Responsible Data Science

Algorithmic Fairness

February 3rd & February 10th, 2025

Instructors:
Lucas Rosenblatt and Dr. Emily Black

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai

Office Hours
Today:
12:40-1:40

Logistics Review

Assignments and grading

Grading: homeworks - 10% x 3 = 30%
project - 25%
final exam - 25%
labs - 10%
quizzes - 10%

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.

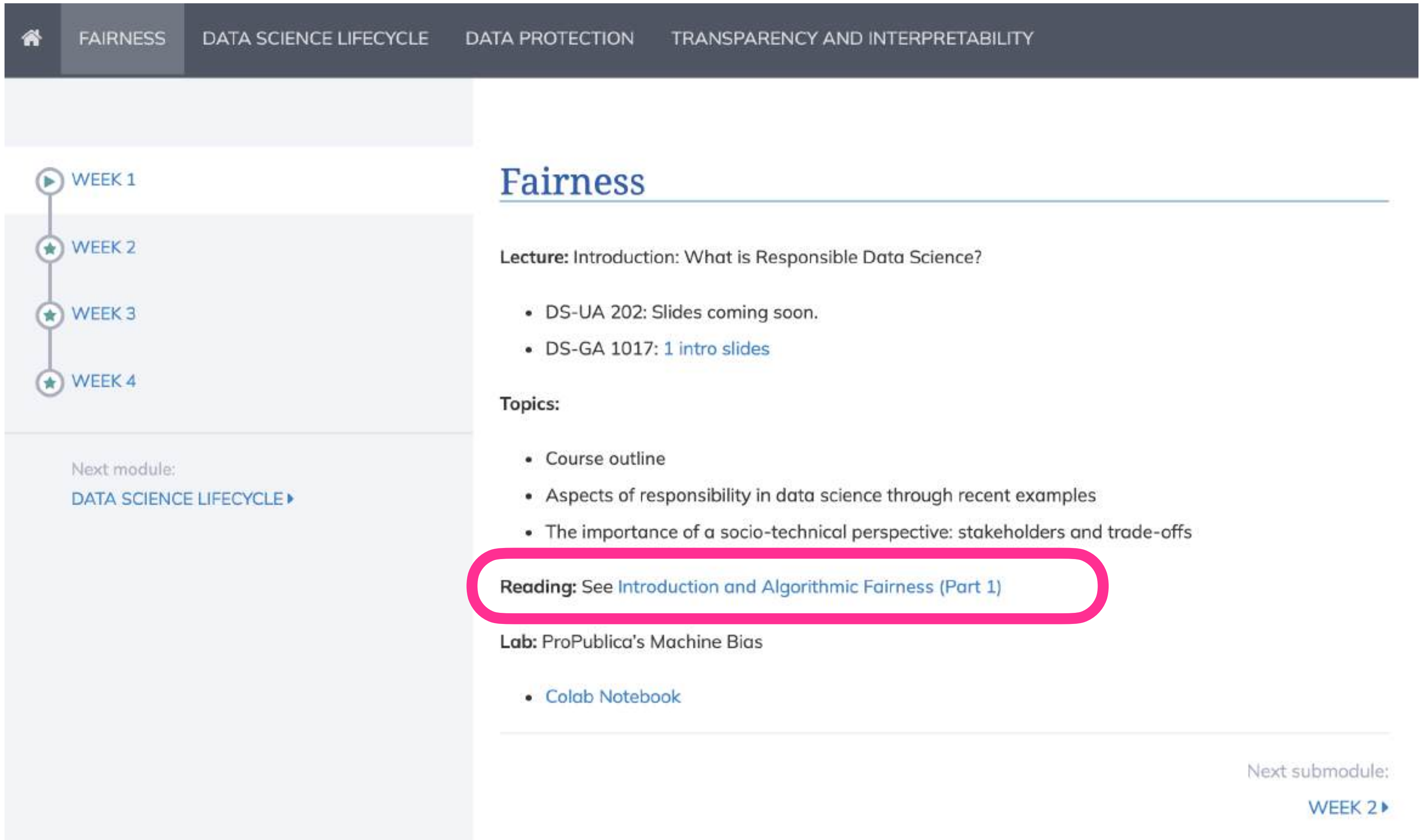
Assignment schedule will be posted to Bright Space (under Course information), subject to change.

Assignments submitted through Gradescope.

Labs + Project submitted through Brightspace

Where to find information

Website: **https://dataresponsibly.github.io/rds/** slides, reading, labs



Fairness

Lecture: Introduction: What is Responsible Data Science?

- DS-UA 202: Slides coming soon.
- DS-GA 1017: 1 intro slides

Topics:

- Course outline
- Aspects of responsibility in data science through recent examples
- The importance of a socio-technical perspective: stakeholders and trade-offs

Reading: See Introduction and Algorithmic Fairness (Part 1)

Lab: ProPublica's Machine Bias

- Colab Notebook

Bright Space: everything assignment-related, Zoom links for lectures and labs, announcements. **Piazza:** discussion board. **Gradescope:** Assignment Submission.

RDS course overview

What is RDS?

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

A technical course, with content drawn from:

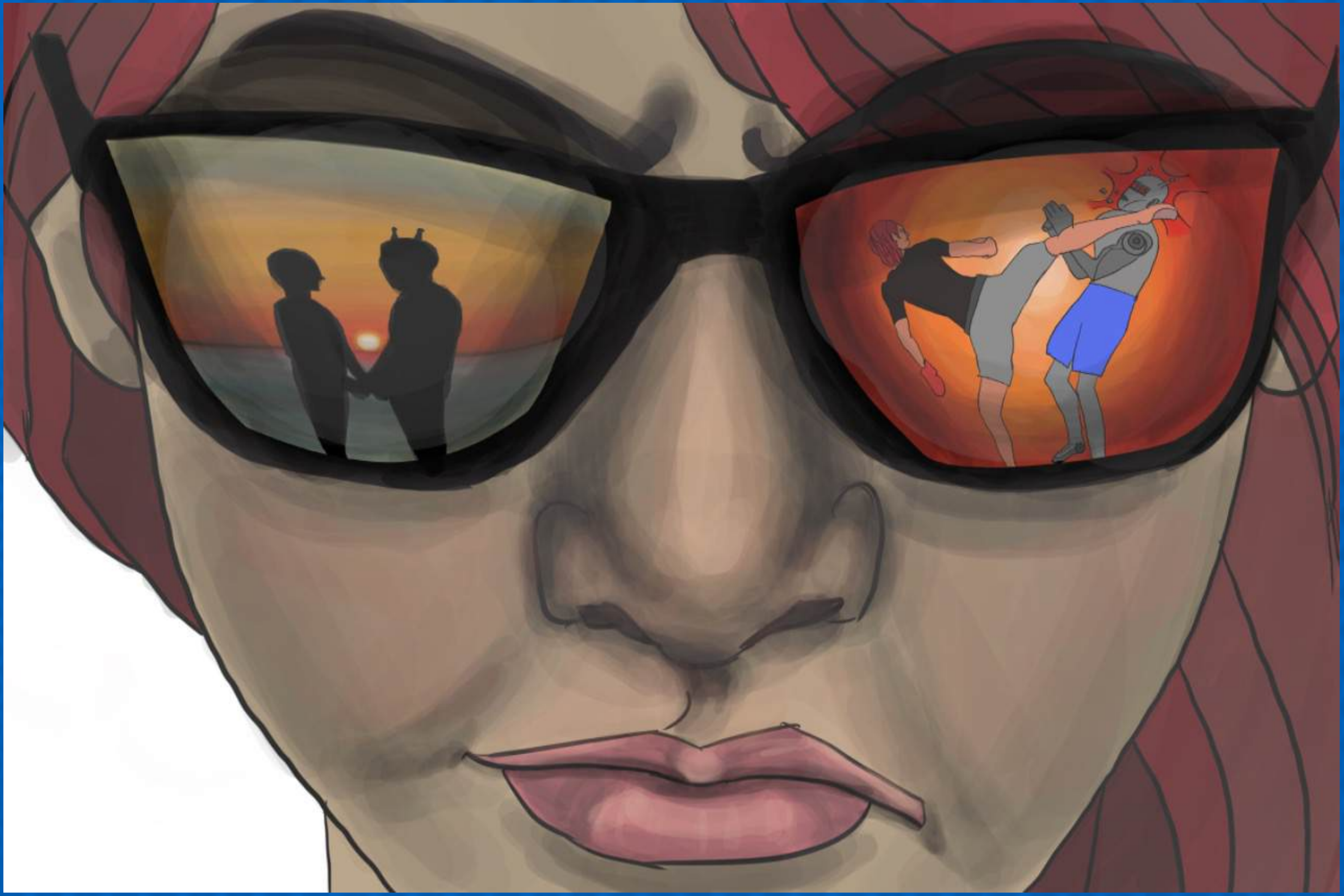
1. fairness, accountability and transparency
2. machine learning
3. privacy & data protection

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

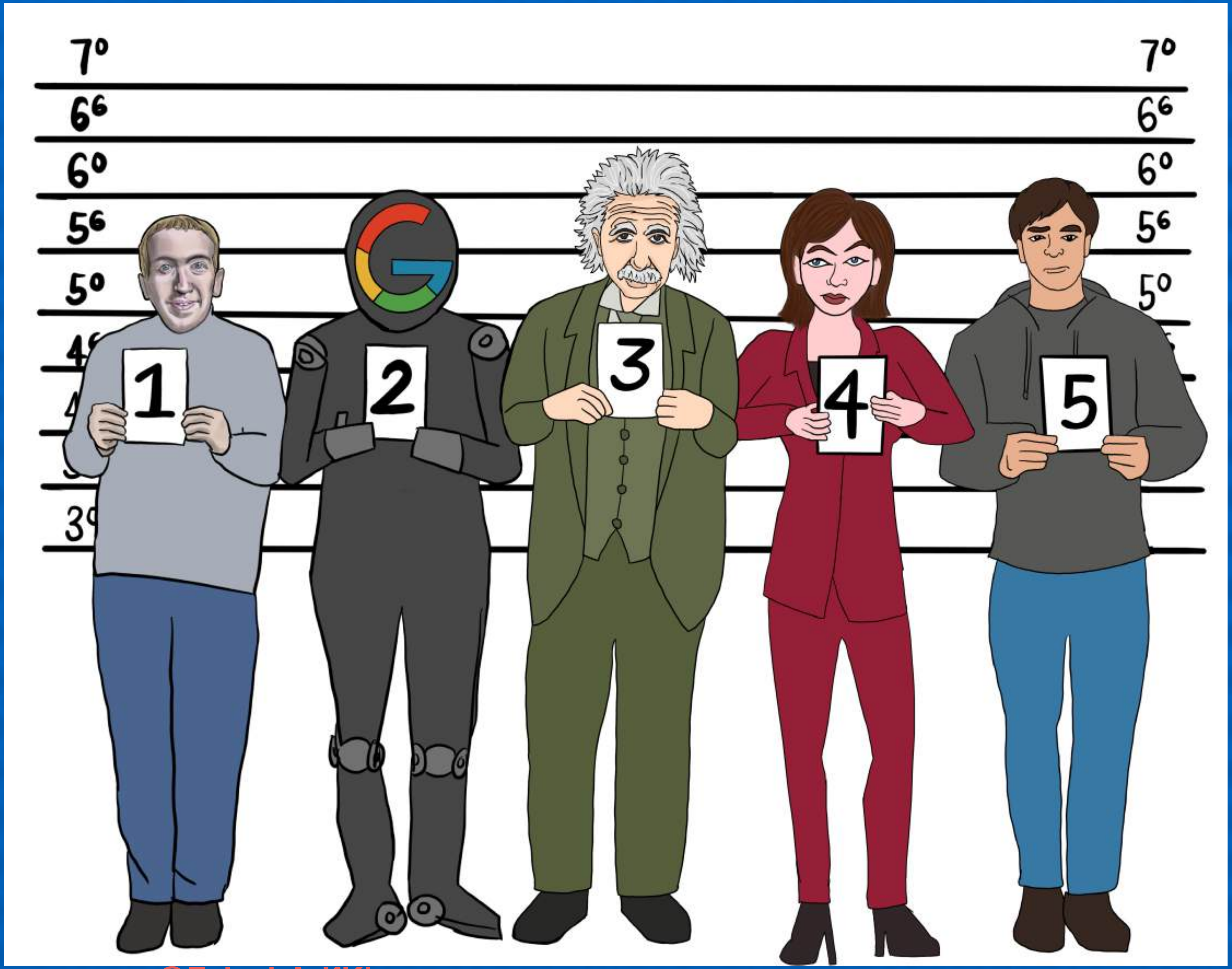
Bottom line: we will learn that many of the problems are **socio-technical**, and so cannot be "solved" with technology alone.

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

Nuance, please!

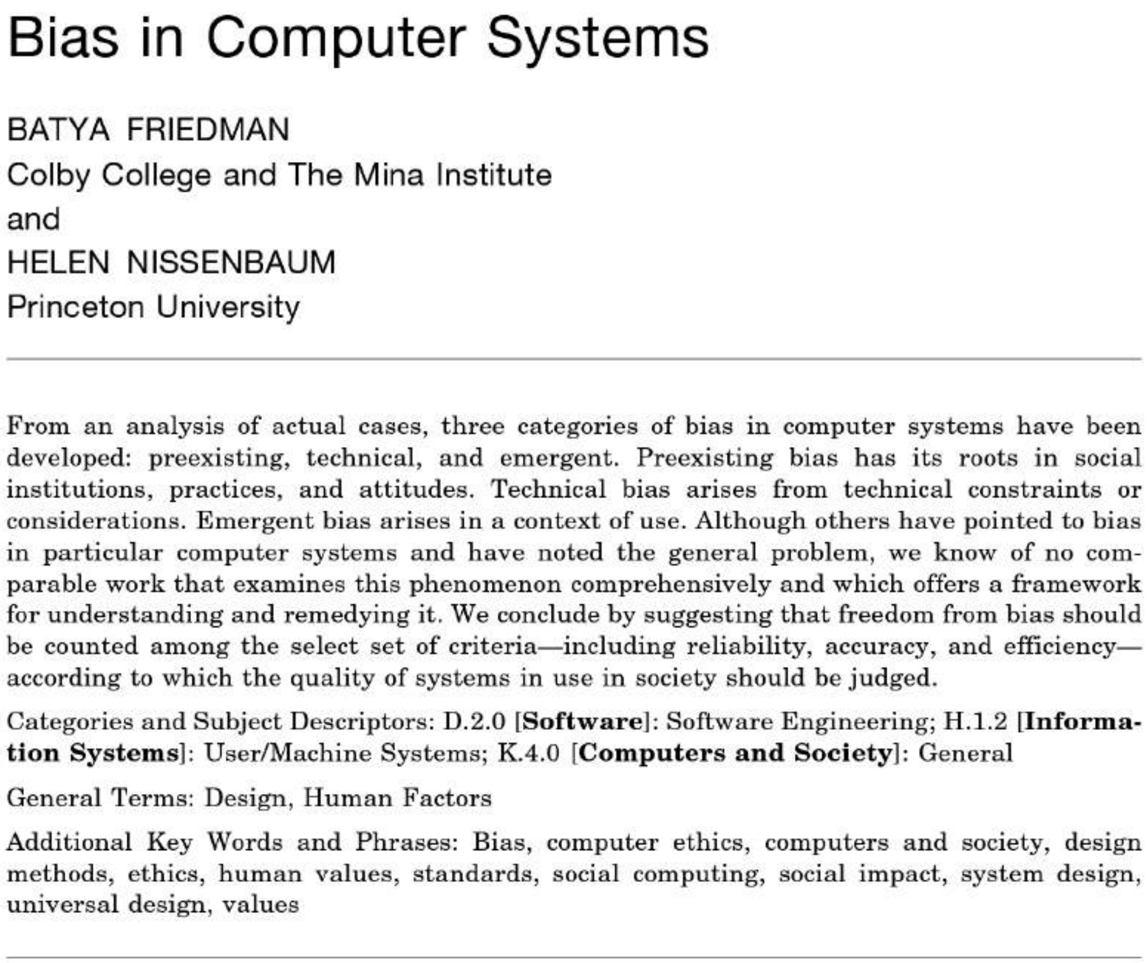


We all are responsible



@FalaahArifKhan

Reading: Algorithmic bias



Bias in Computer Systems

BATYA FRIEDMAN
Colby College and The Mina Institute
and
HELEN NISSENBAUM
Princeton University

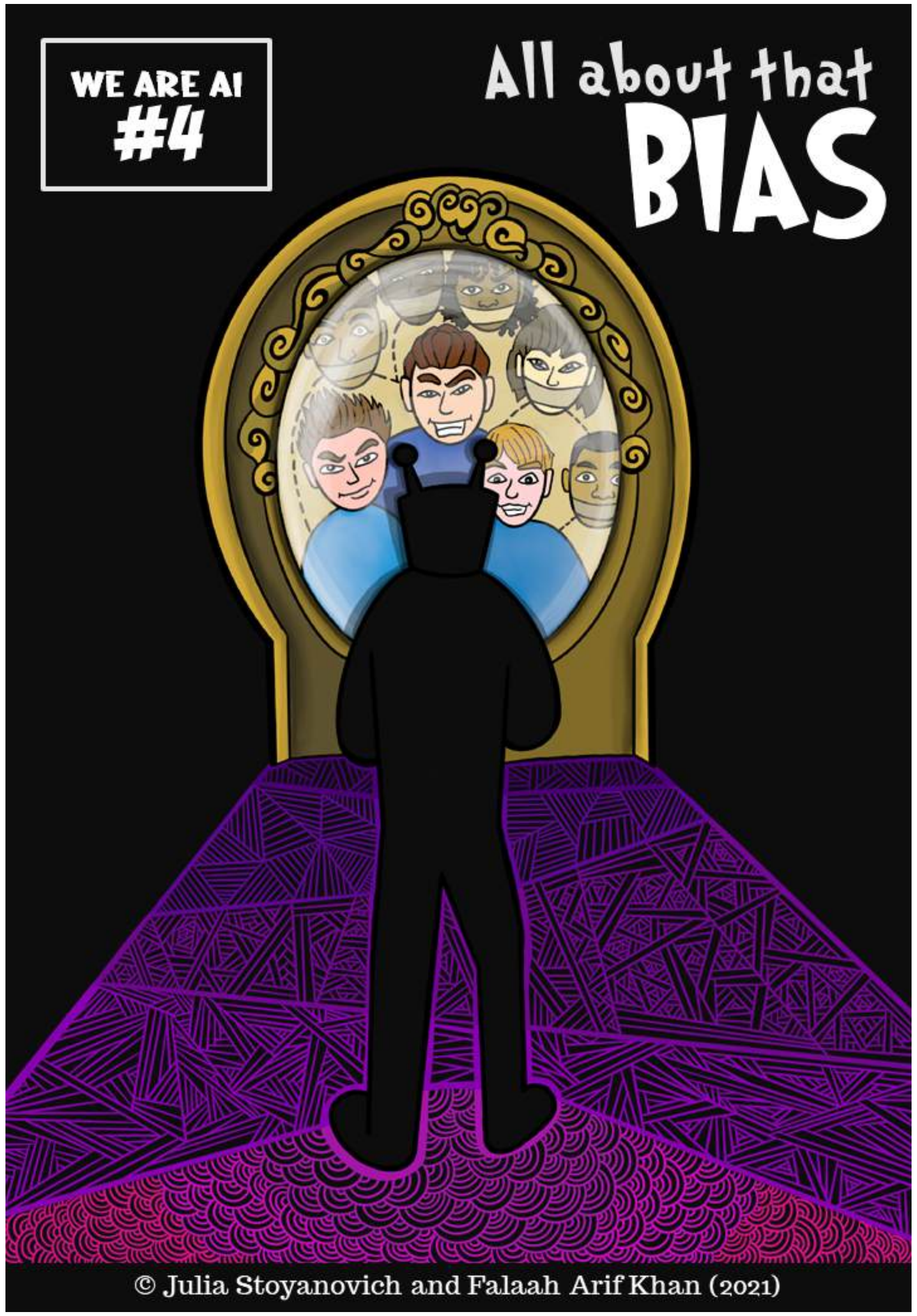
From an analysis of actual cases, three categories of bias in computer systems have been developed: preexisting, technical, and emergent. Preexisting bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias arises in a context of use. Although others have pointed to bias in particular computer systems and have noted the general problem, we know of no comparable work that examines this phenomenon comprehensively and which offers a framework for understanding and remedying it. We conclude by suggesting that freedom from bias should be counted among the select set of criteria—including reliability, accuracy, and efficiency—according to which the quality of systems in use in society should be judged.

Categories and Subject Descriptors: D.2.0 [**Software**]: Software Engineering; H.1.2 [**Information Systems**]: User/Machine Systems; K.4.0 [**Computers and Society**]: General

General Terms: Design, Human Factors

Additional Key Words and Phrases: Bias, computer ethics, computers and society, design methods, ethics, human values, standards, social computing, social impact, system design, universal design, values

[Friedman & Nissenbaum, Comm ACM (1996)]



© Julia Stoyanovich and Falaah Arif Khan (2021)

Reading: Algorithmic fairness

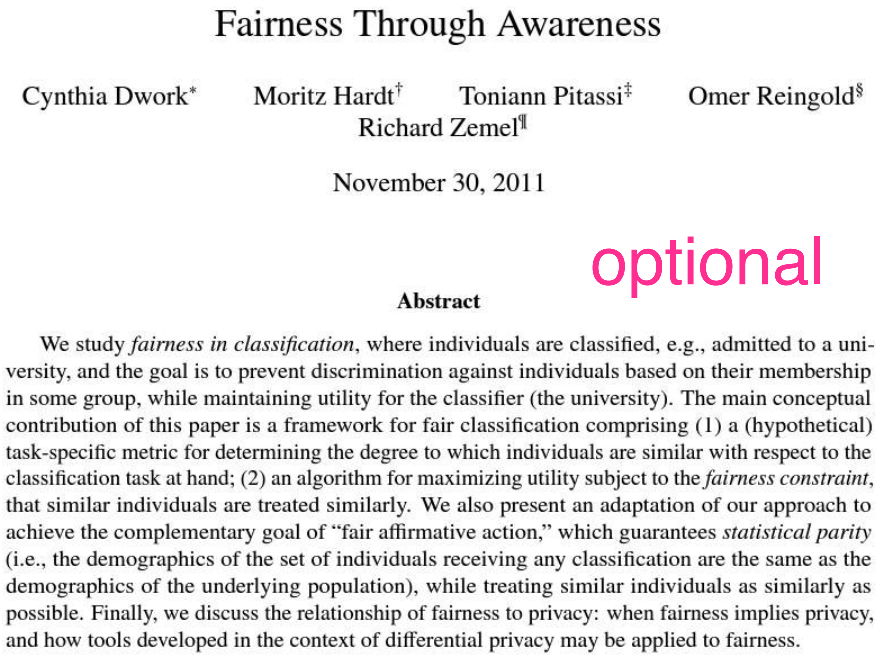
DOI:10.1145/3376898

A group of industry, academic, and government experts convene in Philadelphia to explore the roots of algorithmic bias.

BY ALEXANDRA CHOULDECHOVA AND AARON ROTH

A Snapshot of the Frontiers of Fairness in Machine Learning

[Chouldechova & Roth, Comm ACM (2020)]



Fairness Through Awareness

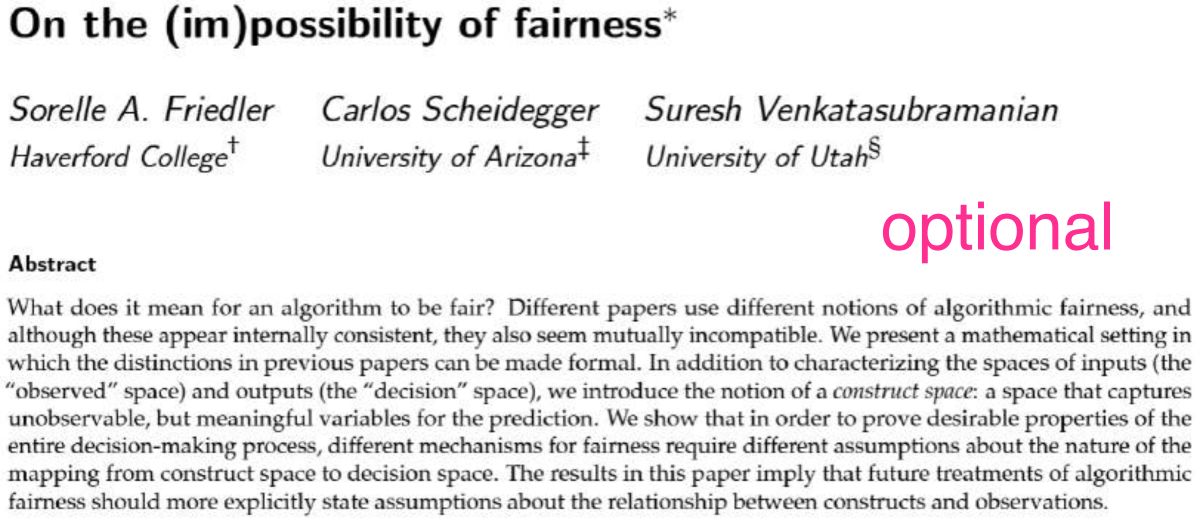
Cynthia Dwork[*] Moritz Hardt[†] Toniann Pitassi[‡] Omer Reingold[§]
Richard Zemel[¶]

November 30, 2011

optional

Abstract

We study *fairness in classification*, where individuals are classified, e.g., admitted to a university, and the goal is to prevent discrimination against individuals based on their membership in some group, while maintaining utility for the classifier (the university). The main conceptual contribution of this paper is a framework for fair classification comprising (1) a (hypothetical) task-specific metric for determining the degree to which individuals are similar with respect to the classification task at hand; (2) an algorithm for maximizing utility subject to the *fairness constraint*, that similar individuals are treated similarly. We also present an adaptation of our approach to achieve the complementary goal of "fair affirmative action," which guarantees *statistical parity* (i.e., the demographics of the set of individuals receiving any classification are the same as the demographics of the underlying population), while treating similar individuals as similarly as possible. Finally, we discuss the relationship of fairness to privacy: when fairness implies privacy, and how tools developed in the context of differential privacy may be applied to fairness.



On the (im)possibility of fairness[*]

Sorelle A. Friedler *Carlos Scheidegger* *Suresh Venkatasubramanian*
Haverford College[†] *University of Arizona*[‡] *University of Utah*[§]

optional

Abstract

What does it mean for an algorithm to be fair? Different papers use different notions of algorithmic fairness, and although these appear internally consistent, they also seem mutually incompatible. We present a mathematical setting in which the distinctions in previous papers can be made formal. In addition to characterizing the spaces of inputs (the "observed" space) and outputs (the "decision" space), we introduce the notion of a *construct space*: a space that captures unobservable, but meaningful variables for the prediction. We show that in order to prove desirable properties of the entire decision-making process, different mechanisms for fairness require different assumptions about the nature of the mapping from construct space to decision space. The results in this paper imply that future treatments of algorithmic fairness should more explicitly state assumptions about the relationship between constructs and observations.

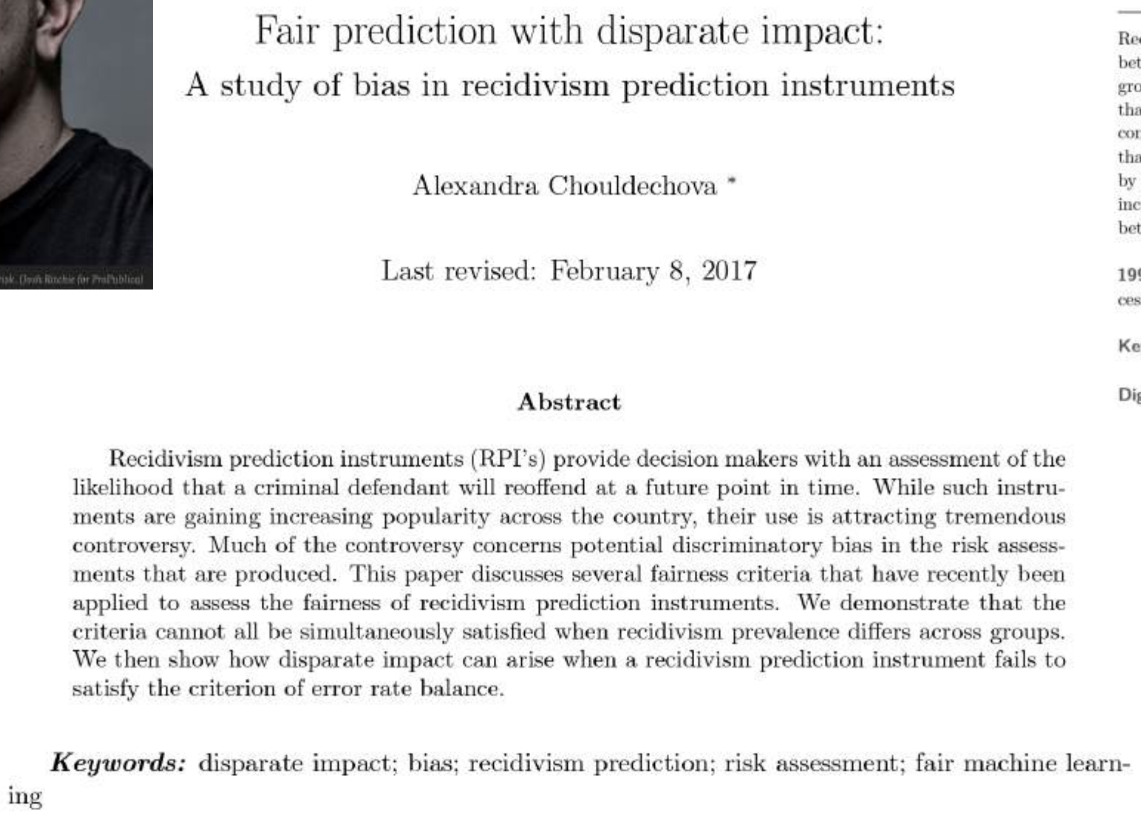
Reading: Fairness in risk assessment

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)



Fair prediction with disparate impact:
A study of bias in recidivism prediction instruments

Alexandra Chouldechova *

Last revised: February 8, 2017

Abstract

Recidivism prediction instruments (RPI's) provide decision makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time. While such instruments are gaining increasing popularity across the country, their use is attracting tremendous controversy. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. This paper discusses several fairness criteria that have recently been applied to assess the fairness of recidivism prediction instruments. We demonstrate that the criteria cannot all be simultaneously satisfied when recidivism prevalence differs across groups. We then show how disparate impact can arise when a recidivism prediction instrument fails to satisfy the criterion of error rate balance.

Keywords: disparate impact; bias; recidivism prediction; risk assessment; fair machine learning

[Chouldechova, BigData (2017)]

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg[1], Sendhil Mullainathan[2], and Manish Raghavan[3]

1 Cornell University, Ithaca, USA
kleinber@cs.cornell.edu
2 Harvard University, Cambridge, USA
mullain@fas.harvard.edu
3 Cornell University, Ithaca, USA
manish@cs.cornell.edu

Abstract

Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

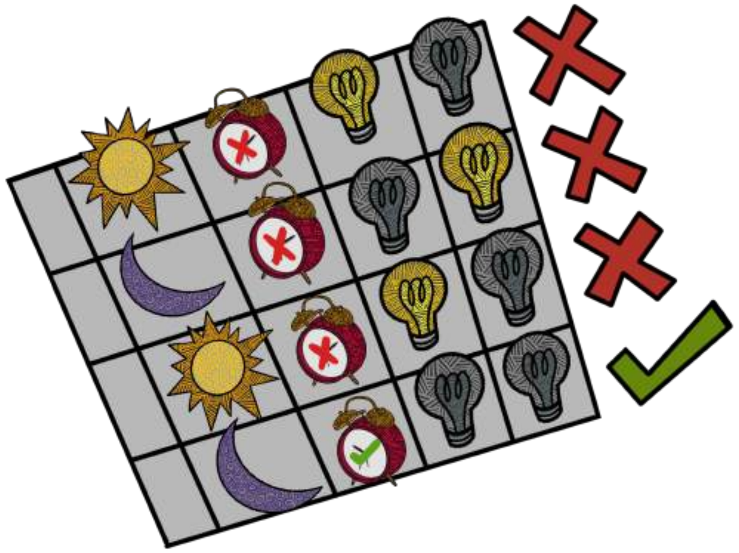
1998 ACM Subject Classification H.2.8 Database Applications, J.1 Administrative Data Processing

Keywords and phrases algorithmic fairness, risk tools, calibration

Digital Object Identifier 10.4230/LIPIcs.ITCS.2017.43

[Kleinberg, Mullainathan & Raghavan, ITCS (2017)]

Individual & cumulative harms

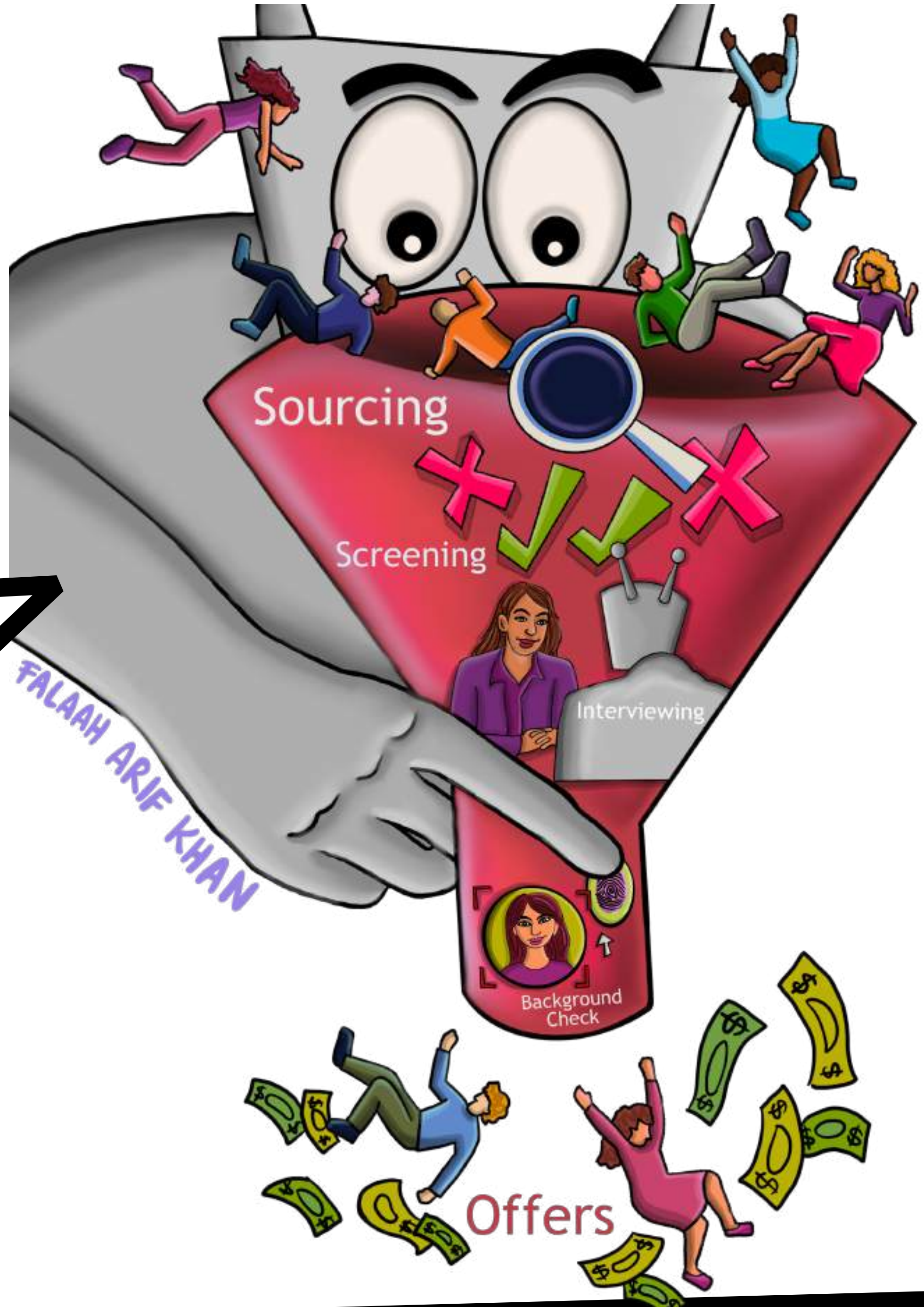


Questions to keep in mind:

what are the **goals** of the AI system?

what are the **benefits** and to **whom**?

what are the **harms** and to **whom**?



technical teaser *c. 2015*: fairness in classification

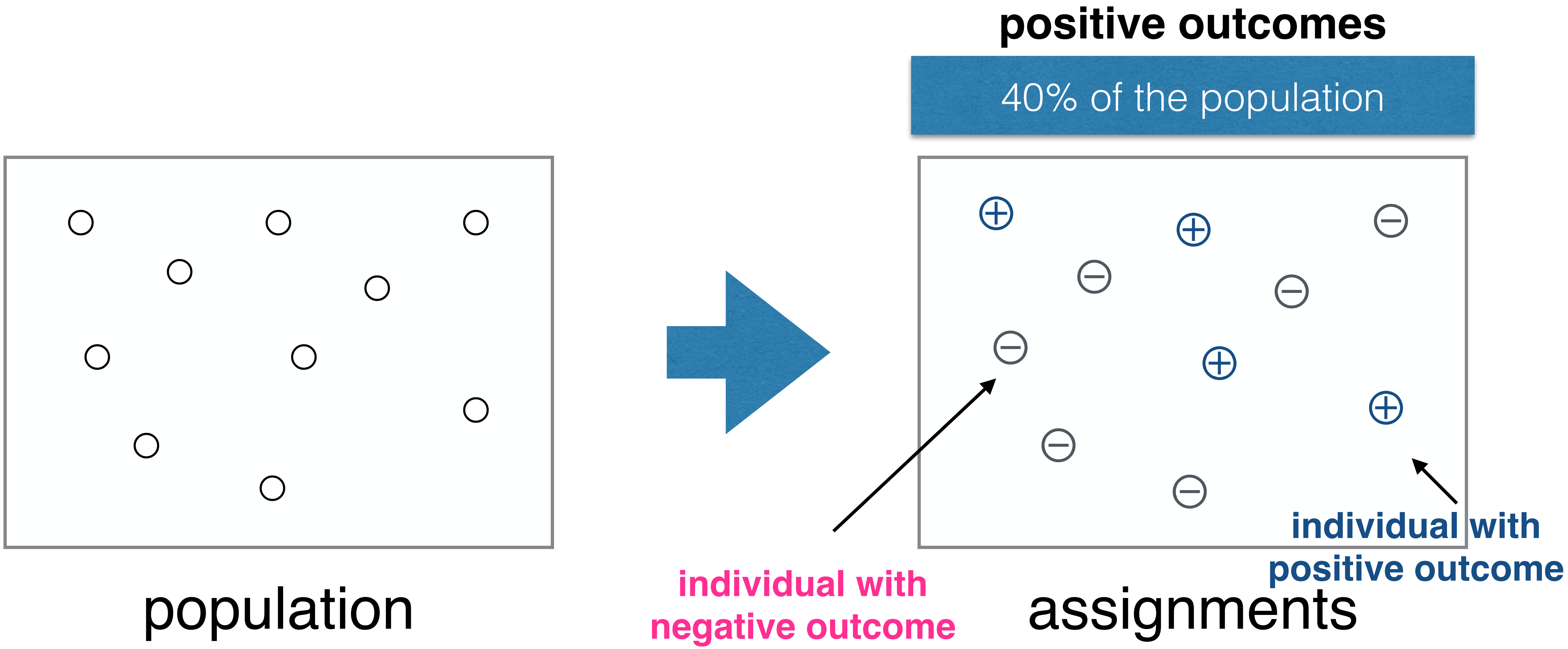
Vendors and outcomes

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
~~shown relevant ad for shoes~~	~~shown irrelevant ad for shoes~~

Fairness in classification

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



positive outcomes

40% of the population

population

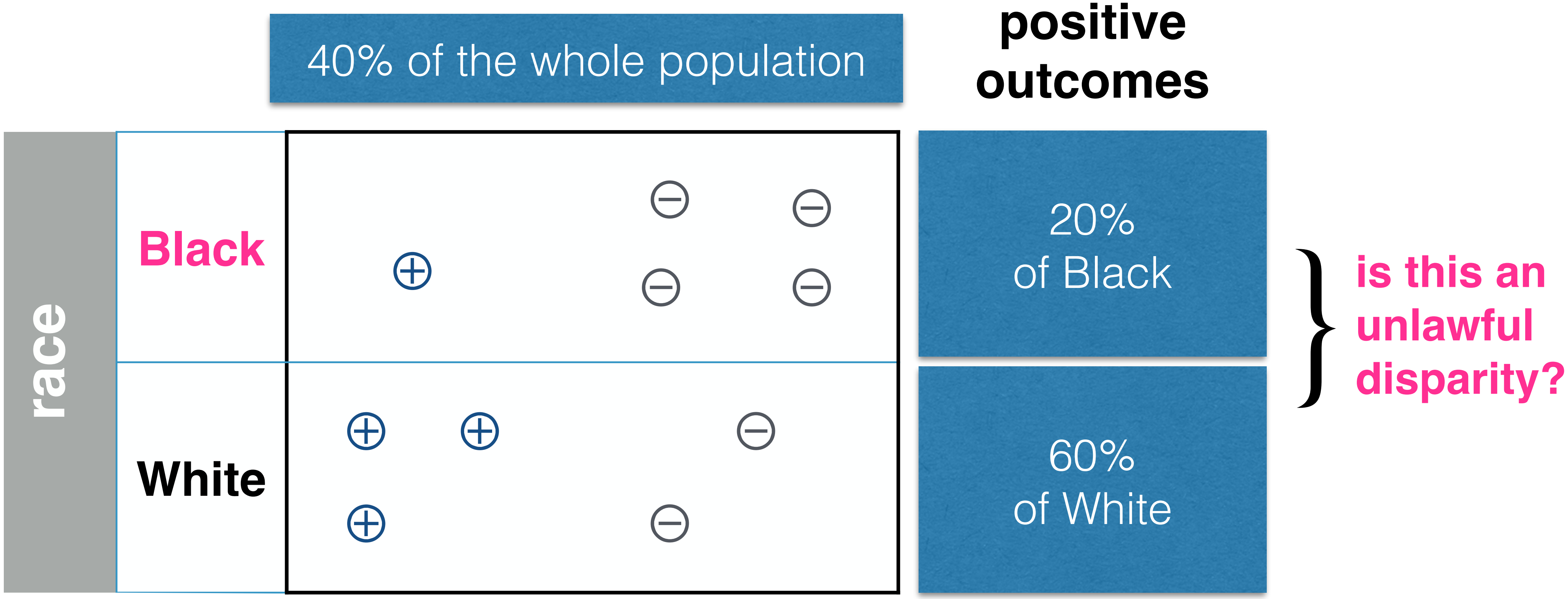
assignments

individual with negative outcome

individual with positive outcome

Fairness in classification

Sub-populations may be treated differently



40% of the whole population

positive outcomes

race

Black ⊕ ⊖ ⊖ ⊖ ⊖

White ⊕ ⊕ ⊖ ⊕ ⊖

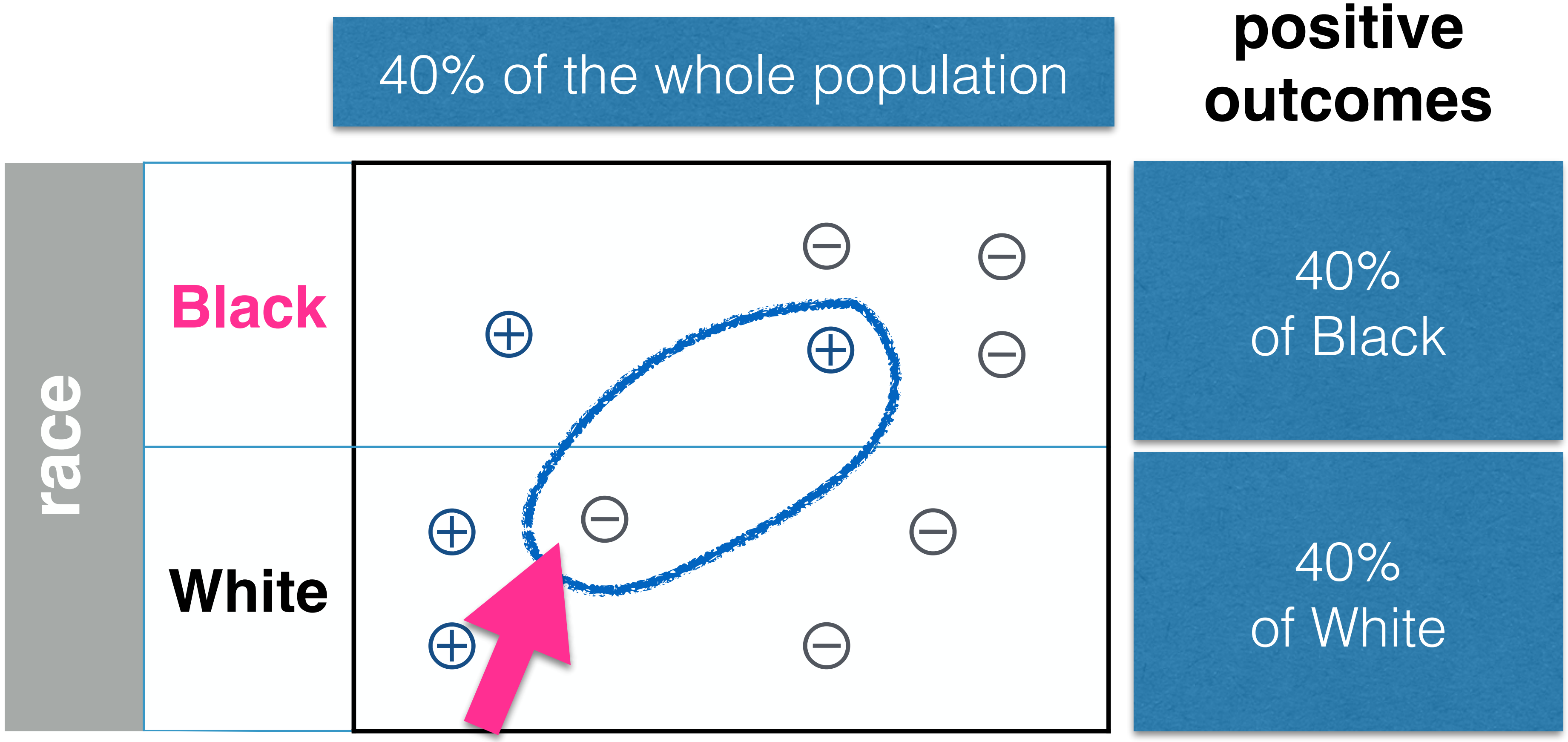
20% of Black

60% of White

is this an unlawful disparity?

Fairness in classification

Sub-populations may be treated differently



40% of the whole population

positive outcomes

race

Black

White

40% of Black

40% of White

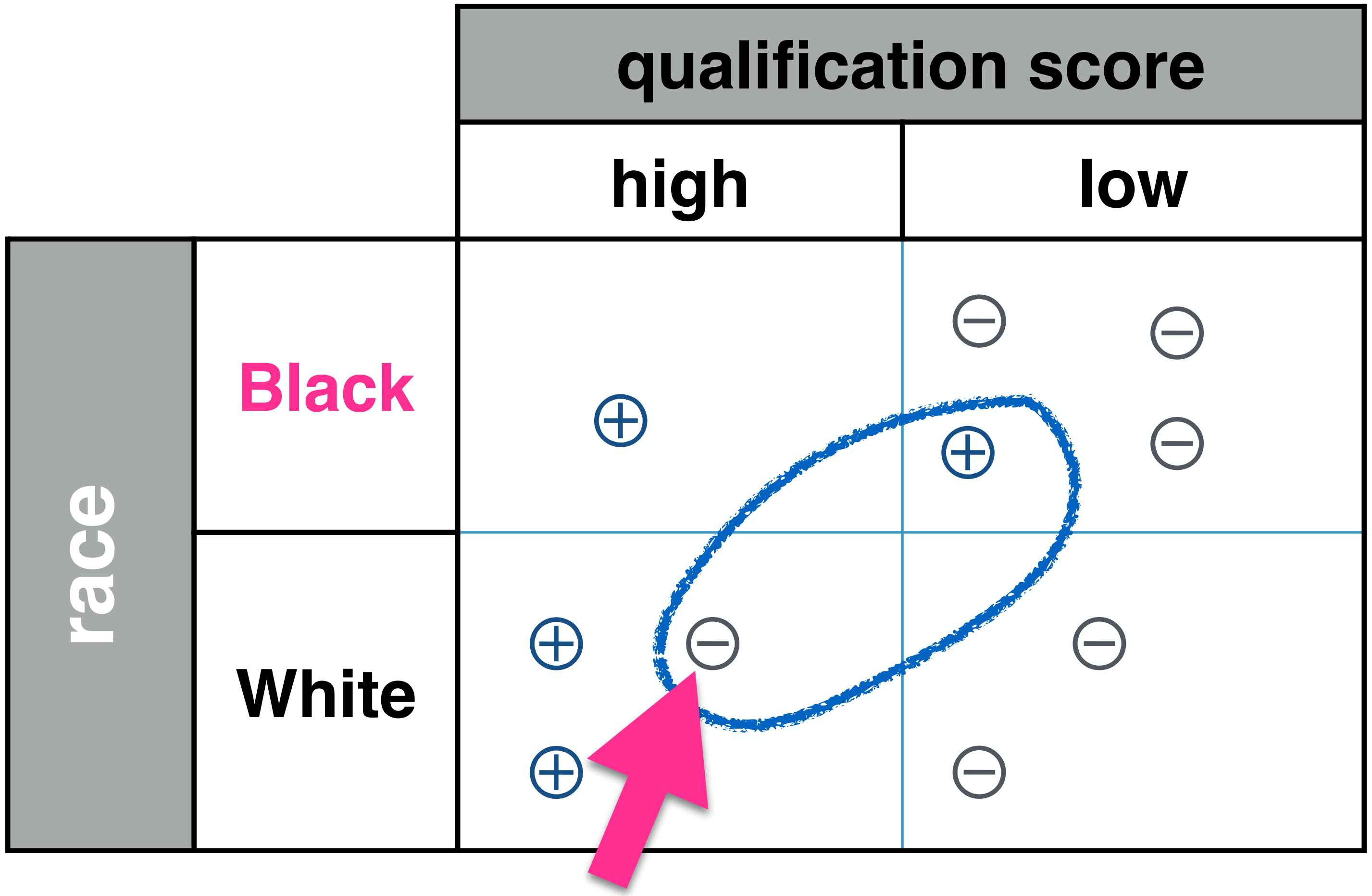
Fairness in classification

Explaining the disparity with proxy variables

		qualification score		positive outcomes
		high	low	
race	Black	⊕	⊖ ⊖ ⊖ ⊖	20% of Black
	White	⊕ ⊕ ⊕	⊖ ⊖	60% of White

quick discussion

Swapping outcomes



		qualification score	
		high	low
race	Black	⊕	⊖ ⊖ ⊕ ⊖
	White	⊕ ⊖ ⊕	⊖ ⊖

positive outcomes

40%
of Black

40%
of White

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

Bias in computer systems

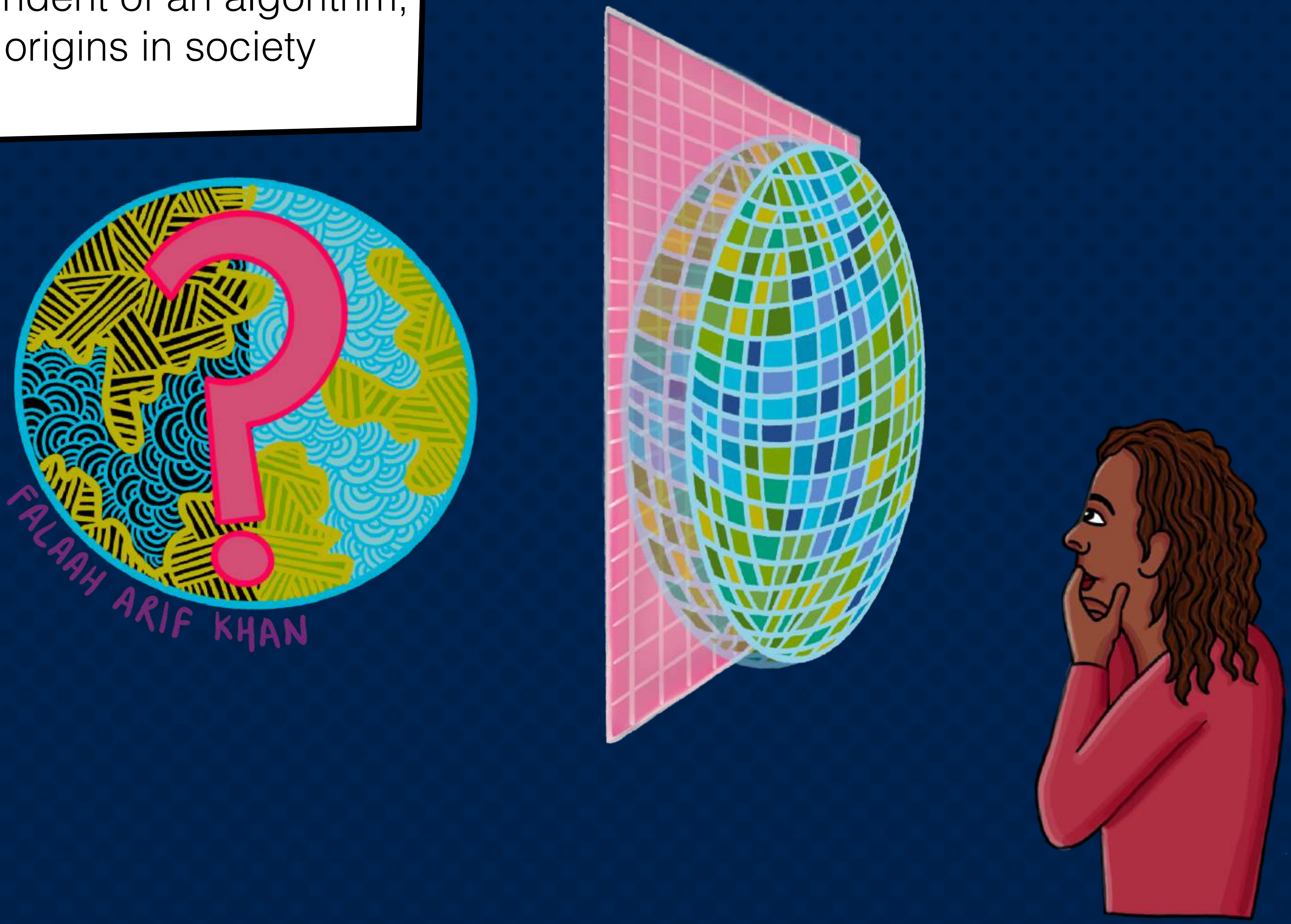
Pre-existing is independent of an algorithm and has origins in society

Technical is introduced or exacerbated by the technical properties of an ADS

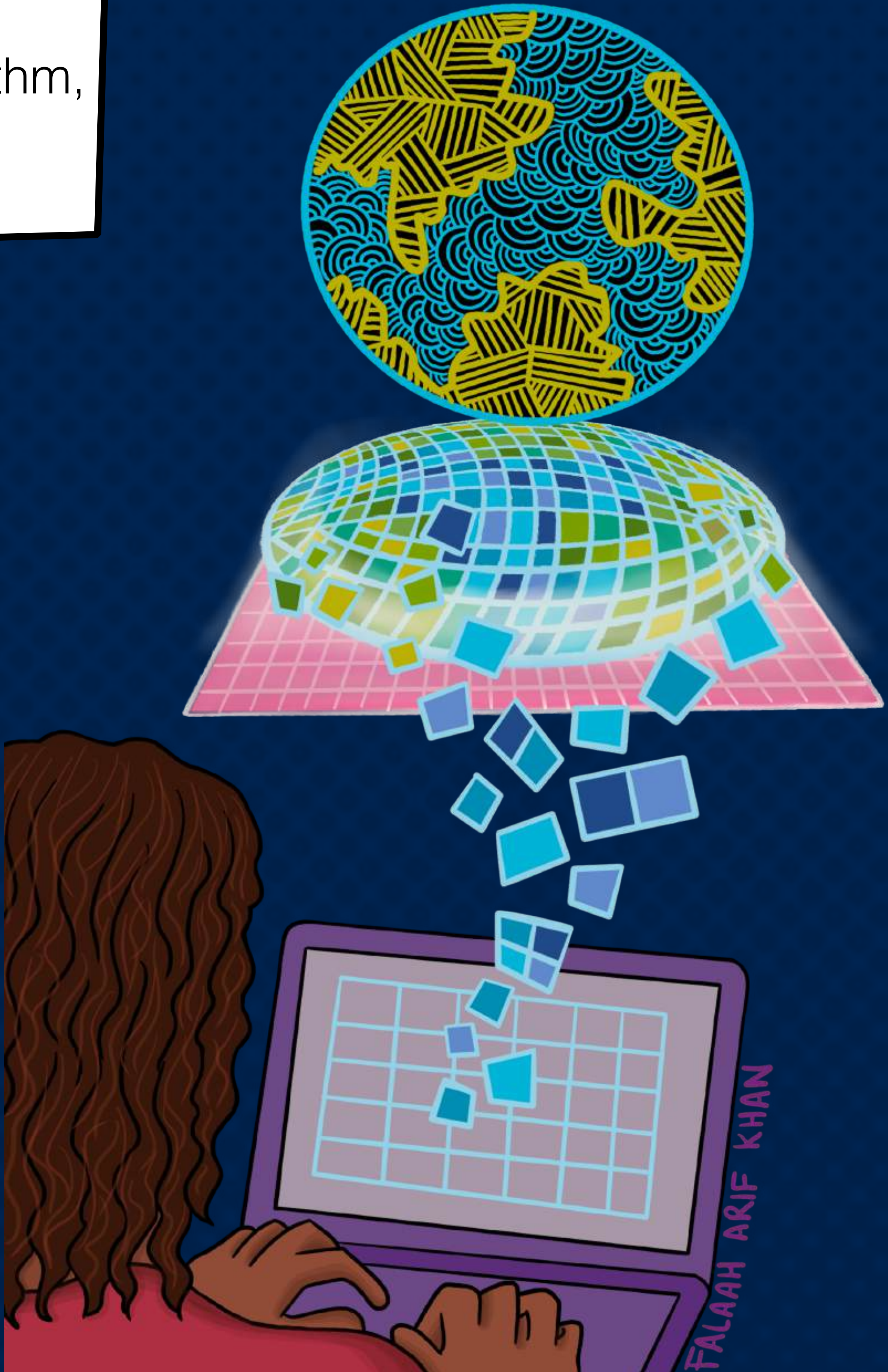
Emergent arises due to context of use

[Friedman & Nissenbaum (1996)]

Pre-existing bias:
independent of an algorithm, has its origins in society



Pre-existing bias:
independent of an algorithm, has its origins in society

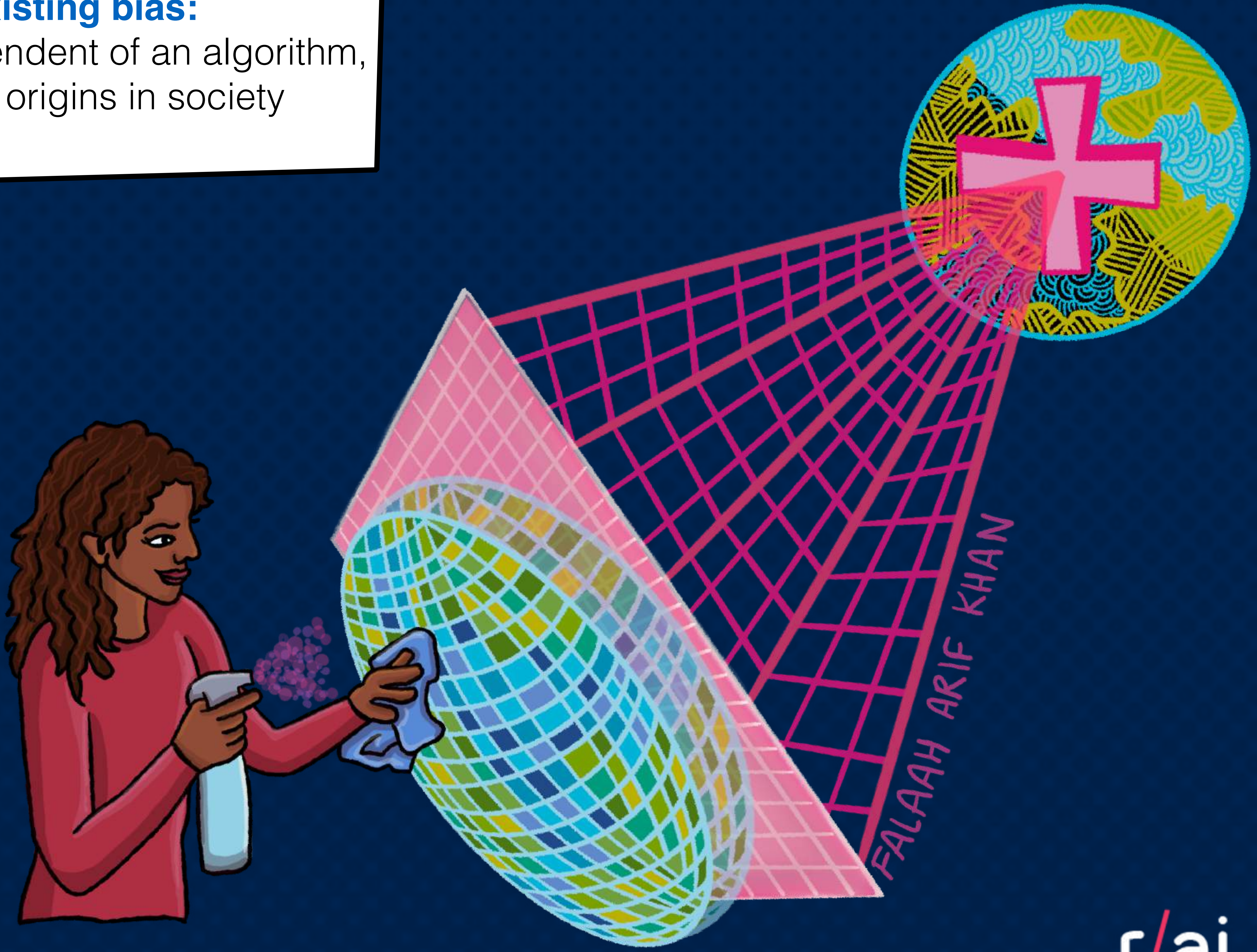


r/ai

Pre-existing bias:
independent of an algorithm,
has its origins in society

r/ai

Pre-existing bias:
independent of an algorithm, has its origins in society



r/ai

bias can lead to discrimination

The evils of discrimination

Disparate treatment

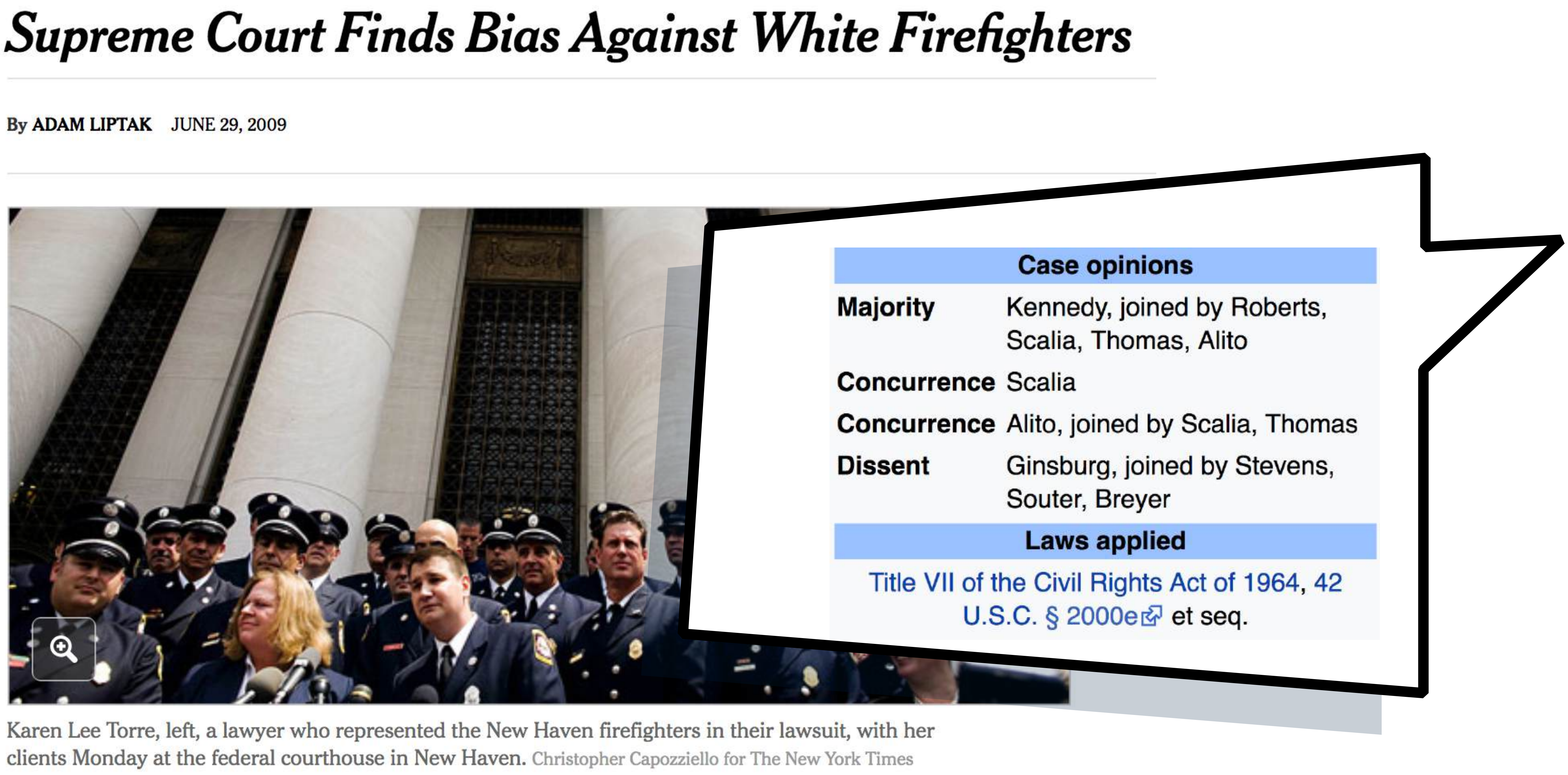
is the illegal practice of treating an entity, such as a job applicant or an employee, differently based on a **protected characteristic** such as race, gender, age, disability status, religion, sexual orientation, or national origin.

Disparate impact

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

r/ai

Ricci v. DeStefano (2009)



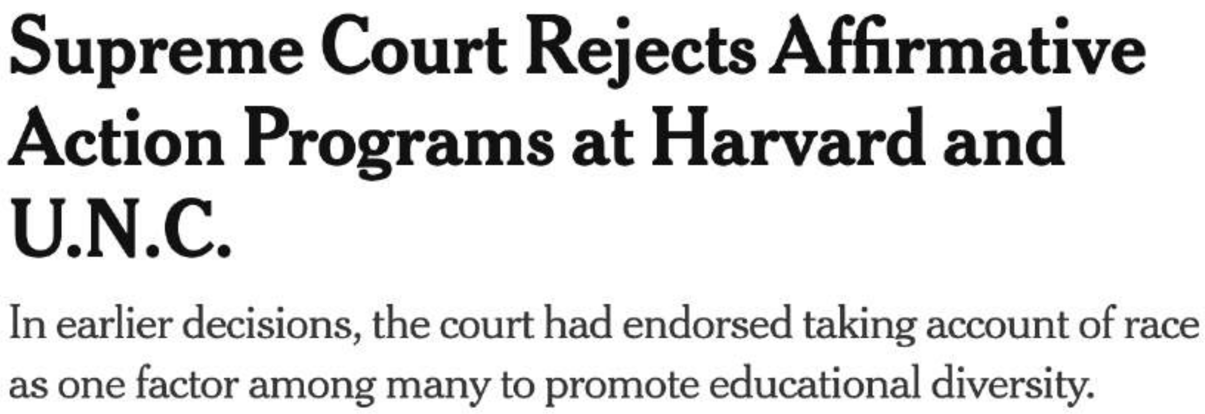
Supreme Court Finds Bias Against White Firefighters

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.	

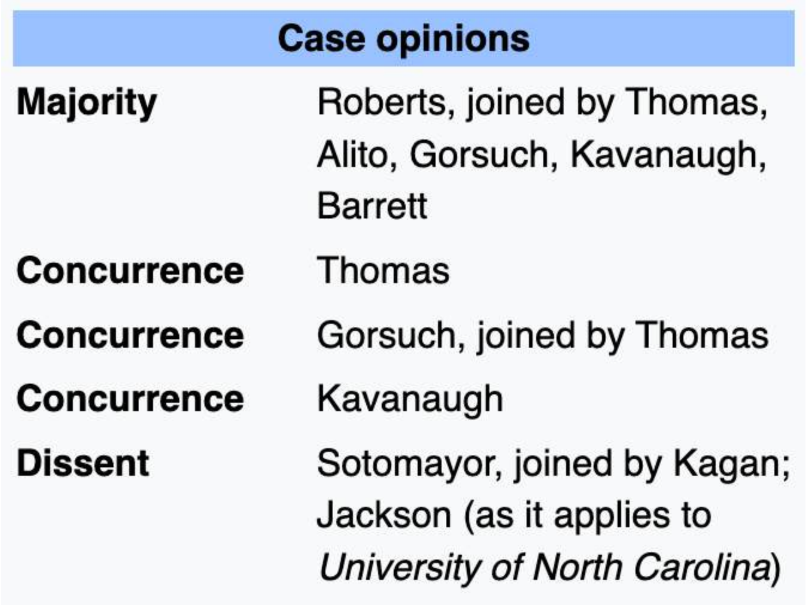
Students for Fair(?) Admissions v. UNC / Harvard (2023)



Supreme Court Rejects Affirmative Action Programs at Harvard and U.N.C.

In earlier decisions, the court had endorsed taking account of race as one factor among many to promote educational diversity.

Harvard's admissions program violates the Equal Protection Clause of the Fourteenth Amendment. United States Court of Appeals for the First Circuit reversed.



Case opinions	
Majority	Roberts, joined by Thomas, Alito, Gorsuch, Kavanaugh, Barrett
Concurrence	Thomas
Concurrence	Gorsuch, joined by Thomas
Concurrence	Kavanaugh
Dissent	Sotomayor, joined by Kagan; Jackson (as it applies to *University of North Carolina*)

Break!

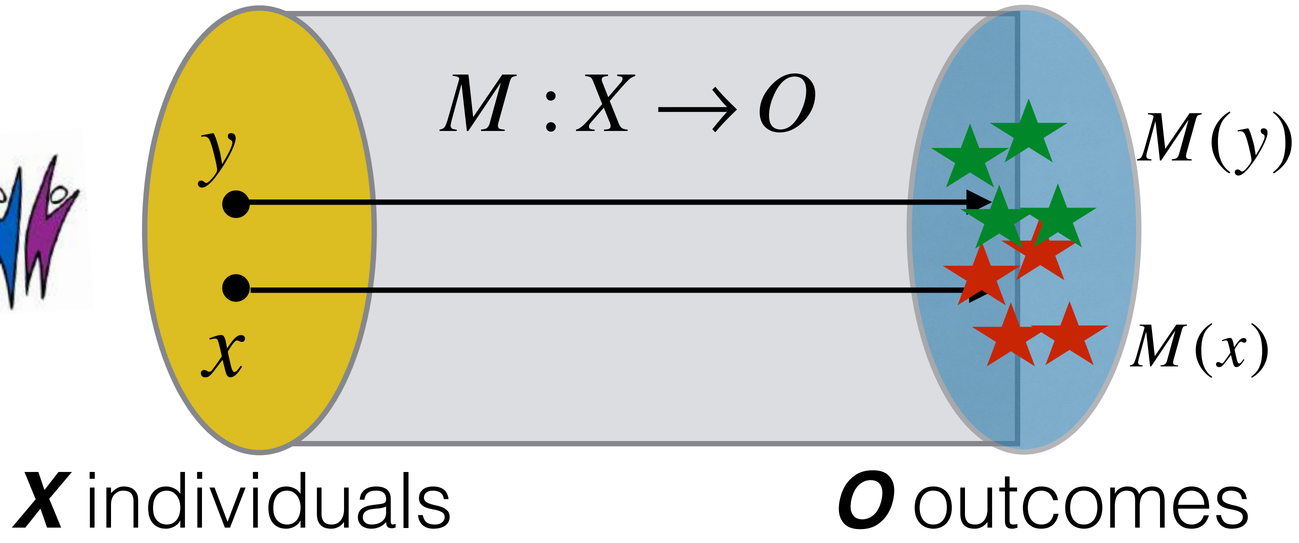
Quiz 1: Machine Bias

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**.



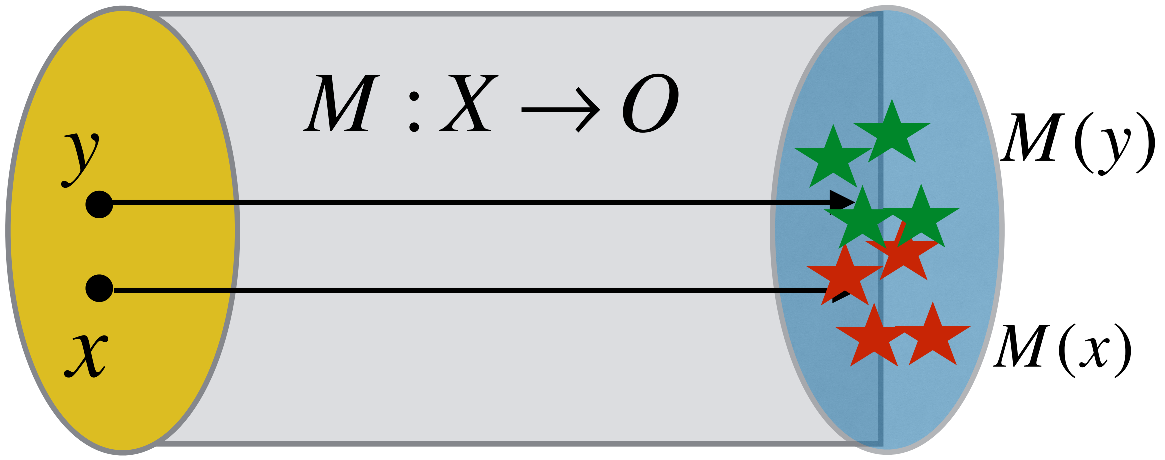
A task-specific distance metric is given $d(x,y)$

$M : X \rightarrow O$ 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 distance metric is given $d(x,y)$

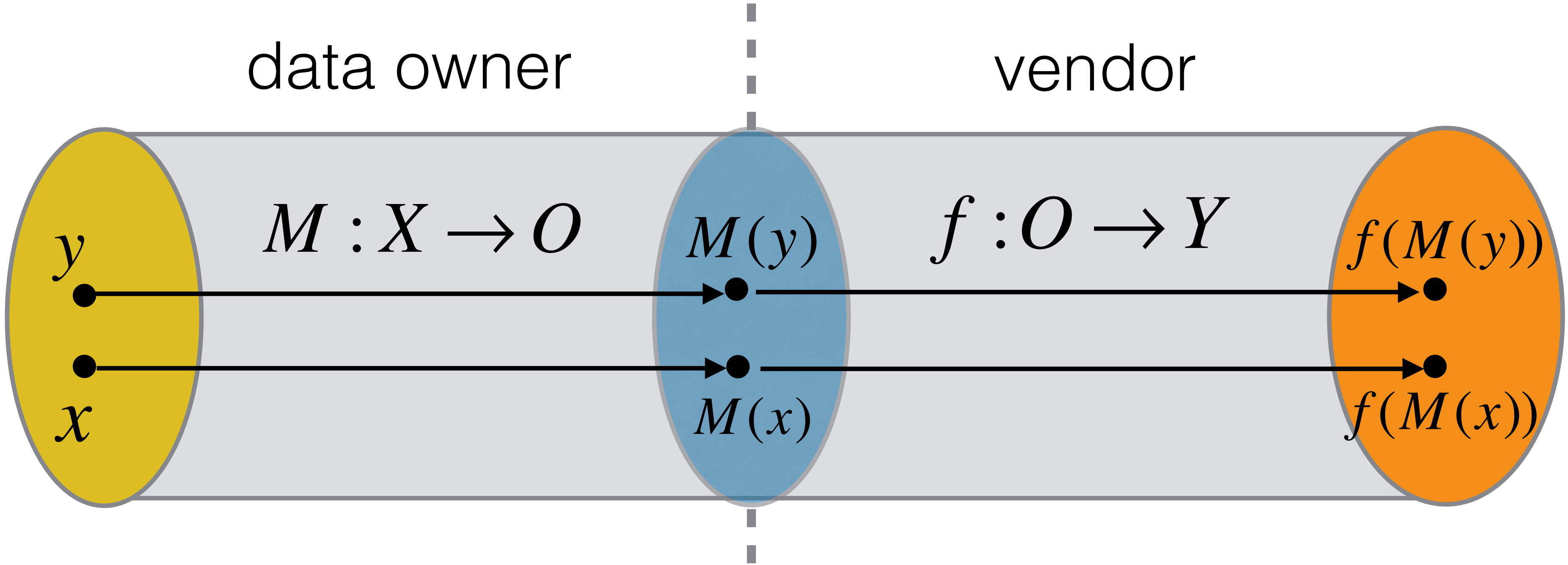
M is a Lipschitz mapping if $\forall x,y \in X \quad \|M(x),M(y)\| \leq d(x,y)$

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*]



simpsons.wikia.com

X individuals

O outcomes

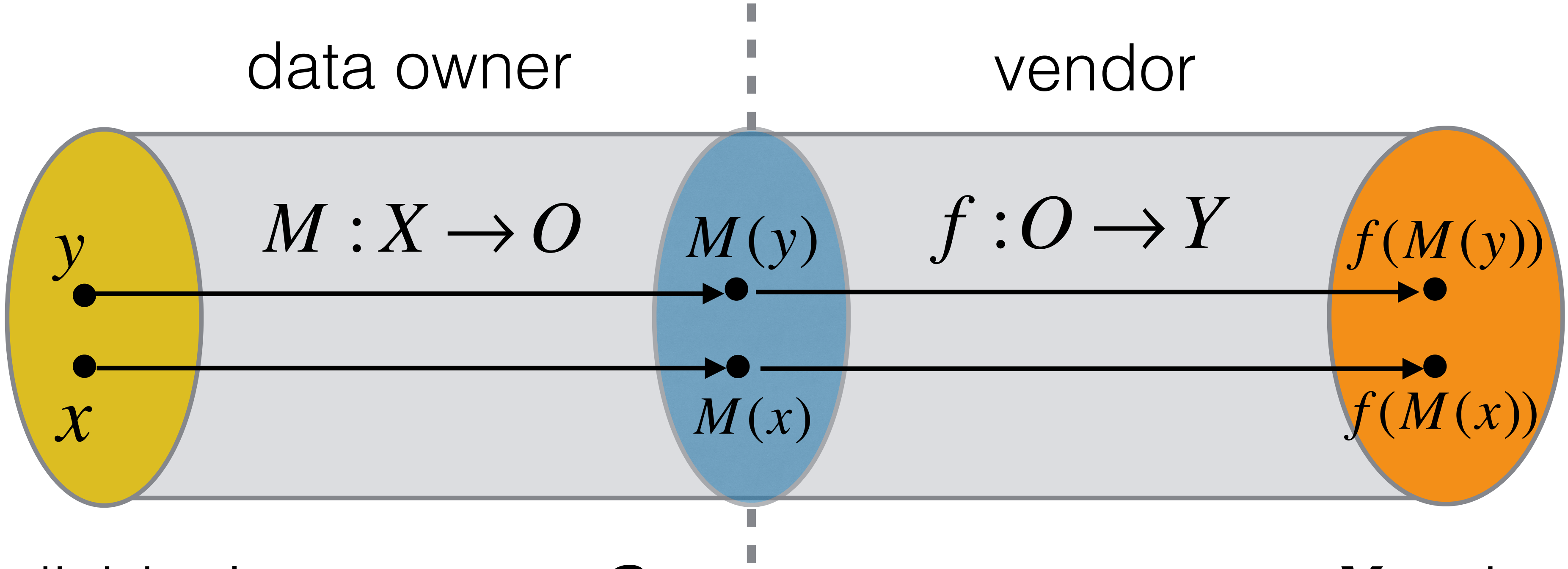
Y actions

fairness enforced at this step

vendor cannot introduce bias

Fairness through a Lipschitz mapping

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



data owner

vendor

$M : X \rightarrow O$

$f : O \rightarrow Y$

y

x

$M(y)$

$M(x)$

$f(M(y))$

$f(M(x))$

X individuals

O outcomes

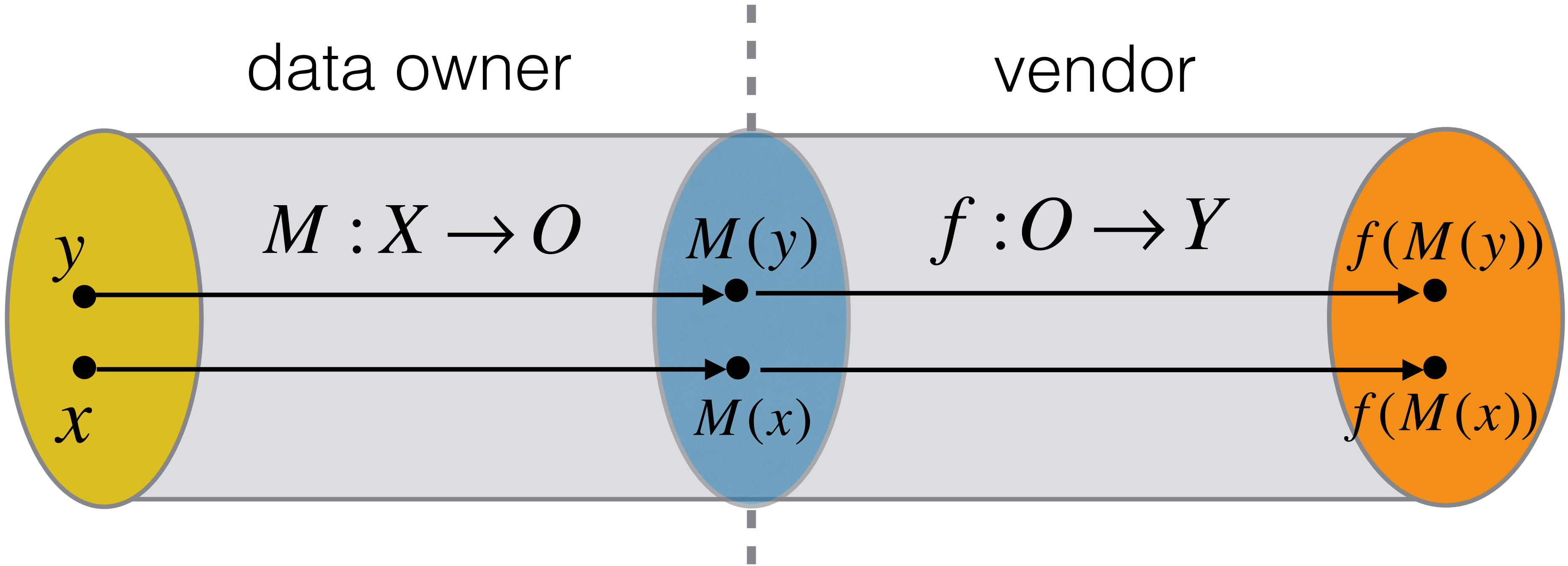
Y actions

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

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

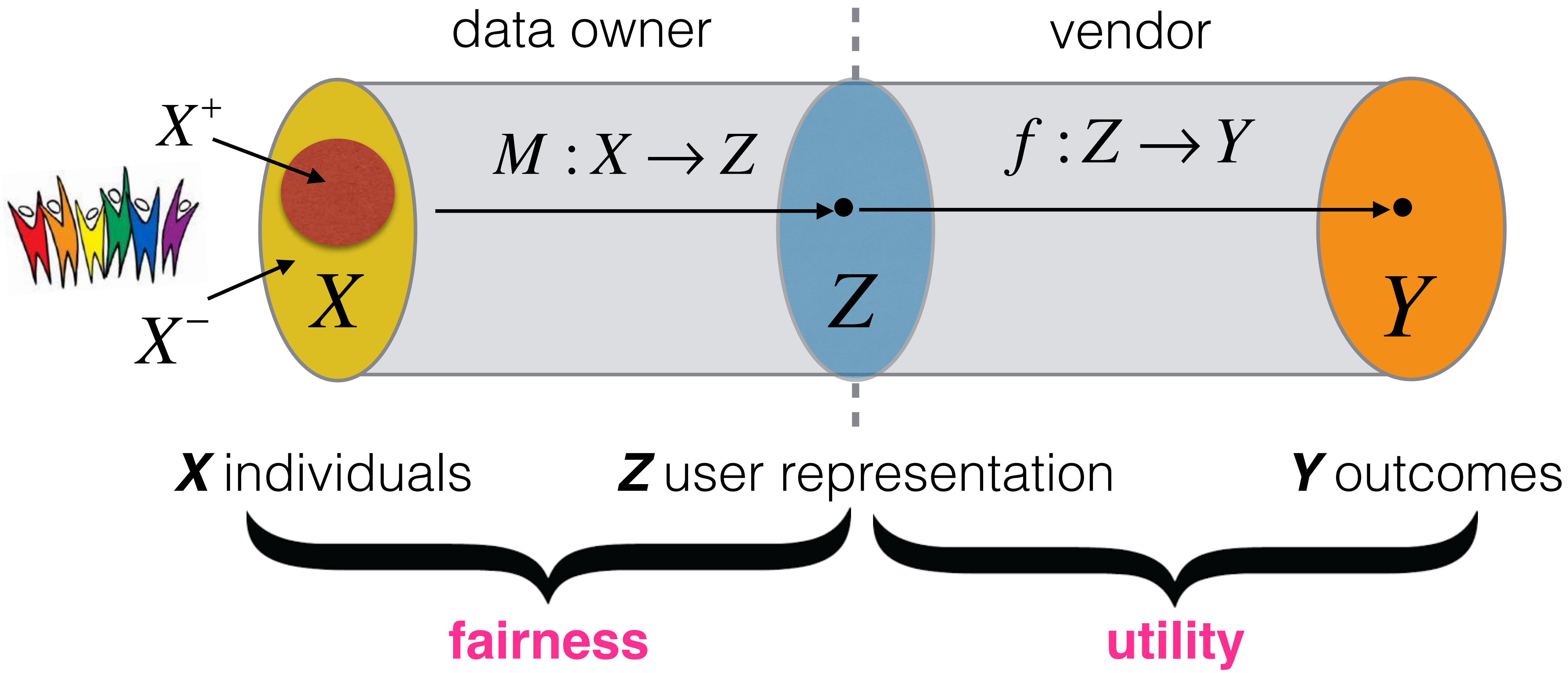
Computed with a linear program of size $poly(|X|,|Y|)$

the same mapping can be used by multiple vendors

learning fair representations

Learning fair representations

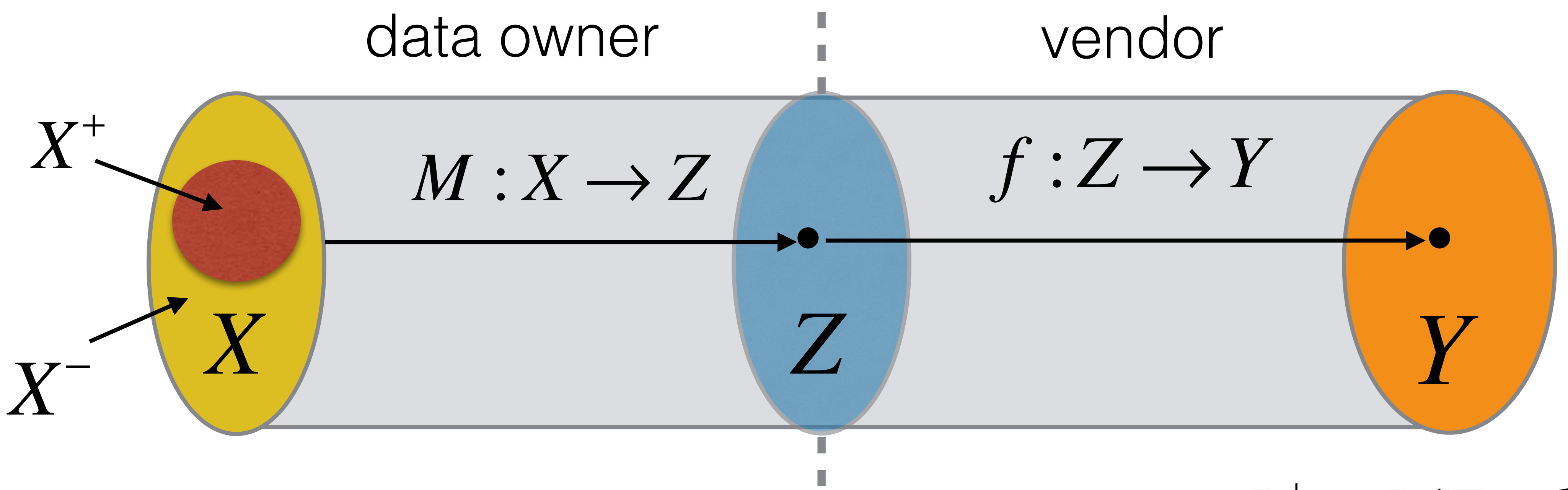
[R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *ICML 2013*]



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

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

M(X) should lose information about membership in S

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

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

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

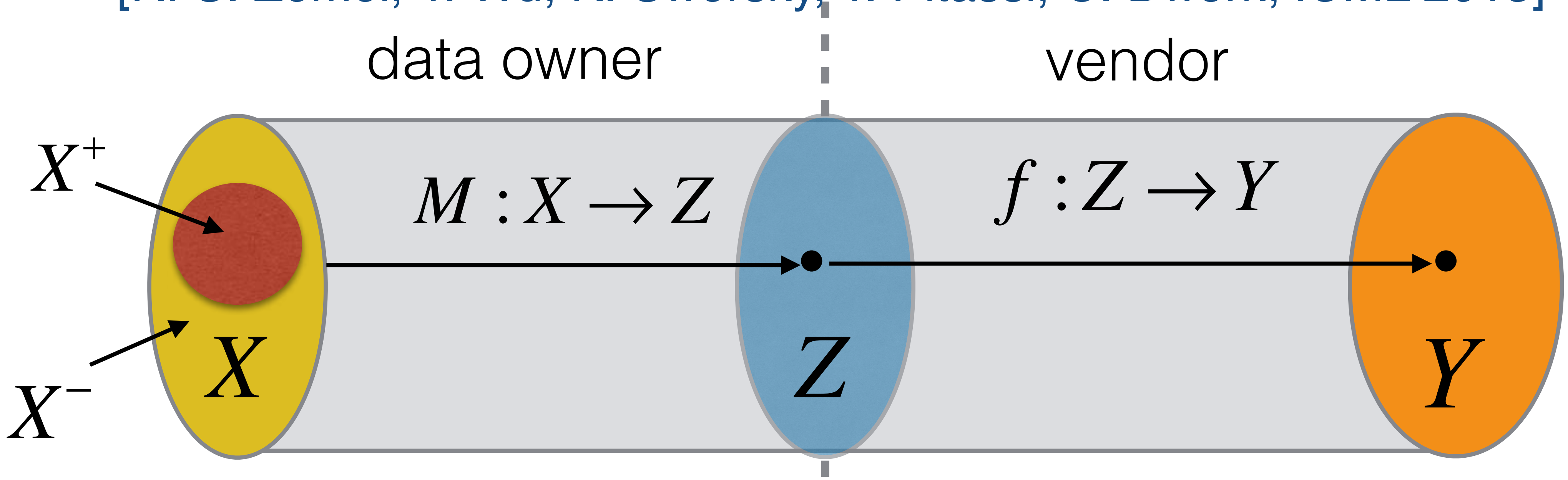
group fairness

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

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

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

$$L_z = \sum_k \left| P_k^+ - P_k^- \right|$$

$$L_x = \sum_n (x_n - \widehat{x_n})^2$$

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

does this make sense?

Responsible Data Science

Algorithmic Fairness

Thank you!

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai