UNIVERSITAT DE BARCELONA

MSc in Fundamental Principles of Data Science

4 Ethical Data Science

Bias and Discrimination I

Jordi Vitrià

Index

1. Bias and Discrimination.
2. The human factor.
3. Automated Discrimination.
4. Case Analysis: Recividism risk.

Bias and discrimination

What do we mean by "data bias"?

The common definition of data **bias** is that the available **data is not representative of the population or phenomenon of study**.

Except for data acquired by a carefully designed randomized sampling process, most organically produced datasets are biased.

But bias also denotes:

- Data includes **content which may contain bias against specific groups of people.**

Ethical Issue!

What do we mean by "algorithmic bias"?

Algorithmic **bias** describes systematic **deviation in output/performance or impact**, relative to some norm or standard.

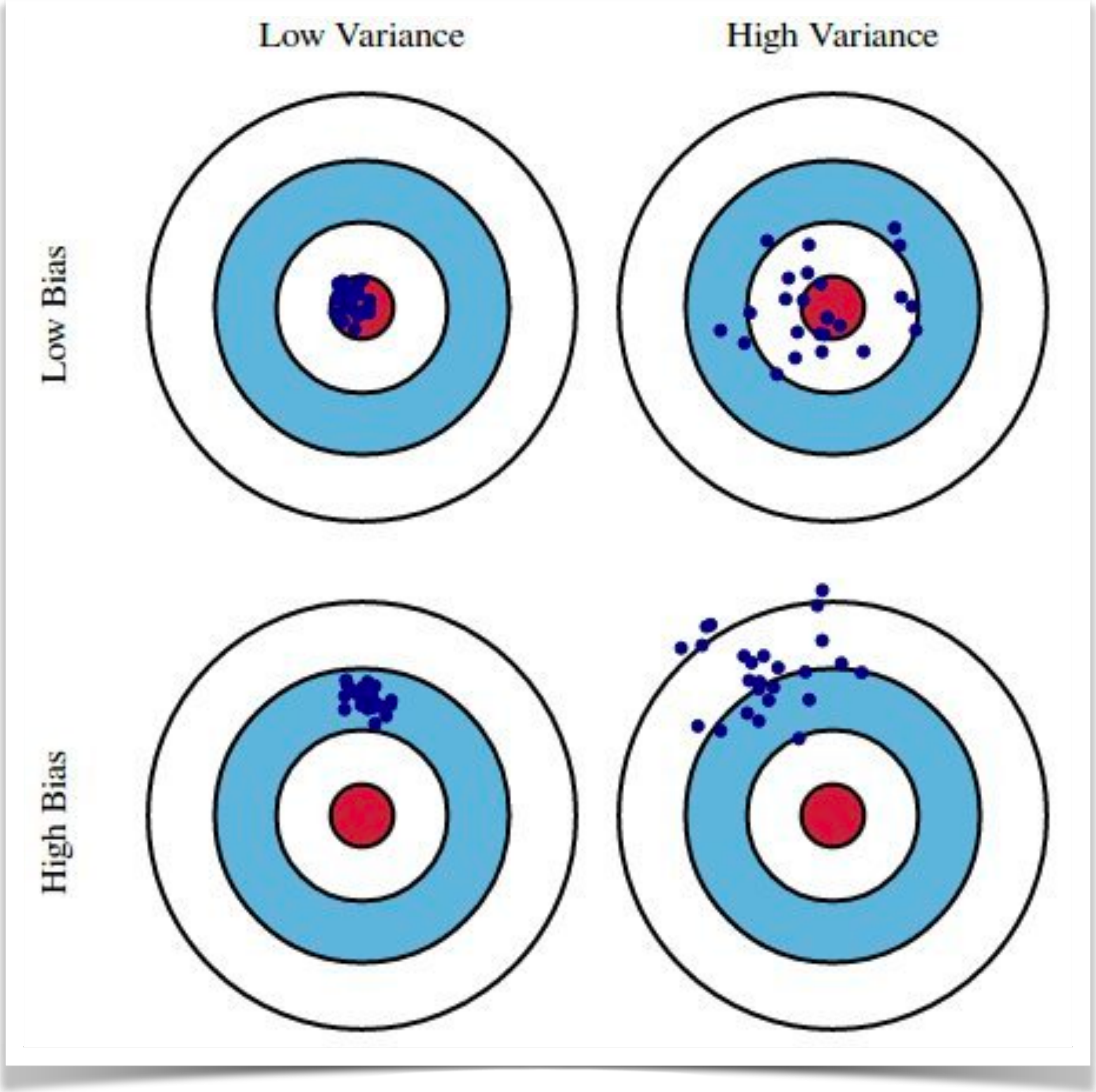
World observation

Example: many universities use **data from past students** to build models for predicting **student success**, where those models can support **informed changes in policies and practices**.

Output

Impact

What do we mean by "algorithmic bias"?



What do we mean by “algorithmic bias”?

Algorithmic bias describes systematic deviation in output/performance or impact, relative to some norm or standard.

An algorithm can be **statistically** XOR **ethically** biased.

Example:
- Our algorithm will be **statistically biased** if predictions differ systematically from previously observed data.
- Our algorithm will be **ethically biased** if predictions depend on the gender of the student.

What do we mean by “algorithmic bias”?

Not all statistically biased behaviors are ethically problematic, while <u>not all statistically unbiased behaviors</u> are ethically acceptable.

Example: An AI system for hiring, trained with historical data, can be statistically unbiased and ethically problematic.

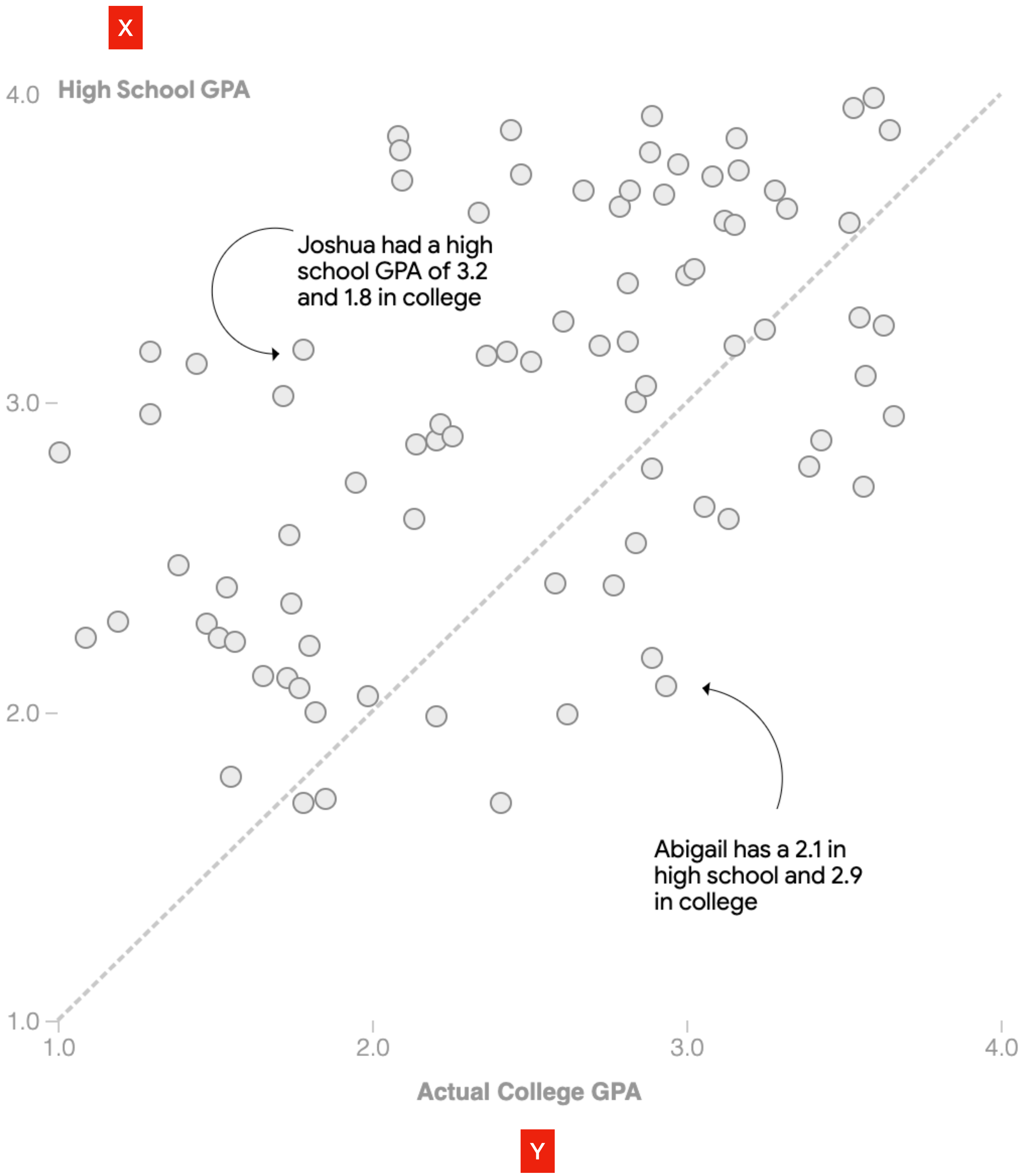
Example

Modeling College GPA

Let's pretend we're college admissions officers trying to predict the GPA students will have in college (in these examples we'll use simulated data).

One simple approach: predict that students will have the same GPA in college as they did in high school.

The Dataset (x, y)



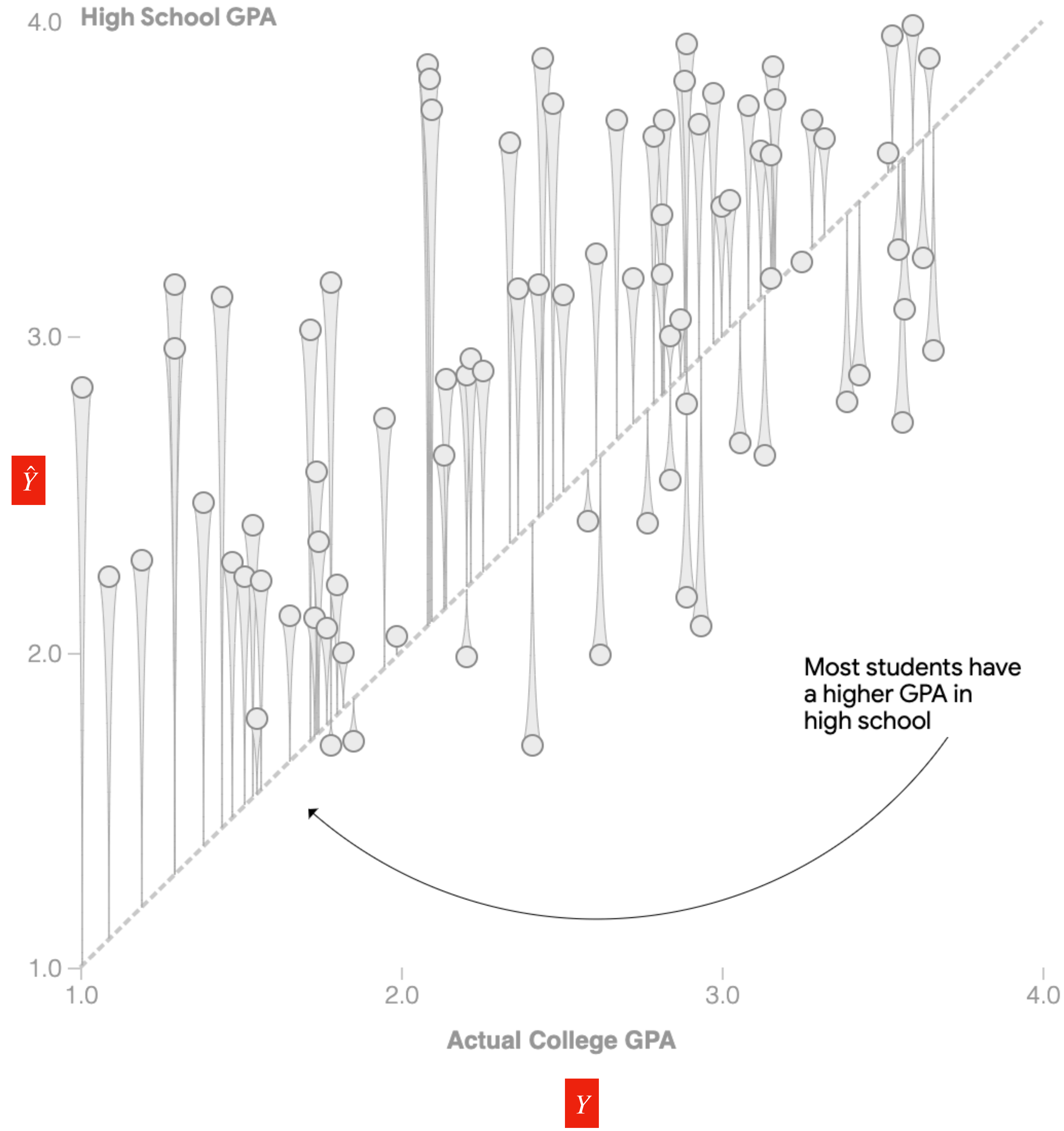
https://pair.withgoogle.com/explorables/hidden-bias/

Example

Naive Predictor $\hat{y} = x$

This is at best a very rough approximation, and it misses a key feature of this data set: students usually have better grades in high school than in college

We're over-predicting college grades more often than we under-predict.



https://pair.withgoogle.com/explorables/hidden-bias/

Example

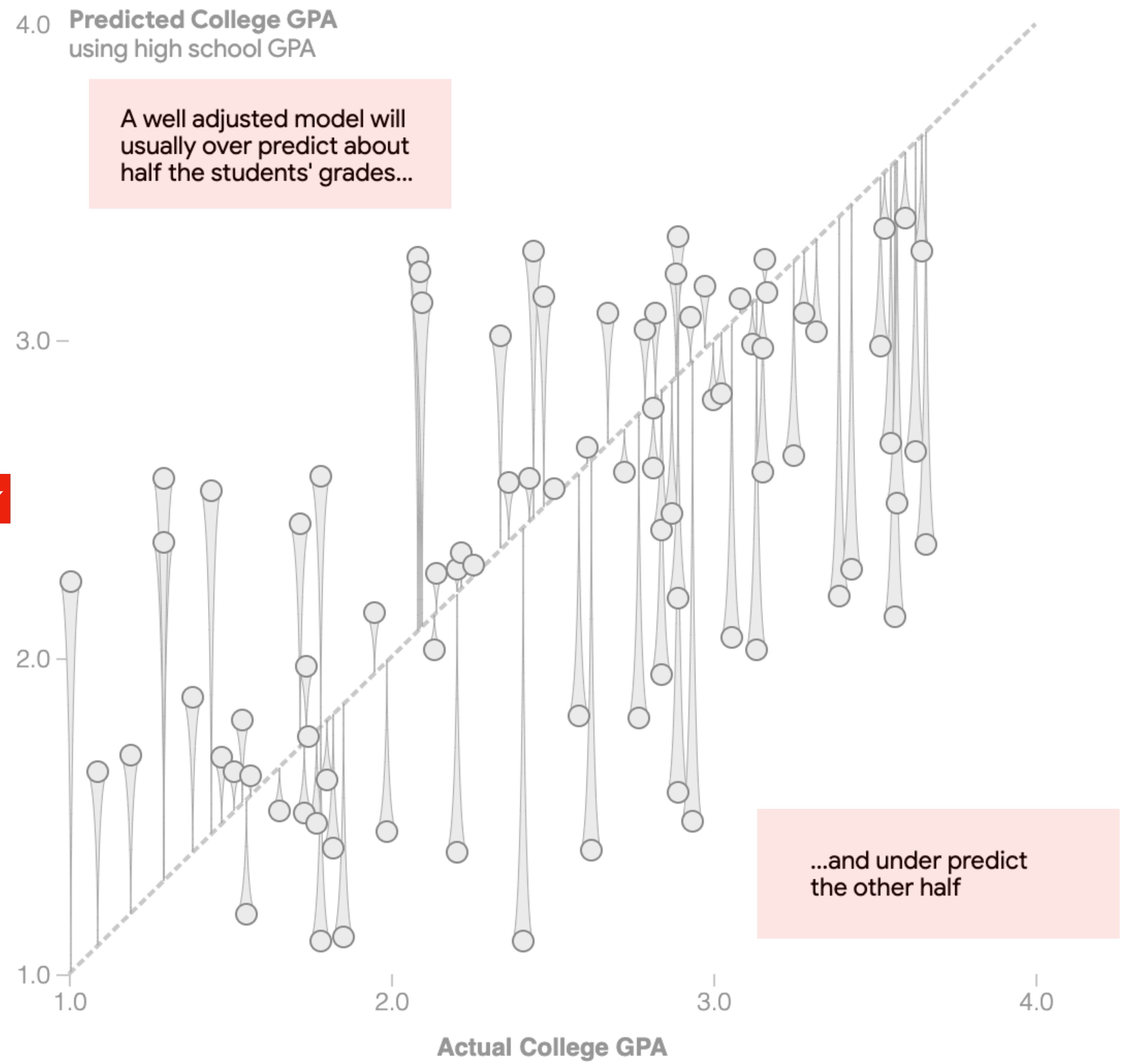
Predictor $\hat{y} = \mathbb{E}(y\,|\,x) = ax + b$

Predicting with ML

If we switched to using a machine learning model and entered these student grades, it would recognize this pattern and adjust the prediction.

The model does this without knowing anything about the real-life context of grading in high school versus college.

$\hat{Y}$



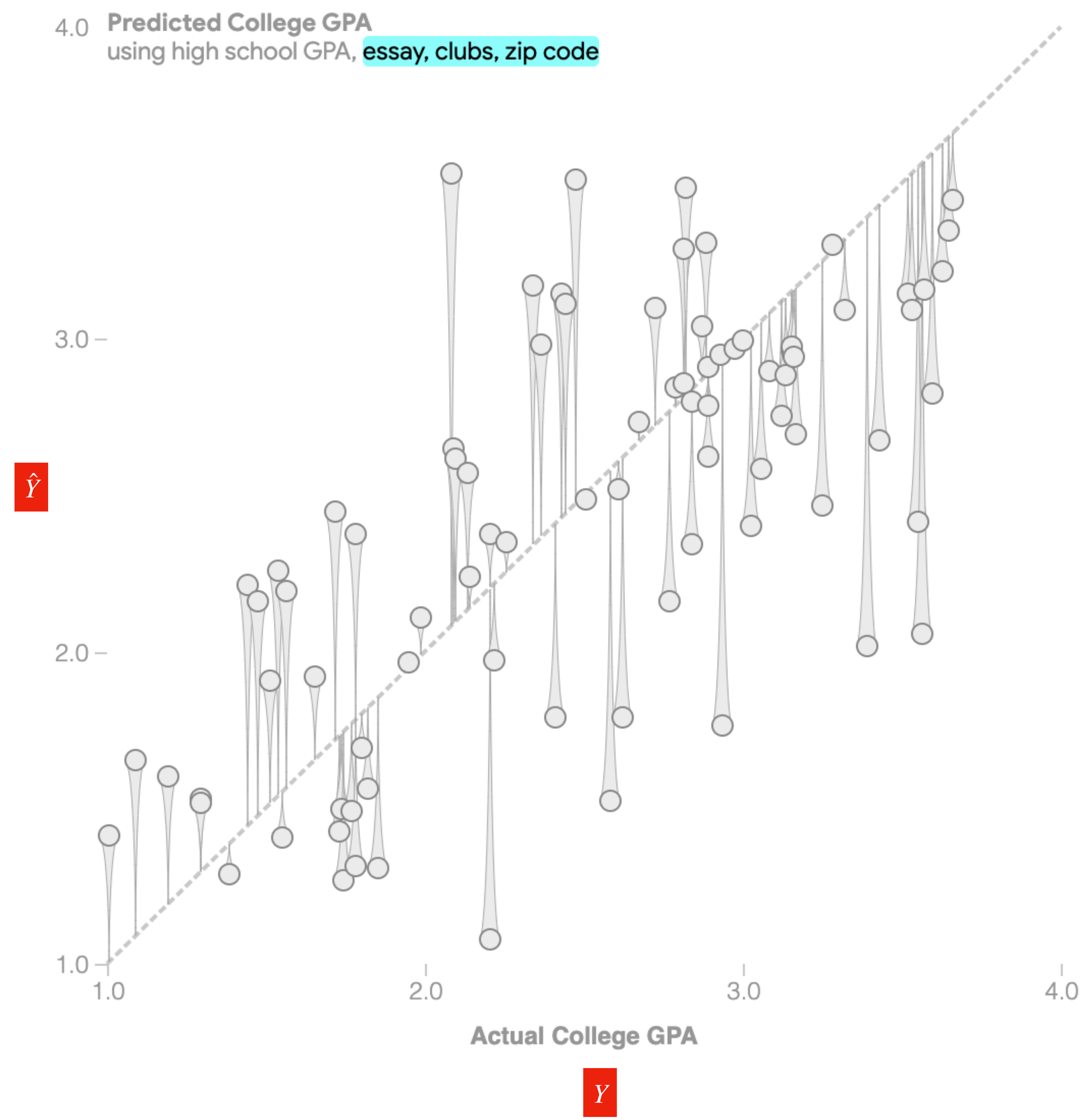
Y

https://pair.withgoogle.com/explorables/hidden-bias/

Example

Predictor $\hat{y} = \mathbb{E}(y \mid x_1, x_2, \ldots, x_K)$

Giving the model more information about students increases accuracy more...

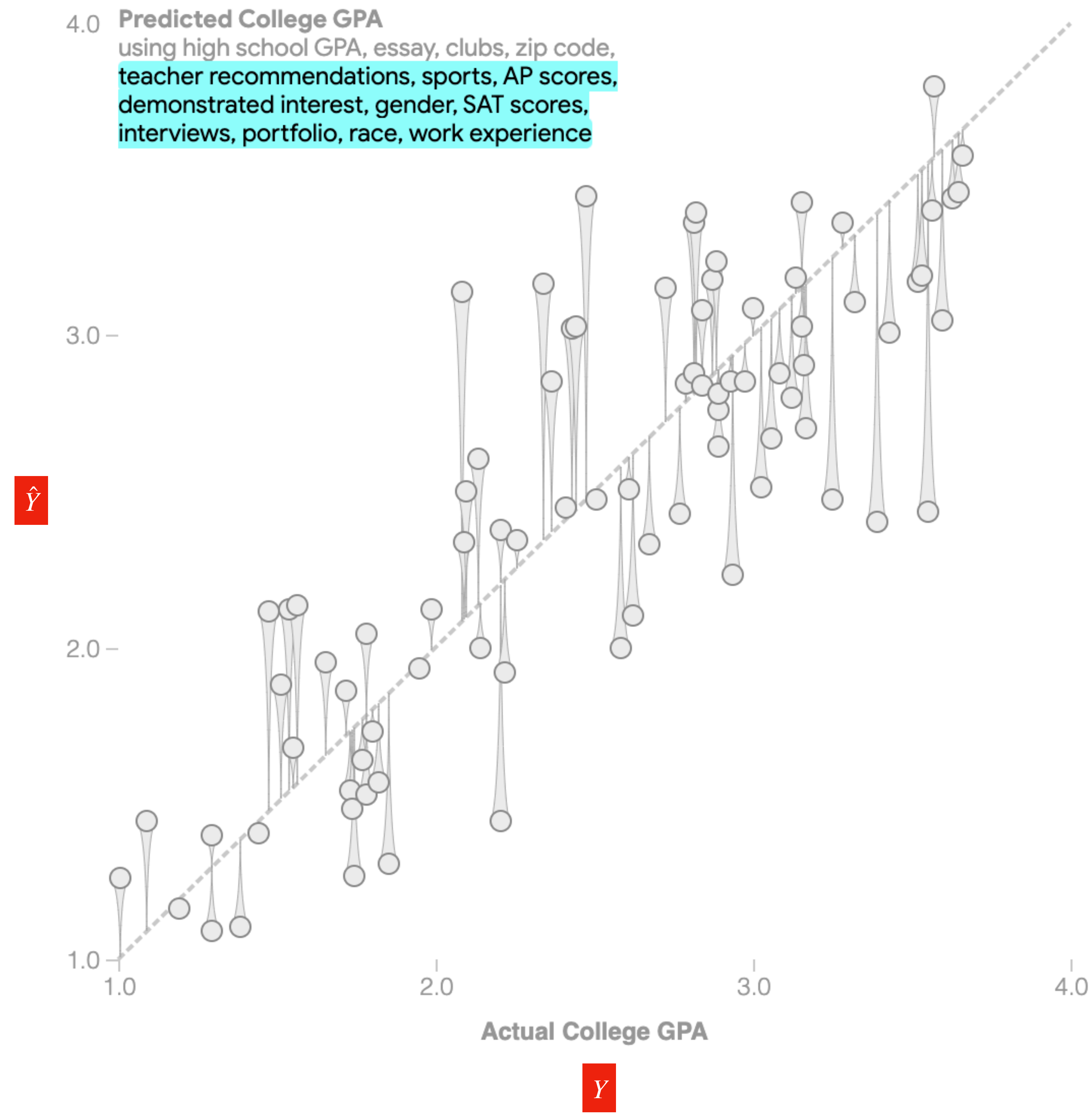


https://pair.withgoogle.com/explorables/hidden-bias/

Example

...and more.

Predictor $\hat{y} = \mathbb{E}(y \mid x_1, x_2, \ldots, x_K, \ldots, x_n)$



https://pair.withgoogle.com/explorables/hidden-bias/

Example

Models can encode previous bias

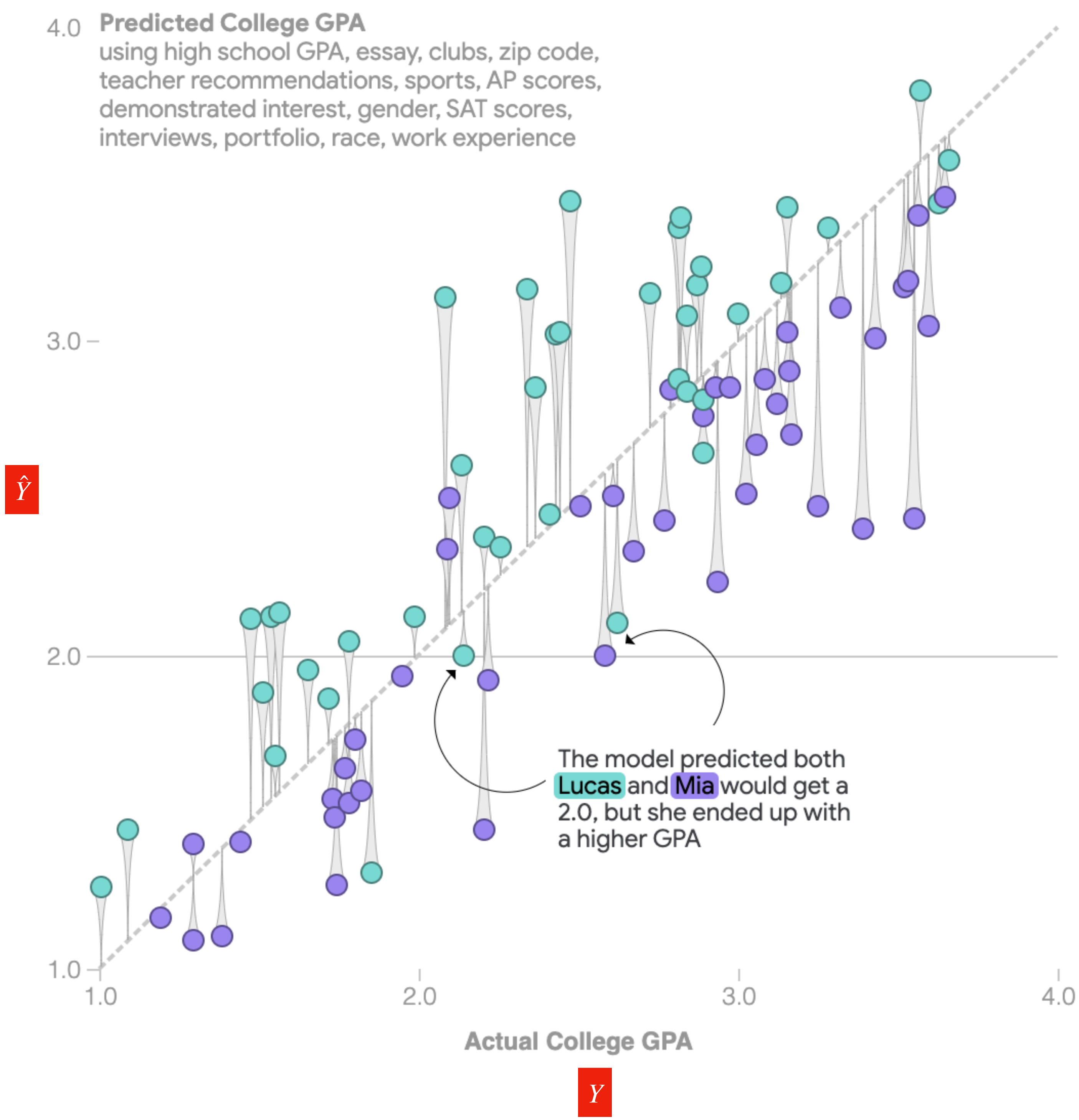
All of this sensitive information about students is just a long list of numbers to model.

If a sexist college culture has historically led to lower grades for female students, the model will pick up on that correlation and predict lower grades for women.

Training on historical data bakes in historical biases. Here the sexist culture has improved, but the model learned from the past correlation and still predicts higher grades for men.

Overpredicting outcomes for men and underpredicting for women can have significant ethical implications, but it is not the only unethical bias we can found.

Higher outcome variance for women can also be an issue from the perspective of "quality of service".



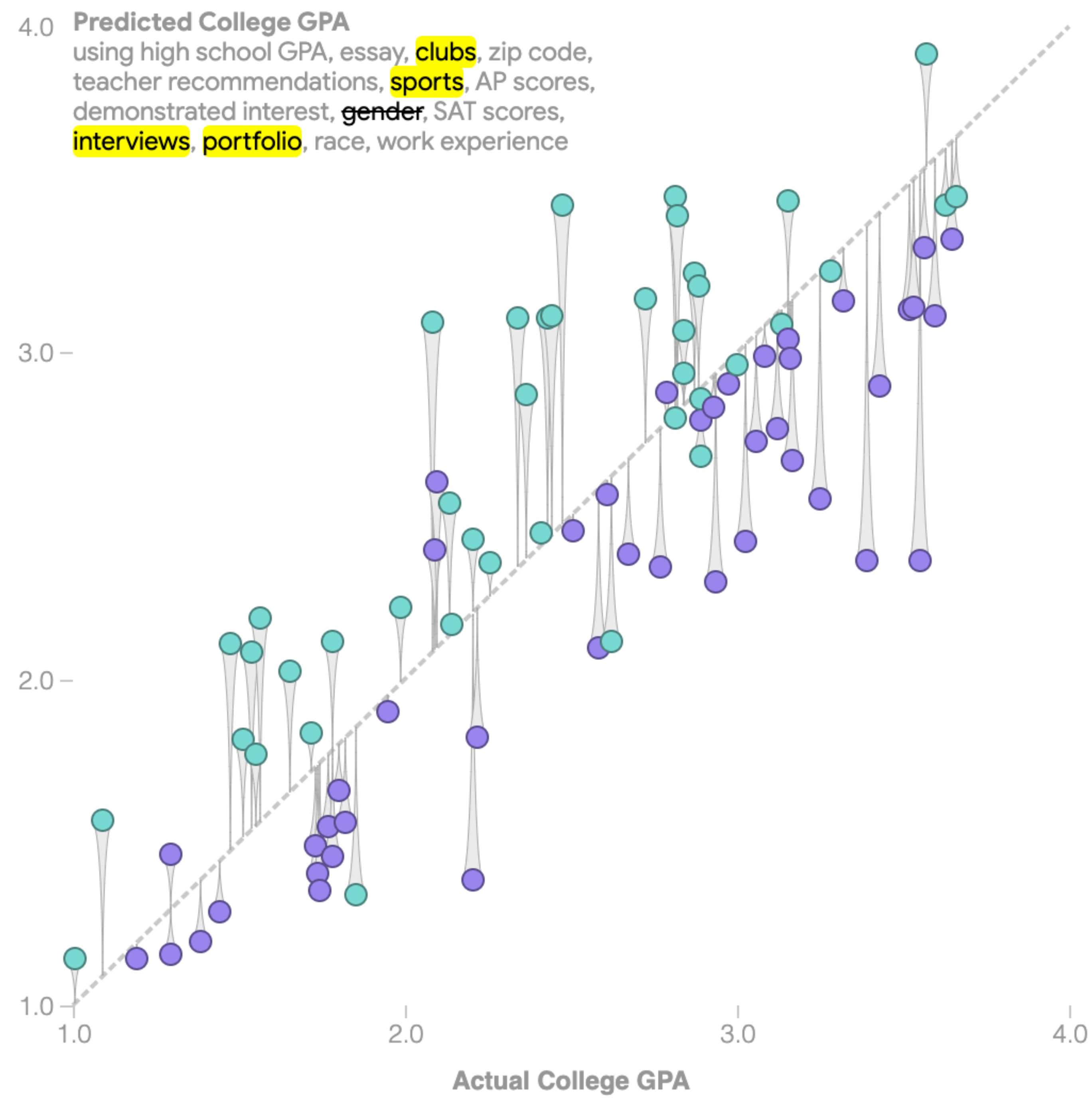
https://pair.withgoogle.com/explorables/hidden-bias/

Example

Hiding protected classes from the model might not stop discrimination

Even if we don't tell the model students' genders, it might still score female students poorly.

With detailed enough information about every student, the model can still synthesize a proxy for gender out of other variables.



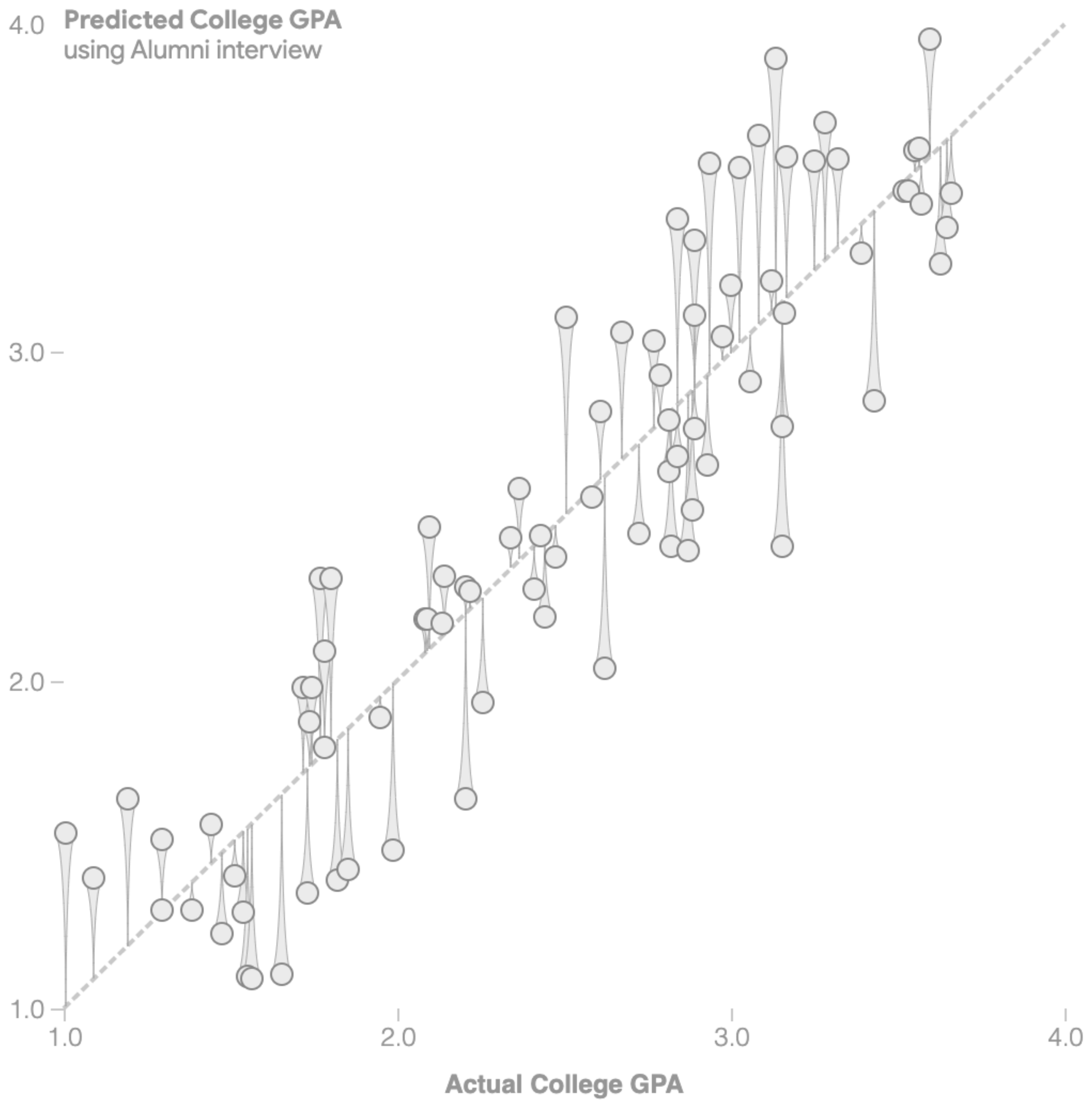
https://pair.withgoogle.com/explorables/hidden-bias/

Example

Predictor $\hat{y} = \mathbb{E}(y \mid x)$

Including a protected attribute may even *decrease* discrimination

Let's look at a simplified model, one only taking into account the recommendation of an alumni interviewer.



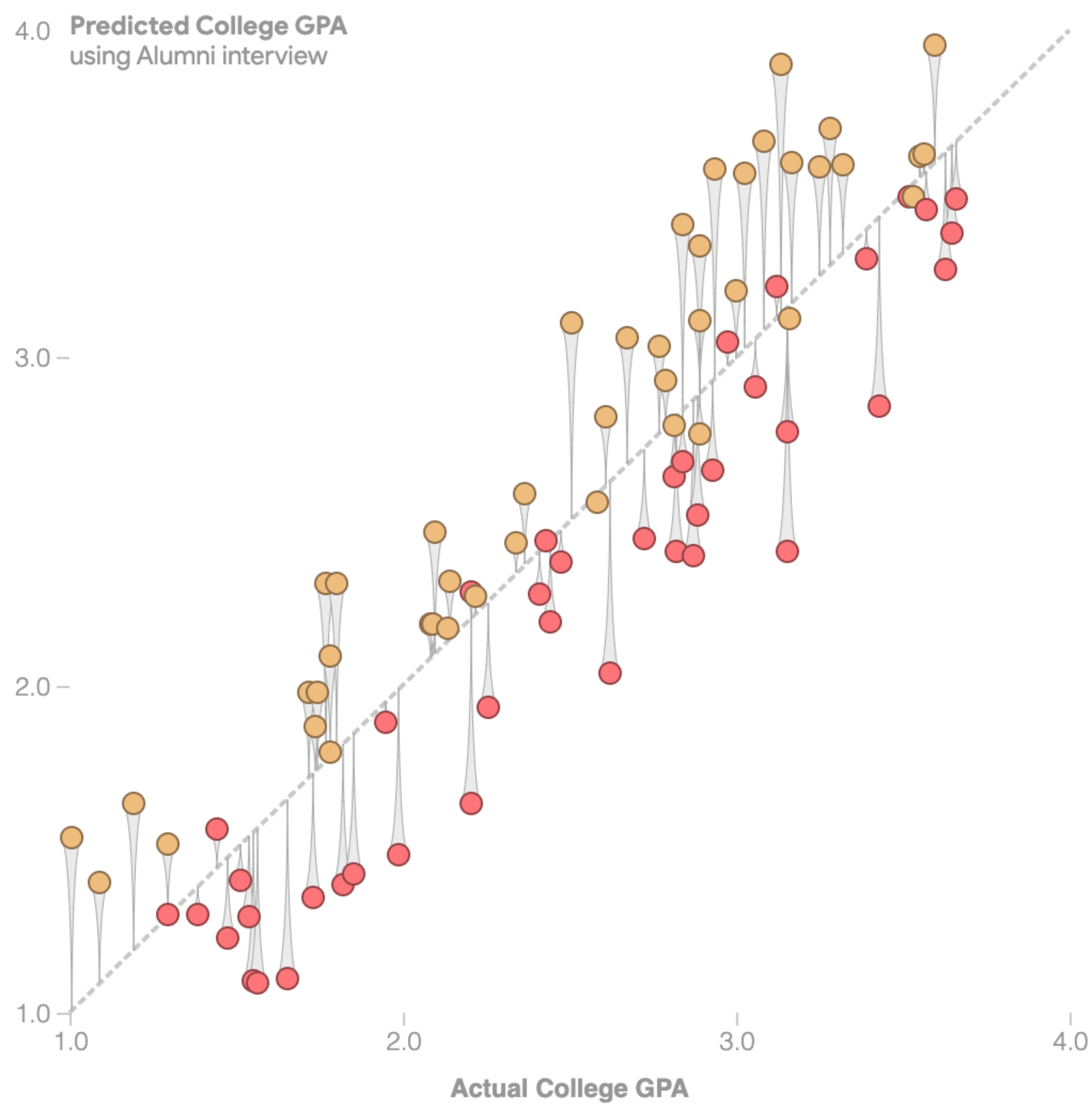
https://pair.withgoogle.com/explorables/hidden-bias/

Example

Predictor $\hat{y} = \mathbb{E}(y \mid x)$

The interviewer is quite accurate, except that they're biased against students with a low household income.

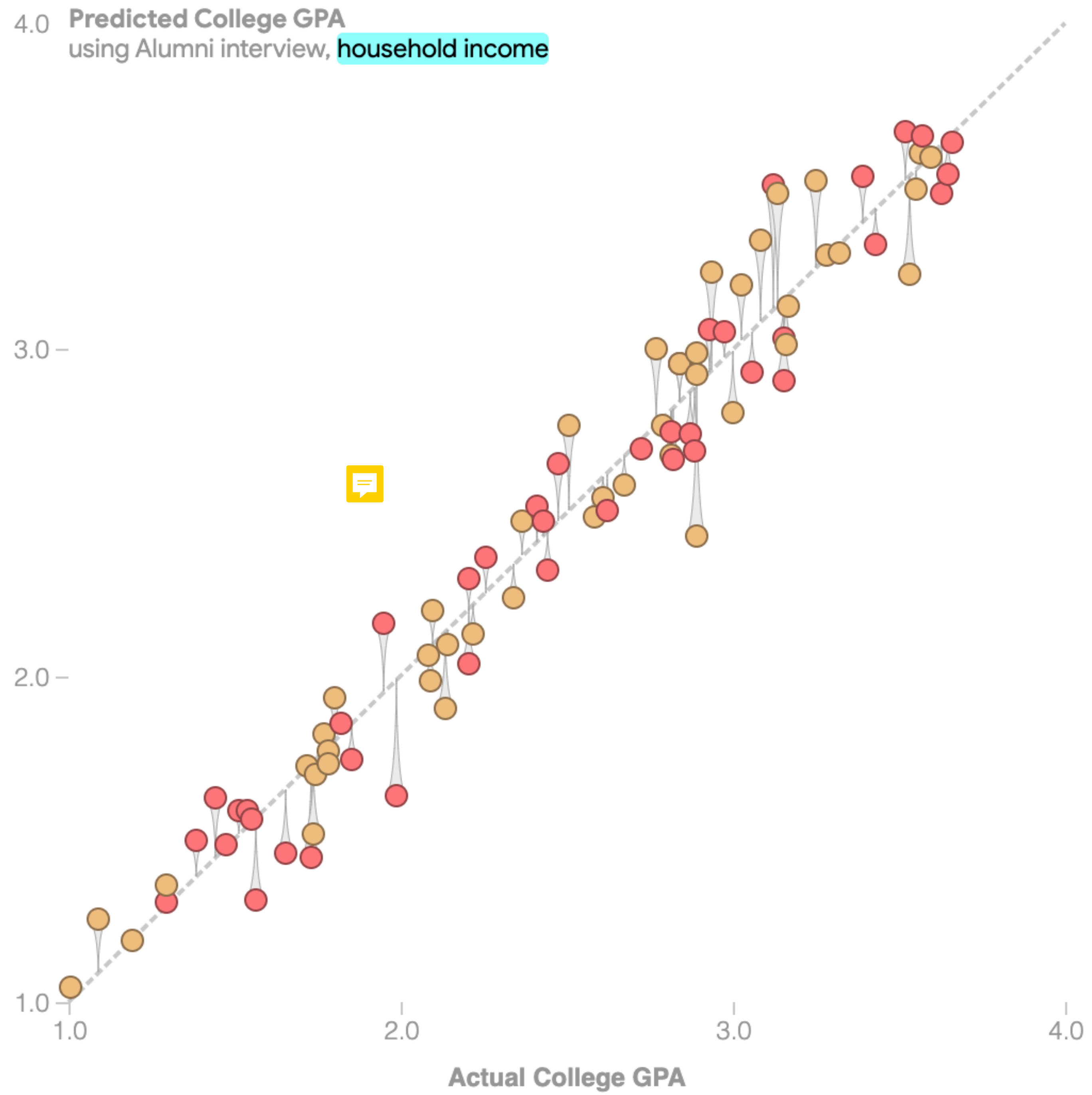
In our toy model, students' grades don't depend on their income once they're in college. In other words, we have biased inputs and unbiased outcomes—the opposite of the previous example, where the inputs weren't biased, but the toxic culture biased the outcomes.



https://pair.withgoogle.com/explorables/hidden-bias/

Example

Predictor $\hat{y} = \mathbb{E}(y \,|\, x_1, x_2, \ldots, x_K)$



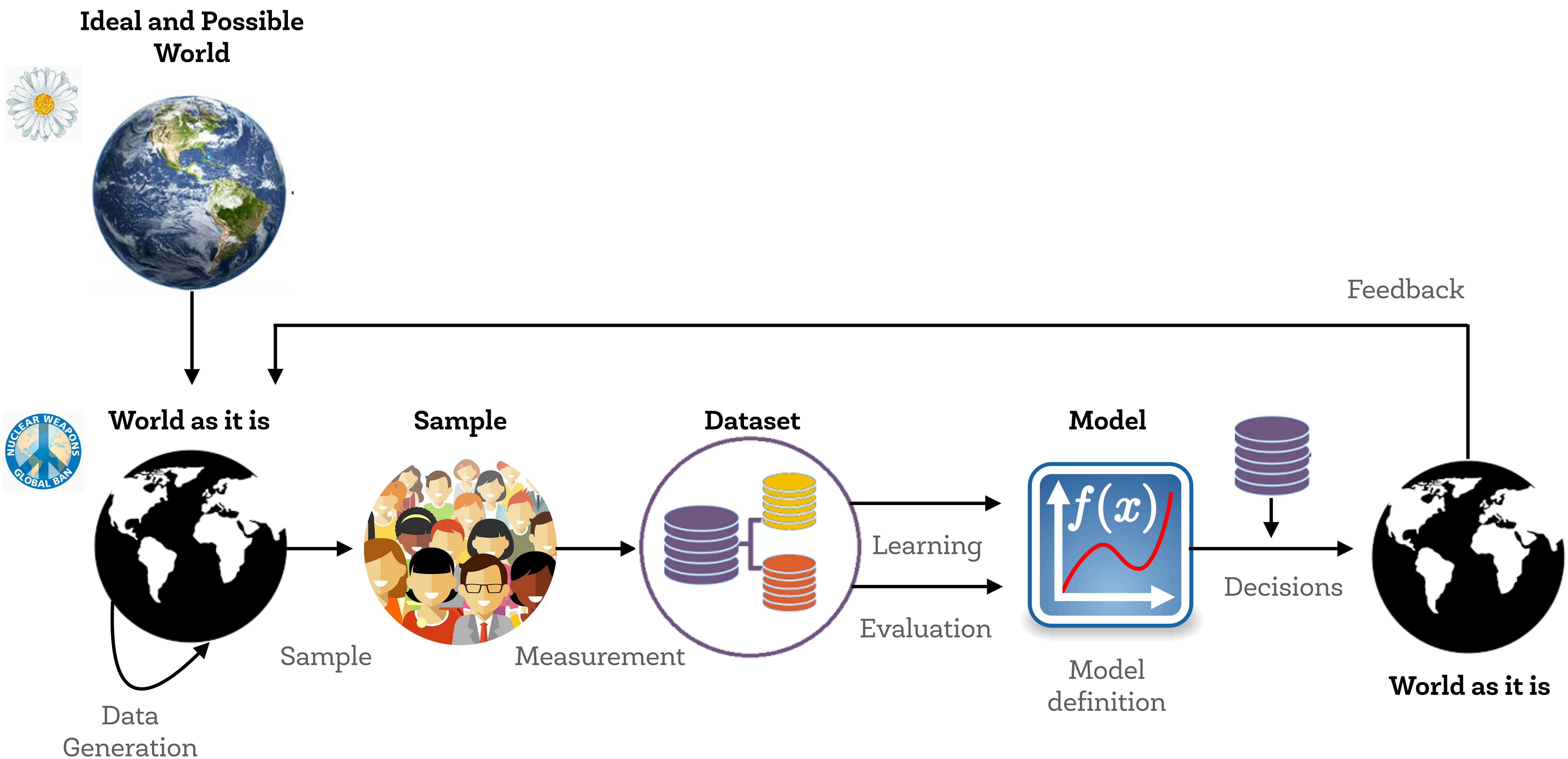
If we also tell the model each student's household income, it will naturally correct for the interviewer's overrating of high-income students just like it corrected for the difference between high school and college GPAs.

By carefully considering and accounting for bias, we've made the model fairer and more accurate. This isn't always easy to do, especially in circumstances like the historically toxic college culture where unbiased data is limited.

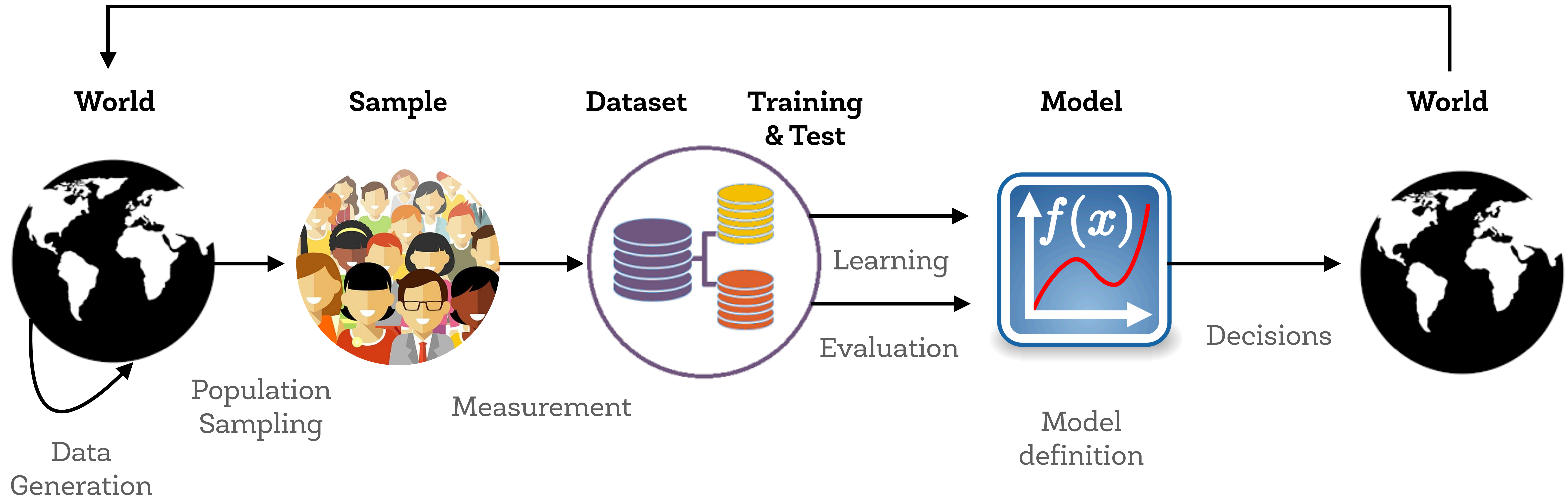
https://pair.withgoogle.com/explorables/hidden-bias/

Sources of Bias

ML model life cycle



Sources of Bias



Structural Bias

Structural bias refers to the social/ institutional **patterns and practices** that confer advantage to some and disadvantage to others based on identity.

Representation Bias

The data we use do not represent the population,

Measurement Bias

The features we use do not represent the phenomenon we are studying,

Aggregation Bias

Aggregation bias occurs when groups are inappropriately combined, resulting in a model that does not perform well for any group.

Simpson's Paradox

Inductive Bias

Inductive bias describes the tendency for a system to prefer a certain set of generalizations over others that are equally consistent with the observed data.

Evaluation Bias

When evaluating a model, metrics calculated against an entire test or validation set don't always give an accurate picture of how the model works.

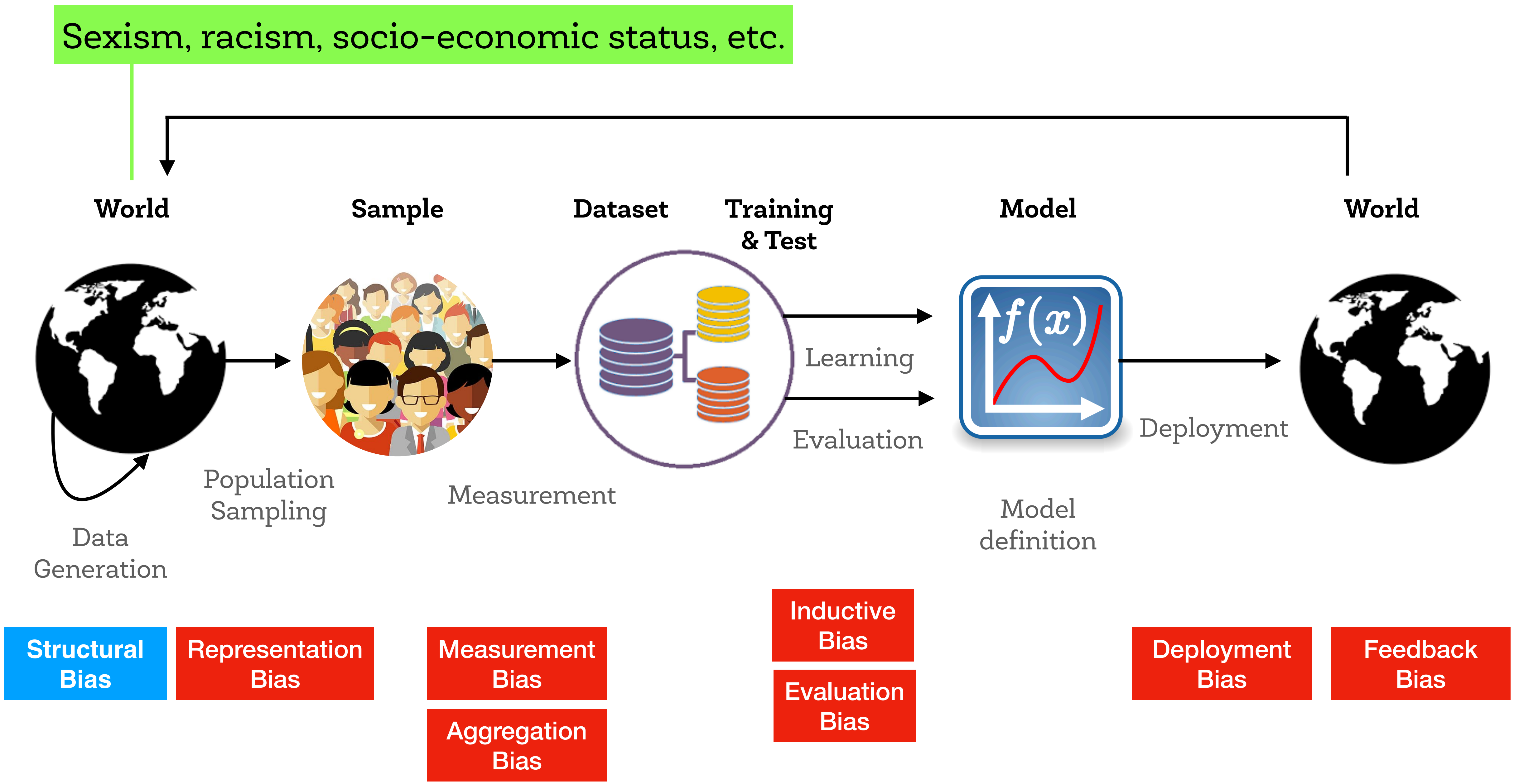
Deployment Bias

Deployment bias refers, generally, **to any bias that arises during deployment**, where a system is used or interpreted in inappropriate ways, perhaps not intended by the designers or developers.

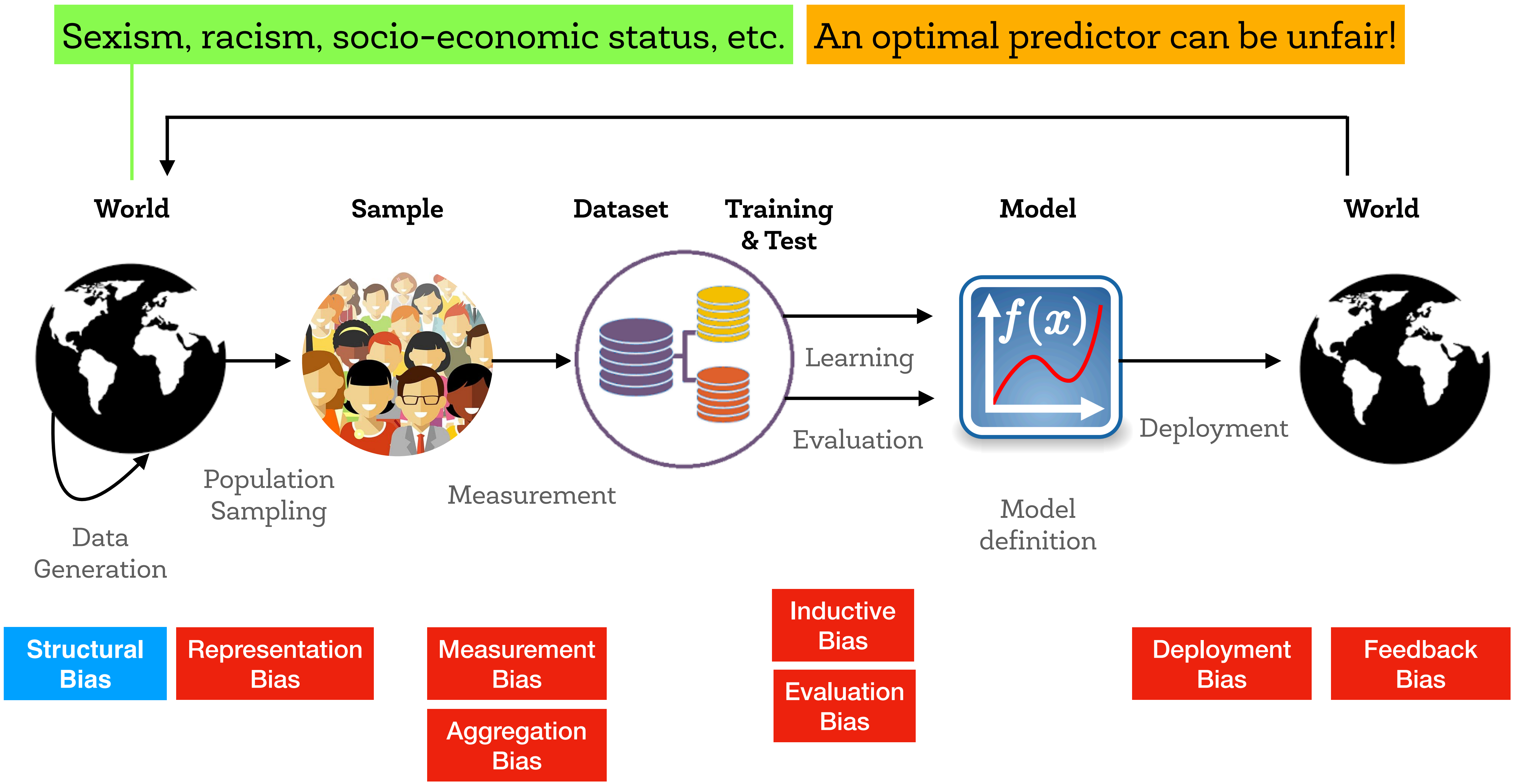
Feedback Bias

This type of bias occurs when a model itself influences the generation of data that is used to train it.

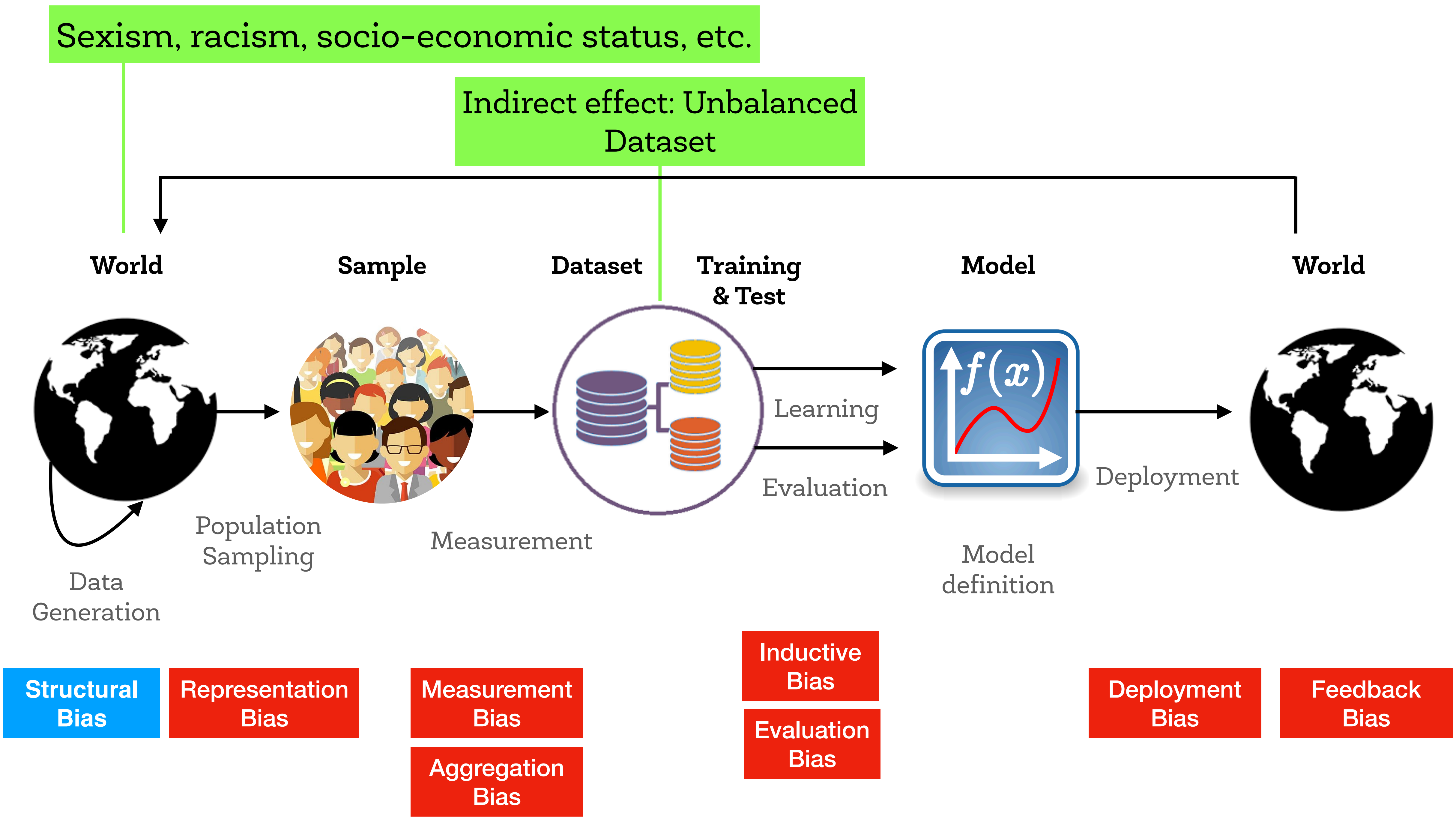
Sources of Bias



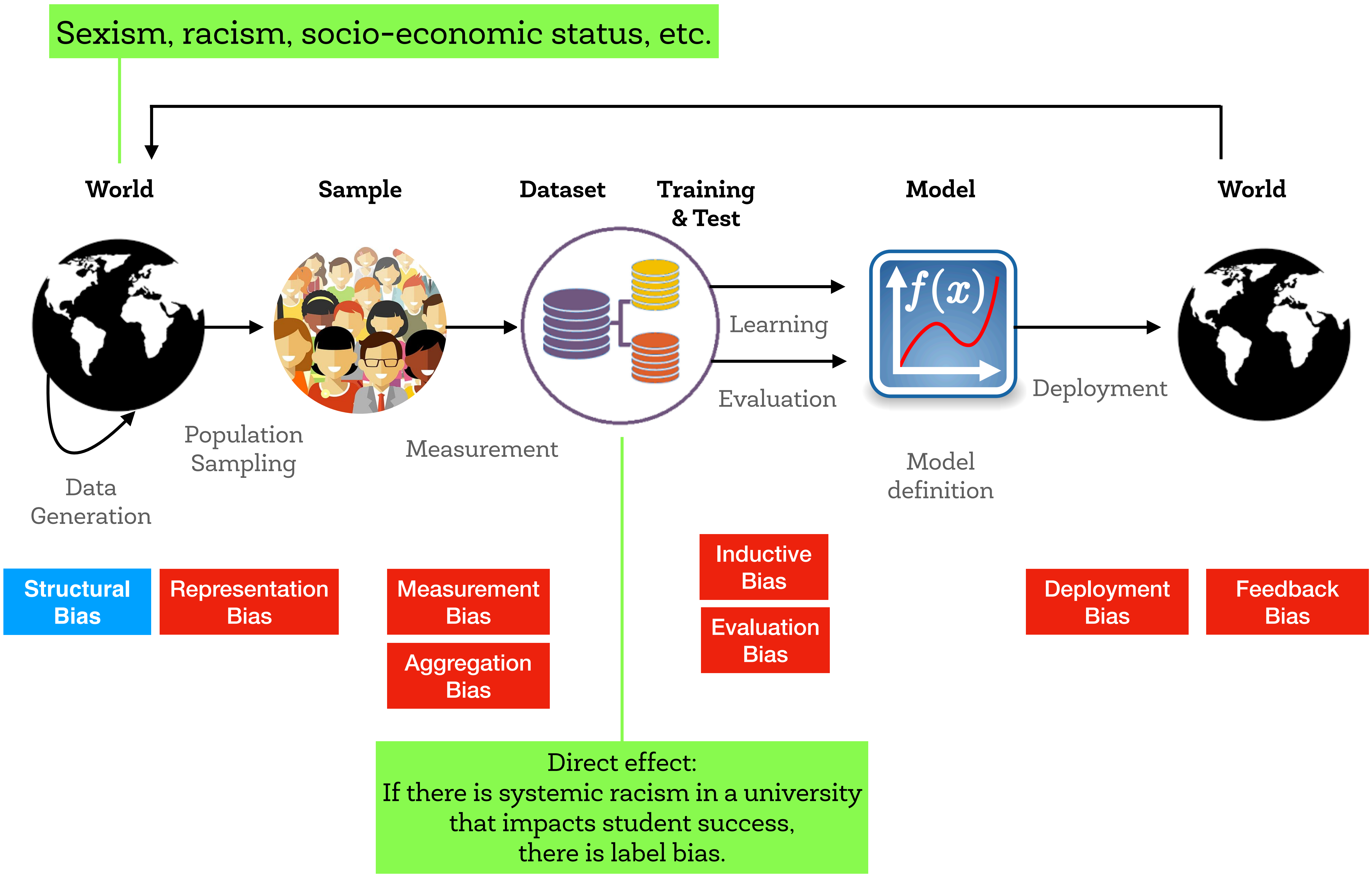
Sources of Bias



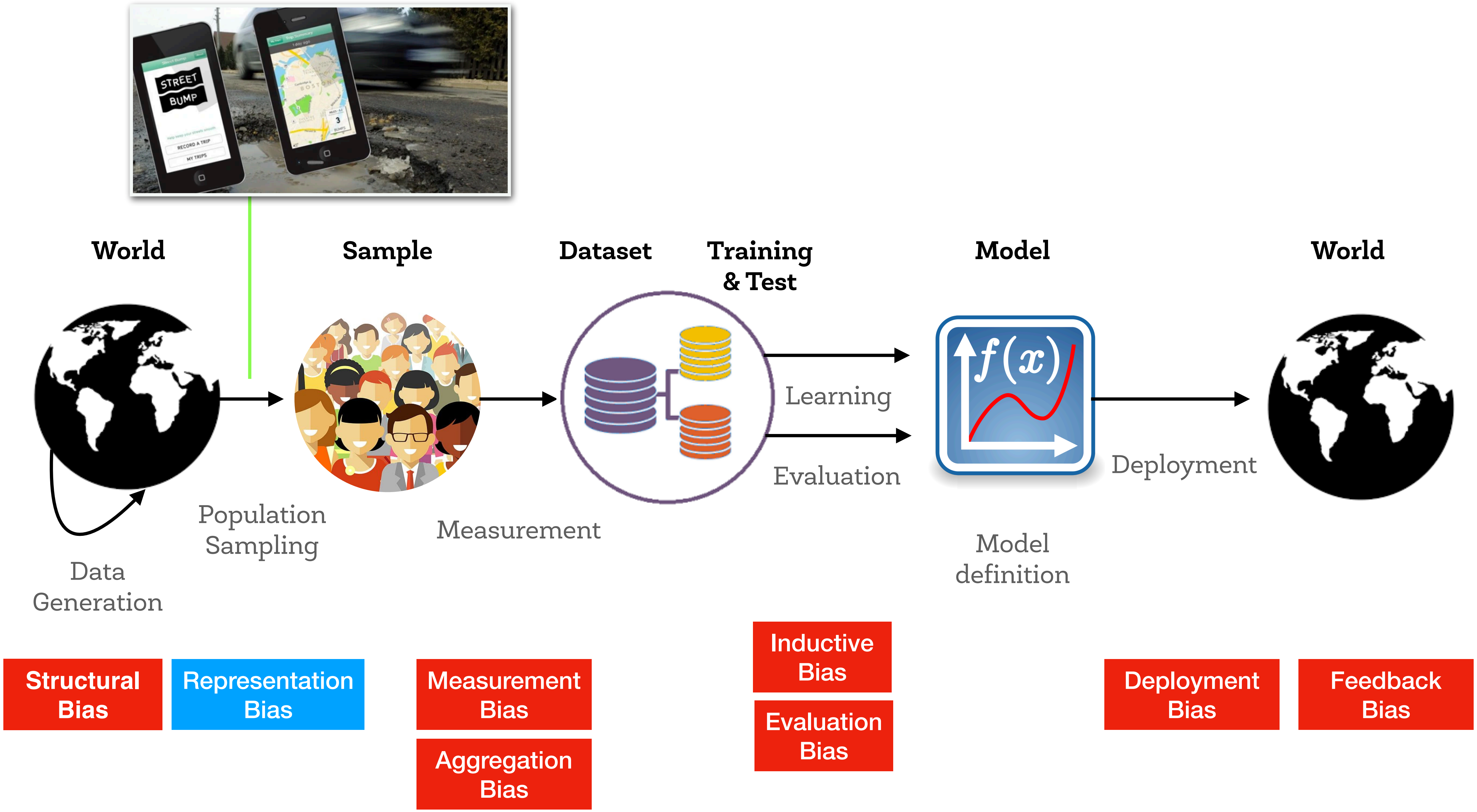
Sources of Bias



Sources of Bias

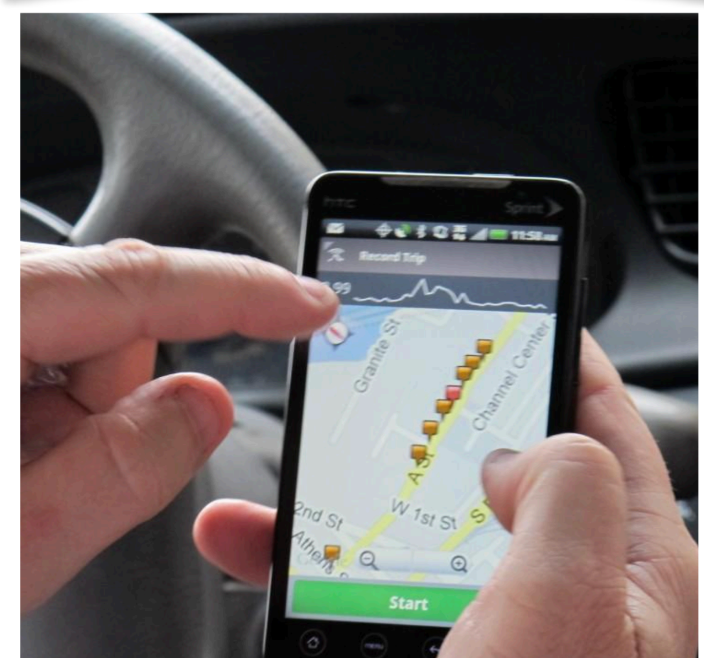


Sources of Bias



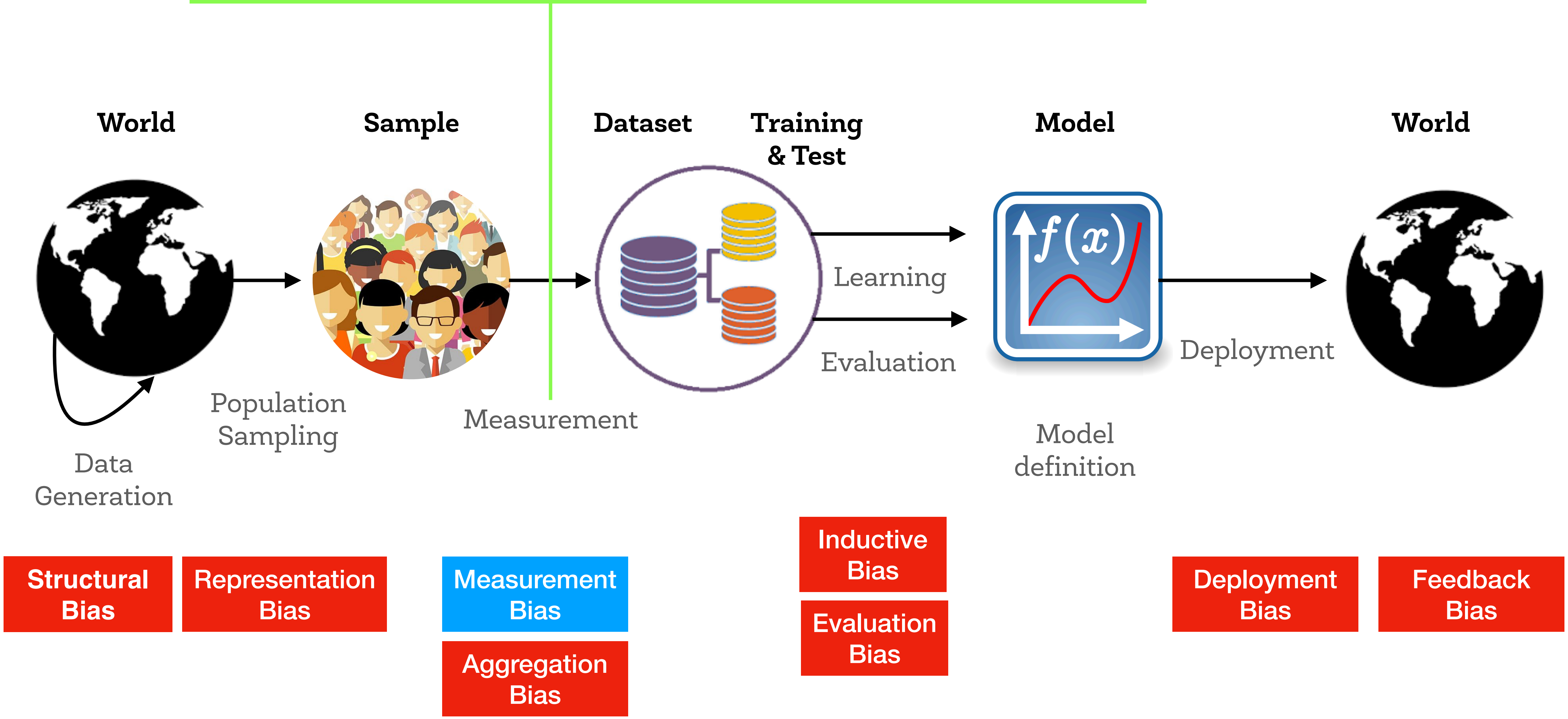
What about applications that aren't about people?

Consider "Street Bump," a project by the city of Boston to crowdsource data on potholes.
The smartphone app automatically detects **pot holes** using data from the smartphone's sensors and sends the data to the city. Infrastructure seems like a comfortably boring application of data-driven decision-making, far removed from the ethical quandaries we've been discussing.
But the data reflects **terms of smartphone ownership**, which are higher in wealthier parts of the city compared to lower-income areas and areas with large elderly populations.



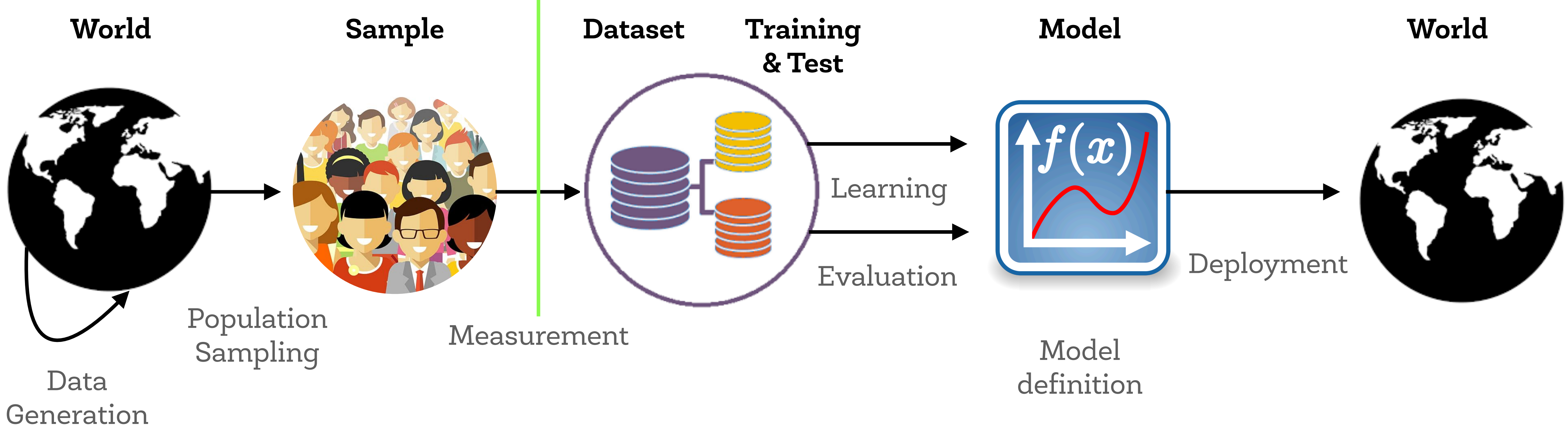
Sources of Bias

Student success can be specified in terms of many different variables that do not represent in a fair way all groups: grades, employer prestige, post-graduate salary, etc.



Sources of Bias

Student success can be specified in terms of many different variables that do not represent in a fair way all groups: grades, employer prestige, post-graduate salary, etc.



Structural Bias

Representation Bias

Measurement Bias

Aggregation Bias

Inductive Bias

Evaluation Bias

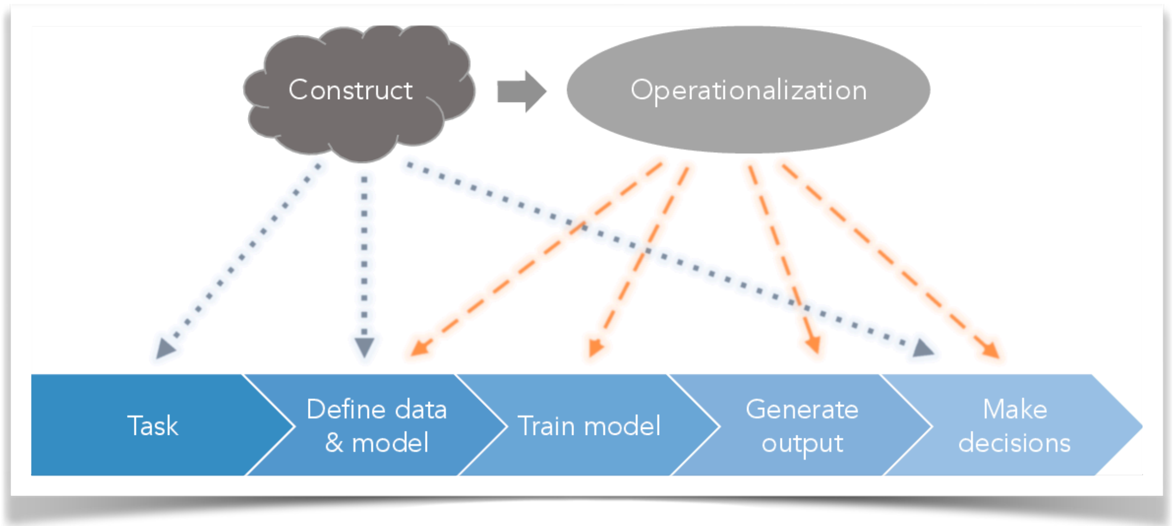
Deployment Bias

Feedback Bias

Reality Gap: "real" data are not necessarily equivalent to "reality"

The features we use determine what real patterns can be detected!

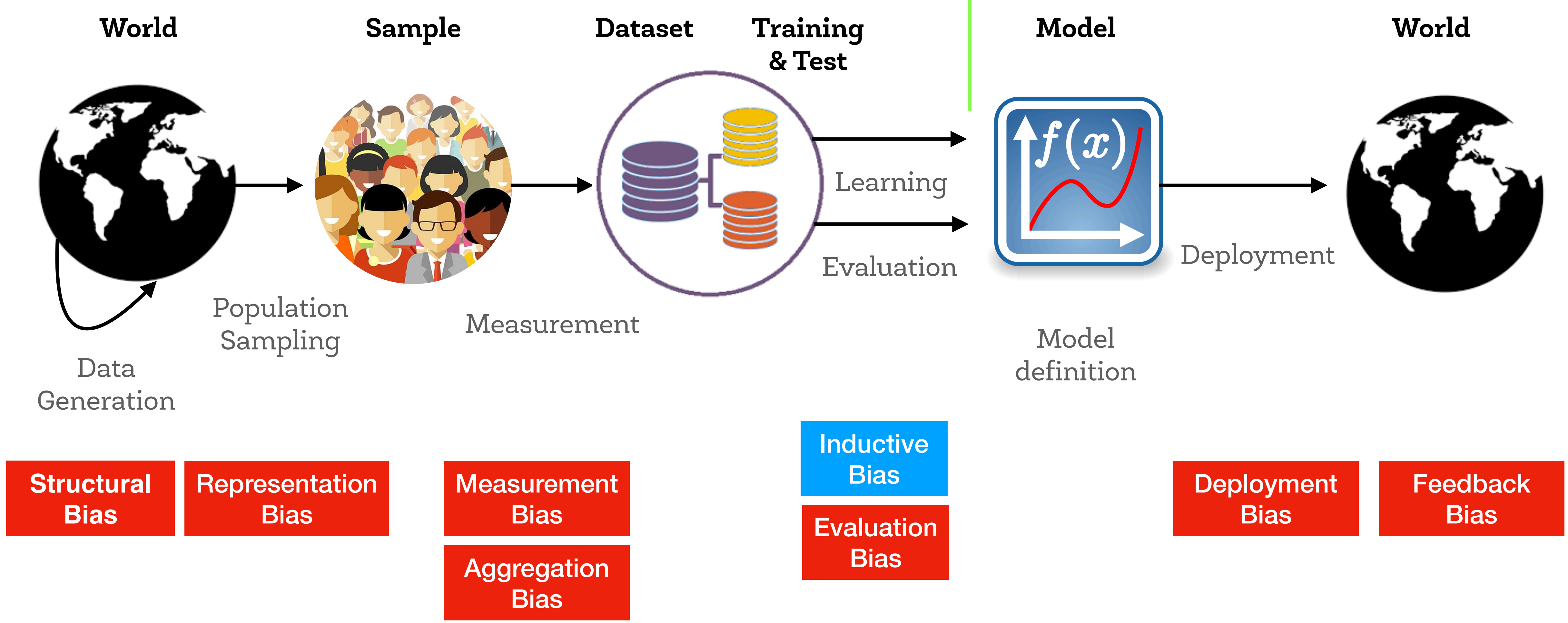
Measuring almost any **attribute about people** is similarly subjective and challenging: *teacher effectiveness*, *economic status*, etc.



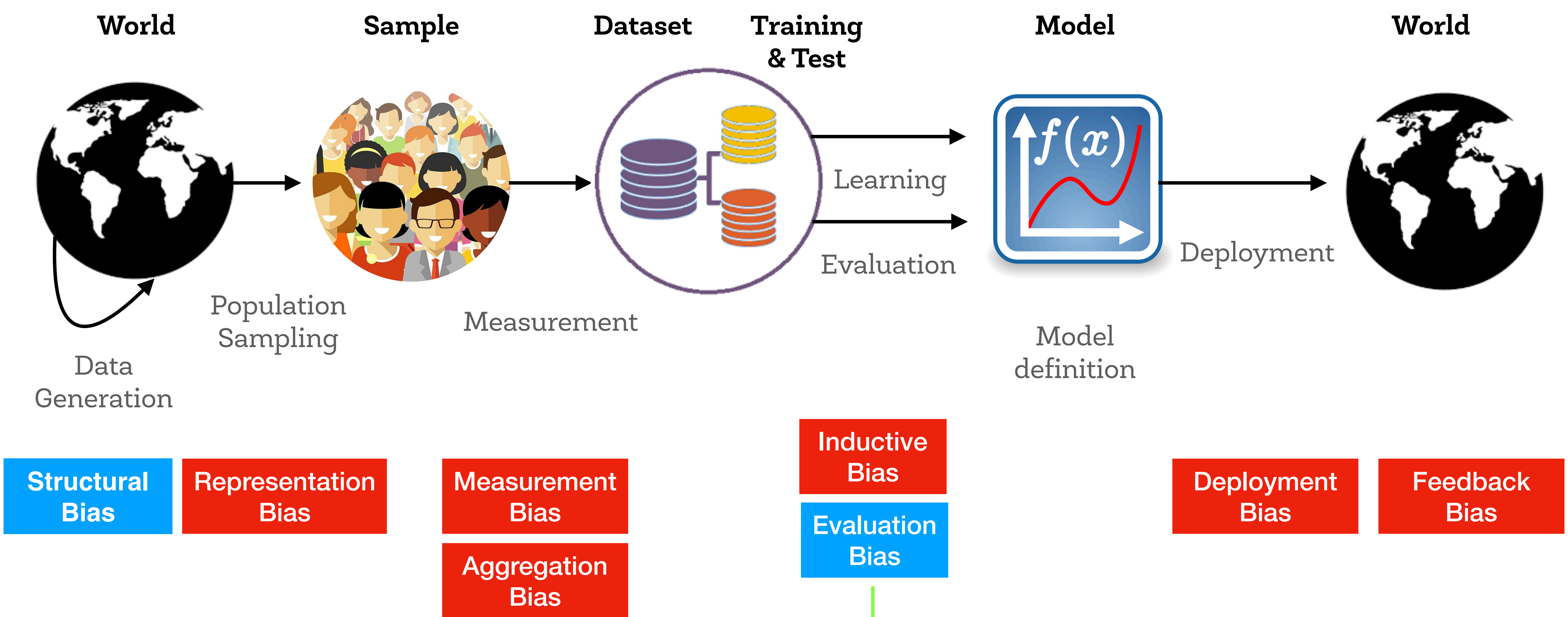
Recommended Reading: **Measurement and Fairness**, by Abigail Z. Jacobs, Hanna Wallach https://arxiv.org/abs/1912.05511

Sources of Bias

The choice of the objective function is not ethically neutral: we could minimize "prediction errors" or minimize "predictions of no-help-needed for students who truly need help".



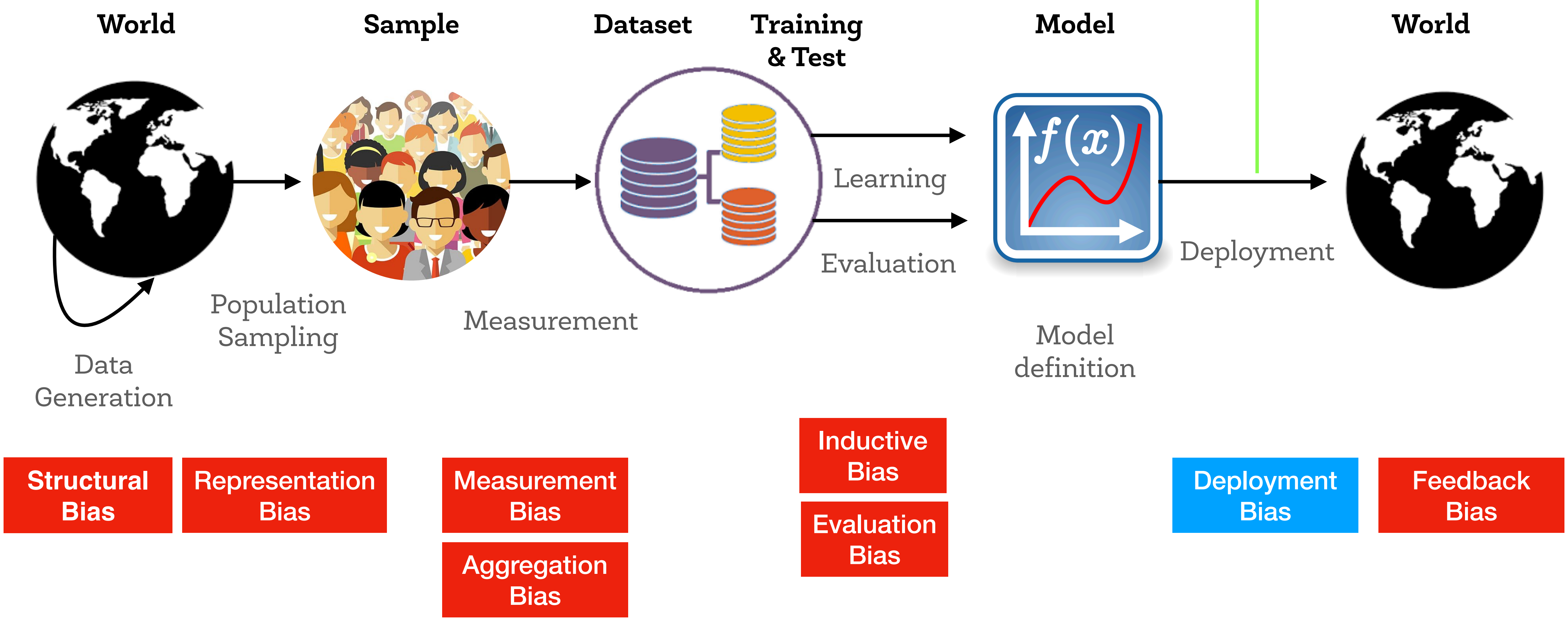
Sources of Bias



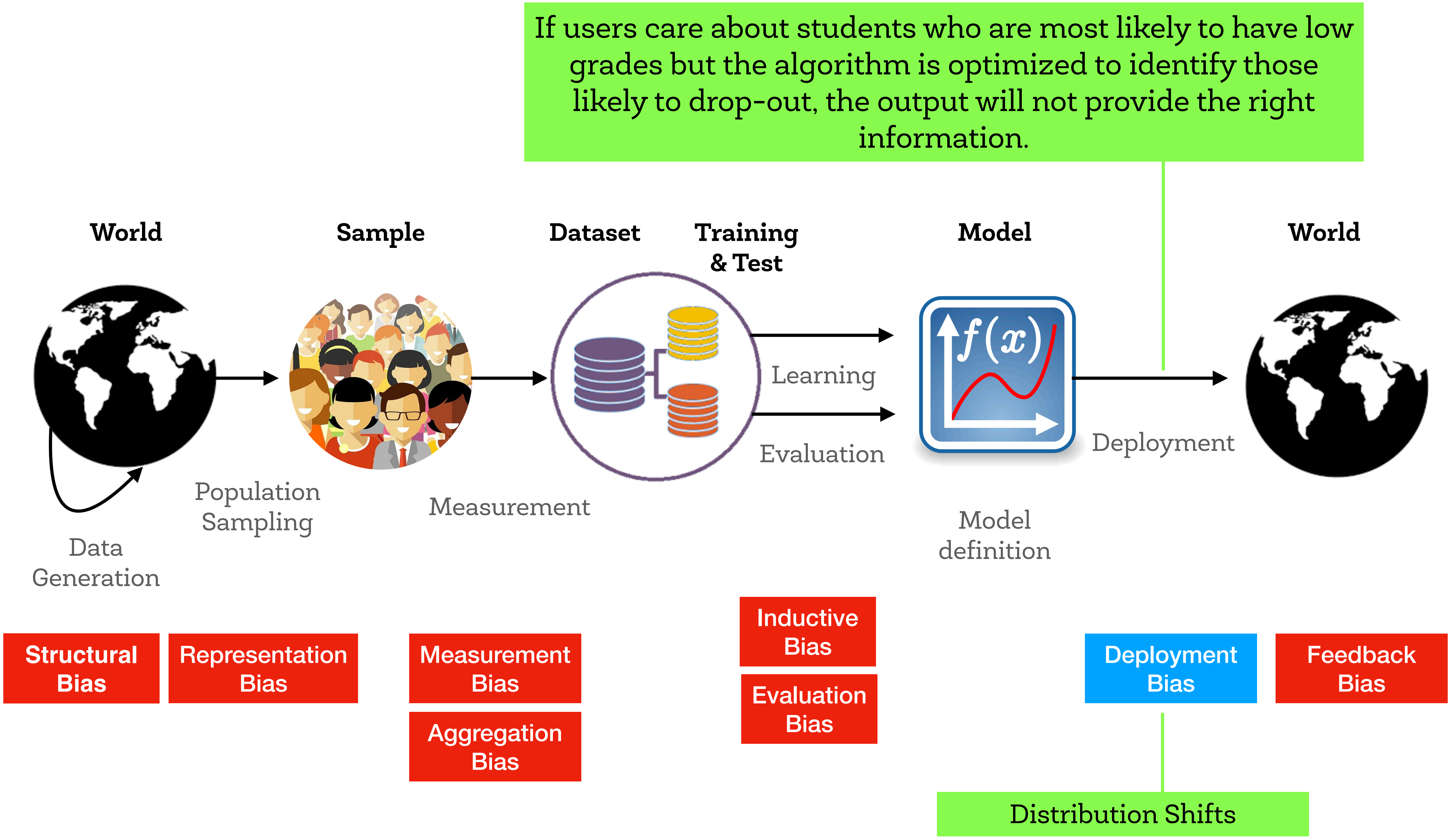
Sometimes, there is no “ground truth” to label data. In many applications the system can do little more than mimic the agreement amongst domain “experts”.

Sources of Bias

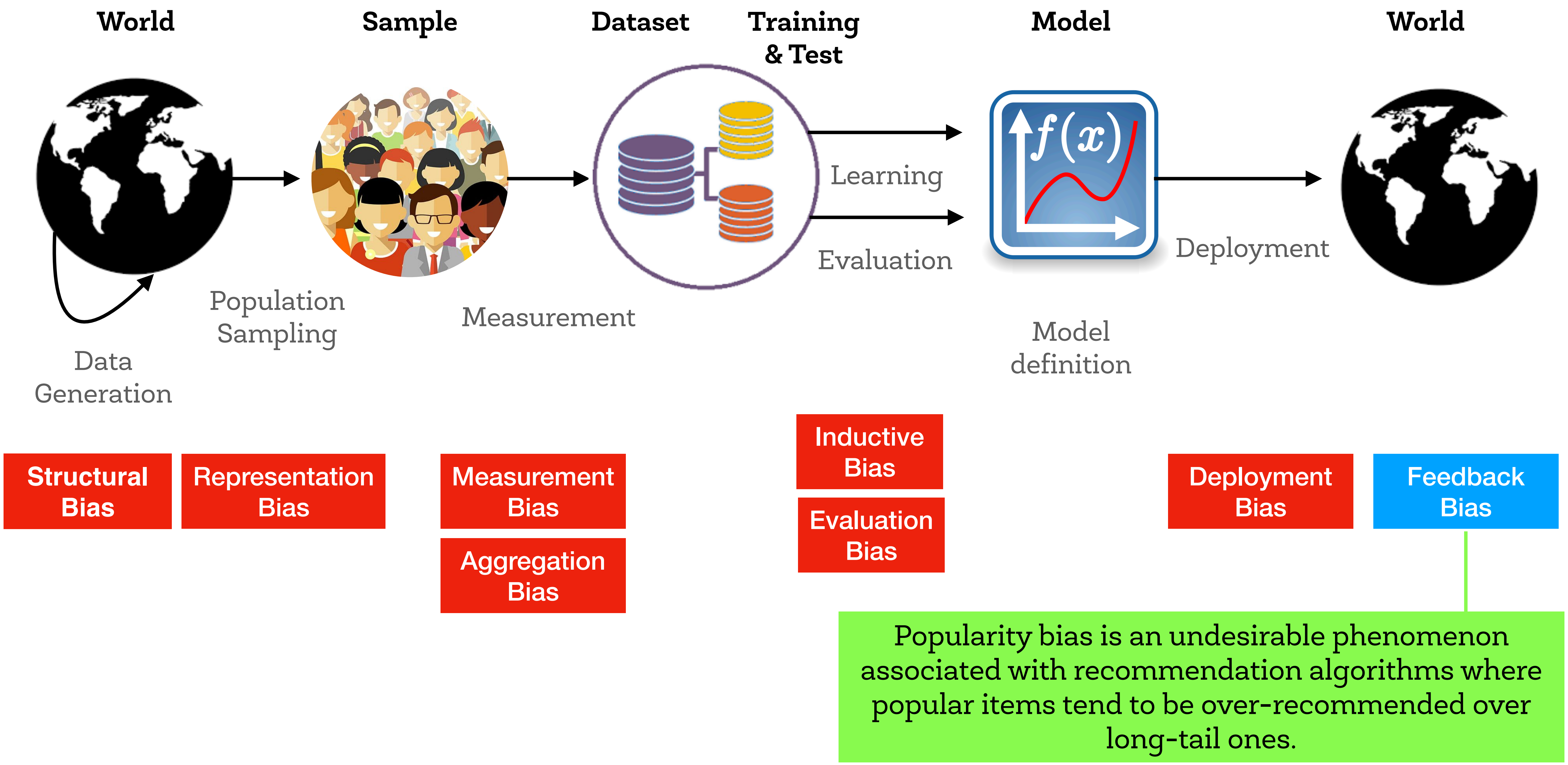
If users care about students who are most likely to have low grades but the algorithm is optimized to identify those likely to drop-out, the output will not provide the right information.



Sources of Bias



Sources of Bias



Every Bias Is Not a Bad Bias

We have seen **bad biases**, biases that are problematic from an ethical point of view because **they configure the distribution of goods, services, risks, and opportunities, or even access to information in ways that are problematic**.

But there are biases that are **inevitable**, that **enable** ML.

Every Bias Is Not a Bad Bias

Bias is a need to generalize!

The Need for Biases in Learning Generalizations

Tom M. Mitchell

1. Introduction

Learning involves the ability to generalize from past experience in order to deal with new situations that are "related to" this experience. The inductive leaap needed to deal with new situations seems to be possible only under certain biases for choosing one generalization of the situation over another. This paper defines precisely the notion of bias in generalization problems, then shows that biases are necessary for the inductive leap. Classes of justifiable biases are considered, and the relationship between bias and domain-independence is considered.

We restrict the scope of this discussion to the problem of generalizing from training instances, defined as follows:

The Generalization Problem
Given:

1. Language of instances.
2. Language of generalizations.
3. Matching predicate for matching generalizations to instances.
4. Sets of positive and negative training instances.

Determine:

⇒ Generalization(s) consistent with the training instances.

As a concrete example of the above generalization problem, consider the task addressed by Winston's program for learning classes of block structures (Winston 1975). Here, the language of instances is the representation used to describe example block structures. The language of generalizations is the language in which learned concepts (e.g., arch, tower) are described. The matching predicate specifies whether a given generalization applies to a given instance (e.g., whether the inferred description of an arch is satisfied by a specific block structure).

This paper addresses a deep difficulty with the generalization problem as defined above: If consistency with the training instances is taken as the sole determiner of appropriate generalizations, then a program can never make the inductive leap necessary to classify instances beyond those it has observed. Only if the program has other sources of information, or biases for choosing one generalization over the other, can it non-arbitrarily classify instances beyond those in the training set.

In this paper, we use the term *bias* to refer to *any basis for choosing one generalization over another, other than strict consistency with the observed training instances*.

Converted to electronic version by: Roby Joehanes, Kansas State University

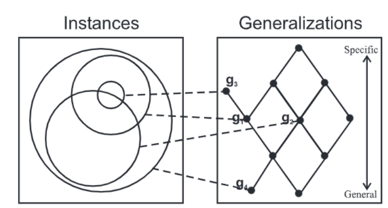


Figure 1: Relationships among Instances and Generalizations (This figure was missing from the original publication and added in 1990.)

2. What is an UNbiased Generalizer

If generalization is the problem of guessing the class of instances to which the positive training instances belong, then an unbiased generalizer is one that makes no a priori assumptions about which classes of instances are most likely, but bases all its choices on the observed data. Two common sources of bias in existing learning systems are (1) the generalization language is not capable of expressing all possible classes of instances, and (2) the generalization procedure that searches through the space of expressible generalizations is itself biased.

2.1. An Unbiased Generalization Language

In considering bias in the generalization language, it is useful to view each generalization as denoting the set of instances that it matches. In figure 1, for example, g_1 and g_2 are two generalizations expressible in some generalization language, and each matches a different subset of the instances.

Relative to a given langauge of instances, an unbiased generalization language is then one which allows describing every possible subset of these instances. In short, an

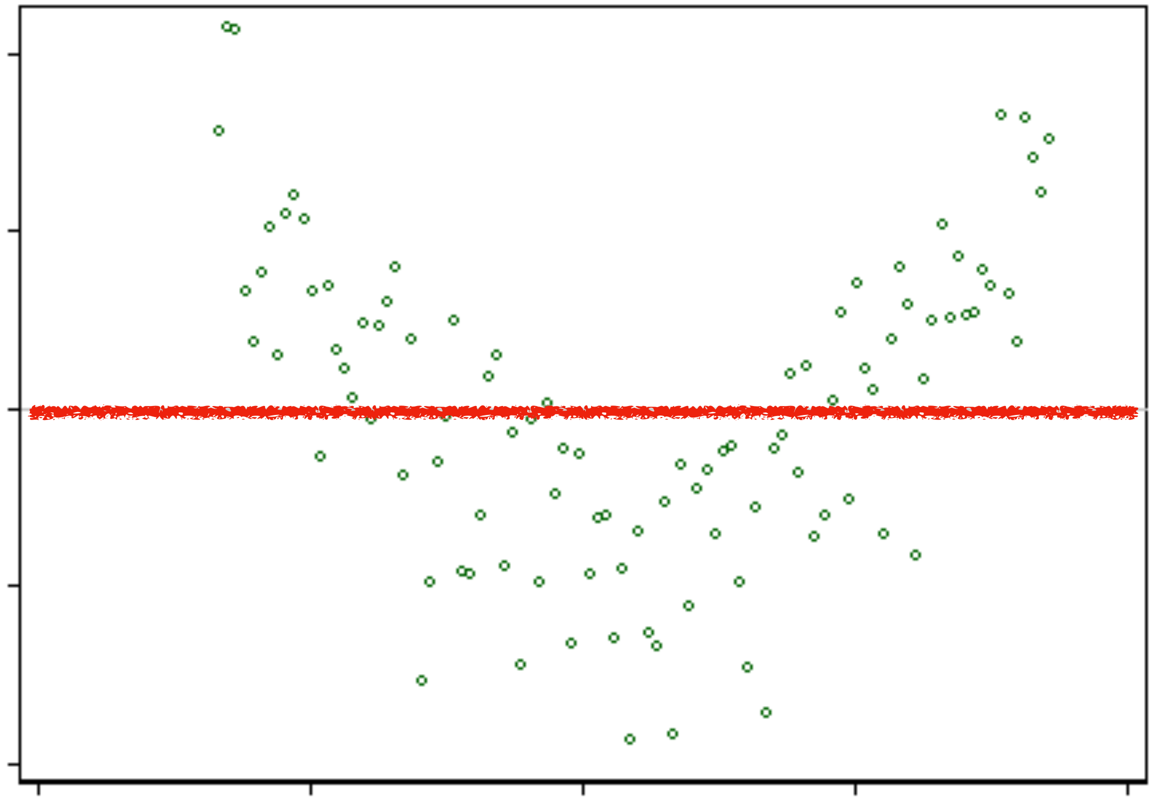
1980:
Bias in ML does help us generalize better and make our model less sensitive to some single data point.

Every Bias Is Not a Bad Bias

Definition: a **hypothesis space** is the set of mathematical functions f_W (hypotheses) that are tested against the training data, based on the **assumption that relevant (real) patters can be expressed by way of a mathematical function**, called the target function.

The learning algorithm cannot uncover patterns that are not described in one of the hypotheses.

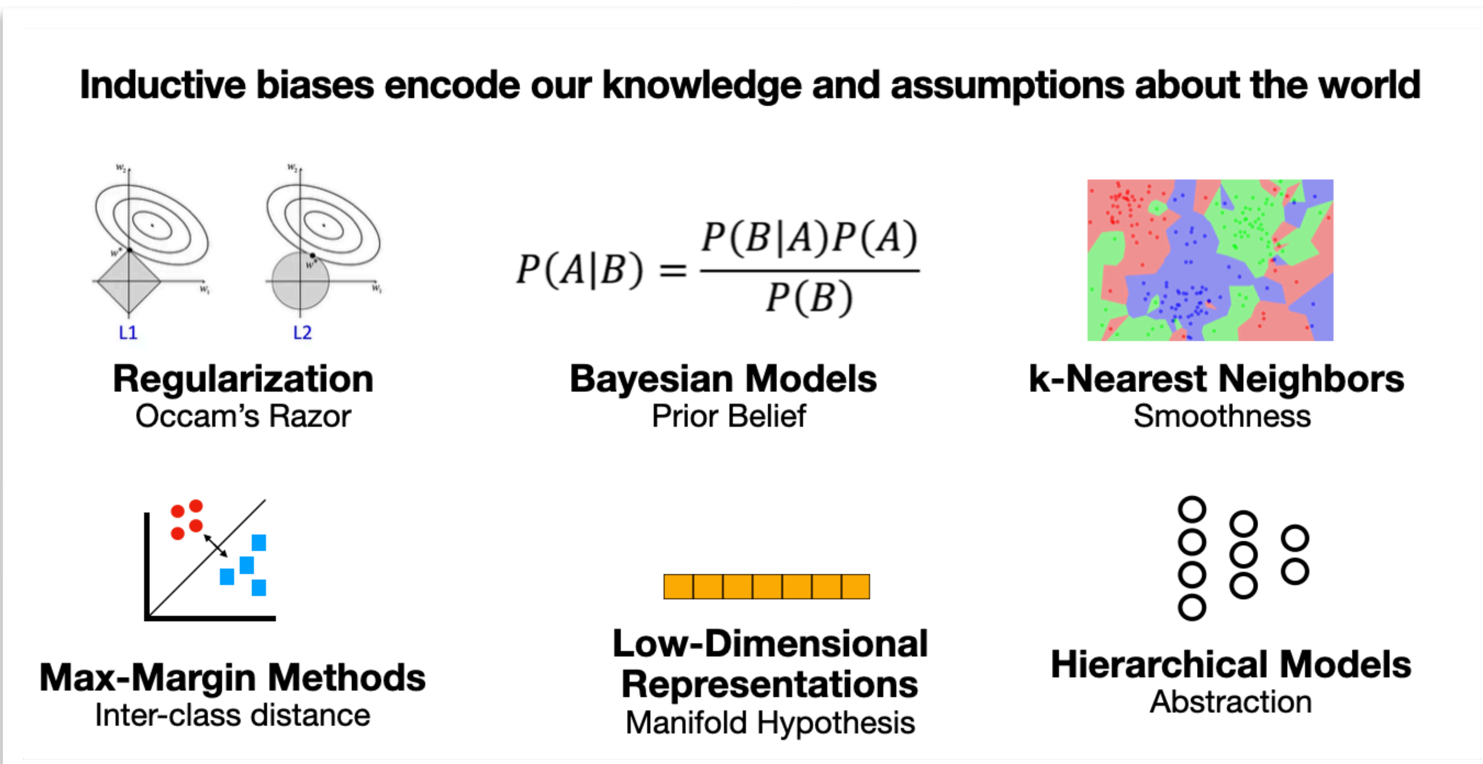
Every Bias Is Not Necessarily a Bad Bias

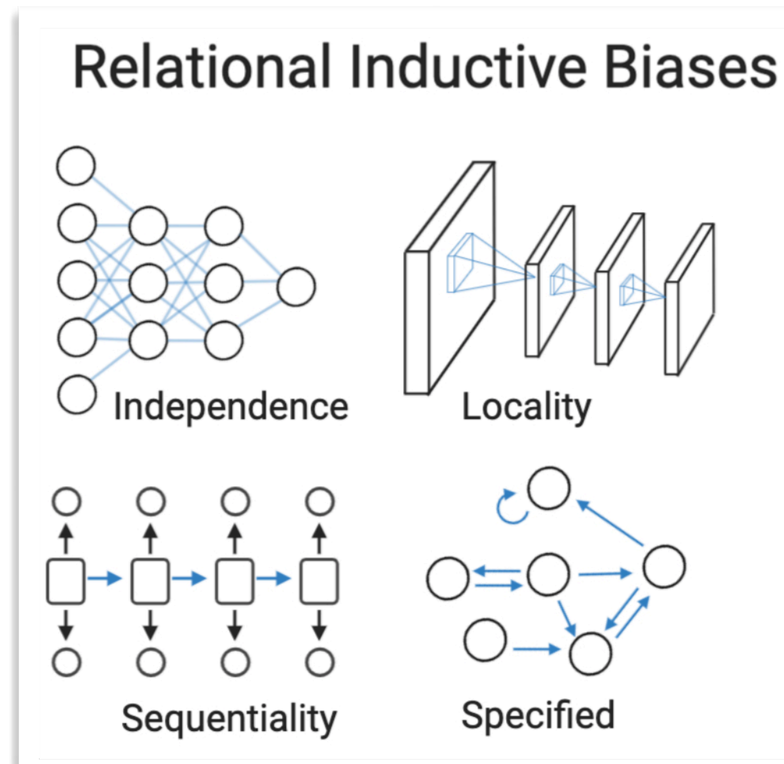


A quadratic pattern cannot be seen by a linear model.

Every Bias Is Not Necessarily a Bad Bias

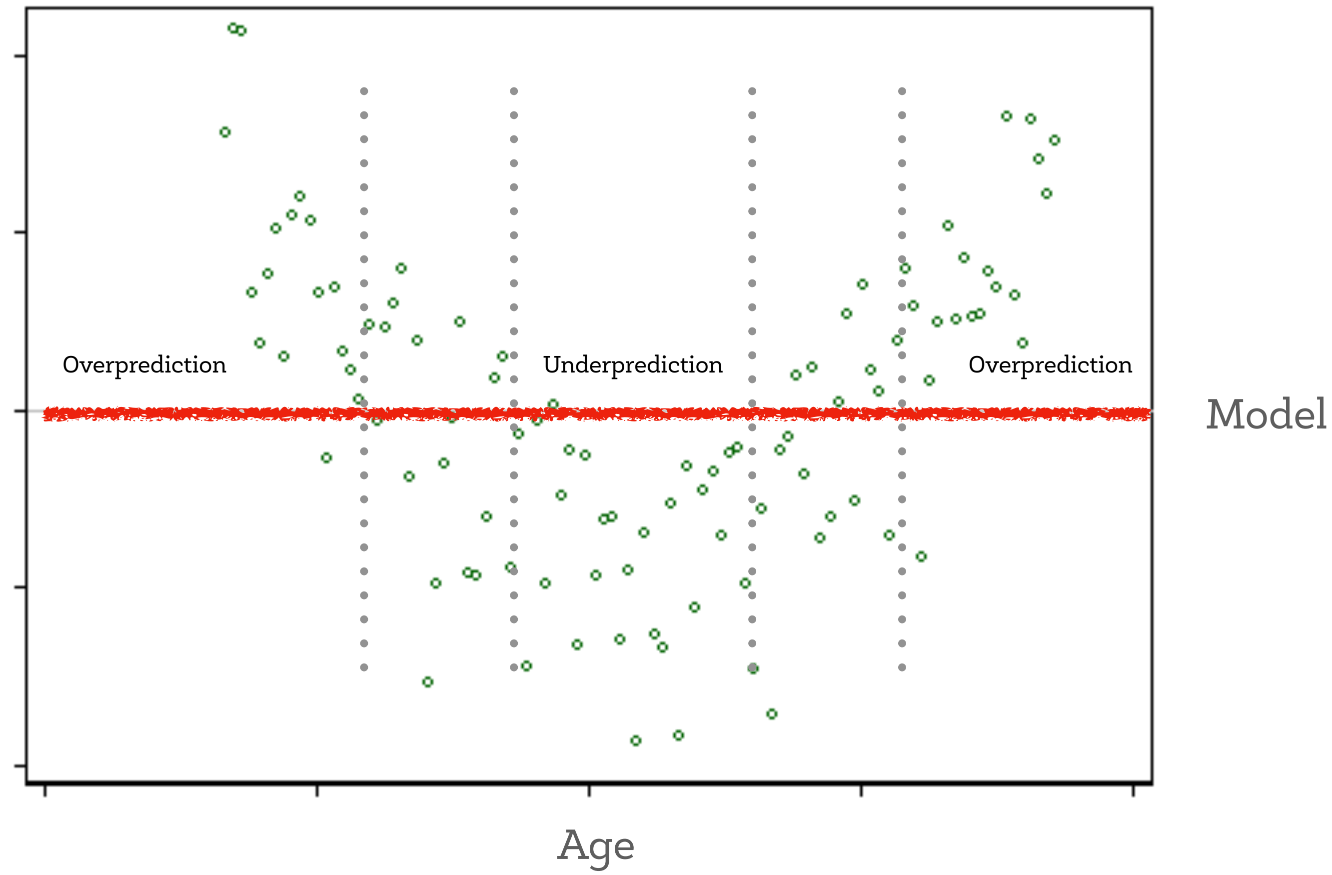
The **inductive bias** (also known as learning bias) of a learning algorithm is the set of **assumptions** (in terms of the hypothesis space, f_W) that **the learner uses** to predict outputs of given inputs that it has not encountered.





Every Bias Is Not Necessarily a Bad Bias

The **inductive bias** is inevitable and, though neither good nor bad in itself, but it is **not neutral** in real world settings: pattern blindness can result in winners and losers!



Algorithm ethical assessment

We've seen that training data reflects the disparities, distortions, and biases from the real world and the measurement process.

Some **patterns** in the training data ("smoking is associated with cancer") represent **knowledge** that we wish to mine using machine learning, while other patterns ("girls like pink and boys like blue") represent **stereotypes** or **bad habits** that we might wish to avoid learning.

Algorithm ethical assessment

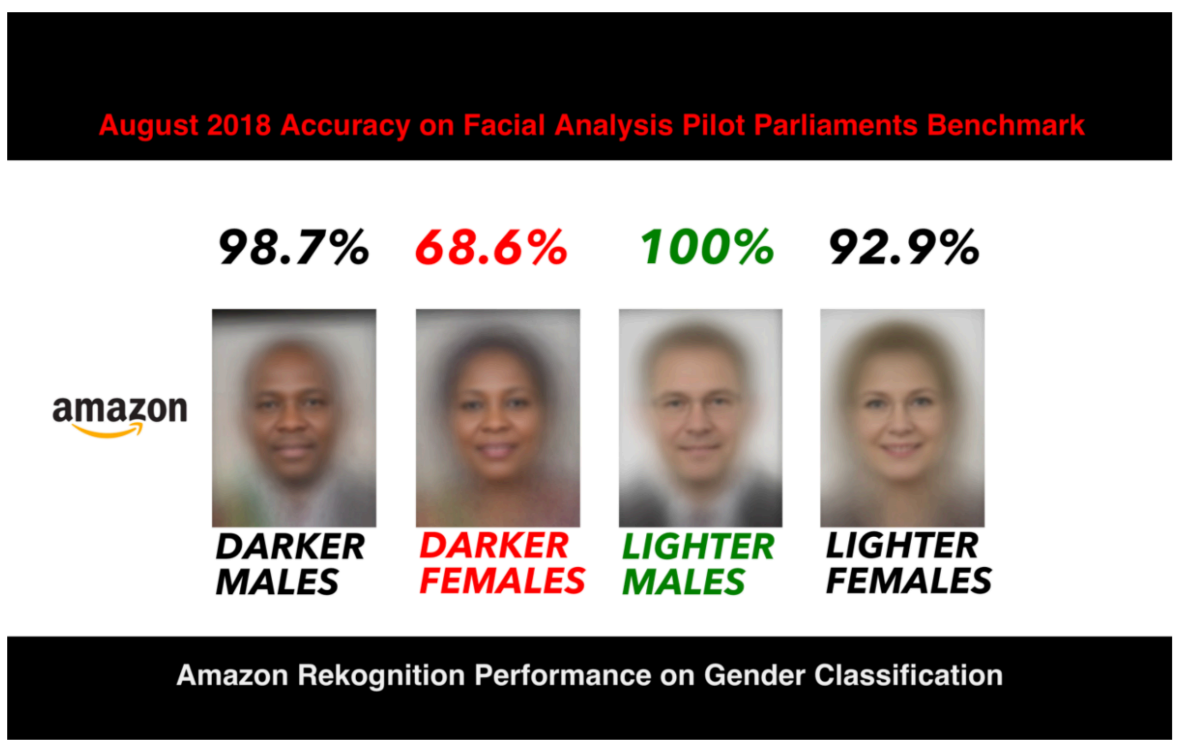
But learning algorithms have no general way to distinguish between these two types of patterns, because they are the result of social norms and moral judgments.

This leads to an obvious question: **when we learn a model from such data, are these disparities preserved, mitigated, or exacerbated?**

Impact of bad biases

Response: Racial and Gender bias in Amazon Rekognition — Commercial AI System for Analyzing Faces.

Joy Buolamwini Jan 25, 2019 · 15 min read



Impact of bad biases

NBER | NATIONAL BUREAU of ECONOMIC RESEARCH

‹ Working Papers

Consumer-Lending Discrimination in the FinTech Era

Robert Bartlett, Adair Morse, Richard Stanton & Nancy Wallace

Discrimination: Unjustified basis of differentiation between individuals

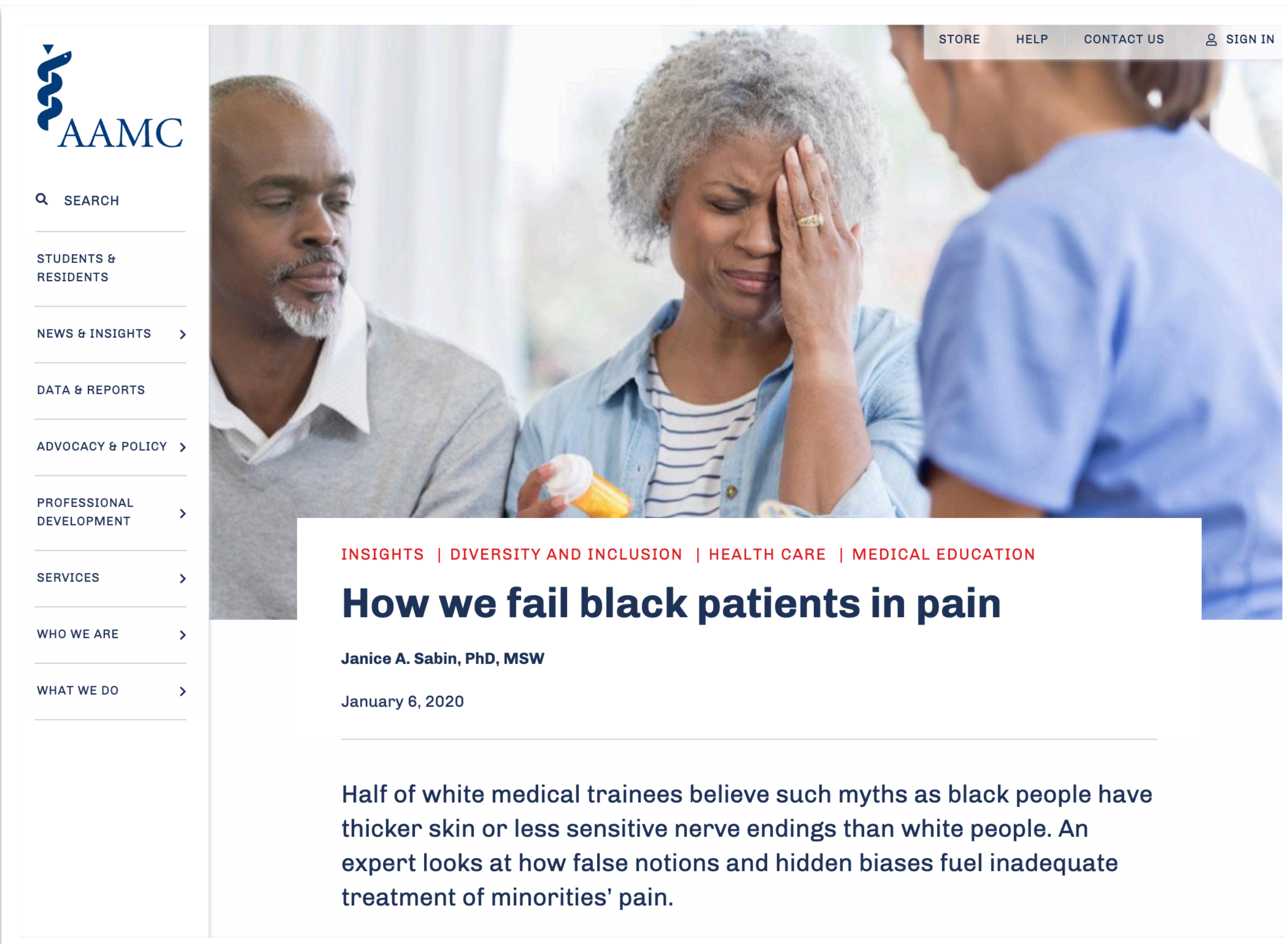
We find that **lenders charge Latin/African-American borrowers 7.9 and 3.6 basis points more for purchase and refinance mortgages respectively**, costing them $765M in aggregate per year in extra interest.

Bad News!

FinTech algorithms also discriminate, but 40% less than face-to-face lenders. The lower levels of price discrimination by algorithms suggests that removing face-to-face interactions can reduce discrimination.

Bias amplification

Impact of bad biases


AAMC

SEARCH

STUDENTS & RESIDENTS

NEWS & INSIGHTS

DATA & REPORTS

ADVOCACY & POLICY

PROFESSIONAL DEVELOPMENT

SERVICES

WHO WE ARE

WHAT WE DO

STORE HELP CONTACT US SIGN IN

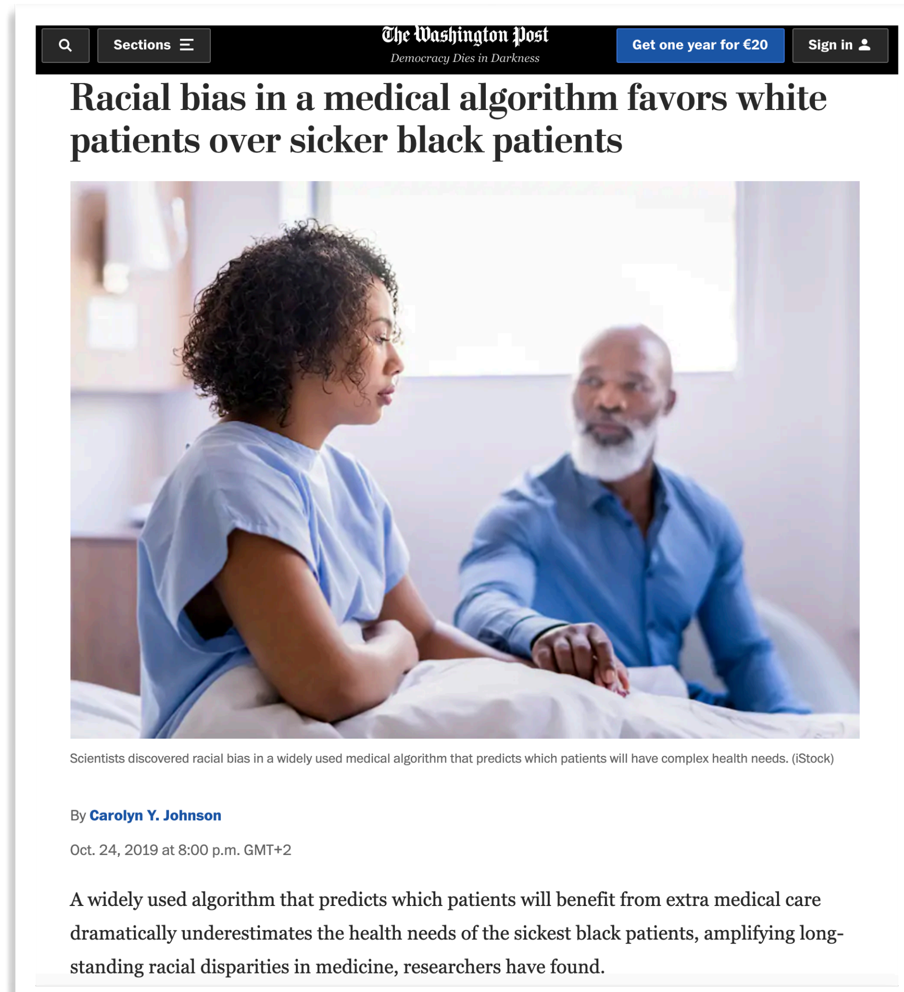
INSIGHTS | DIVERSITY AND INCLUSION | HEALTH CARE | MEDICAL EDUCATION

How we fail black patients in pain

Janice A. Sabin, PhD, MSW

January 6, 2020

Half of white medical trainees believe such myths as black people have thicker skin or less sensitive nerve endings than white people. An expert looks at how false notions and hidden biases fuel inadequate treatment of minorities' pain.


Sections

The Washington Post

Democracy Dies in Darkness

Get one year for €20

Sign in

Racial bias in a medical algorithm favors white patients over sicker black patients

Scientists discovered racial bias in a widely used medical algorithm that predicts which patients will have complex health needs. (iStock)

By **Carolyn Y. Johnson**

Oct. 24, 2019 at 8:00 p.m. GMT+2

A widely used algorithm that predicts which patients will benefit from extra medical care dramatically underestimates the health needs of the sickest black patients, amplifying long-standing racial disparities in medicine, researchers have found.

Impact of bad biases

PNAS Proceedings of the National Academy of Sciences of the United States of America

NEW RESEARCH IN

Physical Sciences

Social Sciences

Biological Sciences

BRIEF REPORT

Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis

Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante

PNAS June 9, 2020 117 (23) 12592-12594; first published May 26, 2020; https://doi.org/10.1073/pnas.1919012117

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved April 30, 2020 (received for review October 30, 2019)

Article | Figures & SI | Info & Metrics | PDF

X-ray image datasets used to diagnose various thoracic diseases

Measuing bias is difficult

and sometimes impossible

Problem: police discrimination analysis

Data:

Which police departments should the feds investigate?

Arrests per 100 residents (2019) and police killings per 100,00 residents (2013-20) by race alongside disparities between those numbers for the police departments with the 37 largest jurisdictions in the U.S.

	ARRESTS/100			KILLINGS/100K		
POLICE DEPARTMENT	WHITE	BLACK	DIS.	WHITE	BLACK	DIS.
Albuquerque, NM	4.0	10.4	2.6	5.5	19.5	3.6
Austin, TX	2.5	9.4	3.8	4.0	7.2	1.8
Baltimore, MD	1.9	5.5	2.9	2.4	7.6	3.2
Boston, MA	0.8	2.4	2.9	0.3	5.8	17.6
Charlotte-Mecklenburg, NC	1.0	4.8	4.8	1.0	3.7	3.7
Chicago, IL*	1.7	6.8	4.1	0.3	7.4	22.1
Columbus, OH	1.0	2.5	2.6	2.5	12.7	5.1
Dallas, TX	2.0	5.0	2.5	3.1	5.1	1.6
Denver, CO	3.6	11.0	3.1	3.0	8.0	2.7
Detroit, MI	1.1	2.0	1.8	1.4	2.5	1.7
El Paso, TX	2.6	5.2	2.0	5.6	8.7	1.6
Fort Worth, TX	1.8	4.1	2.3	1.8	5.7	3.2
Fresno, CA**	5.6	11.2	2.0	3.5	2.7	0.8
Honolulu, HI	2.2	5.0	2.2	2.2	0.0	0.0
Houston, TX	1.1	3.5	3.2	1.6	7.5	4.7
Indianapolis, IN	2.8	6.1	2.2	2.1	7.5	3.5
Jacksonville, FL*	2.4	6.1	2.6	4.2	8.6	2.1
Las Vegas Metro, NV	3.9	13.2	3.4	3.3	5.9	1.8
Los Angeles, CA	1.8	4.4	2.4	1.8	8.2	4.6
Louisville Metro, KY	4.3	10.0	2.3	2.5	9.1	3.7
Memphis, TN	2.3	6.3	2.7	2.4	3.6	1.5
Mesa, AZ	3.1	12.9	4.2	4.6	0.0	0.0
Milwaukee, WI	1.2	4.4	3.8	1.0	7.0	7.3
Nashville Metropolitan, TN	2.7	6.5	2.4	0.8	3.8	4.7
New York, NY*	2.0	5.5	2.7	0.4	2.9	7.9
Oklahoma City, OK	2.1	6.3	3.0	5.0	27.2	5.5
Philadelphia, PA**	2.3	4.6	2.0	0.5	3.9	7.0
Phoenix, AZ	3.5	10.6	3.0	7.2	15.2	2.1
Portland, OR	3.0	12.8	4.3	2.9	11.1	3.9
Sacramento, CA	3.0	8.3	2.8	3.7	9.3	2.5
San Antonio, TX	2.5	9.3	3.7	3.0	10.5	3.5
San Diego, CA	2.8	8.7	3.2	2.0	3.5	1.7
San Francisco, CA	2.0	11.9	5.8	1.4	11.5	8.1
San Jose, CA	2.8	6.7	2.4	2.6	3.4	1.3
Seattle, WA	1.1	7.0	6.1	2.2	12.4	5.7
Tucson, AZ	7.2	20.2	2.8	4.2	7.9	1.9
Washington, D.C.*	0.9	6.4	7.3	0.4	5.4	13.4

*The departments serving Chicago, Jacksonville, New York and Washington, D.C., do not report their arrests disaggregated by race to the FBI, but release their data independently. We excluded arrests for traffic violations to make their data comparable to that released by the FBI.

**Data from the Fresno and Philadelphia police departments are from 2018.

SOURCES: MAPPING POLICE VIOLENCE, FBI UNIFORM CRIME REPORT, U.S. CENSUS BUREAU

Measuing bias is difficult

Problem:
police discrimination analysis

Data:

Which police departments should the feds investigate?

Arrests per 100 residents (2019) and police killings per 100,00 residents (2013-20) by race alongside disparities between those numbers for the police departments with the 37 largest jurisdictions in the U.S.

	ARRESTS/100			KILLINGS/100K		
POLICE DEPARTMENT	WHITE	BLACK	DIS.	WHITE	BLACK	DIS.
Albuquerque, NM	4.0	10.4	2.6	5.5	19.5	3.6
Austin, TX	2.5	9.4	3.8	4.0	7.2	1.8
Baltimore, MD	1.9	5.5	2.9	2.4	7.6	3.2
Boston, MA	0.8	2.4	2.9	0.3	5.8	17.6
Charlotte-Mecklenburg, NC	1.0	4.8	4.8	1.0	3.7	3.7
Chicago, IL*	1.7	6.8	4.1	0.3	7.4	22.1
Columbus, OH	1.0	2.5	2.6	2.5	12.7	5.1
Dallas, TX	2.0	5.0	2.5	3.1	5.1	1.6
Denver, CO	3.6	11.0	3.1	3.0	8.0	2.7
Detroit, MI	1.1	2.0	1.8	1.4	2.5	1.7
El Paso, TX	2.6	5.2	2.0	5.6	8.7	1.6
Fort Worth, TX	1.8	4.1	2.3	1.8	5.7	3.2
Fresno, CA**	5.6	11.2	2.0	3.5	2.7	0.8
Honolulu, HI	2.2	5.0	2.2	2.2	0.0	0.0
Houston, TX	1.1	3.5	3.2	1.6	7.5	4.7
Indianapolis, IN	2.8	6.1	2.2	2.1	7.5	3.5
Jacksonville, FL*	2.4	6.1	2.6	4.2	8.6	2.1
Las Vegas Metro, NV	3.9	13.2	3.4	3.3	5.9	1.8
Los Angeles, CA	1.8	4.4	2.4	1.8	8.2	4.6
Louisville Metro, KY	4.3	10.0	2.3	2.5	9.1	3.7
Memphis, TN	2.3	6.3	2.7	2.4	3.6	1.5
Mesa, AZ	3.1	12.9	4.2	4.6	0.0	0.0
Milwaukee, WI	1.2	4.4	3.8	1.0	7.0	7.3
Nashville Metropolitan, TN	2.7	6.5	2.4	0.8	3.8	4.7
New York, NY*	2.0	5.5	2.7	0.4	2.9	7.9
Oklahoma City, OK	2.1	6.3	3.0	5.0	27.2	5.5
Philadelphia, PA**	2.3	4.6	2.0	0.5	3.9	7.0
Phoenix, AZ	3.5	10.6	3.0	7.2	15.2	2.1
Portland, OR	3.0	12.8	4.3	2.9	11.1	3.9
Sacramento, CA	3.0	8.3	2.8	3.7	9.3	2.5
San Antonio, TX	2.5	9.3	3.7	3.0	10.5	3.5
San Diego, CA	2.8	8.7	3.2	2.0	3.5	1.7
San Francisco, CA	2.0	11.9	5.8	1.4	11.5	8.1
San Jose, CA	2.8	6.7	2.4	2.6	3.4	1.3
Seattle, WA	1.1	7.0	6.1	2.2	12.4	5.7
Tucson, AZ	7.2	20.2	2.8	4.2	7.9	1.9
Washington, D.C.*	0.9	6.4	7.3	0.4	5.4	13.4

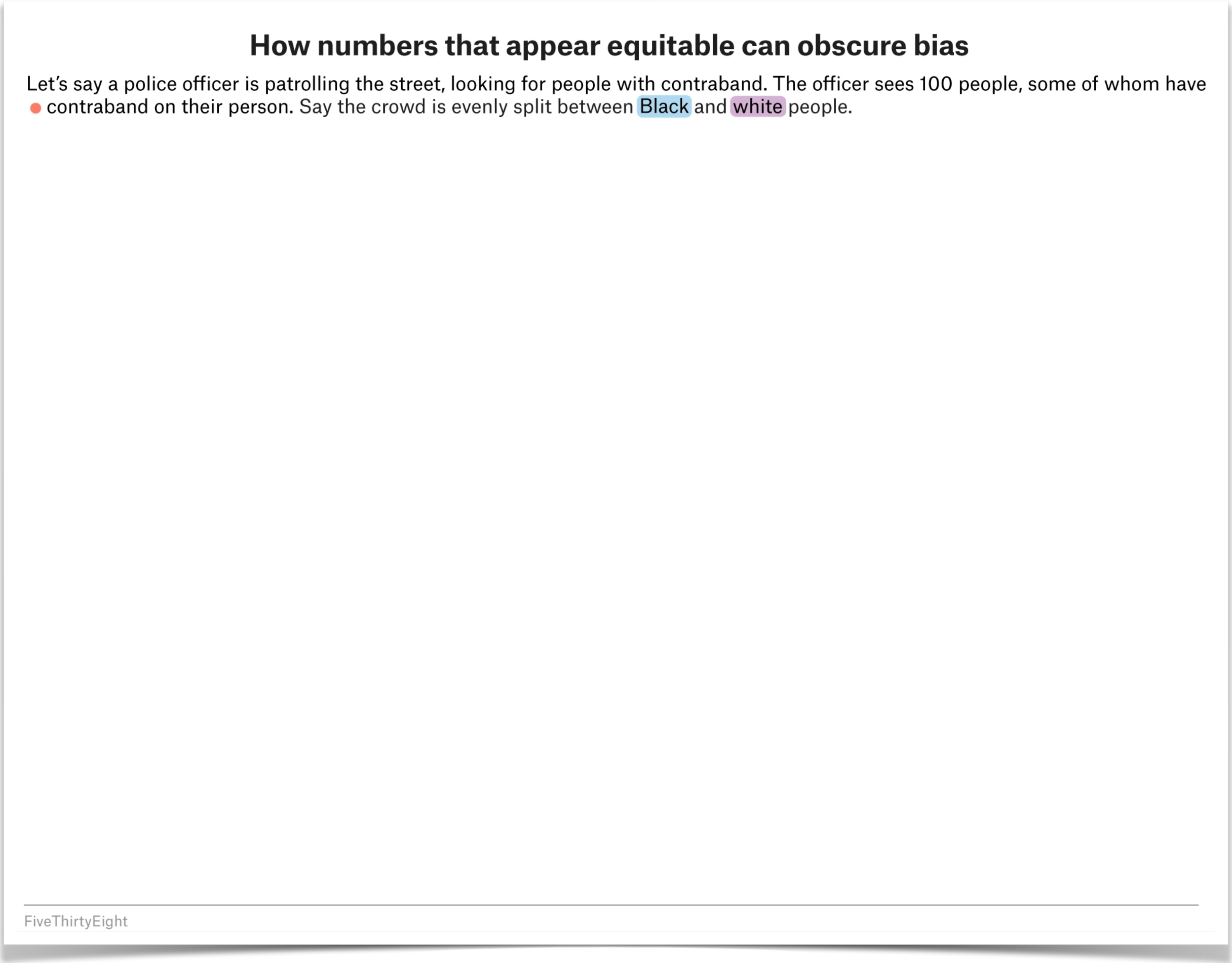
*The departments serving Chicago, Jacksonville, New York and Washington, D.C., do not report their arrests disaggregated by race to the FBI, but release their data independently. We excluded arrests for traffic violations to make their data comparable to that released by the FBI.

**Data from the Fresno and Philadelphia police departments are from 2018.

SOURCES: MAPPING POLICE VIOLENCE, FBI UNIFORM CRIME REPORT, U.S. CENSUS BUREAU

If the rate of using force against stopped Black people and the rate of using force against stopped white people are the same, can we conclude that we are observing a fair behavior?

Measuing bias is difficult



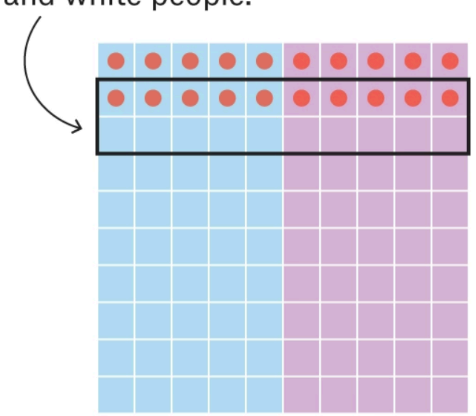
Measuing bias is difficult

How numbers that appear equitable can obscure bias

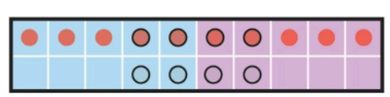
Let's say a police officer is patrolling the street, looking for people with contraband. The officer sees 100 people, some of whom have contraband on their person. Say the crowd is evenly split between Black and white people.

SCENARIO 1

The police officer stops **20** people, pulling aside equal numbers of Black and white people.



Of the **20** people stopped, the officer uses force against **8** of them.



The police officer used force against stopped white people and stopped Black people at the same rate: **40%**.

But that's not the only scenario that can lead to that **40%** number.

FiveThirtyEight

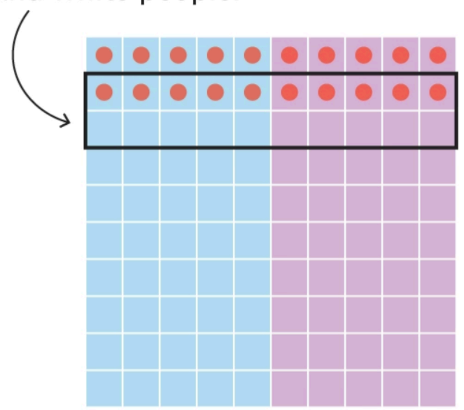
Measuing bias is difficult

How numbers that appear equitable can obscure bias

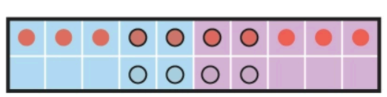
Let's say a police officer is patrolling the street, looking for people with contraband. The officer sees 100 people, some of whom have ● contraband on their person. Say the crowd is evenly split between Black and white people.

SCENARIO 1

The police officer stops **20** people, pulling aside equal numbers of Black and white people.



Of the **20** people stopped, the officer uses ○ force against **8** of them.

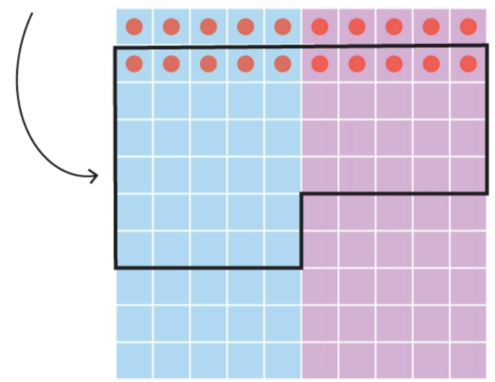


The police officer used force against stopped white people and stopped Black people at the same rate: **40%**.

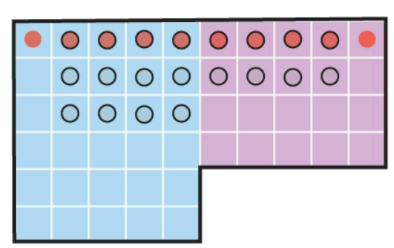
But that's not the only scenario that can lead to that **40%** number.

SCENARIO 2

This time, of the **100** people the officer sees, he stops **50**. But this time he is biased in whom he pulls aside.



The officer uses force against **20** people this time.



This time, like last time, the police officer used force against stopped white people and stopped Black people at the same rate: **40%**.

FiveThirtyEight

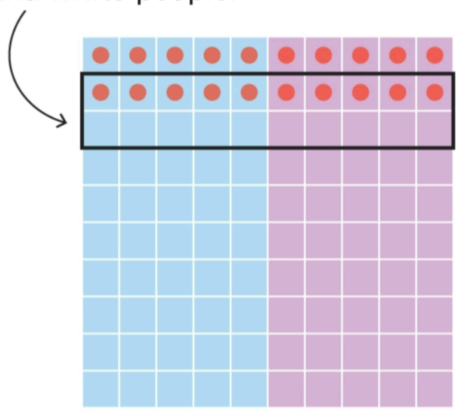
Measuing bias is difficult

How numbers that appear equitable can obscure bias

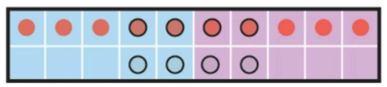
Let's say a police officer is patrolling the street, looking for people with contraband. The officer sees 100 people, some of whom have ● contraband on their person. Say the crowd is evenly split between Black and white people.

SCENARIO 1

The police officer stops **20** people, pulling aside equal numbers of Black and white people.



Of the **20** people stopped, the officer uses ○ force against **8** of them.

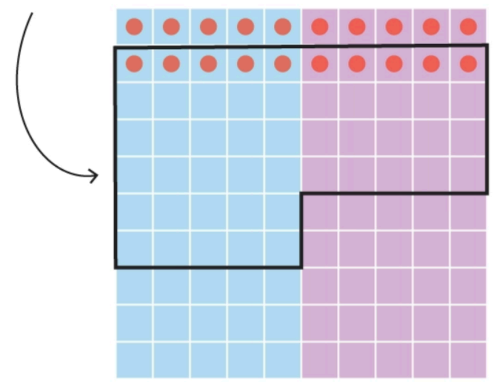


The police officer used force against stopped white people and stopped Black people at the same rate: **40%**.

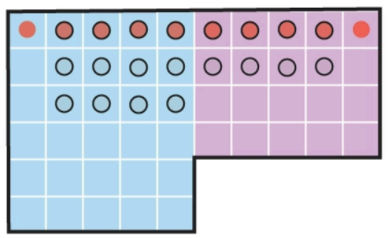
But that's not the only scenario that can lead to that **40%** number.

SCENARIO 2

This time, of the **100** people the officer sees, he stops **50**. But this time he is biased in whom he pulls aside.



The officer uses force against **20** people this time.



This time, like last time, the police officer used force against stopped white people and stopped Black people at the same rate: **40%**.

ANALYSIS

Things might appear equal, but in the second scenario, more Black people were stopped by the police than white people.

While use of force among **stopped** people is equal, use of force among all **observed** people is not:

$\frac{12}{50}$ = **24%** of Black people have force used against them

$\frac{8}{50}$ = **16%** of white people have force used against them

FiveThirtyEight

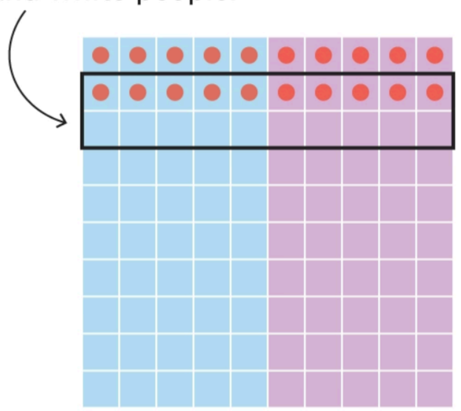
Measuing bias is difficult

How numbers that appear equitable can obscure bias

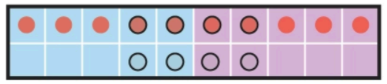
Let's say a police officer is patrolling the street, looking for people with contraband. The officer sees 100 people, some of whom have contraband on their person. Say the crowd is evenly split between Black and white people.

SCENARIO 1

The police officer stops **20** people, pulling aside equal numbers of Black and white people.



Of the **20** people stopped, the officer uses force against **8** of them.

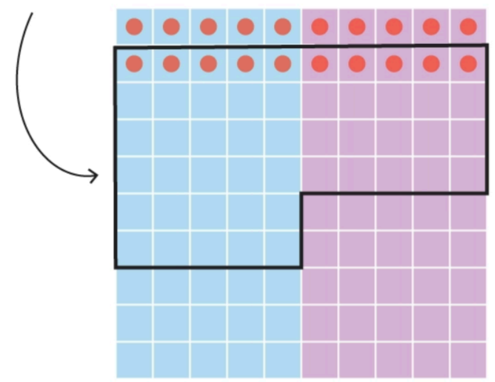


The police officer used force against stopped white people and stopped Black people at the same rate: **40%**.

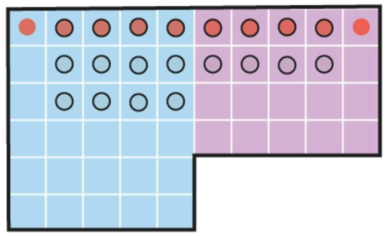
But that's not the only scenario that can lead to that **40%** number.

SCENARIO 2

This time, of the **100** people the officer sees, he stops **50**. But this time he is biased in whom he pulls aside.



The officer uses force against **20** people this time.



This time, like last time, the police officer used force against stopped white people and stopped Black people at the same rate: **40%**.

ANALYSIS

Things might appear equal, but in the second scenario, more Black people were stopped by the police than white people.

While use of force among **stopped** people is equal, use of force among all **observed** people is not:

$\frac{12}{50}$ = **24%** of Black people have force used against them

$\frac{8}{50}$ = **16%** of white people have force used against them

CONCLUSION

This is why knowing how often police use force against people they've stopped is **not enough information** to know whether use of force is racially biased. In real life, we don't have data on everyone who was observed but not stopped, but we need that to know whether use of force is biased overall.

FiveThirtyEight

Discrimination

discriminate

Cambridge Dictionary
Make your words meaningful

verb

UK /dɪˈskrɪm.ɪ.neɪt/ US /dɪˈskrɪm.ə.neɪt/

discriminate ***verb*** **(TREAT DIFFERENTLY)**

C1 [I]

to treat a person or particular group of people differently, especially in a worse way from the way in which you treat other people, because of their skin colour, sex, sexuality, etc.:

C2 [I + adv/prep] formal

to be able to recognize the difference between people or things:

Law & Discrimination

Under the most advanced law systems, everyone is protected from **unlawful behavior** (discrimination) when the cause of this behavior is that they **have** or **are perceived to have** a "*protected characteristic*" or are associated with someone who has a **protected characteristic**:

- Age
- Disability
- Gender
- Civil state
- Pregnancy and maternity
- Race
- Religion and belief
- Sex
- Sexual orientation

Be careful!

In many classification tasks, available data contain **protected characteristics** of an individual.

Some have hoped that **removing or ignoring protected attributes** would somehow ensure the impartiality of the resulting classifier. Unfortunately, this practice is usually somewhere on the spectrum between **ineffective** and **harmful**.

In a typical data set, we have many features that are slightly correlated with the sensitive attribute. However, if numerous such features are available, as is the case in a typical browsing history, the task of predicting gender becomes feasible at high accuracy levels.

https://fairmlbook.org/

ENGINEER POINT OF VIEW: THAT'S NOT MY BUSINESS!

But, isn't discrimination the very point of machine learning?

Yes, but it is not admissible when this discrimination/differentiation is based on unjustified causes, is practically irrelevant or is morally wrong.

WE NEED A CASE BY CASE ANALYSIS
FAIRNESS CANNOT BE AUTOMATED

Discrimination is not a general concept, it's **domain and feature specific**!

Law & Discrimination

There are several types of discrimination:

https://www.equalityhumanrights.com/sites/default/files/ea_legal_definitions_0.pdf

1. **Direct discrimination** . This means treating someone less favorably than someone else because of a protected characteristic.
2. **Direct discrimination by perception**. This means treating one person less favorably than someone else, because you incorrectly think they have a protected characteristic.
3. **Discrimination arising from disability**. This means treating a disabled person unfavorably because of something connected with their disability when this cannot be objectively justified.
4. **Direct discrimination by association.** This means treating someone less favorably than another person because they are associated with a person who has a protected characteristic.
5. **Failing to make reasonable adjustments.** To do this for disabled people is also a form of discrimination.
6. **Harassment.** Harassment is unwanted behavior related to a protected characteristic which has the purpose or effect of violating someone's dignity or which creates a hostile, degrading, humiliating or offensive environment.

Law & Discrimination

Disparate treatment or direct discrimination:
Treatment depends on class membership

Disparate impact or indirect discrimination:
Outcome depends on class membership

Law & Discrimination

1 An employer does not interview a job applicant because of the applicant's ethnic background

2 A hair salon owner has a policy of not employing stylists who cover their hair, believing it is important for them to exhibit their flamboyant haircuts.

3 An employer dismisses a worker because she has had three months' sick leave. The employer is aware that the worker has multiple sclerosis and most of her sick leave is disability-related.

4 An employer offers flexible working to all staff. Requests are supposed to be considered based on business need. A manager allows a man's request to work flexibly to train for a qualification but does not allow another man's request to work flexibly to care for his disabled child.

5 An employer has a policy that designated car parking spaces are only offered to senior managers. A worker who is not a manager, but has a mobility impairment is not given a designated car parking space.

6 A builder addresses abusive and hostile remarks to a customer because of her race after their business relationship has ended.

1. **Direct discrimination** . This means treating someone less favourably than someone else because of a protected characteristic.
2. **Direct discrimination by perception**. This means treating one person less favourably than someone else, because you incorrectly think they have a protected characteristic.
3. **Discrimination arising from disability**. This means treating a disabled person unfavourably because of something connected with their disability when this cannot be objectively justified.
4. **Direct discrimination by association.** This means treating someone less favourably than another person because they are associated with a person who has a protected characteristic.
5. **Failing to make reasonable adjustments.** To do this for disabled people is also a form of discrimination.
6. **Harassment.** Harassment is unwanted behaviour related to a protected characteristic which has the purpose or effect of violating someone's dignity or which creates a hostile, degrading, humiliating or offensive environment.

https://www.equalityhumanrights.com/sites/default/files/ea_legal_definitions_0.pdf

Law & Discrimination

Algorithmic discrimination scenarios:

- Access to employment
- Access to education
- Access to government/companies benefits
- Access to penitentiary alternatives
- Etc.

Anti-discrimination legislation typically seeks **equal access/ treatment** (mitigation of direct discrimination) to employment, working conditions, education, social protection, goods, and services, but in some cases, **equal outcome** is also sought (mitigation of indirect discrimination).

Law & Discrimination

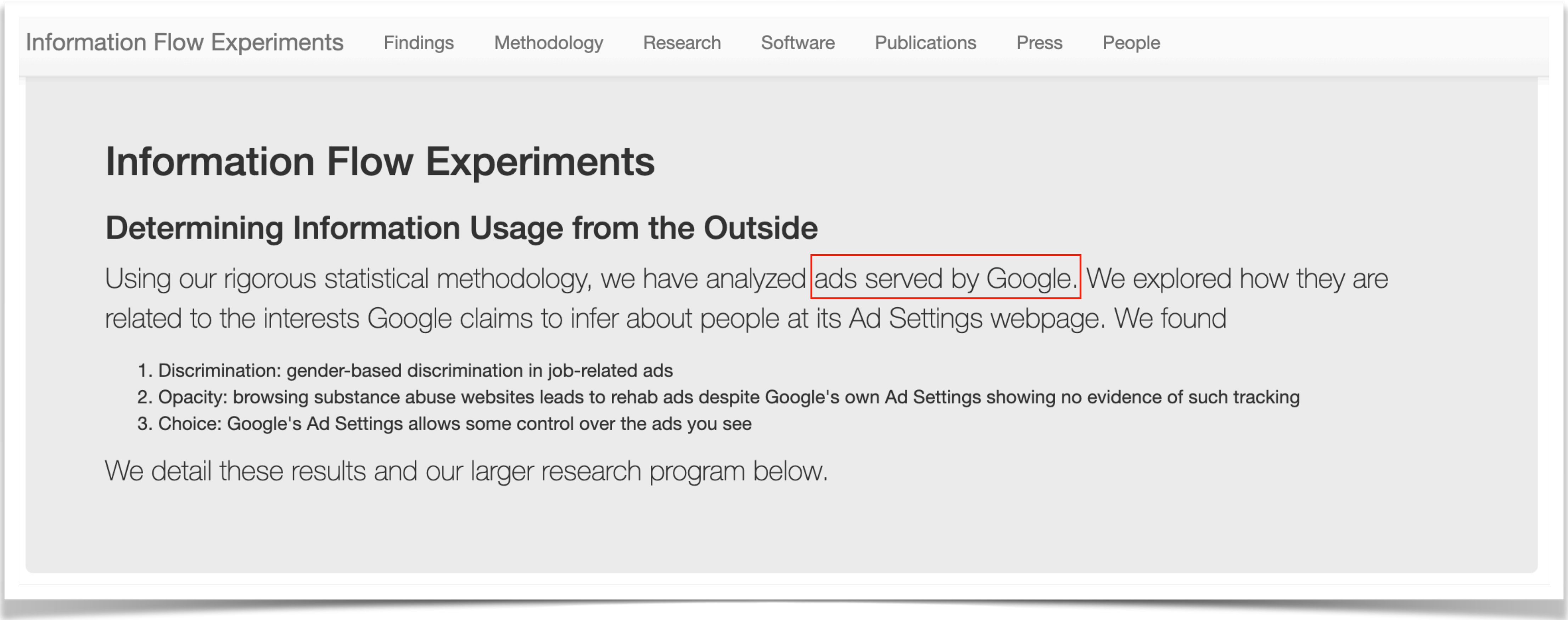
In general, anti-discrimination laws aim to achieve **equality of opportunity**.

Narrow notions of equality of opportunity are concerned with ensuring that decision-making **treats similar people similarly on the basis of relevant features**, given their current degree of similarity.

Broader notions of equality of opportunity are concerned with organizing society in such a way that **people of equal talents and ambition can achieve equal outcomes** over the course of their live.

Somewhere in between is a notion of equality of opportunity that forces decision-making to **treat seemingly dissimilar people similarly**, **on the belief that their current dissimilarity is the result of past injustice**.

Example



Information Flow Experiments Findings Methodology Research Software Publications Press People

Information Flow Experiments

Determining Information Usage from the Outside

Using our rigorous statistical methodology, we have analyzed ads served by Google. We explored how they are related to the interests Google claims to infer about people at its Ad Settings webpage. We found

1. Discrimination: gender-based discrimination in job-related ads
2. Opacity: browsing substance abuse websites leads to rehab ads despite Google's own Ad Settings showing no evidence of such tracking
3. Choice: Google's Ad Settings allows some control over the ads you see

We detail these results and our larger research program below.

https://www.cs.cmu.edu/~mtschant/ife/

Example

Over hundreds of browsers, we randomly edited the profile to be either "female" or "male" and visited job-related websites. We found that the "male" instances were much more likely to receive ads promoting high paying jobs than the "female" instances.

Top ads for identifying the female group

Ad Title	Ad URL	Times shown to Females	Times shown to Males
Jobs (Hiring Now)	www.jobsinyourarea.co	45	8
4Runner Parts Service	www.westernpatoyotaservice.com	36	5
Criminal Justice Program	www3.mc3.edu/Criminal+Justice	29	1
Goodwill - Hiring	goodwill.careerboutique.com	121	39
UMUC Cyber Training	www.umuc.edu/cybersecuritytraining	38	30

Top ads for identifying the male group

Ad Title	Ad URL	Times shown to Females	Times shown to Males
$200k+ Jobs - Execs Only	careerchange.com	311	1816
Find Next $200k+ Job	careerchange.com	7	36
Become a Youth Counselor	www.youthcounseling.degreeleap.com	0	310
CDL-A OTR Trucking Jobs	www.tadrivers.com/OTRJobs	0	8
Free Resume Templates	resume-templates.resume-now.com	8	10

The human factor

Human Biases

Our brains are evolved to help us survive. That means they take a lot of shortcuts to help us get through the day. These shortcuts, or **heuristics, are vital. But they come at a cost**.

Our world is much more complex than the world our brains developed these heuristics. Unconscious brains can be unreliable in this environment.

Our unconscious can helps us in some situations, but it is not always the right tool. We must be sure that it will not hurt others.

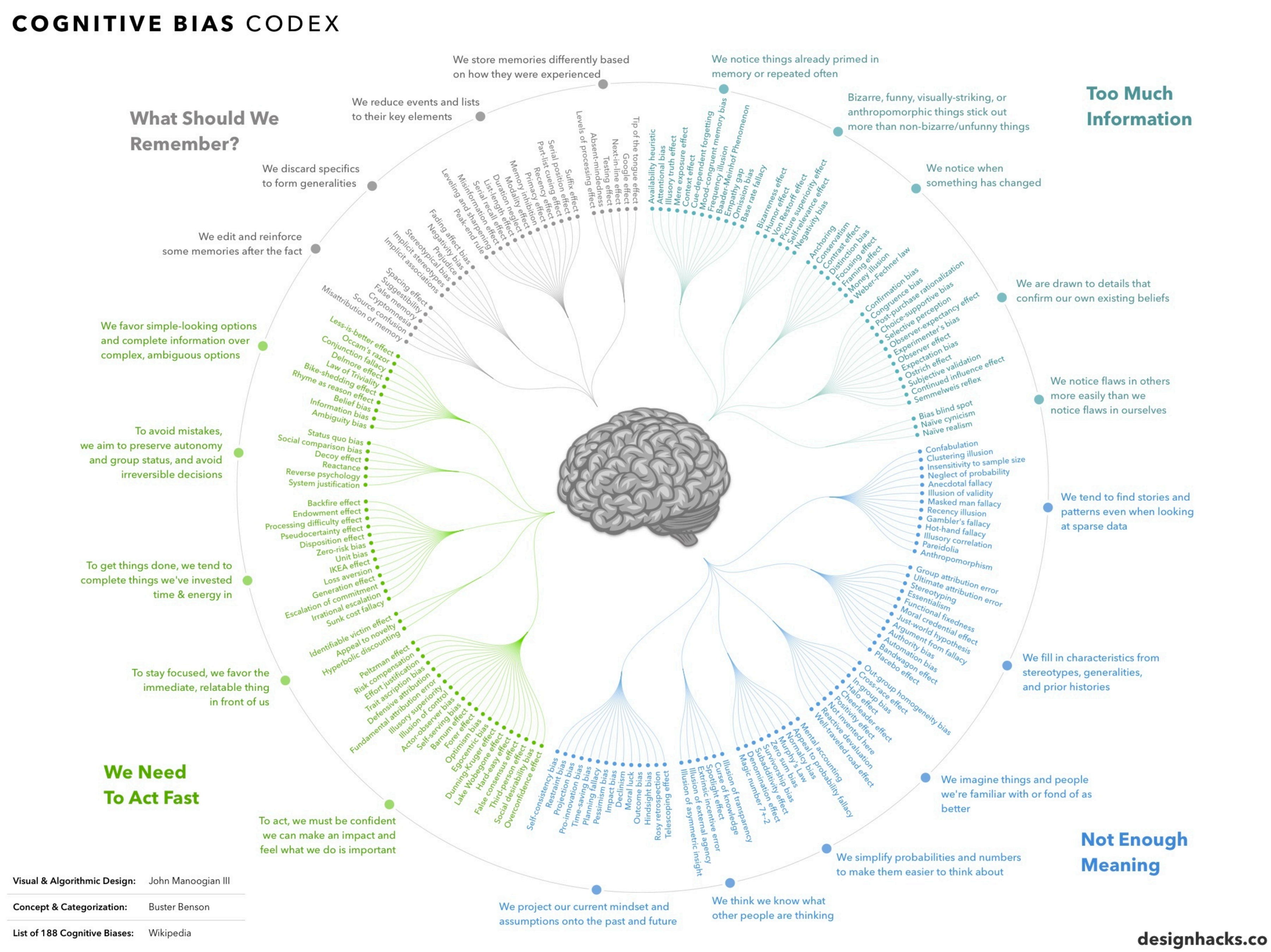
The halo effect

People who looks healthy or attractive are also competent and good.

Reading:

Physiognomy's New Clothes, by Blaise Agüera y Arcas, Margaret Mitchell and Alexander Todorov.

Unconscious Human Biases



https://www.visualcapitalist.com/wp-content/uploads/2017/09/cognitive-bias-infographic.html

Human decision making

We know that human decision-making is affected by:

- Unconscious thoughts, biases, etc. MIND
- Unthinking custom and practice, or unconsciously absorving beliefs of our friends, family, society, etc. PERSONAL HISTORY
- Personal ethical decision making profile. F.e. you prioritize relationships in your decicion-making. DEFAULT SETTING
- Reflective practice, to consider context and the people who will be affected by your decisions.

The role of ethics is to have a toolkit to do reflective practice, and to able of making and justifiying our decisions

Automated Discrimination

Algorithmic Fairness

Algorithm fairness is the field of research aimed at understanding and correcting unwanted biases.

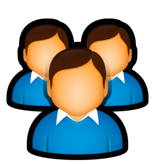
Specifically, it includes:

- Researching the **causes of bias** in data and algorithms
- Defining and applying **measurements of fairness**
- Developing data collection and modelling methodologies aimed at creating **fair algorithms**.

How to measure fairness

We can distinguish between two approaches to formalizing fairness:

– **Individual fairness** definitions are based on the premise that similar entities should be treated similarly.



– **Group fairness** definitions are based on the definition of group entities and ask that all groups are treated similarly.

To operationalize both approaches to fairness, we need to define **similarity for the input and the output** of an algorithm.

For group fairness, the challenge lies in determining **how to partition entities into groups (protected attributes)**

How to measure fairness

Individual fairness. X discriminates against Y in relation to Z if:

- Y has property P and Z does not have P.
- X treats worse Y than she treats Z and this is because Y has P and Z does not have P.

How to measure fairness

Group fairness. X group-discriminates against someone, Y, in relation to another, Z, by Φ-ing (e.g., hiring Z rather than Y), if:

- There is a **property**, P, such that Y has P or X believes that Y has P, and Z does not have P or X believes that Z does not have P,
- X treats Y worse than he treats or would treat Z by Φ-ing.
- It is because (X believes that) Y has P and (X believes that) Z does not have P that X treats Y worse than Z by Φ-ing" and,
- P is the property of being a member of a certain socially salient group (to which Z does not belong).

Lippert-Rasmussen K (2014) Born free and equal?: a philosophical inquiry into the nature of discrimination. In: Oxford university press, Oxford, New York

Fairness Definitions

One way of formulating **individual fairness** is a distance-based one.

Given a **distance measure** d between two entities and a distance measure D between the outputs of an algorithm, we would like the distance between the output of the algorithm for two entities to be small, when the entities are similar.

Fairness Definitions

Another form of **individual fairness** is *counterfactual fairness*.

An **output is fair toward an entity if it is the same in both the actual world and a counterfactual world where the entity belonged to a different group**.

> *Given that Alice did not get promoted in her job, and given that she is a woman, and given everything else we can observe about her circumstances and performance, what is the probability of her getting a promotion if she was a man instead?*

Causal inference is used to formalize this notion of fairness.

Fairness Definitions

For simplicity, let us assume two **groups**, namely the protected group G^+ (f.e. women) and the non-protected (or, privileged) group G^- (f.e. men) and a binary classifier.

We will start by presenting statistical approaches commonly used in **classification**. Assume that Y is the actual and $\hat{Y}$ the predicted output of the binary classifier, that is, Y is the "ground truth", and $\hat{Y}$ the output of the algorithm.

There are equivalent frameworks for regression and ranking.

Let 1 be the positive class that leads to a **favorable decision**, e.g., someone getting a loan, or being admitted at a competitive school, and S be the predicted probability for a certain classification.

Fairness Definitions

Statistical approaches to **group fairness** can be distinguished as:

- **Base rates** approaches: that use only the output $\hat{Y}$ of the algorithm,
- **Accuracy-based** approaches: that use both the output $\hat{Y}$ of the algorithm and the ground truth Y, and
- **Calibration** approaches: that use the predicted probability S and the ground truth Y.

Base rate fairness

Base rate fairness compares (ratio or difference)

- the probability $P(\hat{Y} = 1 \,|\, X \in G^{+})$ that an entity X receives the favorable outcome when X belongs to the protected group
- with the corresponding probability $P(\hat{Y} = 1 \,|\, X \in G^{-})$ that X receives the favorable outcome when X belongs to the non-protected group.

When the probabilities of a favorable outcome are equal for the two groups, we have a special type of fairness termed **demographic, or statistical parity**:
$P(\hat{Y} = 1 \,|\, X \in G^{+}) \sim P(\hat{Y} = 1 \,|\, X \in G^{-})$

Base rate fairness

In a more general setting we can define demographic parity in terms of statistical **independence**: the protected characteristic must be statistically independent of the outcome.

$\hat{Y}$ independent of the protected characteristic for all groups a, b and all values d:

$$p(\hat{Y} = d \mid X \in a) = p(\hat{Y} = d \mid X \in b)$$

Base rate fairness

Base rate fairness ignores the actual output.

For example, assume that the classification task is getting or not a job and the protected group G^+ is based on gender.

Statistical parity asks for a specific ratio of women in the positive class, **even when there are not that many women in the input who are well qualified for the job**.

Base rate fairness

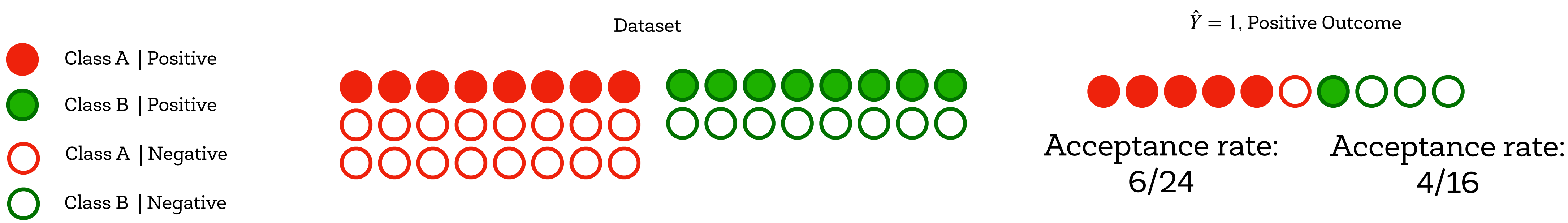
What is the 80% Rule?

The rule states that **companies should be hiring protected groups at a rate that is at least 80% of that of white men**. The 80% rule was created to help companies determine if they have been unwittingly discriminatory in their hiring process.

Base rate fairness

Let's assume we're building an application to select promising candidates for a job. Our model will aim to learn the typical profile of those who can be hired.

In this example **we get demographic parity**:



Reminder: $P(\hat{Y} = 1 \mid X \in G) = \dfrac{P(\hat{Y} = 1, X \in G)}{P(X \in G)}$

We must take into account that:

- Demographic parity can reject the optimal classifier.

Warning!

Decisions based on a classifier that satisfies independence can have **undesirable properties** (and similar arguments apply to other statistical criteria).

Imagine a company that in group A hire diligently selected applicants at some rate $p > 0$.

In group B, the company hires carelessly selected applicants at the same rate p.

Eventhough the acceptance rates in both groups are identical, it is far more likely that unqualified applicants are selected in one group than in the other.

As a result, it will appear in hindsight that members of group B performed worse than members of group A, thus establishing a negative track record for group B.

Accuracy-based fairness

Accuracy-based fairness warrants that various types of classification **errors** (e.g., true positives, false positives) are equal across groups.

Depending on the type of classification errors considered, the achieved type of fairness takes different names.

Accuracy-based fairness

		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**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = Σ False positive / Σ Predicted condition positiv	
	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 (NP = Σ True negative / Σ Predicted condition negativ	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	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) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score = 2 · Precision · Reca / Precision + Reca
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Accuracy-based fairness

		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**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positiv}}$
	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 (NP = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negativ}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	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) $= \frac{\text{LR+}}{\text{LR−}}$ / F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Reca}}{\text{Precision} + \text{Reca}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

$\hat{Y} = 1 \quad Y = 1$ True positive rate, recall

$\hat{Y} = 0 \quad Y = 1$ False negative rate

$\hat{Y} = 1 \quad Y = 0$ False positive rate

$\hat{Y} = 0 \quad Y = 0$ True negative rate

Accuracy-based fairness

The case in which we ask that

$$P(\hat{Y} = 1 \mid Y = 1, X \in G^{+}) = P(\hat{Y} = 1 \mid Y = 1, X \in G^{-})$$

(same *True Positive Rate*) is called **equal opportunity**.

Reminder: $P(\hat{Y} = 1 \mid Y = 1, X \in G) = \dfrac{P(\hat{Y} = 1, Y = 1, X \in G)}{P(Y = 1, X \in G)}$

Comparing equal opportunity with statistical parity, again the members of the two groups have the same chance of getting the favorable outcome, but <u>only when these members qualify</u>.

Accuracy-based fairness

In the general case, this method can be called **separation**: $\hat{Y}$ must be independent of the protected characteristic, conditional on Y.

Separation acknowledges that in many scenarios, the **sensitive characteristic may be correlated with the target variable**.

A bank might argue that it is a matter of business necessity to therefore have different lending rates for these groups.

For example, one group might have a higher default rate on loans than another.

Roughly speaking, the separation criterion allows correlation between the score and the sensitive attribute to the extent that it is justified by the target variable.

Accuracy-based fairness

The case in which we ask that

$$p(\hat{Y} = 1 \mid Y = 1, X \in G^{+}) = p(\hat{Y} = 1 \mid Y = 1, X \in G^{-})$$

$$p(\hat{Y} = 1 \mid Y = 0, X \in G^{+}) = p(\hat{Y} = 1 \mid Y = 0, X \in G^{-})$$

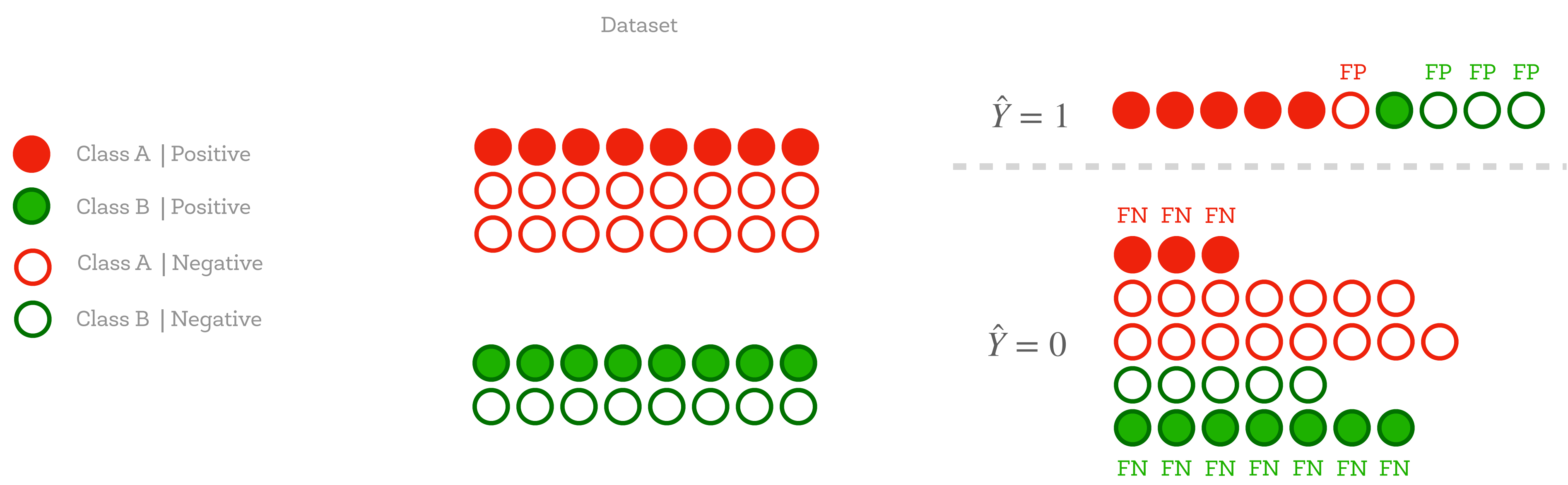
is called **equalized odds**.

All groups experience the same ***true positive rate*** and the same ***false positive rate***.

Accuracy-based fairness

Equalized odds requires both the fraction of non-defaulters that qualify for loans and the fraction of defaulters that qualify for loans to be constant across groups.

Accuracy-based fairness



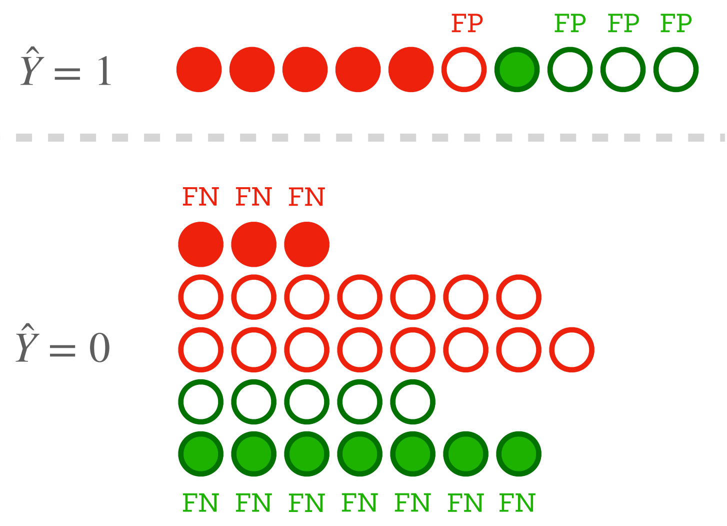
$$TPR_A = \frac{5}{8} \qquad TPR_B = \frac{1}{8}$$

$$FPR_A = \frac{1}{16} \qquad FPR_B = \frac{3}{8}$$

Accuracy-based fairness

In many applications (e.g. hiring), people care more about the true positive rate than false positive rate so many works focus on **equal of opportunity**:

$$p(\hat{Y} = 1 \mid Y = 1, X \in G^+) = p(\hat{Y} = 1 \mid Y = 1, X \in G^-)$$



$$TPR_A = \frac{5}{8}$$

$$TPR_B = \frac{1}{8}$$

Accuracy-based fairness

Separation may not help closing the gap between two groups in the real world.

For example, imagine group A has 100 applicants and 58 of them are qualified while group B also have 100 applicants but only 2 of them are qualified.

If the company decides to accept 30 applicants and satisfies equality of opportunities, 29 offers will be conferred to group A while only 1 offer will be conferred to group B.

If the job is a well-paid job, group A tends to have a better living condition and affords better education for their kids, and thus enable them to be qualified for such well-paid jobs when they grow up. **The gap between group A and group B will tend to be enlarged over time.**

Relationships between criteria

The criteria we reviewed **constrain the joint distribution** $P(X, Y, \hat{Y})$ **in non-trivial ways**. We should therefore suspect that imposing any two of them simultaneously over-constrains the space to the point where only degenerate solutions remain.

It can be shown that if we assume that Y is binary, the protected feature is not independent of Y, and $\hat{Y}$ is not independent of Y, then, **independence and separation cannot both hold.**

It is impossible to satisfy all definitions of group fairness, meaning that the data scientists need to choose one to refer to when starting a fairness analysis.

Relationships between criteria

Incompatibility of fairness metrics doesn't imply that fairness efforts are fruitless.

Instead, it suggests that fairness must be defined **contextually** for a given ML problem, with the goal of preventing **harms specific to its use cases**.

Fairness for decisions

For binary decision procedures, we can summarize a procedure with the confusion matrix, which illustrates match and mismatch between decision $\hat{Y}$ and true status Y.

	Positive Status $Y = 1$	**Negative Status** $Y = 0$	**Prevalence ("base rate")** $P[Y = 1]$	
Positive Decision $d = 1$	**True Positive (TP)**	**False Positive (FP)**	**Positive Predictive Value (PPV), aka precision** $P[Y = 1 \mid d = 1]$	**False Discovery Rate (FDR)** $P[Y = 0 \mid d = 1]$
Negative Decision $d = 0$	**False Negative (FN)**	**True Negative (TN)**	**False Omission Rate (FOR)** $P[Y = 1 \mid d = 0]$	**Negative Predictive Value (NPV)** $P[Y = 0 \mid d = 0]$
Positive Decision Rate $P[d = 1]$	**True Positive Rate (TPR), aka recall, aka sensitivity** $P[d = 1 \mid Y = 1]$	**False Positive Rate (FPR)** $P[d = 1 \mid Y = 0]$	**Accuracy** $P[d = Y]$	
	False Negative Rate (FNR) $P[d = 0 \mid Y = 1]$	**True Negative Rate (TNR), aka specificity** $P[d = 0 \mid Y = 0]$		

Confusion Matrix

Demographic Parity:

$$p(\hat{Y} = 1 \mid X \in G^{+}) = p(\hat{Y} = 1 \mid X \in G^{-})$$

Fairness for decisions

For binary decision procedures, we can **summarize** a procedure with the confusion matrix, which illustrates match and mismatch between decision $\hat{Y}$ and true status Y.

	Positive Status $Y = 1$	Negative Status $Y = 0$	Prevalence ("base rate") $P[Y = 1]$	
Positive Decision $d = 1$	True Positive (TP)	False Positive (FP)	Positive Predictive Value (PPV), aka precision $P[Y = 1 \mid d = 1]$	False Discovery Rate (FDR) $P[Y = 0 \mid d = 1]$
Negative Decision $d = 0$	False Negative (FN)	True Negative (TN)	False Omission Rate (FOR) $P[Y = 1 \mid d = 0]$	Negative Predictive Value (NPV) $P[Y = 0 \mid d = 0]$
Positive Decision Rate $P[d = 1]$	True Positive Rate (TPR), aka recall, aka sensitivity $P[d = 1 \mid Y = 1]$	False Positive Rate (FPR) $P[d = 1 \mid Y = 0]$	Accuracy $P[d = Y]$	
	False Negative Rate (FNR) $P[d = 0 \mid Y = 1]$	True Negative Rate (TNR), aka specificity $P[d = 0 \mid Y = 0]$		

Confusion Matrix

For any box in the confusion matrix involving the decision d, we can define fairness as equality across groups.

For example, **Equal False Omission Rates**:

$p(Y = 1 \mid \hat{Y} = 0, X \in G^{+}) = p(Y = 1 \mid \hat{Y} = 0, X \in G^{-})$

Fairness for scores

Some machine learning systems produce **scores** instead of **labels**, f.e. probabilistic classifiers, recommenders, etc.

Some of the measures we have seen can be generalized to scores.

Fairness for scores

For score outputs, we can consider the following initial definitions of fairness based on equal metrics across groups:

- **Balance for the Positive Class**: the average score assigned to positive members, $\mathbb{E}(S \mid Y = 1)$, should be the same across groups.
- **Balance for the Negative Class**: the average score assigned to negative members, $\mathbb{E}(S \mid Y = 0)$, should be the same across groups.
- **Calibration**: the fraction of those marked with a given score who are actually positive, $\mathbb{E}(Y = 1 \mid S = d)$, should be the same across groups.
- **AUC (Area Under Curve) Parity**: the area under the receiver operating characteristic (ROC) curve should be the same across groups. The AUC can be interpreted as the probability that a randomly chosen positive individual $Y = 1$ is scored higher than a randomly chosen negative individual.

Fairness for scores

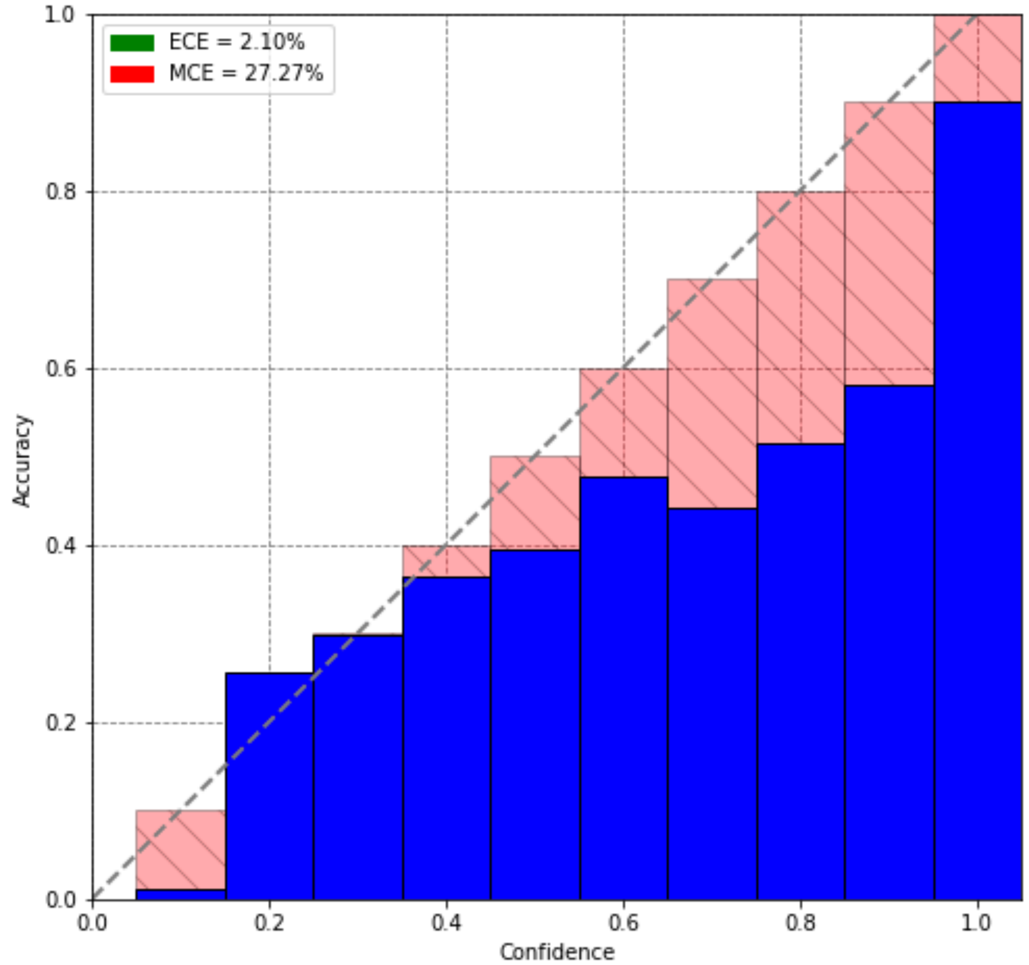
Calibration-based fairness considers probabilistic classifiers that predict a probability p for each class.

In general, a classification algorithm is considered to be **well-calibrated** if: **when the algorithm predicts a set of individuals as having probability p of belonging to the positive class, then approximately a p fraction of this set is actual members of the positive class**.

In terms of fairness, intuitively, we would like the classifier to be **equally well calibrated for both groups**.

Calibration

To get an intuitive understanding of how well a specific model performs in this regard, **Realiability Diagramms** are often used.



Accuracy

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i)$$

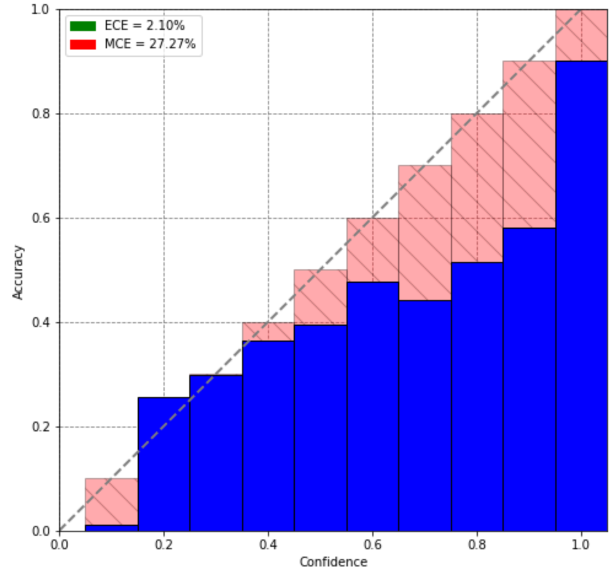
Confidence

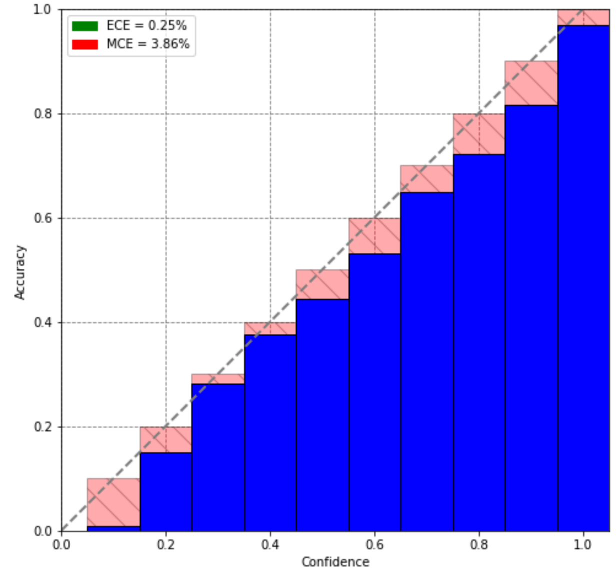
$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$$

Calibration

The **Expected Calibration Error** (ECE) simply takes a weighted average over the absolute accuracy/confidence difference.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$$





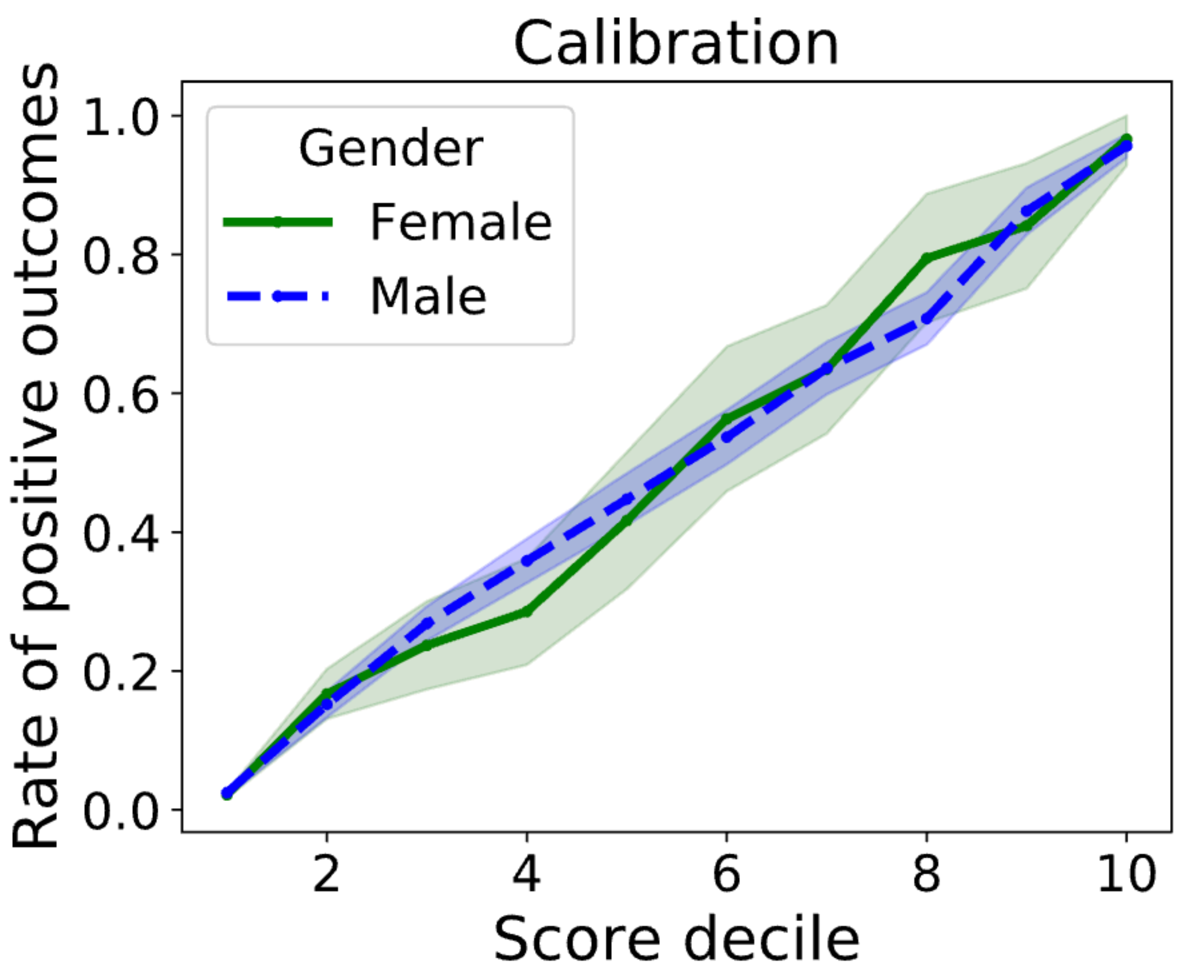
Calibration-based fairness

Calibration-based fairness is asking that for any predicted probability score $p \in [0,1]$, the probability of positives among those with a given score is equal for both groups, i.e.,

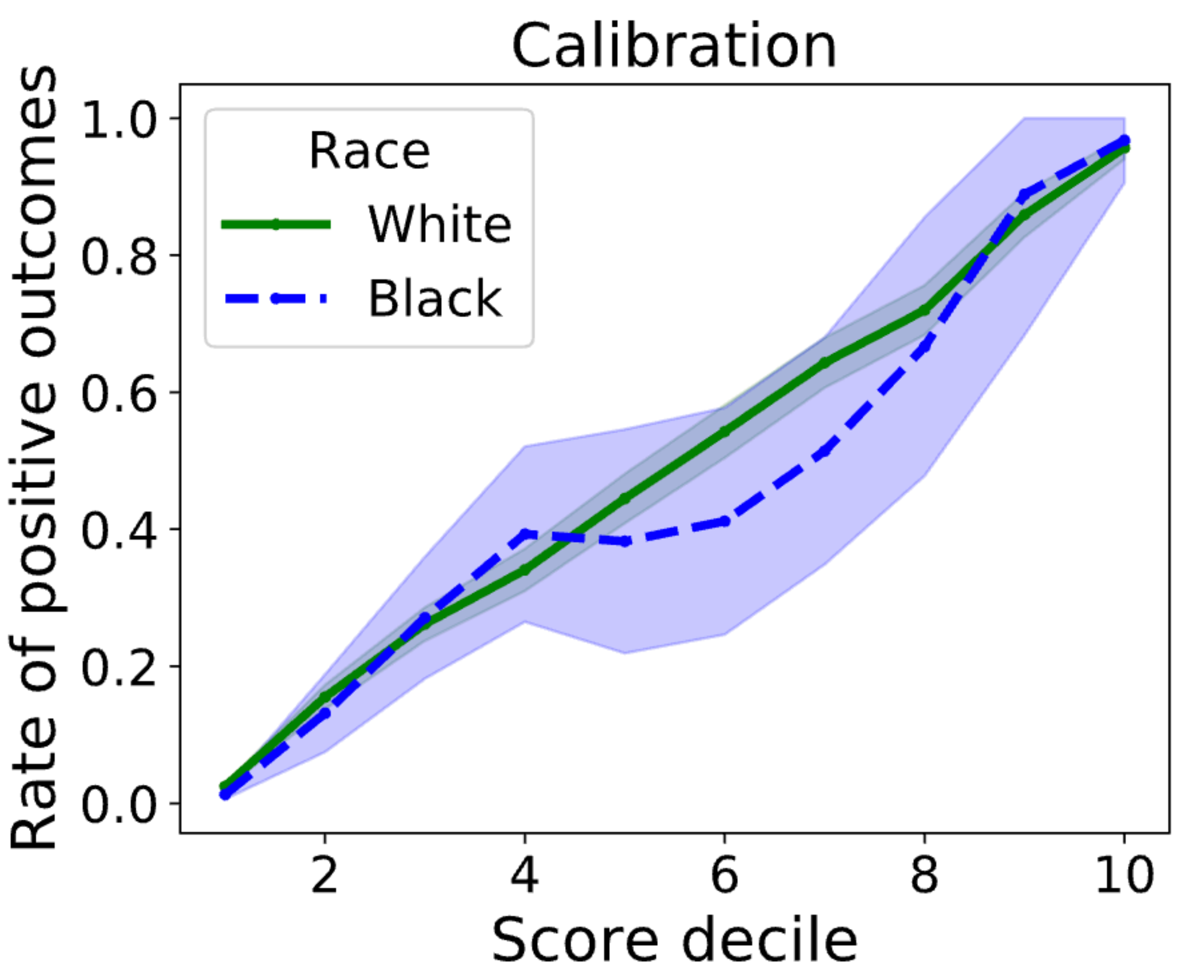
$$P(Y = 1 \mid S = p, X \in G^{+}) = P(Y = 1 \mid S = p, X \in G^{-})$$

Calibration-based fairness

The fraction of those marked with a given score who are actually positive should be the same across groups.



Calibration by gender on UCI adult data. A straight diagonal line would correspond to perfect calibration.



Calibration by race on UCI adult data.

Online Example

Attacking discrimination with smarter machine learning or why fairness is part of a multi-objective task.

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Limitations

Group-based measures in general tend to **ignore the merits of each individual in the group**.

Some individuals in a group may be better for a given task than other individuals in the group, which is not captured by some group-based fairness definitions.

Limitations

This issue may lead to two problematic behaviors, namely,

(a) the **self-fulfilling prophecy** where by deliberately choosing the less qualified members of the protected group we aim at building a bad track record for the group,

and

(b) **reverse tokenism** where by not choosing a well qualified member of the non-protected group we aim at creating convincing refutations for the members of the protected group that are also not selected.

Bias preservation or transformation?

'**Bias preserving**' fairness metrics seek to **reproduce historic performance** in the outputs of the target model with **equivalent error rates** for each group as reflected in the training data (or status quo).
F.e. Equal FPR

$$p(\hat{Y} = 1 \mid Y = 0, X \in G^{+}) = p(\hat{Y} = 1 \mid Y = 0, X \in G^{-})$$

In contrast, '**bias transforming**' metrics do not blindly accept social bias as a given or neutral starting point that should be preserved, but instead require people to make an **explicit decision as to which biases the system should exhibit**.

F.e. Demographic parity

$$p(\hat{Y} = 1 \mid X \in G^{+}) = p(\hat{Y} = 1 \mid X \in G^{-})$$

Bias preservation or transformation?

Bias **preserving** criteria are **always satisfied by a perfect classifier** that exactly predicts its target labels with zero error, **replicating bias present in the data**.

Bias **transforming** metrics are **not necessarily satisfied** by a perfect classifier.

Bias preservation or transformation?

Bias Preservation in Machine Learning:
The Legality of Fairness Metrics Under EU Non-Discrimination Law

Sandra Wachter[1], Brent Mittelstadt[2] and Chris Russell[3]

Fairness metric	Bias preserving?
1. Group fairness, Statistical (demographic) parity	✗
2. Conditional statistical (demographic) parity, Conditional independence	✗
3. Predictive parity, outcome test	✓
4. False positive error rate balance	✓
5. False negative error rate balance, Equal opportunity	✓
6. Equalized odds	✓
7. Conditional use accuracy equality	✓
8. Overall accuracy equality	✓
9. Treatment equality	✓
10. Test-fairness or calibration	✓
11. Well-calibration	✓
12. Balance for positive class	✓
13. Balance for negative class	✓
14. Causal discrimination (direct discrimination)	*
15. Fairness through unawareness	*
16. Fairness through awareness	✗
17. Counterfactual fairness	✗
18. No unresolved discrimination	✗
19. No proxy discrimination	✗
20. Path based causal reasoning	✗

Table 1 – Bias preserving fairness metrics

Bias transforming

DEMOGRAPHIC DISPARITY (DD)

Is the disadvantaged class a **bigger proportion of the rejected outcomes** than the proportion of accepted outcomes for the same class?

$$DD = P(X \in G^+ \mid \hat{Y} = 0) - P(X \in G^+ \mid \hat{Y} = 1)$$

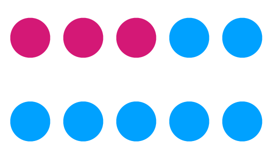
For example, in the case of college admissions, if women applicants comprised 40% of the rejected applicants and comprised only 30% of the accepted applicants, we say that there is **demographic disparity** because the rate at which women were rejected exceeds the rate at which they were accepted.

This applies to cases were we can accept different a priori preferences between groups.

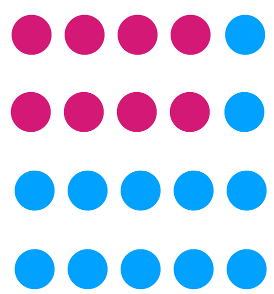
Bias transforming

Woman
Man

$$DD = P(X \in a \mid \hat{Y} = 0) - P(X \in a \mid \hat{Y} = 1)$$



Accepted = 3/10 = 30%



Rejected = 8 /20 = 40%

$$DD = 0.4 - 0.3 = 0.1$$

Bias transforming

CONDITIONAL DEMOGRAPHIC DISPARITY (CDD)

We can condition DD on attributes that define a strata of subgroups on the dataset.

This is necessary to rule out **Simpson's paradox** (a problem that appears when aggregating data).

Example:

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	**44%**	4,321	35%

Graduate school admissions to University of California, Berkeley.

44% of the male applicants were accepted compared to only 35% of female applicants… Is there discrimination?

Bias transforming

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	**44%**	4,321	35%

Graduate school admissions to University of California, Berkeley.

Accepted women: 35% of 4231 = 1481
Accepted applicants: 44% of 8442 + 35% of 4231 = 5195

Non accepted women: 65% of 4231 = 2750
Non accepted applicants: 56% of 8442 + 65% of 4231 = 7478

DD = 2750/7478 - 1481/5195 = 0.08

There is evidence of (small) **demographic disparity**.

Bias transforming

However, **when examining the individual departments**, it appeared that 6 out of 85 departments were significantly biased against men, while 4 were significantly biased against women.

The issue was that women were much more likely to apply to more competitive departments (such as English) that were much more likely to reject graduates of any gender, whereas other departments (such as Engineering) were more lenient.

In the language of demographic parity: although Berkeley's pattern of admission exhibited evidence of demographic disparity, once we condition according to "department applied for", the apparent bias disappears.

Bias transforming

Let's consider these dataset:

	Admitted			Rejected		
Department	*Male*	*Female*	*Total*	*Male*	*Female*	*Total*
A	512	89	601	313	19	332
B	313	17	330	207	8	215
C	120	202	322	205	391	596
D	138	131	269	279	244	523
E	53	94	147	138	299	437
F	22	24	46	351	317	668
Total	**1158**	**557**	**1715**	**1493**	**1278**	**2771**

Table 1 – Berkeley admissions data by department and gender

	Admitted		Rejected	
Department	*Male*	*Female*	*Male*	*Female*
A	85%	15%	94%	6%
B	95%	5%	96%	4%
C	37%	63%	34%	66%
D	51%	49%	53%	47%
E	36%	64%	32%	68%
F	48%	52%	53%	47%
Total	**68%**	**32%**	**54%**	**46%**

Table 2 – Admissions and rejections by gender

$$DD = 0.46 - 0.32 = 0.14$$

Bias against women.

Bias transforming

Let's consider these tables:

	Admitted			Rejected		
Department	*Male*	*Female*	*Total*	*Male*	*Female*	*Total*
A	512	89	601	313	19	332
B	313	17	330	207	8	215
C	120	202	322	205	391	596
D	138	131	269	279	244	523
E	53	94	147	138	299	437
F	22	24	46	351	317	668
Total	**1158**	**557**	**1715**	**1493**	**1278**	**2771**

Table 1 – Berkeley admissions data by department and gender

Simpson's Paradox
is a statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into subpopulations.

	Admitted		Rejected	
Department	*Male*	*Female*	*Male*	*Female*
A	85%	15%	94%	6%
B	95%	5%	96%	4%
C	37%	63%	34%	66%
D	51%	49%	53%	47%
E	36%	64%	32%	68%
F	48%	52%	53%	47%
Total	**68%**	**32%**	**54%**	**46%**

Table 2 – Admissions and rejections by gender

Bias in favour of women

Bias in favour of men

Bias transforming

The **Conditional Demographic Disparity** metric gives a single measure for all the disparities found in the subgroups defined by an attribute (f.e. department) by averaging (each subgroup weighted in proportion to the number of observations it contains) them.

$$CDD = \frac{1}{n} \sum_i n_i DD_i$$

n : Total number of observations

n_i : Number of observations for each subgroup

Conditionally Admitted		Conditionally Rejected	
Male	*Female*	*Male*	*Female*
58%	42%	60%	40%

Table 3 – Admissions data conditioned on department

Small bias in **favour** of women!

Fairness Metrics Summary

Individual Fairness: Distance, Counterfactual.

Demographic parity:

- $P(\hat{Y} = 1 \mid X \in G^+) \sim P(\hat{Y} = 1 \mid X \in G^-)$

G^+ : Protected group
G^- : Non-protected group

Equal opportunity,

- $P(\hat{Y} = 1 \mid Y = 1, X \in G^+) \sim P(\hat{Y} = 1 \mid Y = 1, X \in G^-)$

Equalized odds:

- $p(\hat{Y} = 1 \mid Y = 1, X \in G^+) \sim p(\hat{Y} = 1 \mid Y = 1, X \in G^-)$
- $p(\hat{Y} = 1 \mid Y = 0, X \in G^+) \sim p(\hat{Y} = 1 \mid Y = 0, X \in G^-)$

Calibration:

- $P(\hat{Y} = 1 \mid S = p, X \in G^+) \sim P(\hat{Y} = 1 \mid S = p, X \in G^-)$

(Conditional) Demographic disparity:

- $P(X \in G^+ \mid \hat{Y} = 0) \sim P(X \in G^+ \mid \hat{Y} = 1)$

Case Analysis: Recidivism risk

The criminal **justice** system needs to evaluate a diverse set of **risks**:

- **The risk of committing a new crime after an arrest (recidivism),**
- The risk of committing a new violent crime (violent recidivism),
- The risk of committing an act of violence against another inmate or penitentiary personnel in jail (intra-penitentiary violence),
- The risk of committing an administrative violation such as breaking the conditions of a permit.
- Etc.

Measuring recidivism risk

Structured risk assessment corresponds to a family of methodologies for evaluating these risks using a systematic process, typically in which a number of different items are evaluated.

We can train a ML system to make automatic decisions based on the scores in each item, but most often, a professional makes a decision based on his/her own evaluation of a defendant and the result of a series of items.

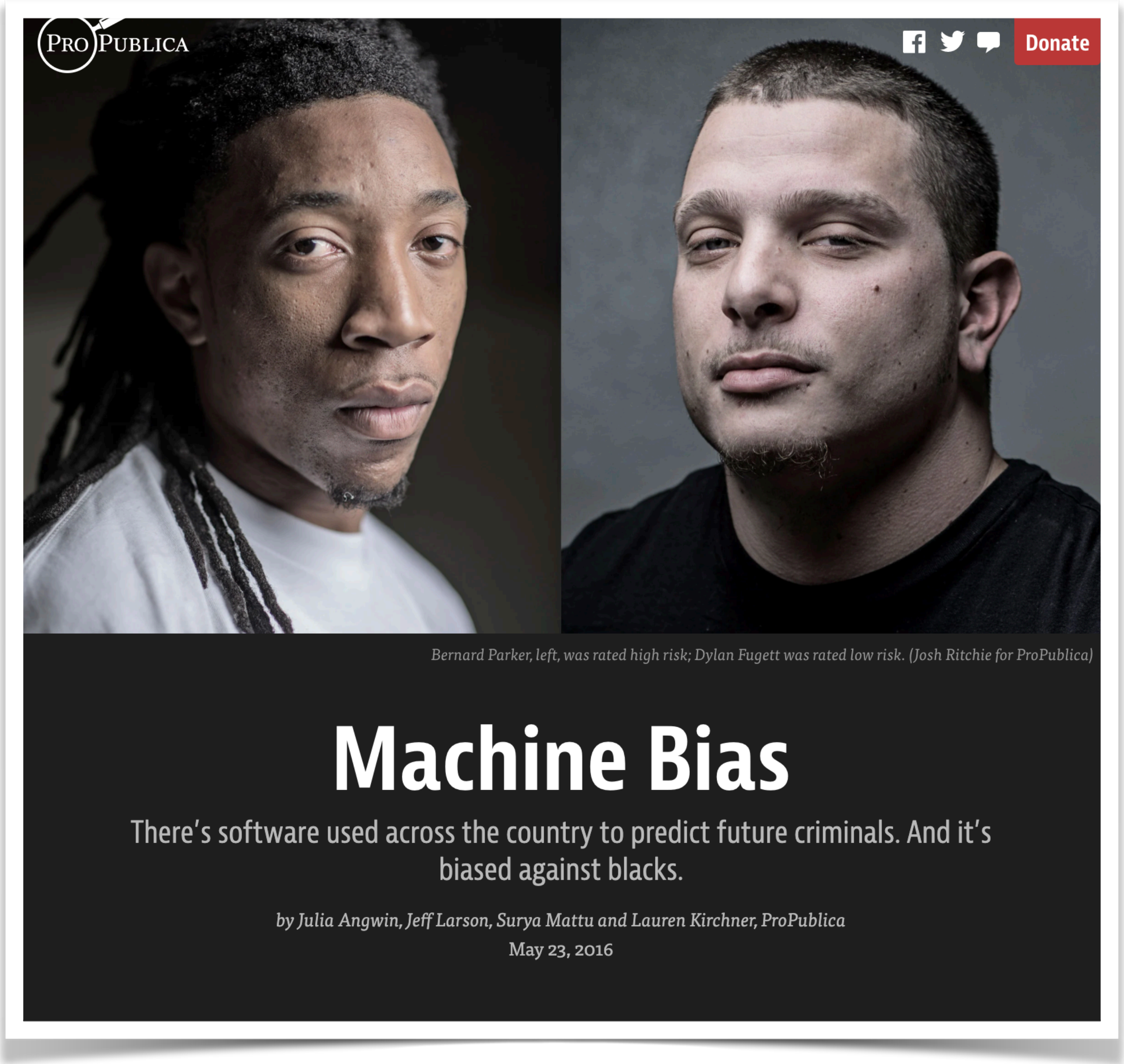
Measuring recidivism risk

COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) is an automated tool that outputs numerical scores, which are labeled, for example, "*risk of recidivism*", "*risk of violent recidivism*", or "*risk of failure to appear*".

These scores are then used in an **unspecified way** to make decisions of *jail*, *bail*, *home arrest*, *release*, etc.

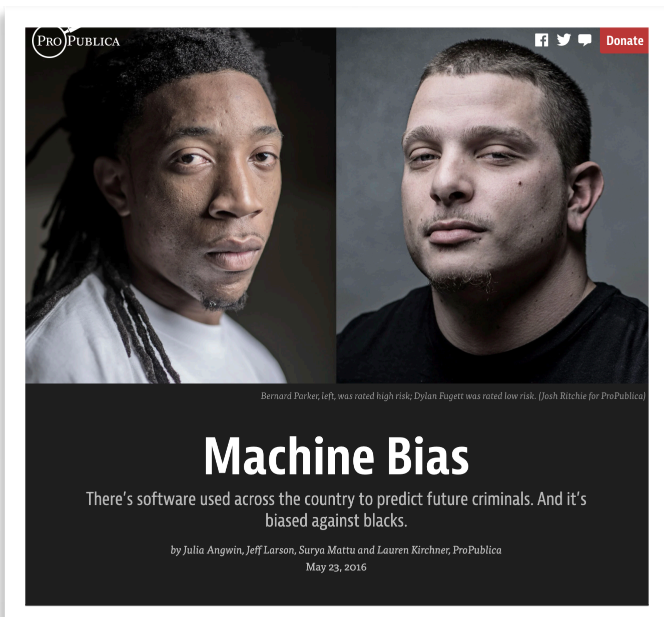
Bail: the temporary release of an accused person awaiting trial.

Measuring recidivism risk



In May 2016, ProPublica, an investigative journal, published a piece called "Machine Bias", in which **COMPAS**, was found **to be biased** against blacks.

Measuring recidivism risk



"We obtained the risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over the next two years, the same benchmark used by the creators of the algorithm."

"...the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years."

"In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

- The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants."

Measuring recidivism risk: the debate

Two of their findings of ProPublica can be phrased in our language as follows:

- COMPAS does not satisfy **equal false negative rates**, in fact, white defendants who did get rearrested ($Y = 1$) were nearly twice as likely to be misclassified as low risk ($\hat{Y} = 0$).

- COMPAS does not satisfy **equal false positive rates**, in fact, black defendants who did not get rearrested ($Y = 0$) were nearly twice as likely to be misclassified as higher risk ($\hat{Y} = 1$).

Accuracy-based fairness

		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**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positiv}}$	
	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 (NP = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negativ}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	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) $= \frac{\text{LR+}}{\text{LR}-}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Reca}}{\text{Precision} + \text{Rec}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Measuring recidivism risk

But the developers did not agree…

Monkey Cage

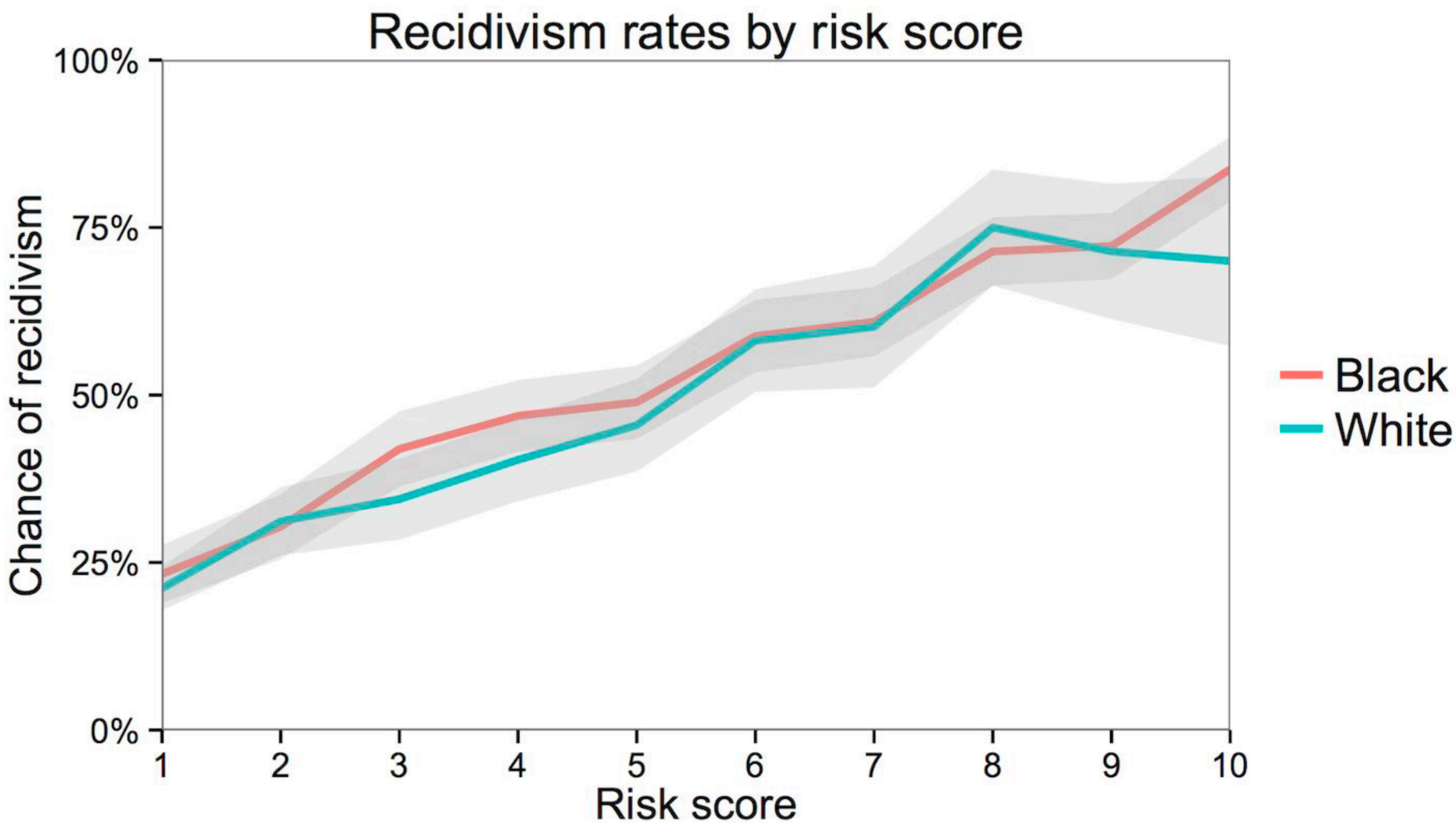
A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel

October 17, 2016

(Rich Pedroncelli/Associated Press)

Measuring recidivism risk



Measuring recidivism risk: the debate

In their response, Equivant/Northpointe, the developers of COMPAS, cited two articles finding that:

- COMPAS satisfies **calibration**: scores mean the same thing regardless of the defendant's race. For example, among defendants with a score of 7, 60 percent of white defendants were rearrested and 61 percent of black defendants were rearrested.

- It can be shown that calibration implies **equal (positive and negative) predictive values** (but not the other way around):

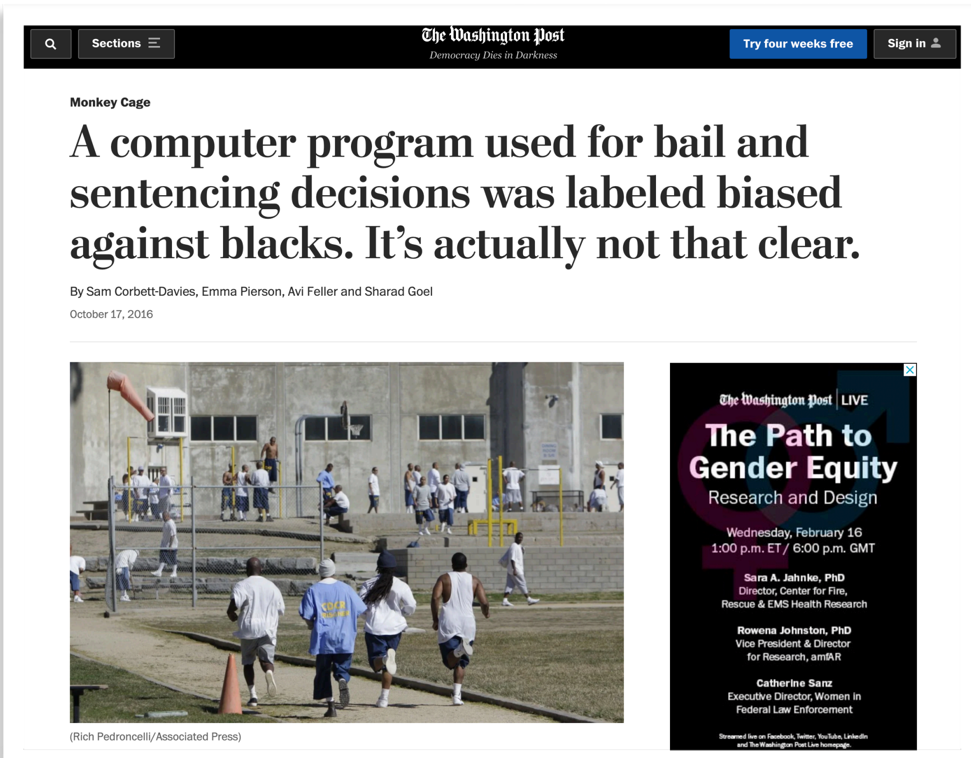
  - among those labeled higher risk ($\hat{Y} = 1$), the proportion of defendants who got rearrested ($Y = 1$) is approximately the same regardless of race.

  - among those labeled lower risk ($\hat{Y} = 0$), the proportion of defendants who did not get rearrested ($Y = 0$) is approximately the same regardless of race.

Accuracy-based fairness

		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**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positiv}}$	
	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 (NP $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negativ}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	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) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Reca}}{\text{Precision} + \text{Reca}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Accuracy-based fairness


Sections
The Washington Post
Democracy Dies in Darkness
Try four weeks free
Sign in

Monkey Cage

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel

October 17, 2016

(Rich Pedroncelli/Associated Press)

WaPo: "Here's the problem: it's actually impossible for a risk score to satisfy both fairness criteria at the same time."

You can't have

1. Equal predictive values (PPV and NPV) and
2. Equal error rates (FPR and FNR, specificity and sensitivity)

[If prevalence is not equal.]

Measuring recidivism risk: analysis

RecidivismCaseStudy

Case study on evaluating statistical tools that predict recidivism.

View the Project on GitHub
AllenDowney/RecidivismCaseStudy

Recidivism Case Study

This case study is based on two articles that were published in 2016:

- "Machine Bias", by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, and published by ProPublica.
- A response by Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel: "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.", published in the Washington Post.

Both articles are about COMPAS, a statistical tool used in the justice system to assign defendants a "risk score" that is intended to reflect the risk that they will commit another crime if released.

The ProPublica article evaluates COMPAS as a binary classifier and compares its error rates for black and white defendants. It concludes that COMPAS is unfair to black defendants because they are more likely to be misclassified as high risk.

In response, the Washington Post article shows that COMPAS has the same predictive value for black and white defendants. And they explain that the test cannot have the same predictive value and the same error rates at the same time.

The purpose of this case study is to understand these conflicting claims, to learn about classification algorithms and the metrics we use to evaluate them, and to think about fairness and the ethics of data science.

The notebooks

- In the first notebook I replicate the analysis from the ProPublica article and define the basic metrics we use to evaluate classification algorithms, including error rates and predictive values.
- In the second notebook I replicate the analysis from the WaPo article and define the calibration curve, the ROC curve, and a related metric, concordance.
- In the third notebook I use the same methods to evaluate the performance of COMPAS for male and female defendants, and lay out the fundamental conflict between two definitions of fairness.

This project is maintained by AllenDowney

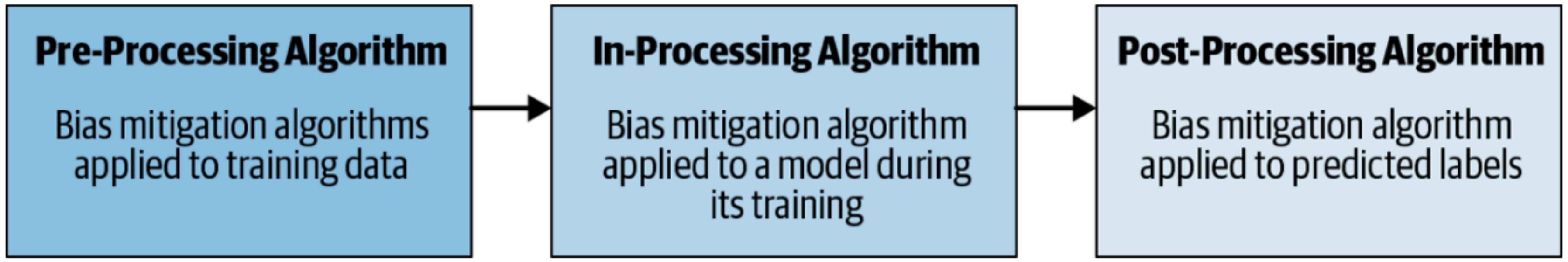
Hosted on GitHub Pages — Theme by orderedlist

https://allendowney.github.io/RecidivismCaseStudy/

Bias Mitigation

The field of **bias mitigation** strategies can be categorised into three types:

- Pre-processing methods manipulate the data to eliminate bias **before** a machine learning (ML) model is able to incorporate these biases based on the data.
- In-processing bias mitigation strategies manipulate the model to mitigate bias that appears **during** the training process.
- Post-processing methods **alter the outcomes** of a model, preying on bias present in the output.



Bias Mitigation

Pre-processing techniques

- Reweighing Pre-Processing: **Generates weights** for the training samples in each (group, label) combination differently to ensure fairness before classification. It does not change any feature or label values, so this is ideal if you are unable to make value changes.

- Optimized Pre-Processing: Learns a probabilistic transformation that **edits the features and labels** in the data with group fairness, individual distortion, and data fidelity constraints and objectives.

- Learning Fair Representations: Finds a **latent representation** that encodes the data well but obfuscates information about protected attributes.

Bias Mitigation

In-processing techniques

- Adversarial Debiasing: **Learns a classifier to maximize prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions.** This approach leads to a fair classifier because the predictions can't carry any group discrimination information that the adversary can exploit.

- Prejudice Remover: Adds a **discrimination-aware regularization** term to the learning objective.

- Meta Fair Classifier: Takes the **fairness metric as part of the input** and returns a classifier optimized for the metric.

Bias Mitigation

Post-processing techniques

- Equalized Odds: Solves a linear program to find probabilities with which to change output labels to optimize equalized odds.

- Calibrated Equalized Odds: Optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective.

- Reject Option Classification: Gives favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band aroundthe decision boundary with the highest uncertainty.

Bias Mitigation

IBM Research Trusted AI | Home | **Demo** | Resources | Events | Videos | Community

AI Fairness 360 - Demo

Data — Check — Mitigate — Compare

Next

1. Choose sample data set

Bias occurs in data used to train a model. We have provided three sample datasets that you can use to explore bias checking and mitigation. Each dataset contains attributes that should be protected to avoid bias.

◯ **Compas (ProPublica recidivism)**

Predict a criminal defendant's likelihood of reoffending.
Protected Attributes:
- **Sex**, privileged: ***Female***, unprivileged: ***Male***
- **Race**, privileged: ***Caucasian***, unprivileged: ***Not Caucasian***

Learn more

◯ **German credit scoring**

Predict an individual's credit risk.
Protected Attributes:
- **Sex**, privileged: ***Male***, unprivileged: ***Female***
- **Age**, privileged: ***Old***, unprivileged: ***Young***

Learn more

◯ **Adult census income**

Predict whether income exceeds $50K/yr based on census data.
Protected Attributes:
- **Race**, privileged: ***White***, unprivileged: ***Non-white***
- **Sex**, privileged: ***Male***, unprivileged: ***Female***

Learn more

Bias Mitigation

Supported bias mitigation algorithms

- Optimized Preprocessing (Calmon et al., 2017)
- Disparate Impact Remover (Feldman et al., 2015)
- Equalized Odds Postprocessing (Hardt et al., 2016)
- Reweighing (Kamiran and Calders, 2012)
- Reject Option Classification (Kamiran et al., 2012)
- Prejudice Remover Regularizer (Kamishima et al., 2012)
- Calibrated Equalized Odds Postprocessing (Pleiss et al., 2017)
- Learning Fair Representations (Zemel et al., 2013)
- Adversarial Debiasing (Zhang et al., 2018)
- Meta-Algorithm for Fair Classification (Celis et al., 2018)
- Rich Subgroup Fairness (Kearns, Neel, Roth, Wu, 2018)
- Exponentiated Gradient Reduction (Agarwal et al., 2018)
- Grid Search Reduction (Agarwal et al., 2018, Agarwal et al., 2019)
- Fair Data Adaptation (Plečko and Meinshausen, 2020, Plečko et al., 2021)

Supported fairness metrics

- Comprehensive set of group fairness metrics derived from selection rates and error rates including rich subgroup fairness
- Comprehensive set of sample distortion metrics
- Generalized Entropy Index (Speicher et al., 2018)
- Differential Fairness and Bias Amplification (Foulds et al., 2018)
- Bias Scan with Multi-Dimensional Subset Scan (Zhang, Neill, 2017)

Bias Mitigation

http://aif360.mybluemix.net/

Reweighting

Sampling, massaging, reweighing and suppression are among different pre-processing bias mitigation techniques proposed from academic literature.

The advantage of reweighting is, instead of modifying the labels, **it assigns different weights to the examples based upon their categories of protected attribute and outcome such that bias is removed from the training dataset**.

The weights are based on frequency counts.

Reweighting

Reweighting works by postulating that a fair data set D would show no conditional dependence of the outcome on a protected attribute.

Hence, it postulates **group membership and outcome should be statistically independent**.

$$P(Y = a, X \in G) = P(Y = a)P(X \in G) = \frac{|\{Y = a\}|}{|D|} \times \frac{|\{X \in G\}|}{|D|}$$

Reweighting **adjusts the data point weights** to make this so.

Reweighting: Adult dataset

The binary target in our example is whether an individual has an income higher or lower than $50k.

It contains several features that are protected by the law, but for simplicity, we will focus on sex.

As can be seen in the table, **Male is the privileged group** with a 31% probability of having a positive outcome (>$50k) compared to an 11% probability of having a positive outcome for the Female group.

Sex	Salary	Count	Class Probability
Female	<=$50k	6,680	0.89
Female	>$50k	828	0.11
Male	<=$50k	10,605	0.69
Male	>$50k	4,679	0.31

$-np$ Negative non privileged

$+np$ Positive non privileged

$-p$ Negative privileged

$+p$ Positive privileged

Reweighting: Adult dataset

Using the frequency counts in the table, the reweighing technique will assign weights as follows:

$$w_{+p} = \frac{n_p \times n_+}{n \times n_{+p}} \qquad w_{+np} = \frac{n_{np} \times n_+}{n \times n_{+np}}$$

$$w_{-p} = \frac{n_p \times n_-}{n \times n_{-p}} \qquad w_{-np} = \frac{n_{np} \times n_-}{n \times n_{-np}}$$

Assignment:
Recividism Analysis in Catalonia

The dataset corresponds to a set of juvenile offenders in Catalonia who were evaluated using **SAVRY**, a structured risk assessment tool. The data on recidivism indicates if the same people committed a new offence in 2013-2015.

Objectives of the assignment:

- To compare the performance of SAVRY and ML-based methods, in terms of both **accuracy** and **fairness** metrics.
- To analyze the **causes of unfairness**.
- To explore a **mitigation** strategy.

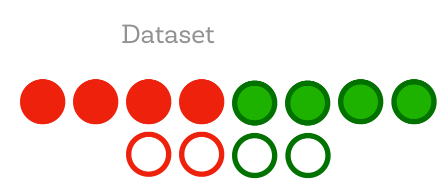
Equalized Base Rates

Let's suppose we have a binary decision problem $Y \in \{-1,1\}$ and my protected feature is $X \in \{A, B\}$. My dataset D is:

I have Equal Base Rates if $P_D(A) = P_D(B)$ and $P_{D'}(Y|X) = P_D(Y|X)$, which is not the case.

randomly resample the training dataset

In this case I need to **oversample** class B in order to get $6 - 3 = 3$ additional samples! Two of these samples will be oversampled from the positive pool. The other one must be sampled from the negative pool. The result is a new dataset D':



Now $P_{D'}(A) = P_{D'}(B)$ and $P_{D'}(Y|X) = P_D(Y|X)$.