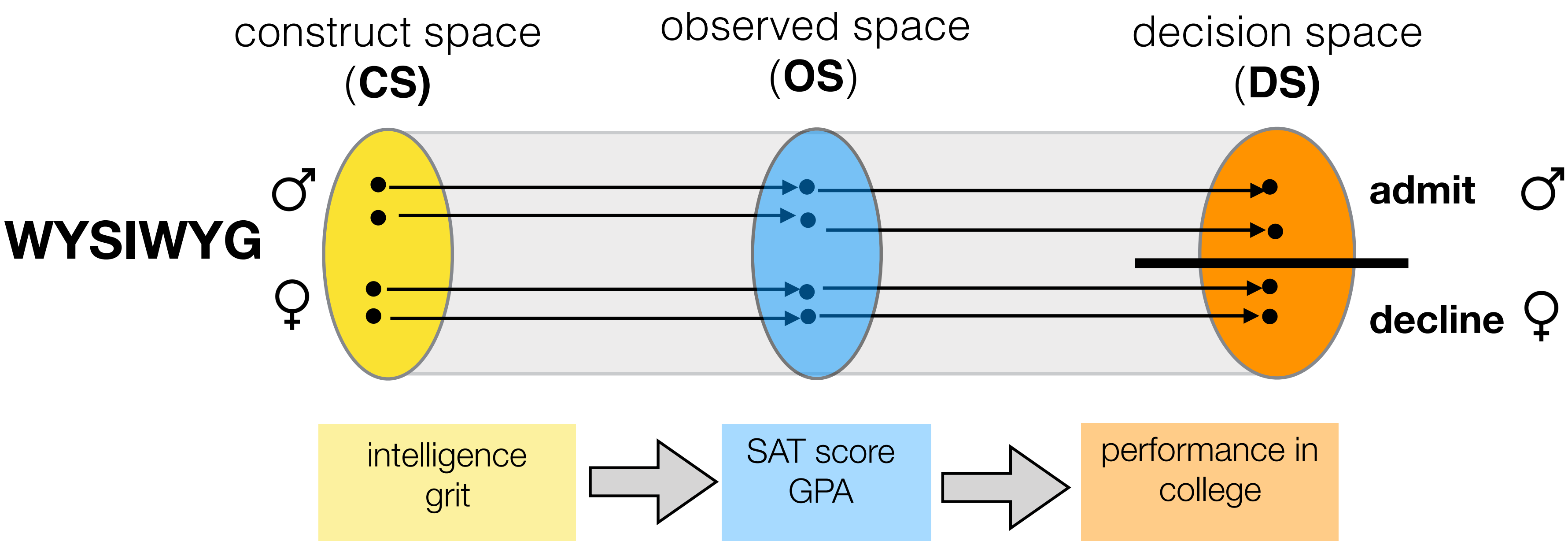
$$f : CS \rightarrow DS \qquad\qquad d_{CS}(x,y) < \varepsilon \Rightarrow d_{DS}(f(x), f(y)) < \varepsilon'$$



let's focus on this portion

WYSWYG

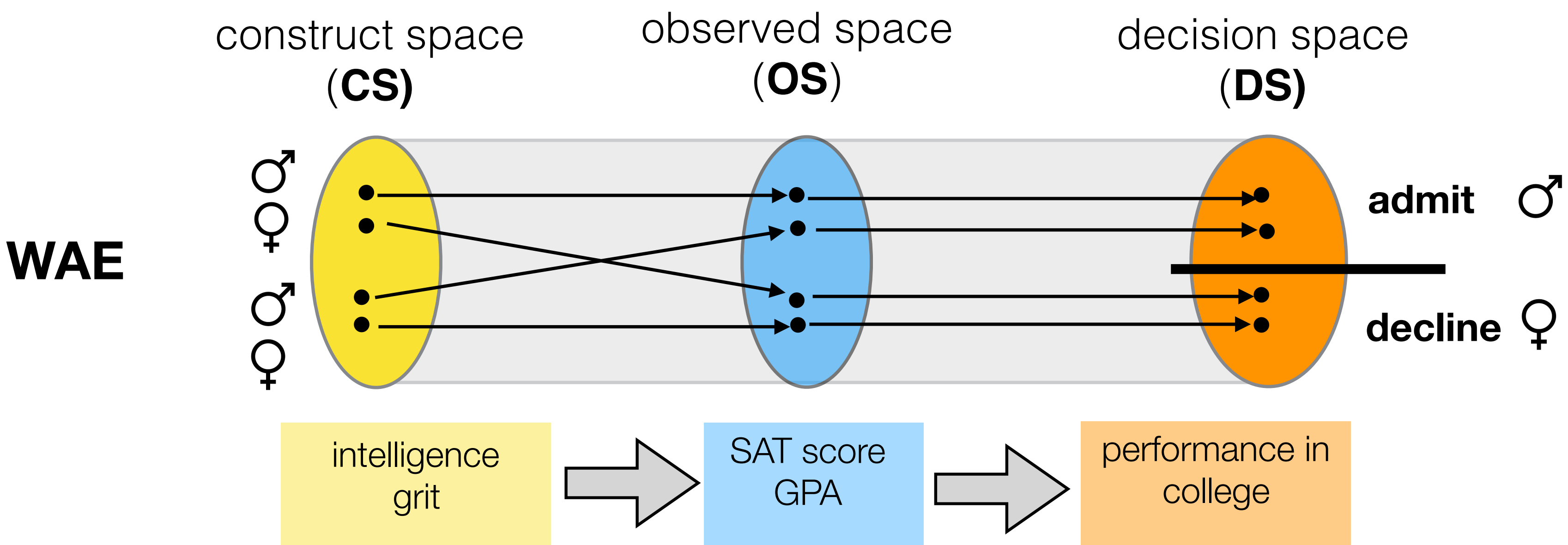
[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]



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.**

WAE

[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?

Fairness and worldviews

individual fairness

equality

group fairness

equity

The evils of discrimination

Disparate treatment

is the illegal practice of treating an entity, such as a job applicant or an 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**.

Ricci v. DeStefano (2009)

Supreme Court Finds Bias Against White Firefighters

By ADAM LIPTAK JUNE 29, 2009

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.	

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 & diversity in selection

state beliefs & assumptions

cannot fully automate responsibility!

Goals and trade-offs

Goals

diversity: pick k=4 candidates, including 2 of each gender, and at least one per race

utility: maximize the total score of selected candidates

score = 373

	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

Problem

fairness: picked the best White and male candidates (A, B) but did not pick the best Black (E, F), Asian (I, J), or female (C, D) candidates

Beliefs

scores are more informative within a group than across groups - **effort is relative to circumstance**

it is important to **reward effort**

@FalaahArifKhan

[Yang, Gkatzelis, Stoyanovich (2019)]

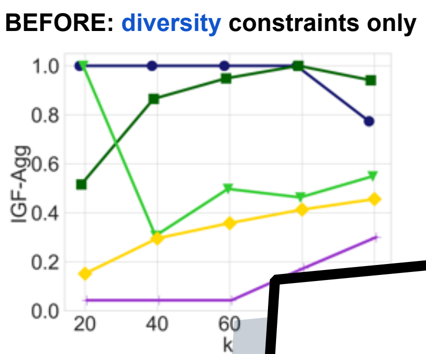
From beliefs to interventions

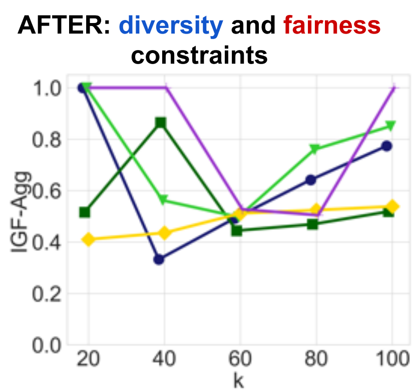
Fairness for female candidates **83 / 95 = 0.91**

C	D	G	H	K	L
95	95	90	86	83	83

highest-scoring skipped

lowest-scoring selected





Beliefs

scores are more informative within a group than across groups - **effort is relative to circumstance**

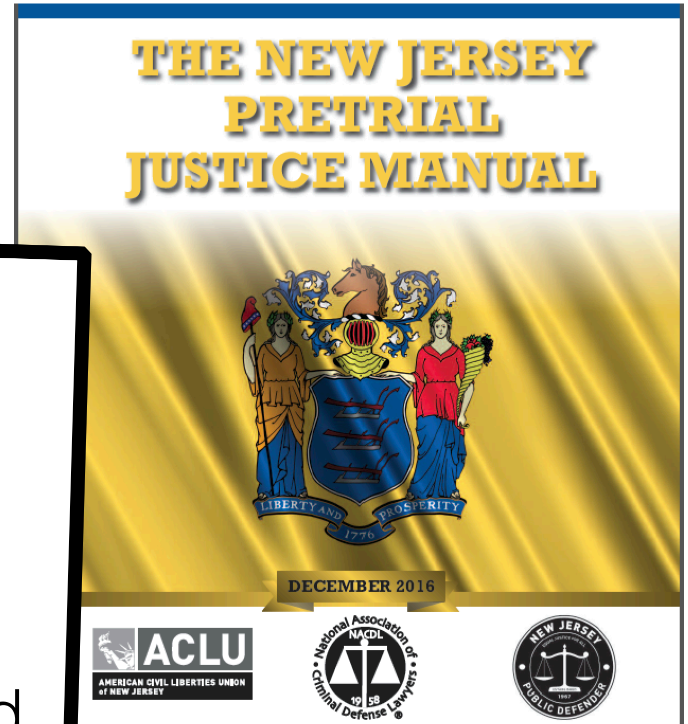
it is important to **reward effort**

[Yang, Gkatzelis, Stoyanovich (2019)]

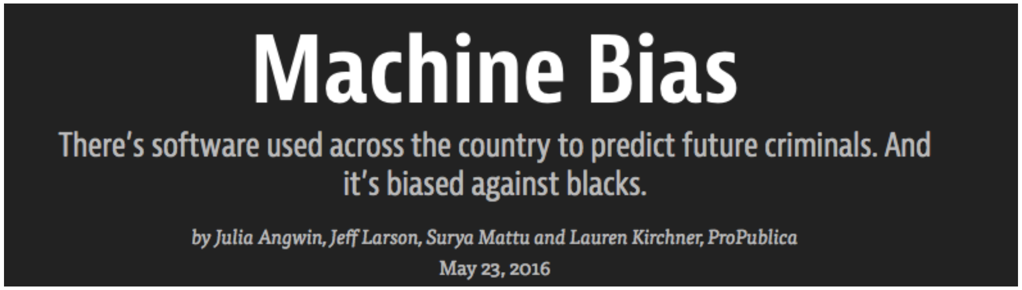
fairness in risk assessment

New Jersey bail reform

Switching from a system based solely on instinct and experience […] to one in which judges have access to **scientific, objective risk assessment** tools could further the criminal justice system's central goals of increasing public safety, reducing crime, and making the most effective, fair, and efficient use of public resources.



ProPublica's COMPAS study



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

A commercial tool **COMPAS** **May 2016** 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

ProPublica's COMPAS study

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.

May 2016

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** in risk assessment is concerned with **how different kinds of error are distributed among sub-populations**

Calibration

positive outcomes: do recidivate

	risk score		
	0.2	**0.6**	**0.8**
White	⊖ ⊖ ⊖ ⊖ ⊕	⊖ ⊖ ⊕ ⊕ ⊕	⊖ ⊕ ⊕ ⊕ ⊕
Black	⊖ ⊖ ⊖ ⊖ ⊕	⊖ ⊖ ⊖ ⊖ ⊕ ⊕ ⊕ ⊕ ⊕ ⊕	⊖ ⊕ ⊕ ⊕ ⊕

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

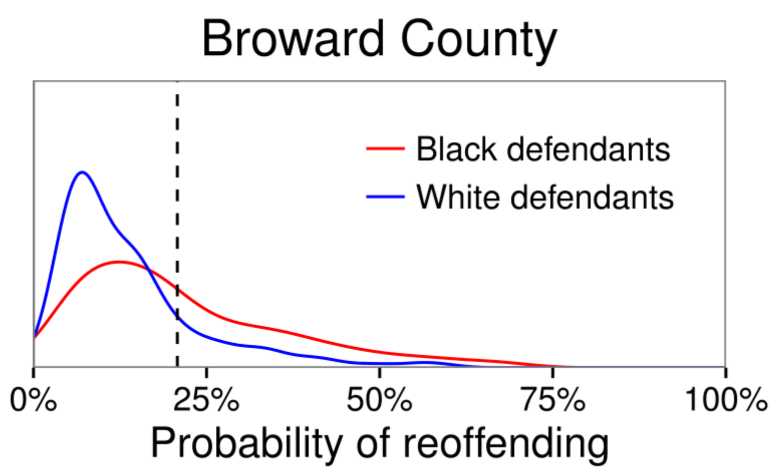
0.6 means 0.6 for any defendant - likelihood of recidivism

why do we want calibration?

COMPAS as a predictive instrument

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*]

An impossibility result

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

...efendants

	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.)

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

A more general statement: Balance

- 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

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

Desiderata, re-stated

- For each group, a $\mathbf{v_b}$ fraction in each bin **b** is positive
- Average score of positive class same across groups
- Average score of negative class same across groups

can we have all these properties?

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

Achievable only in trivial cases

- 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

a negative result, need tradeoffs

proof sketched out in (starts 12 min in)

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

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

Fairness for whom?

Decision-maker: of those labeled low-risk, how many will recidivate?

Defendant: how likely will I be incorrectly labeled high-risk?

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

based on a slide by Arvind Narayanan

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 = ?