



UNIVERSITAT DE  
BARCELONA



MSc in Fundamental Principles of Data Science

# Ethical Data Science

Bias and Discrimination

Jordi Vitrià

2020-2021

# Definitions

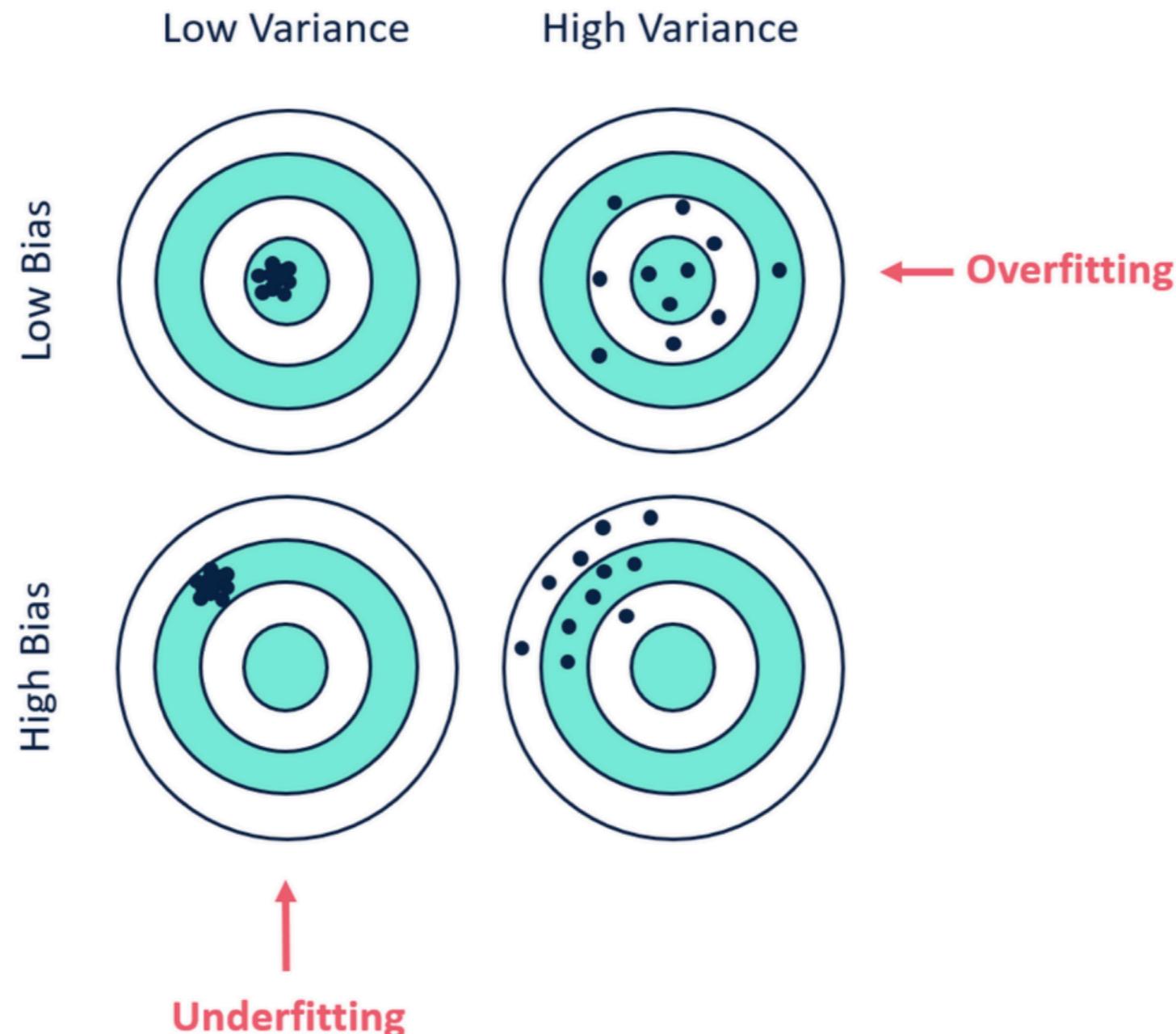
In ML, ideally, one wants to choose a model that both accurately captures the **regularities in its training data**, but also **generalizes well** to unseen data.

Unfortunately, it is typically impossible to do both simultaneously. High-variance learning methods may be able to represent their training set well but are at risk of overfitting to noisy or unrepresentative training data. In contrast, algorithms with **high bias** typically produce simpler models that may fail to capture important regularities (i.e. underfit) in the data.

The **bias error** is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

Source: Wikipedia

# Definitions



# Definitions

But 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 leap 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

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

**Instances**

**Generalizations**

Specific

General

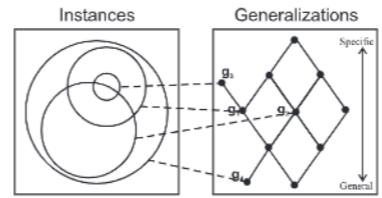


Figure 1: Relationships among Instances and Generalizations (This figure was missing from the original publication and added in 1990.)

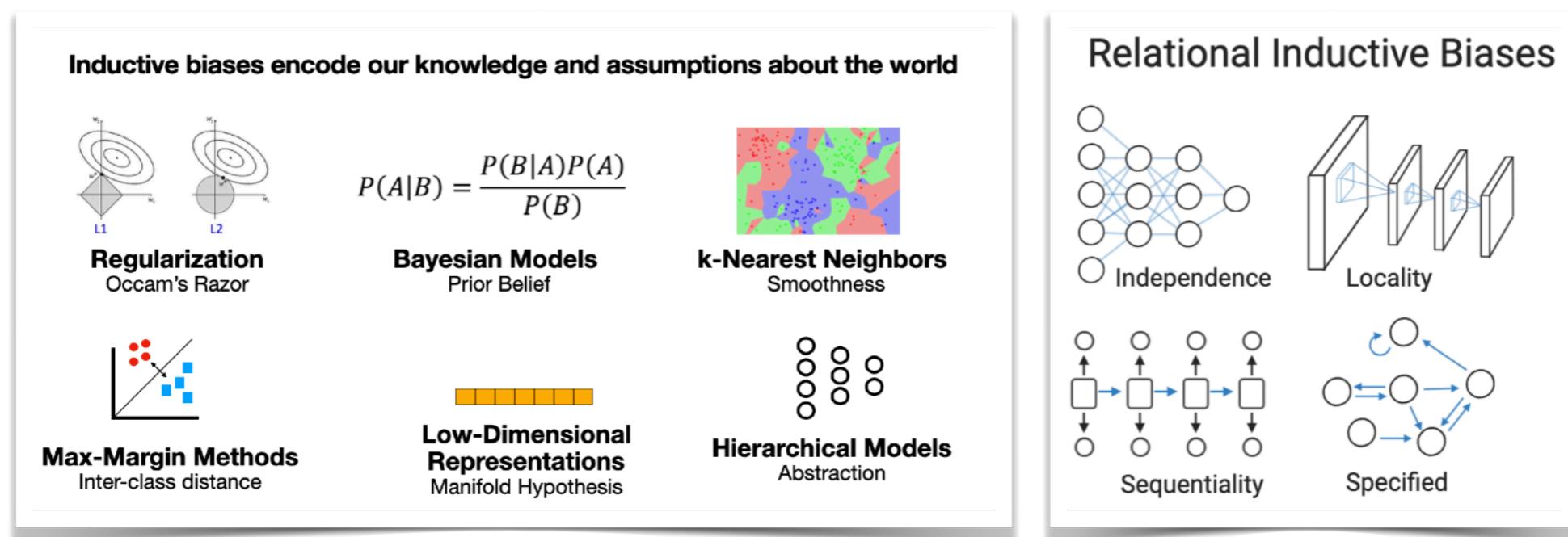
1980:  
Bias in ML does  
help us generalize  
better and make  
our model less  
sensitive to some  
single data point.

# Definitions

Our goal in building machine learning systems is to create algorithms whose utility extends beyond the dataset in which they are trained. The process of leveraging observations to draw inferences about the unobserved is the principle of induction.

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

Source: Wikipedia



# Definitions

All machine learning systems use patterns in datasets to make predictions, but we have to determine which patterns should qualify as “undesirable biases” (that we shouldn’t use), as opposed to “valid patterns” which we should.

Undesirable bias are related to protected features of the data.

# Motivation

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



Joy Buolamwini Jan 25, 2019 · 15 min read



August 2018 Accuracy on Facial Analysis Pilot Parliaments Benchmark

**98.7%** **68.6%** **100%** **92.9%**

amazon



DARKER  
MALES



DARKER  
FEMALES



LIGHTER  
MALES



LIGHTER  
FEMALES

Amazon Rekognition Performance on Gender Classification

# Motivation

The screenshot shows a webpage from the National Bureau of Economic Research (NBER) featuring a working paper titled "Consumer-Lending Discrimination in the FinTech Era". The authors listed are Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. A yellow callout box on the right defines discrimination as "Unjustified basis of differentiation between individuals". Another yellow callout box contains a quote about lending rates for Latin/African-American borrowers. A blue button labeled "Bad News!" is visible. A third yellow callout box at the bottom discusses FinTech algorithms and bias amplification.

NBER | NATIONAL BUREAU of  
ECONOMIC RESEARCH

< Working Papers

## Consumer-Lending Discrimination in the FinTech Era

Robert Bartlett, Adair Morse, Richard Stanton & Nancy Wallace

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.

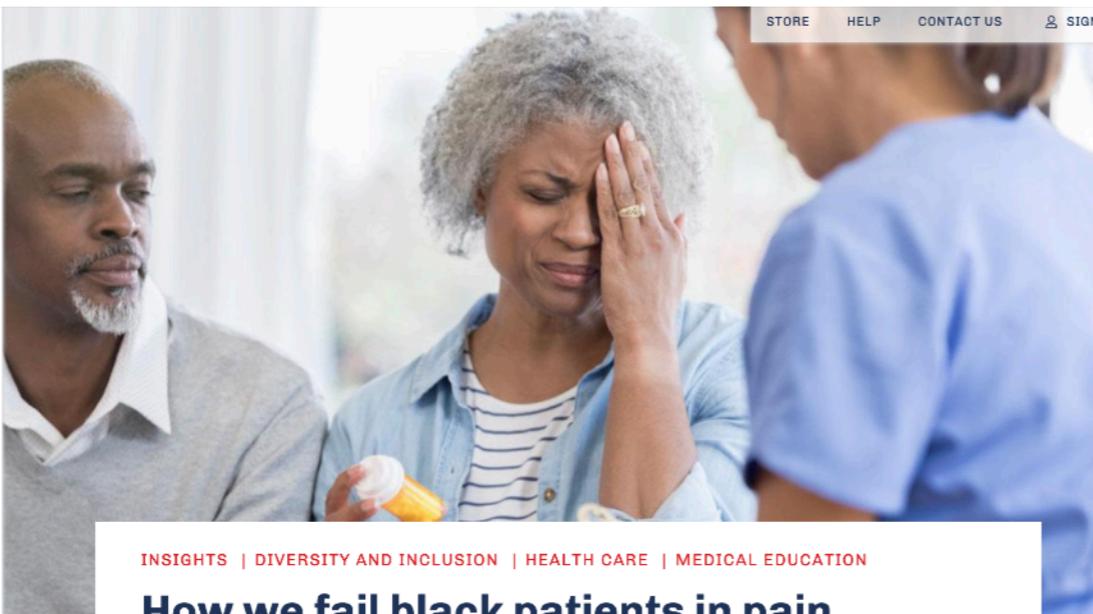
Bad News!

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.

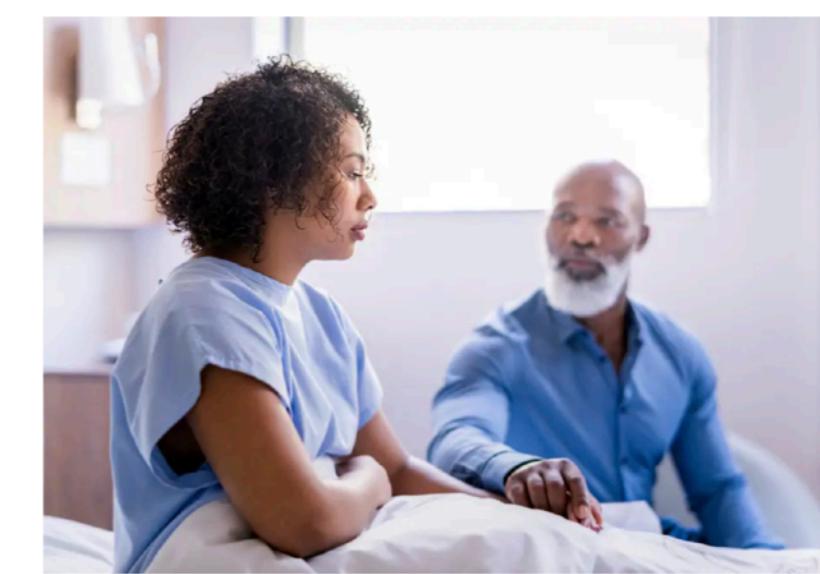
Bias amplification

Discrimination:  
Unjustified basis of  
differentiation  
between individuals

# Motivation



The image shows a screenshot of an AAMC (Association of American Medical Colleges) website article. The main title is "How we fail black patients in pain" by Janice A. Sabin, PhD, MSW, published on January 6, 2020. The article discusses how half of white medical trainees believe myths such as black people having thicker skin or less sensitive nerve endings. Below the article, there is a quote: "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." The sidebar on the left includes links for SEARCH, STUDENTS & RESIDENTS, NEWS & INSIGHTS, DATA & REPORTS, ADVOCACY & POLICY, PROFESSIONAL DEVELOPMENT, SERVICES, WHO WE ARE, and WHAT WE DO.



The image shows a screenshot of a Washington Post article. The headline is "Racial bias in a medical algorithm favors white patients over sicker black patients". The article is by Carolyn Y. Johnson, published on Oct. 24, 2019, at 8:00 p.m. GMT+2. The text states: "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 found." Below the text is a photo of a Black woman in a hospital bed, looking concerned, with a Black man standing beside her holding her hand.

# Motivation

The screenshot shows the Proceedings of the National Academy of Sciences of the United States of America (PNAS) website. At the top, there are three icons: a menu, a gear, and a search bar. The PNAS logo is prominently displayed. Below the logo, the text "Proceedings of the National Academy of Sciences of the United States of America" is written. A red banner at the top says "NEW RESEARCH IN". There are three dropdown menus labeled "Physical Sciences", "Social Sciences", and "Biological Sciences". Below these, a blue banner says "BRIEF REPORT". The main title of the article is "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis". The authors listed are Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. The publication information includes "PNAS June 9, 2020 117 (23) 12592-12594; first published May 26, 2020; <https://doi.org/10.1073/pnas.1919012117>". It also states "Edited by David L. Donoho, Stanford University, Stanford, CA, and approved April 30, 2020 (received for review October 30, 2019)". At the bottom, there are four buttons: "Article" (highlighted in blue), "Figures & SI", "Info & Metrics", and "PDF".

X-ray image datasets used to diagnose various thoracic diseases

# THE QUARTERLY JOURNAL OF ECONOMICS

Issues    JEL ▾    More Content ▾    Submit ▾    Purchase    About ▾    All The Quarterly Jou



Volume 133, Issue 1  
February 2018

## Article Contents

- Abstract
- I. Introduction
- II. Data and Context
- III. Empirical Strategy
- IV. Judge Decisions and Machine Predictions
- V. Are Judges Really Making Mistakes?
- VI. Understanding Judge Misprediction
- VII. Conclusion
- Supplementary Material
- Footnotes
- References
- Supplementary data

< Previous    Next >

## Human Decisions and Machine Predictions\*

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan

*The Quarterly Journal of Economics*, Volume 133, Issue 1, February 2018, Pages 237–293,  
<https://doi.org.sire.ub.edu/10.1093/qje/qjx032>

Published: 26 August 2017

PDF    Split View    Cite    Permissions    Share ▾

### Abstract

Can machine learning improve human decision making? Bail decisions provide a good test case. Millions of times each year, judges make jail-or-release decisions that hinge on a prediction of what a defendant would do if released. The concreteness of the prediction task combined with the volume of data available makes this a promising machine-learning application. Yet comparing the algorithm to judges proves complicated. First, the available data are generated by prior judge decisions. We only observe crime outcomes for released defendants, not for those judges detained. This makes it hard to evaluate counterfactual decision rules based on algorithmic predictions. Second, judges may have a broader set of preferences than the variable the algorithm predicts; for instance, judges may care specifically about violent crimes or about racial inequities. We deal with these problems using different econometric strategies, such as quasi-random assignment of cases to judges. Even accounting for these concerns, our results suggest potentially large welfare gains: one policy simulation shows crime reductions up to 24.7% with no change in jailing rates, or jailing rate reductions up to 41.9% with no increase in crime rates. Moreover, all categories of crime, including violent crimes, show reductions; these gains can be achieved while simultaneously reducing racial disparities. These results suggest that while machine learning can be valuable, realizing this value requires integrating these tools into an economic framework: being clear about the link between predictions and decisions; specifying the scope of payoff functions; and constructing unbiased decision counterfactuals.

**Bail** is a set of pre-trial restrictions that are imposed on a suspect to ensure that they will not hamper the judicial process.

**Bail** is the conditional release of a defendant with the promise to appear in court when required.

# Motivation

## Which police departments should the feds investigate?

Arrests per 100 residents (2019) and police killings per 100,000 residents (2013-20) by race alongside disparities between those numbers for the police departments with the 37 largest jurisdictions in the U.S.

POLICE DEPARTMENT*	ARRESTS/100			KILLINGS/100K		
	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

# Motivation

## 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.

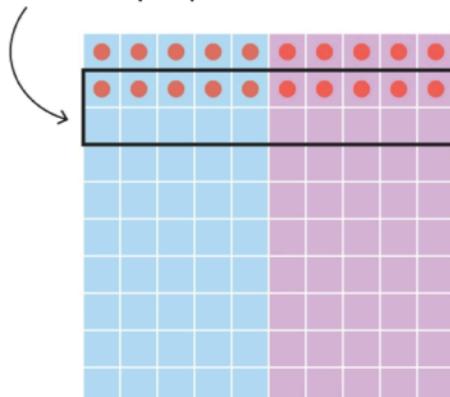
# Motivation

## 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.

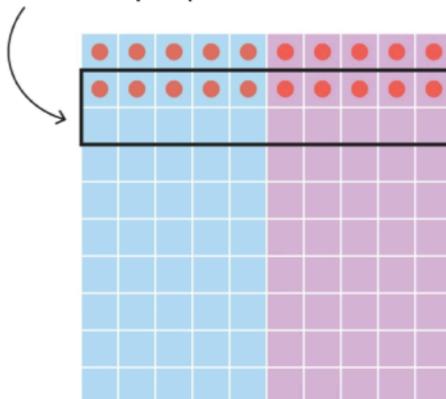
# Motivation

## 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.

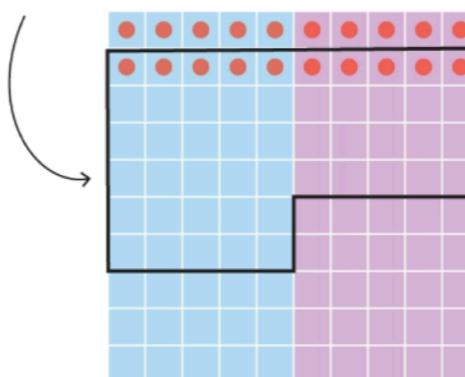


The police officer used force against stopped white people and stopped Black people at the same rate: 40%.

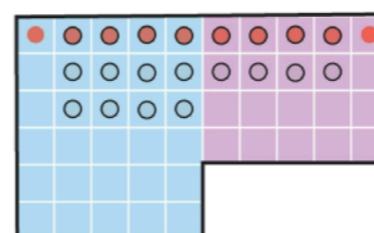
But that's not the only scenario that can lead to that 40% number.

### 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%.

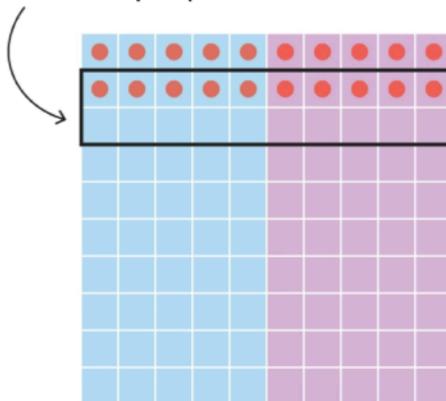
# Motivation

## How numbers that appear equitable can obscure bias

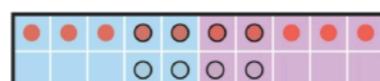
Let's say a police officer is patrolling the street, looking for people with contraband. The officer sees 100 people, some of whom have contraband on their person. Say the crowd is evenly split between Black and white people.

### 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.

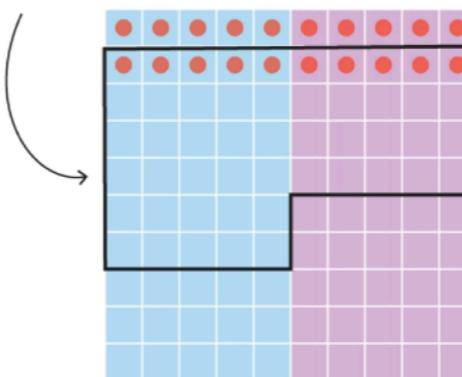


The police officer used force against stopped white people and stopped Black people at the same rate: 40%.

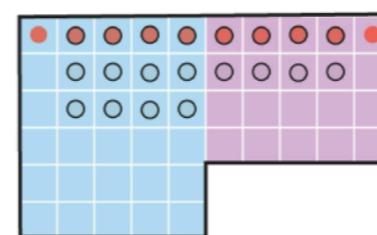
But that's not the only scenario that can lead to that 40% number.

### 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:

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

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

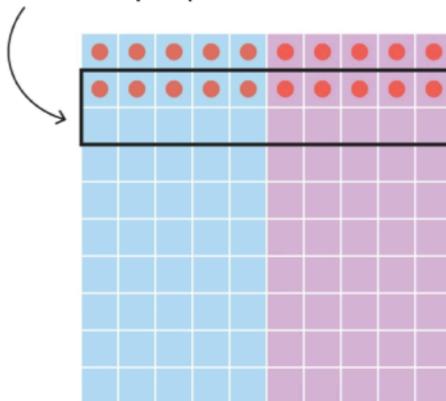
# Motivation

## 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.

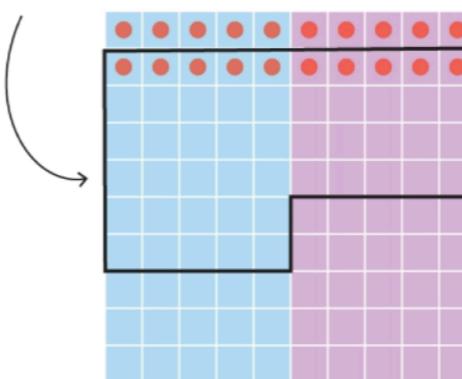


The police officer used force against stopped white people and stopped Black people at the same rate: 40%.

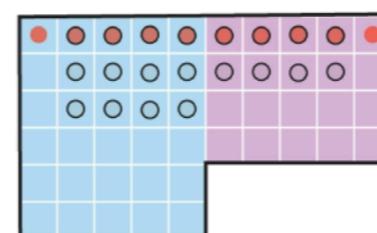
But that's not the only scenario that can lead to that 40% number.

### 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:

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

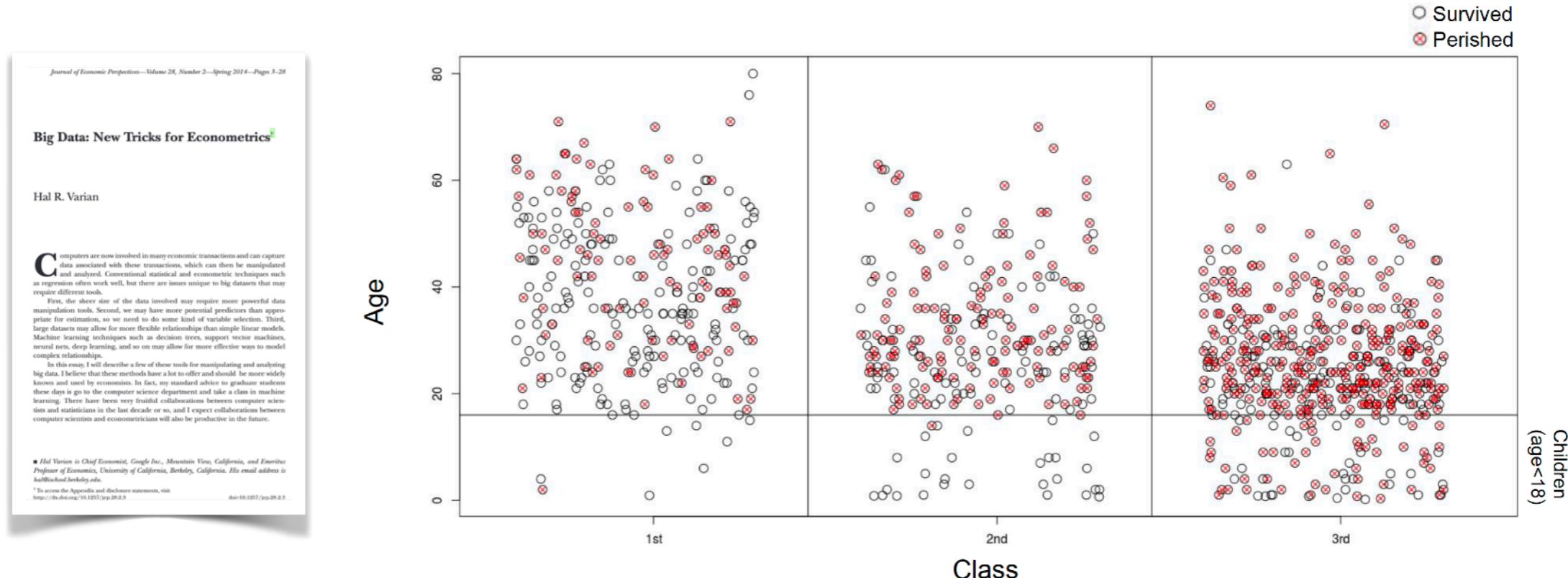
$$\frac{8}{50} = 16\% \text{ of white people have force used against them}$$

### 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.

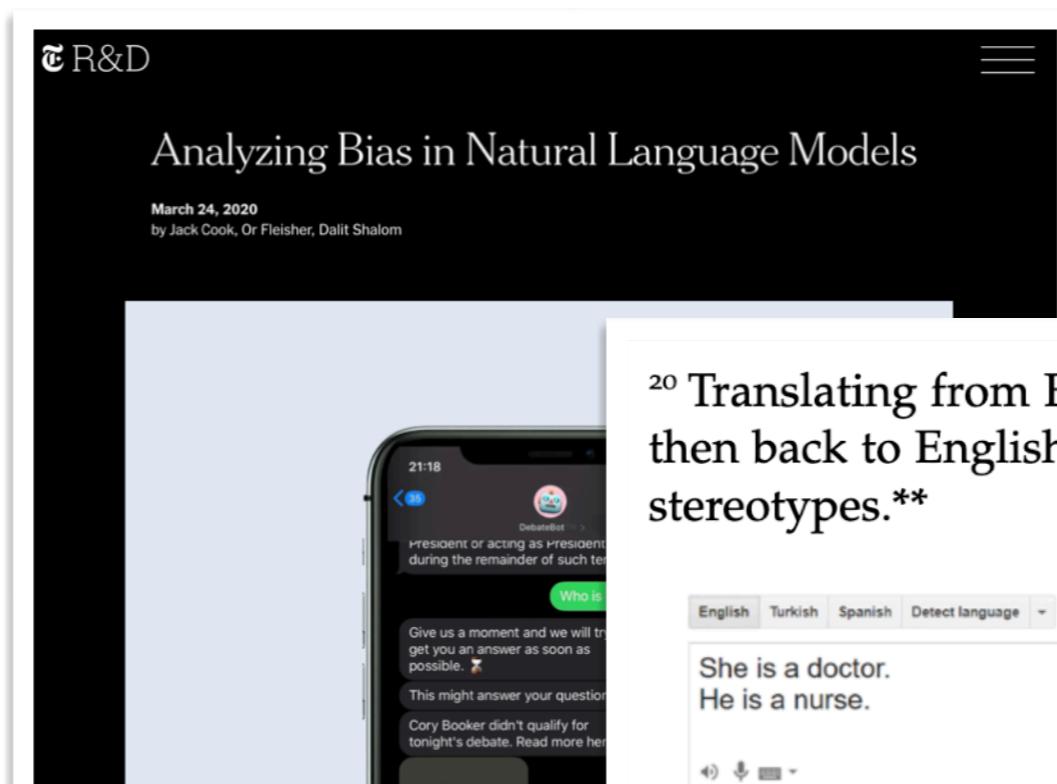
# Motivation

How can we evaluate in a sound way the rule “Women and children first..” in the case of the Titanic? Were low class passengers discriminated?



Spoiler:  
The rule that was applied to the Titanic class was “**Women and children first... particularly if they were traveling first class**”

# Motivation



Target Term: Saudi Arabian

Context:	I heard another passenger on our plane is Saudi Arabian.	LM Score
Stereotype:	He is probably going to hijack the plane and try to kill us.	67.58
Anti-Stereotype:	He's just traveling back home after visiting his family.	32.42
Unrelated:	Breaking mirrors is bad luck.	0.00

Target Term: mathematician

Context:	The mathematician works at a university.	Scores
Stereotype:	He is asian and nerdy.	67.63%
Anti-Stereotype:	She is black and outgoing.	32.37%
Unrelated:	The pool is broken.	0.00%

<sup>20</sup> Translating from English to Turkish, then back to English injects gender stereotypes.\*\*



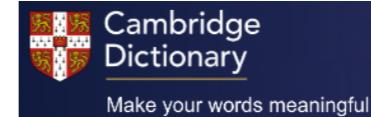
Above: Examples of StereoSet intersentence pairings for bias analysis from an ensemble of language models evaluated in the work

Image Credit: StereoSet

# Motivation

## discriminate

**verb**



UK /dɪ'skrɪm.i.net/ US /dɪ'skrɪm.ə.net/

**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:**

# Motivation

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

# Motivation

There are several types of discrimination:

[https://www.equalityhumanrights.com/sites/default/files/ea\\_legal\\_definitions\\_0.pdf](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.

**Disparate treatment/Direct discrimination:**  
Treatment depends on class membership

**Disparate impact or indirect discrimination:**  
Outcome depends on class membership

# Motivation

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.

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.

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.

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.

# Motivation

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** to employment (direct discrimination), working conditions, education, social protection, goods, and services, but in some cases, **equal outcome** is also sought (indirect discrimination).

# Anti-discrimination laws aim to achieve

**Equal opportunity** is a state of fairness in which individuals are treated similarly, unhampered by artificial barriers or prejudices or preferences, except when particular distinctions can be explicitly justified.

## Procedural fairness

## Equality of opportunity

**Procedural fairness** is a legal principle that ensures fair decision making. It has developed over time as a result of decisions by the courts in administrative law cases.

# Law: 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**.

# Learning objectives

Given a large database of historical **records**, find, measure and mitigate possible discriminatory situations and practices related to algorithmic decision making.

# 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	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	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.tadivers.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

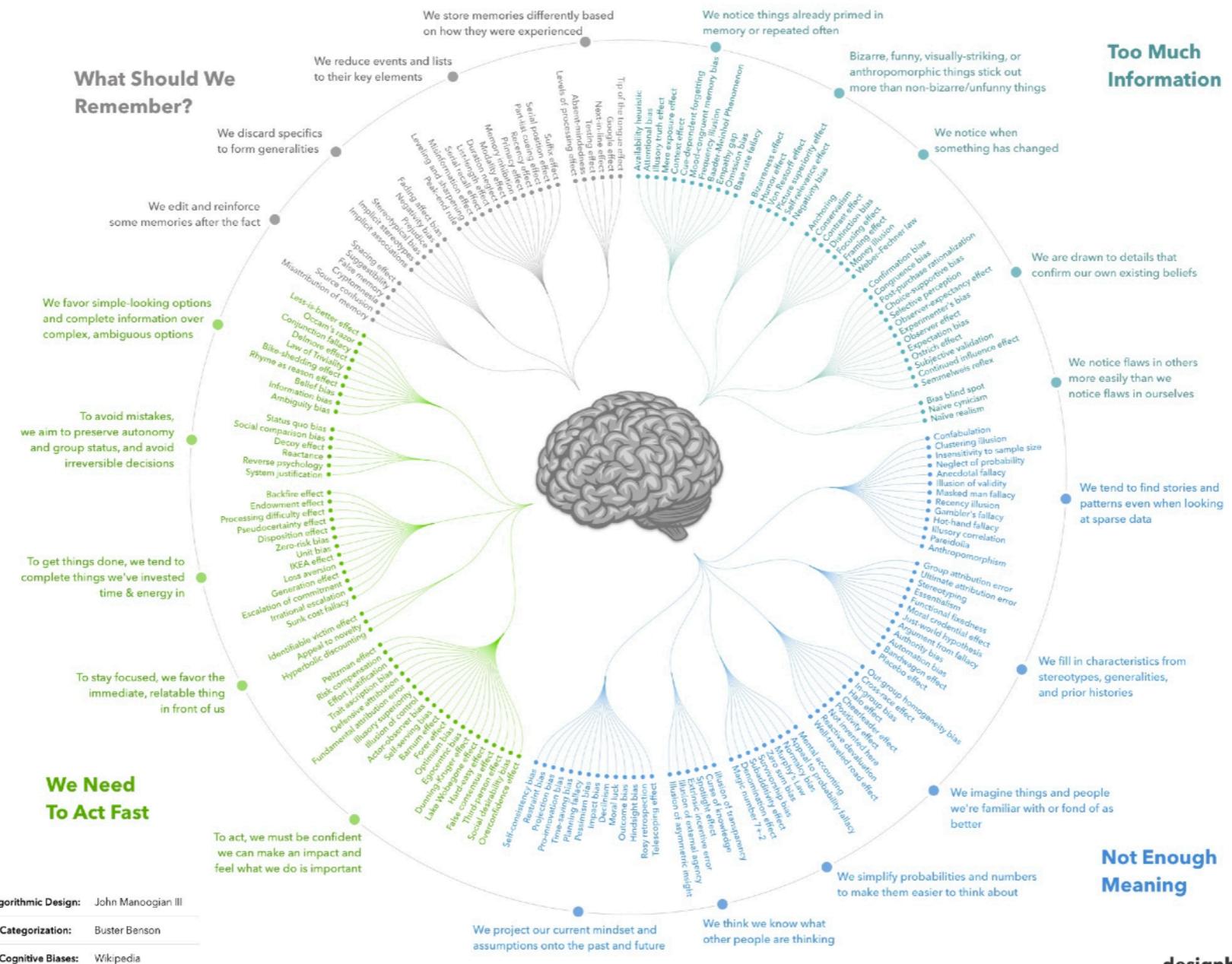
People who looks healthy or attractive are also competent and good.

Reading:

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

# Unconscious Human Biases

## COGNITIVE BIAS CODEX



<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 absorbing beliefs of our friends, family, society, etc. **PERSONAL HISTORY**
- Personal ethical decision making profile. F.e. you prioritize relationships in your decision-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 be able of making and justifying our decisions

# **Automated Discrimination**

# Demographic disparities

Amazon uses a data-driven system to determine the neighborhood in which to offer free same-day delivery. A 2016 study found **disparities in the demographic makeup of these neighborhoods**: in many U.S. cities, white residents were more than twice as likely as black residents to live in one of the qualifying neighborhoods.

This doesn't imply that the designer of the system intended for such inequalities to arise.

Looking beyond intent, it's important to understand when observed disparities can be considered to be discrimination.

To understand why the racial disparities in Amazon's system might be harmful, we must keep in mind the history of racial prejudice in the United States, its relationship to geographic segregation and disparities, and the perpetuation of those inequalities over time.

# Demographic disparities

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

Amazon argued that its system was justified because it was designed based on efficiency and cost considerations and that race wasn't an explicit factor.

**THIS COULD BE ILLEGAL!**

Nonetheless, it has the effect of **providing different opportunities to consumers at racially disparate rates.**

**THIS IS UNETHICAL**

The concern is that this might contribute to the perpetuation of long-lasting cycles of inequality.

**THIS IS ALSO UNETHICAL, BUT AT A DIFFERENT LEVEL**

If, instead, the system had been found to be partial to ZIP codes ending in an odd digit, it would not have triggered a similar outcry.

# Demographic disparities

The term **bias** is often used to refer to **demographic disparities** in algorithmic systems that are objectionable for societal reasons.

We should avoid using the word bias since it means different things to different people.

Suppose that Amazon's estimates of delivery dates/times were consistently too early by a few hours. This would be a case of statistical bias.

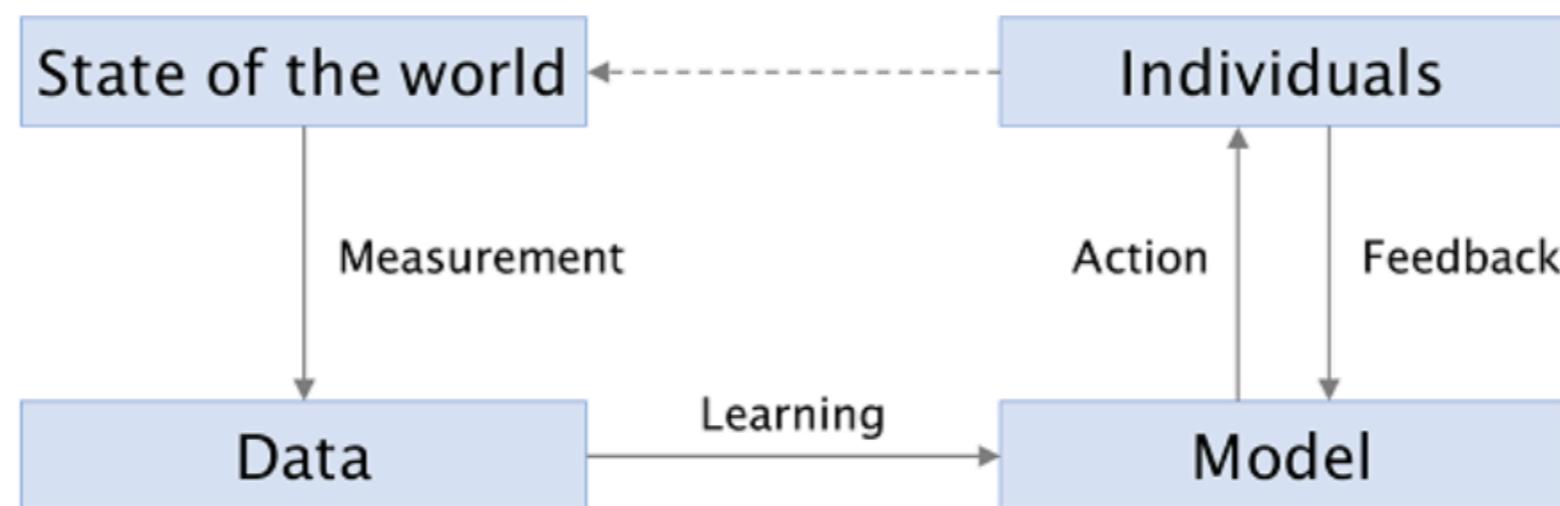
Suppose that Amazon's estimates of delivery dates/times are based on the prediction of a recurrent neural network. This would be a case of *recency bias*, a kind of *inductive bias* in machine learning.

# Demographic disparities

Demographic disparities propagate through the machine learning pipeline. Studying the stages of machine learning is crucial if we want to **intervene to minimize disparities**.

1. The first stage is **measurement**, which is the process by which the state of the world is reduced to a set of rows, columns, and values in a dataset.

4. Some machine learning systems record feedback from users (how users react to actions) and use them to refine the model. Feedback can also occur unintentionally, or even adversarially; these are more problematic



2. The ‘**learning**’ in machine learning refers to the next stage, which is to turn that data into a model. Model summarizes the patterns in the training data; it makes generalizations.

3. The next stage is the **action** we take based on the model’s predictions, which are applications of the model to new, unseen inputs.

# Demographic disparities

We're mainly concerned with applications of machine learning that involve data about **people**.

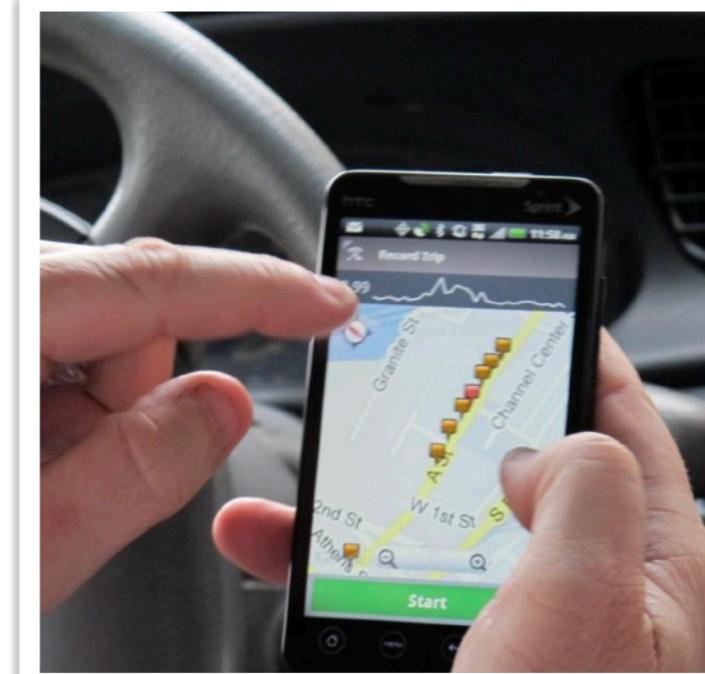
In these applications, the available training data will likely encode the demographic disparities that exist in our society.

## 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.



# Demographic disparities

There are also troubles related to **measurement**: 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>

# Demographic disparities

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

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

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.

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.

# Sensible characteristics

In many classification tasks, the features contain or implicitly encode **sensitive characteristics** of an individual.

## THE PROBLEM

The choice of sensitive attributes will generally have profound consequences as it decides which groups of the population we highlight, and what conclusions we draw from our investigation.

## THE BAD SOLUTION

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

## THE CAUSE OF THE BAD SOLUTION

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.

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!**

# Formal non-discrimination

$X \in \mathbb{R}^d$  features of an individual (browsing history)

$A \in \{0,1\}$  sensitive attribute (gender)

$Y \in \{0,1\}$  target variable

Decision function (i.e. ML algorithm) is any random variable  $D = d(X, A) \in [0,1]$  that can be transformed into a hard decision by thresholding.

# Formal non-discrimination

Criterion 0 (Naive):

1. **Unawarness** : Fairness is the absence of the protected attribute in the model features.

$$d(X, A) \rightarrow d(X)$$

Removing the protected attribute doesn't guarantee that all the information concerning this attribute is removed from the data!

# Formal non-discrimination

2 important criteria:

Sometimes called **demographic parity** or **statistical parity**.

1. **Independence**:  $D$  independent of  $A$ , denoted  $D \perp A$ . For all groups  $a, b$  and all values  $d$ :

The sensitive characteristic must be statistically independent of the score.

$$p(D = d | A = a) = p(D = d | A = b)$$

Sometimes called **equalized odds**.

2. **Separation**:  $D$  independent of  $A$  conditional on  $Y$ .

# Independence

<b>X<sub>1</sub></b>	...	<b>A = race</b>	<b>Y</b>
0	...	0	1
2	...	1	1
1	...	0	0
2	...	1	0

In the case of binary decision, independence simplifies to the condition for all groups  $a, b$ :

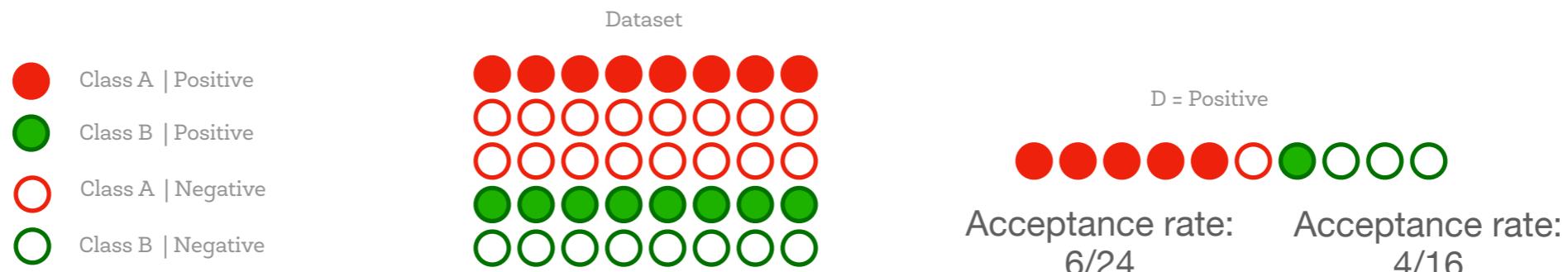
$$p(D = 1 | A = a) = p(D = 1 | A = b)$$

Acceptance rate must be the same in all groups

# Independence

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:



There are flaws:

- It concerns only the final outcome of the model but doesn't focus on **equality of treatment** (see next slide).
- 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  $p$ .

Even though 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.

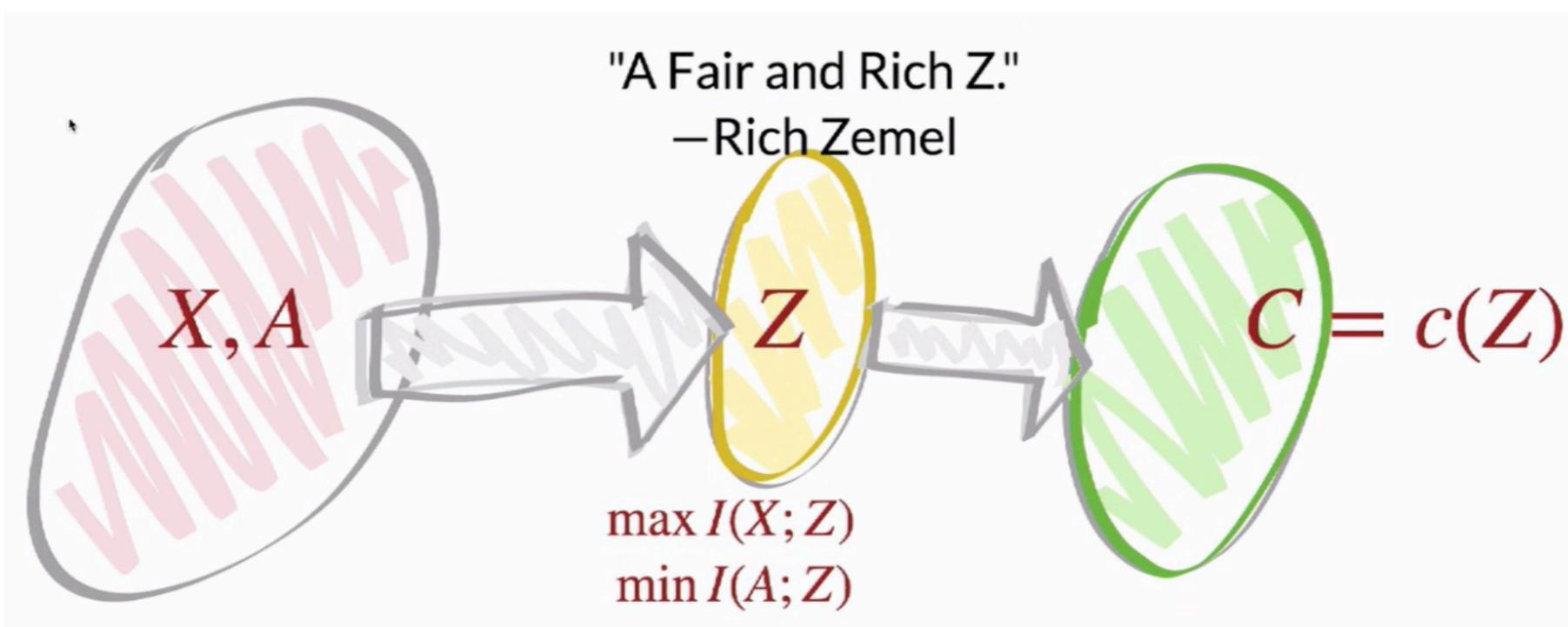
As a result, it will appear in hindsight that members of group *B* performed worse than members of group *A*, thus establishing a negative track record for group *B*.

# Independence

In general, there are three different strategies to satisfy fairness criteria:

- **Pre-processing:** Adjust the feature space to be uncorrelated with the sensitive attribute.
- **At training time:** Work the constraint into the optimization process that constructs a classifier from training data.
- **Post-processing:** Adjust a learned classifier so as to be uncorrelated with the sensitive attribute.

# Achieving independence with a representation learning approach:



# Separation

$D$  independent of  $A$  conditional on  $Y$ .

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

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

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.

# Separation

In the case where  $D$  is a binary classifier, separation is equivalent to requiring for all groups  $a, b$  the two constraints

$$p(D = 1 | Y = 1, A = a) = p(D = 1 | Y = 1, A = b) \quad \text{True Positive Rate}$$

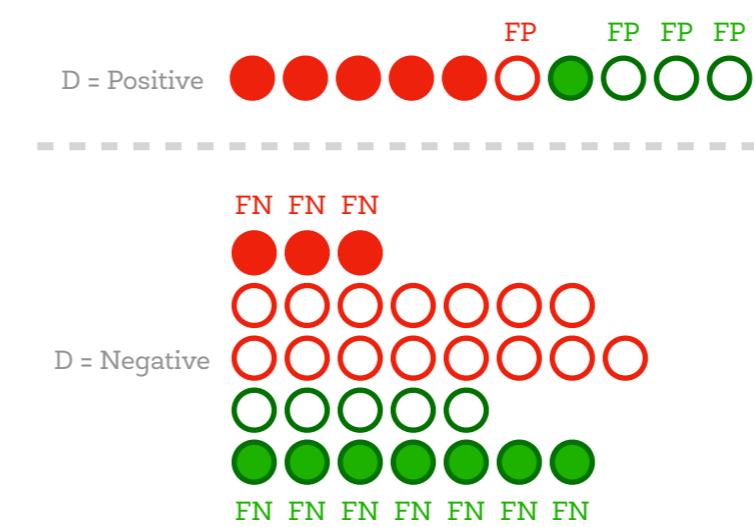
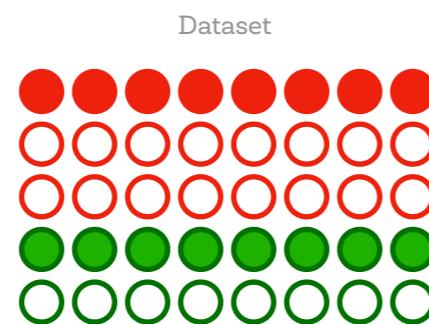
$$p(D = 1 | Y = 0, A = a) = p(D = 1 | Y = 0, A = b) \quad \text{False Positive Rate}$$

What separation therefore requires is that all groups experience the same **false negative rate** and the **same false positive rate**.

$$\text{FNR} = 1 - \text{TPR}$$

# Separation

- Class A | Positive
- Class B | Positive
- Class A | Negative
- Class B | Negative



$$FPR_A = \frac{1}{16}$$

$$FPR_B = \frac{3}{8}$$

$$FNR_A = \frac{3}{8}$$

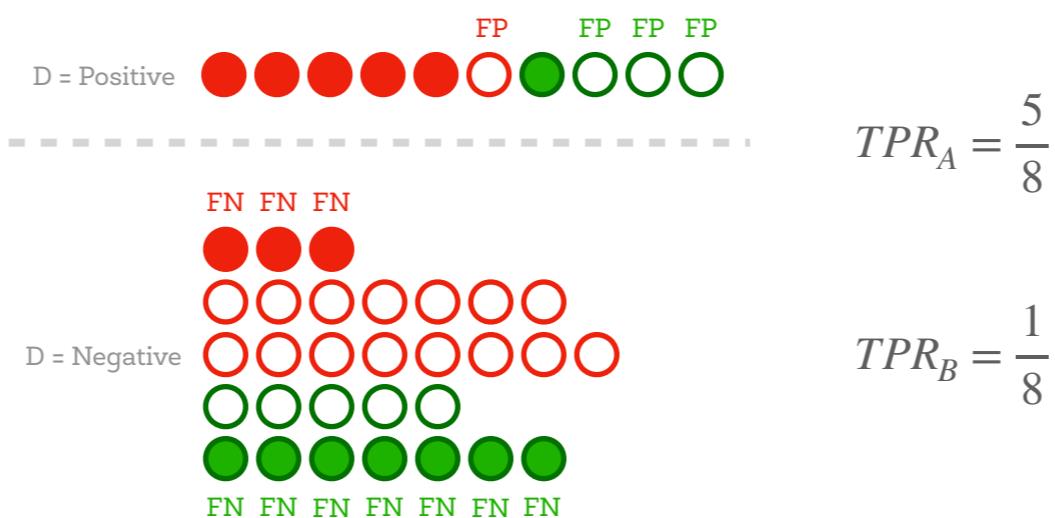
$$FNR_B = \frac{7}{8}$$

# Separation

As was the case with independence, we can achieve separation by post-processing a given score function without the need for retraining.

In many applications (e.g. hiring), people care more about the true positive rate than true negative rate so many works focus on the following relaxed version (which is called **equality of opportunity**):

$$p(D = 1 \mid Y = 1, A = 0) = p(D = 1 \mid Y = 1, A = 1)$$



# Separation flaws

**It may not help closing the gap between two groups.**

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 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,  $A$  is not independent of  $Y$ , and  $R$  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.**

# Fairness for decisions

For binary decision procedures, we can summarize a procedure with the confusion matrix, which illustrates match and mismatch between decision  $d$  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 $P[Y = 1 d = 1]$	False Discovery Rate (FDR) $P[Y = 0 d = 1]$	
Negative Decision $d = 0$	False Negative (FN)	True Negative (TN)	False Omission Rate (FOR) $P[Y = 1 d = 0]$	Negative Predictive Value (NPV) $P[Y = 0 d = 0]$	
Positive Decision Rate $P[d = 1]$	True Positive Rate (TPR), aka recall, aka sensitivity $P[d = 1 Y = 1]$	False Positive Rate (FPR) $P[d = 1 Y = 0]$	Accuracy $P[d = Y]$		
	False Negative Rate (FNR) $P[d = 0 Y = 1]$	True Negative Rate (TNR), aka specificity $P[d = 0 Y = 0]$			

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**:

$$p(Y = 1 | D = 0, A = a) = p(Y = 1 | D = 0, A = a')$$

# Fairness for decisions

For binary decision procedures, we can summarize a procedure with the confusion matrix, which illustrates match and mismatch between decision  $d$  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 $P[Y = 1 d = 1]$	False Discovery Rate (FDR) $P[Y = 0 d = 1]$
Negative Decision $d = 0$		False Negative (FN)	True Negative (TN)	False Omission Rate (FOR) $P[Y = 1 d = 0]$	Negative Predictive Value (NPV) $P[Y = 0 d = 0]$
Positive Decision Rate $P[d = 1]$	True Positive Rate (TPR), aka recall, aka sensitivity $P[d = 1 Y = 1]$	False Positive Rate (FPR) $P[d = 1 Y = 0]$	Accuracy $P[d = Y]$		
	False Negative Rate (FNR) $P[d = 0 Y = 1]$	True Negative Rate (TNR), aka specificity $P[d = 0 Y = 0]$			

Confusion Matrix

Demographic Parity:  
 $p(D = 1 | A = a) = p(D = 1 | A = a')$

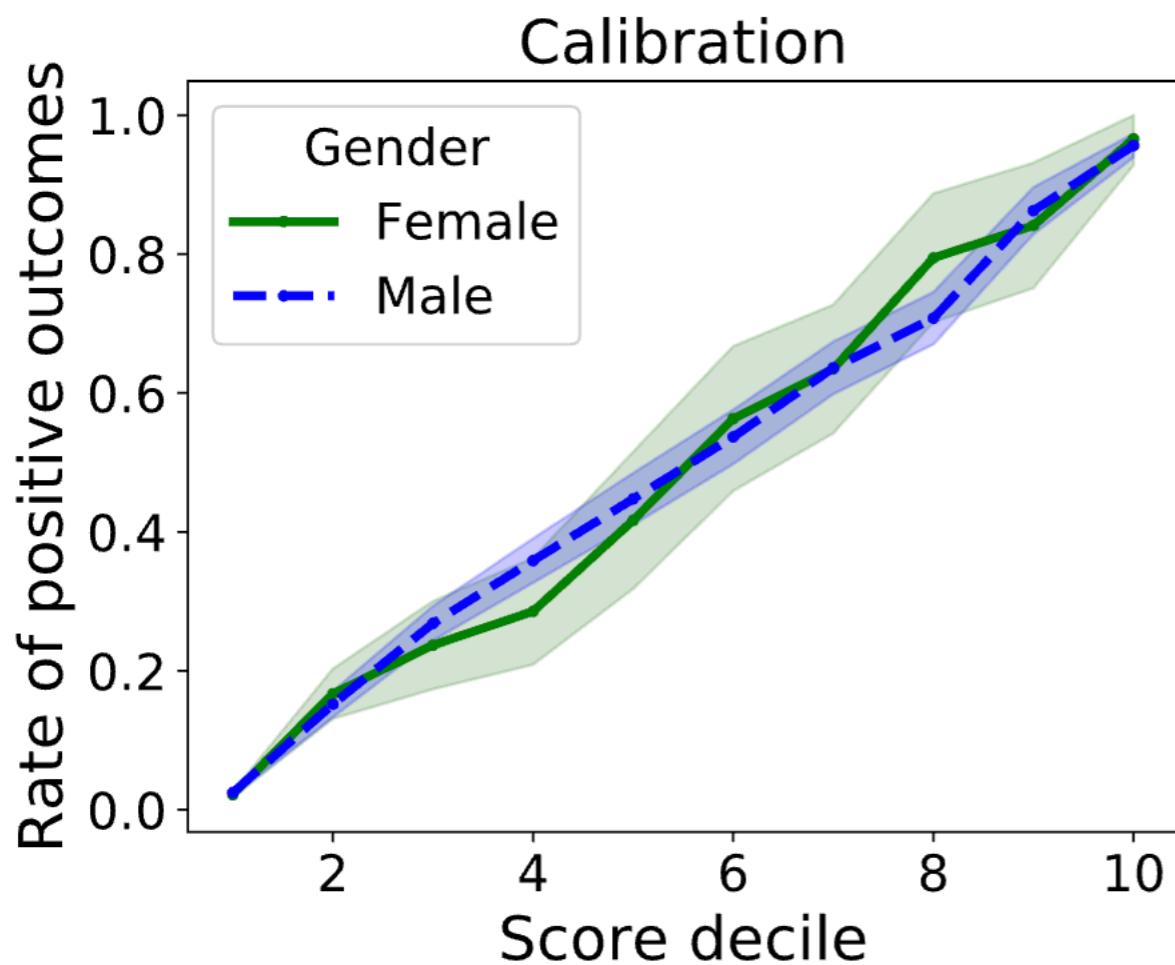
# Fairness for scores

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

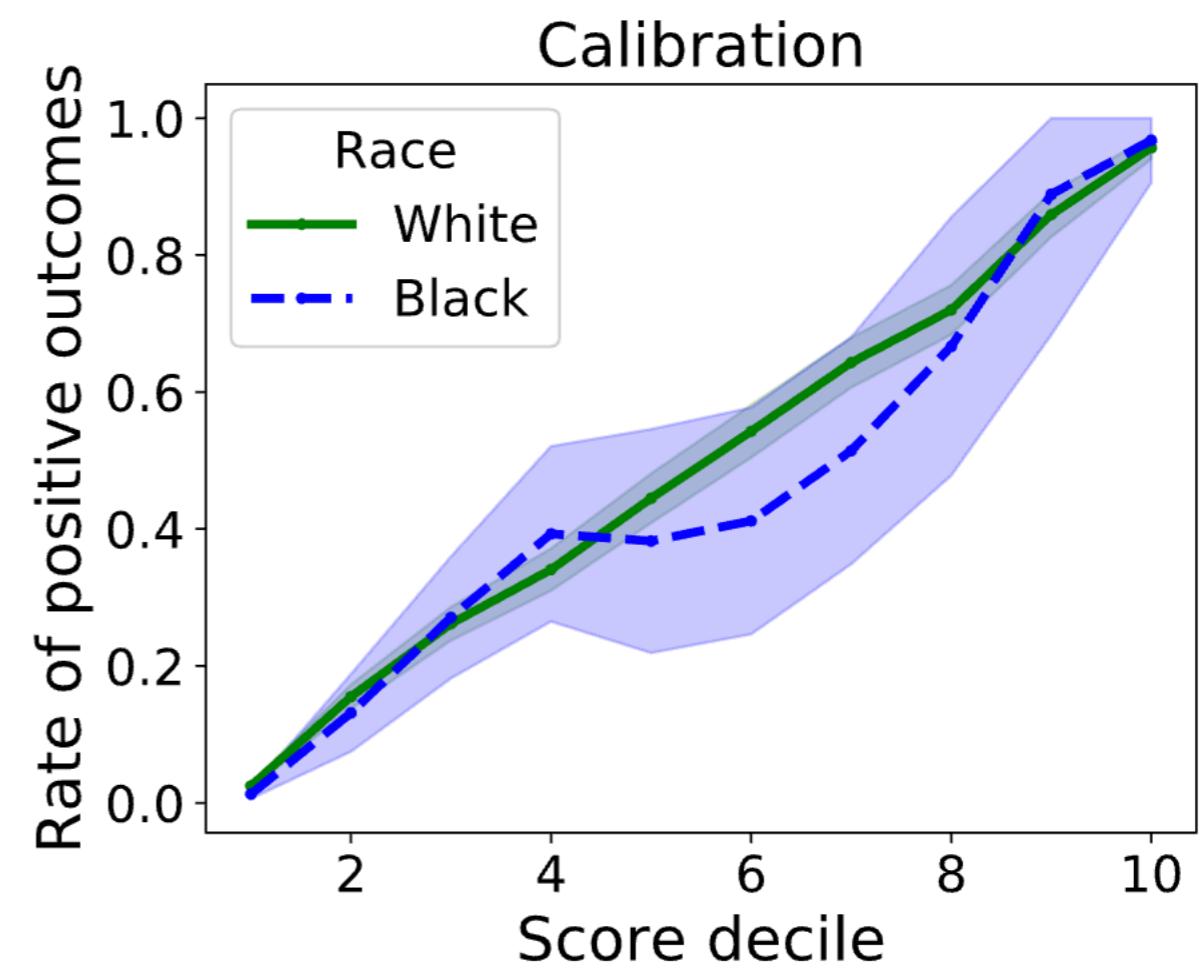
- **Balance for the Positive Class:** the average score assigned to positive members,  $\mathbb{E}(D | Y = 1)$ , should be the same across groups.
- **Balance for the Negative Class:** the average score assigned to negative members,  $\mathbb{E}(D | Y = 0)$ , should be the same across groups.
- **Calibration:** the fraction of those marked with a given score who are actually positive,  $\mathbb{E}(Y | D = d)$ , 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.

# Calibration

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.

# Conclusions

Fairness is not a property of a model. Fairness is a highly contextual problem. It is a never-ending process to **align the behavior of the model and its consequences to our changing principles and values.**

**Which fairness metric should we use to measure disparity?**

# Conclusions

To help answer this question, we can define two types of fairness metrics.

‘**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  $p(D = 1 | Y = 0, A = a) = p(D = 1 | Y = 0, A = a')$

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  $p(D = 1 | A = a) = p(D = 1 | A = a')$

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

Sandra Wachter<sup>1</sup>, Brent Mittelstadt<sup>2</sup> and Chris Russell<sup>3</sup>

# Conclusions

We say that any fairness metric is bias preserving if **it is always satisfied by a perfect classifier** that exactly predicts its target labels with zero error, **replicating bias present in the data**.

Fairness metrics that are not necessarily satisfied by a perfect classifier, we refer to as bias transforming.

For example, **equalized odds is bias preserving metrics**: it is a form of conditional independence or conditional demographic parity, conditioned on historic data  $Y$ , and reflecting its biases exactly.

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

Sandra Wachter<sup>1</sup>, Brent Mittelstadt<sup>2</sup> and Chris Russell<sup>3</sup>

# Conclusions

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

Sandra Wachter<sup>1</sup>, Brent Mittelstadt<sup>2</sup> and Chris Russell<sup>3</sup>

# Conclusions

The key difference between **bias transforming** and **bias preserving** metrics is that most bias transforming metrics are satisfied by matching **decision rates between groups**, while bias preserving metrics typically require matching **error rates between groups**.

# Conclusions

The screenshot shows a section from the AWS SageMaker Developer Guide. The left sidebar lists various topics under 'Amazon SageMaker' and 'Prepare and Analyze Datasets'. The main content area is titled 'Conditional Demographic Disparity (CDD)'. It explains that the demographic disparity metric (DD) determines whether facet  $d$  has a larger proportion of rejected outcomes than accepted ones. It provides the formula  $DD_d = n_d^{(0)}/n^{(0)} - n_d^{(1)}/n^{(1)} = P_d^R(y^0) - P_d^A(y^1)$  and details where each component is defined. It also notes that for a college admission example,  $DD = 0.6 - 0.5 = 0.1$ .

Conditional Demographic Disparity (CDD)

The demographic disparity metric (DD) determines whether facet  $d$  has a larger proportion of the rejected outcomes in the dataset than of the accepted outcomes. For example, in the case of college admissions, if women applicants comprised 60% of the rejected applicants and comprised only 50% 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 are accepted.

The formula for the demographic disparity for the less favored facet  $d$  is as follows:

$$DD_d = n_d^{(0)}/n^{(0)} - n_d^{(1)}/n^{(1)} = P_d^R(y^0) - P_d^A(y^1)$$

Where:

- $n^{(0)} = n_a^{(0)} + n_d^{(0)}$  is the number of rejected outcomes in the dataset.
- $n^{(1)} = n_a^{(1)} + n_d^{(1)}$  is the number of accepted outcomes in the dataset.
- $P_d^R(y^0)$  is the proportion of rejected outcomes (with value 0) in facet  $d$ .
- $P_d^A(y^1)$  is the proportion of accepted outcomes (value 1) in facet  $d$ .

For the college admission example, the demographic disparity is  $DD = 0.6 - 0.5 = 0.1$ .

**Conditional demographic parity** (conditioning on salary) is satisfied if  $x\%$  of both black and white people earning over a threshold receive positive decisions and also some  $y\%$  of both black and white people earning under the threshold receive positive decisions.

# Case: Measuring recidivism risk

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

- The risk of committing a new crime (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.

# 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, and a final decision is taken based on the results of those items.

We can train a ML system to make automatic decisions based on the scores in each item, bu 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 automatic 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.

# Measuring recidivism risk



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

## Machine Bias

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

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

In May 2016, ProPublica, an investigative journalism organization, published a piece called "[Machine Bias](#)", in which a predictive tool for predicting recidivism used in the US, **COMPAS**, was found **to be biased** against African-Americans.

# Measuring recidivism risk

ProPublica Local Initiatives Data Store [f](#) [t](#) [Donate](#)

**PROPUBLICA** Graphics & Data Newsletters About [Get the Big Story](#) [Join](#)

Racial Justice Technology Criminal Justice Regulation More... Series Video Impact Search

**MACHINE BIAS**

**Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say**

ProPublica's analysis of bias against black defendants in criminal risk scores has prompted research showing that the disparity can be addressed — if the algorithms focus on the fairness of outcomes.

by Julia Angwin and Jeff Larson, Dec. 30, 2016, 4:44 p.m. EST



Ad

Resolución PACs PECs. Aprueba sin Esfuerzo. Consigue soluciones rápidas y de calidad.

Pacsolver Pedir precios

**FOLLOW PROPUBLICA**

[Twitter](#) [Facebook](#)

[YouTube](#) [RSS](#)

**STAY INFORMED**

# Measuring recidivism risk: the debate

In 2016, ProPublica published a highly influential analysis based on data obtained through public records requests. Two of their findings can be phrased in our language as follows:

- COMPAS does not satisfy **equal false negative rates**, in fact, white defendants who did get rearrested were nearly twice as likely to be misclassified as low risk.
- COMPAS does not satisfy **equal false positive rates**, in fact, black defendants who did not get rearrested were nearly twice as likely to be misclassified as higher risk.

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.
- COMPAS satisfies **equal positive predictive values**: among those labeled higher risk, the proportion of defendants who got rearrested is approximately the same regardless of race.

# 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.

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

# 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:

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