Fairness in Machine Learning

Fernanda Viégas @viegasf
Martin Wattenberg @wattenberg
Google Brain


As AI touches high-stakes aspects of everyday life,
fairness becomes more important

How can an algorithm even be unfair?
Aren't algorithms beautiful neutral pieces of mathematics?

Scholium.

It may be observed that the foregoing process includes the arithmetical rule for finding the greatest common factor of two numbers: which is to *divide the greater number by the lesser, and find the remainder; the lesser by the remainder, and find the second remainder, if there be one; the preceding remainder by this, and find the third remainder; and so on, until a remainder be found which is contained an exact number of times in the next preceding; this last remainder will be the greatest common factor required.*

The Euclidean algorithm (first discovered in 300 BCE) as described in *Geometry, plane, solid and spherical*, Pierce Morton, 1847.

"Classic" non-ML problem: implicit cultural assumptions

Example: names are complex.

1. People have exactly one canonical full name.
2. People have exactly one full name which they go by.
3. People have, at this point in time, exactly one canonical full name.
4. People have, at this point in time, one full name which they go by.
5. People have exactly N names, for any value of N.
6. People's names fit within a certain defined amount of space.
7. People's names do not change.
8. People's names change, but only at a certain enumerated set of events.
9. People's names are written in ASCII.
10. People's names are written in any single character set.

...

37. Two different systems containing data about the same person will use the same name for that person.
38. Two different data entry operators, given a person's name, will by necessity enter bitwise equivalent strings on any single system, if the system is well-designed.
39. People whose names break my system are weird outliers. They should have had solid, acceptable names, like 田中太郎.
40. People have names.

Kalzumeus Archive Greatest Hits Standing Invitation Start Here About me

Falsehoods Programmers Believe About Names

June 17, 2010 in Uncategorized

[This post has been translated into Japanese by one of our readers: 和訳もあります。]

John Graham-Cumming wrote an article today complaining about how a computer system he was working with described his last name as having invalid characters. It of course does not, because anything someone tells you is their name is — by definition — an appropriate identifier for them. John was understandably vexed about this situation, and he has every right to be, because **names are central to our identities**, *virtually by definition*.

WHO AM I?

My name is Patrick McKenzie (better known as patio11 on the Internets.)

Twitter: @patio11 HN: patio11

I ALSO EMAIL ESSAYS.

Get essays via email on marketing and selling software.

Brilliant, fun article.
Read it! :)

Patrick McKenzie
http://www.kalzumeus.com/2010/06/17/falsehoods-programmers-believe-about-names/

What's different with machine learning?

Algorithm, 300 BCE

Scholium.

It may be observed that the foregoing process includes the arithmetical rule for finding the greatest common factor of two numbers: which is to *divide the greater number by the lesser, and find the remainder; the lesser by the remainder, and find the second remainder, if there be one; the preceding remainder by this, and find the third remainder; and so on, until a remainder be found which is contained an exact number of times in the next preceding; this last remainder will be the greatest common factor required.*

Classical algorithms don't rely on data

What's different with machine learning?

Algorithm, 300 BCE

Scholium.

It may be observed that the foregoing process includes the arithmetical rule for finding the greatest common factor of two numbers: which is to *divide the greater number by the lesser, and find the remainder; the lesser by the remainder, and find the second remainder, if there be one; the preceding remainder by this, and find the third remainder; and so on, until a remainder be found which is contained an exact number of times in the next preceding; this last remainder will be the greatest common factor required.*

Classical algorithms don't rely on data

Algorithm, 2017 CE

```
with tf.Session() as sess:
  # Restore variables from disk.
  saver.restore(sess, "/tmp/model.ckpt")
  print("Model restored.")
  # Do some work with the model
  ...
```

ML systems rely on real-world data and can pick up biases from data

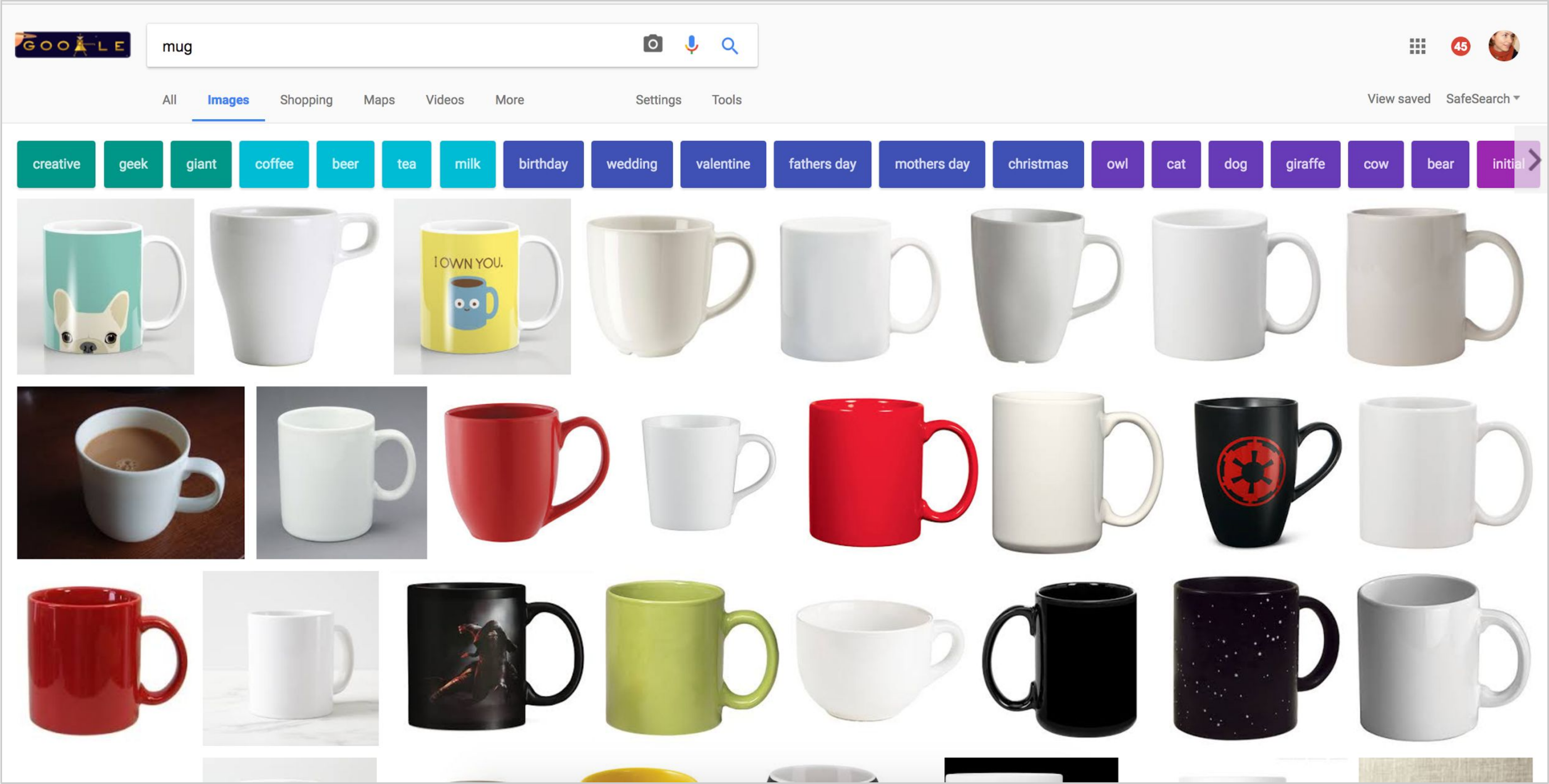
Sometimes bias starts before an algorithm ever runs...

It can start with the data

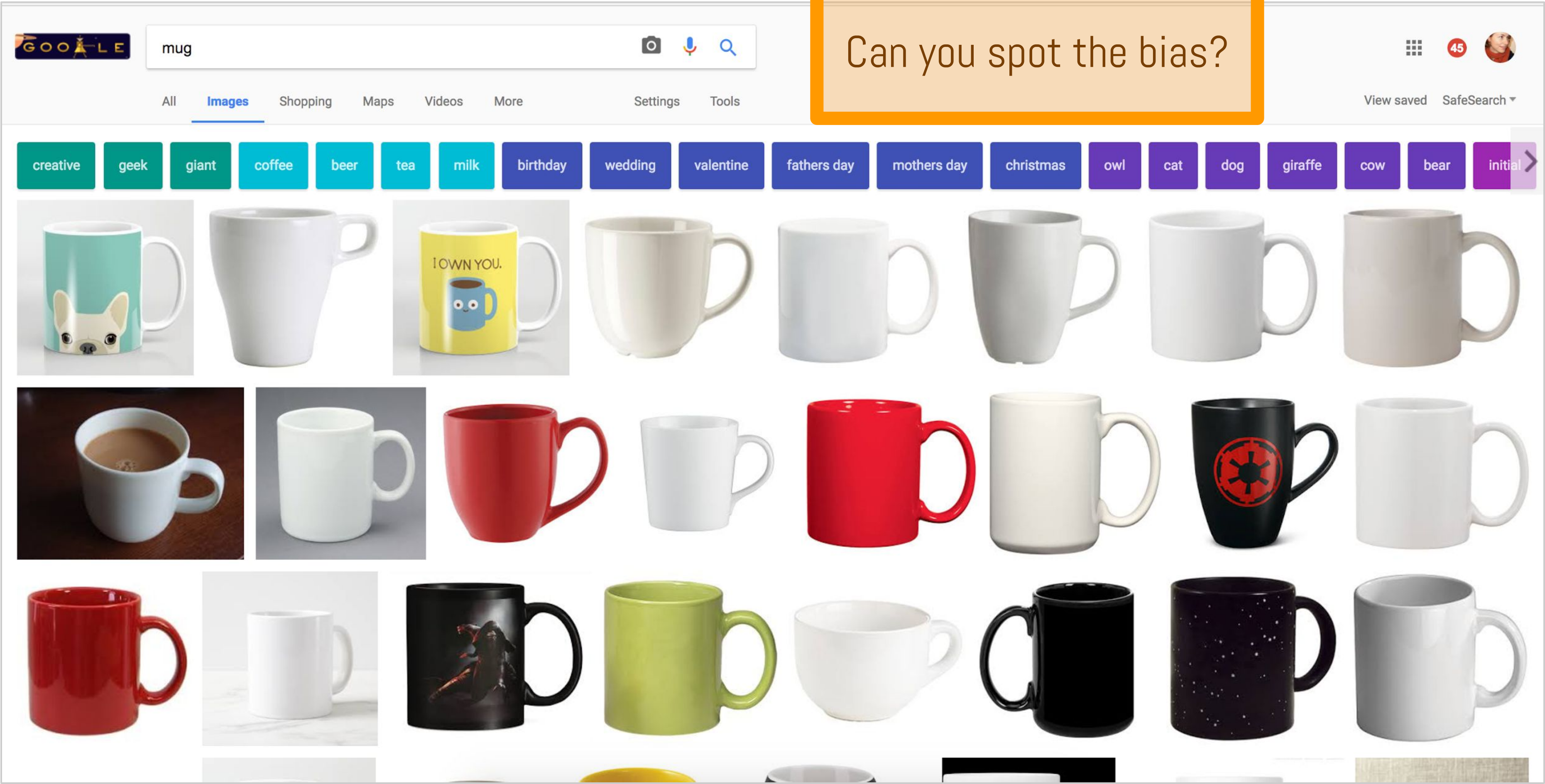
Sometimes bias starts before an algorithm ever runs...

It can start with the data

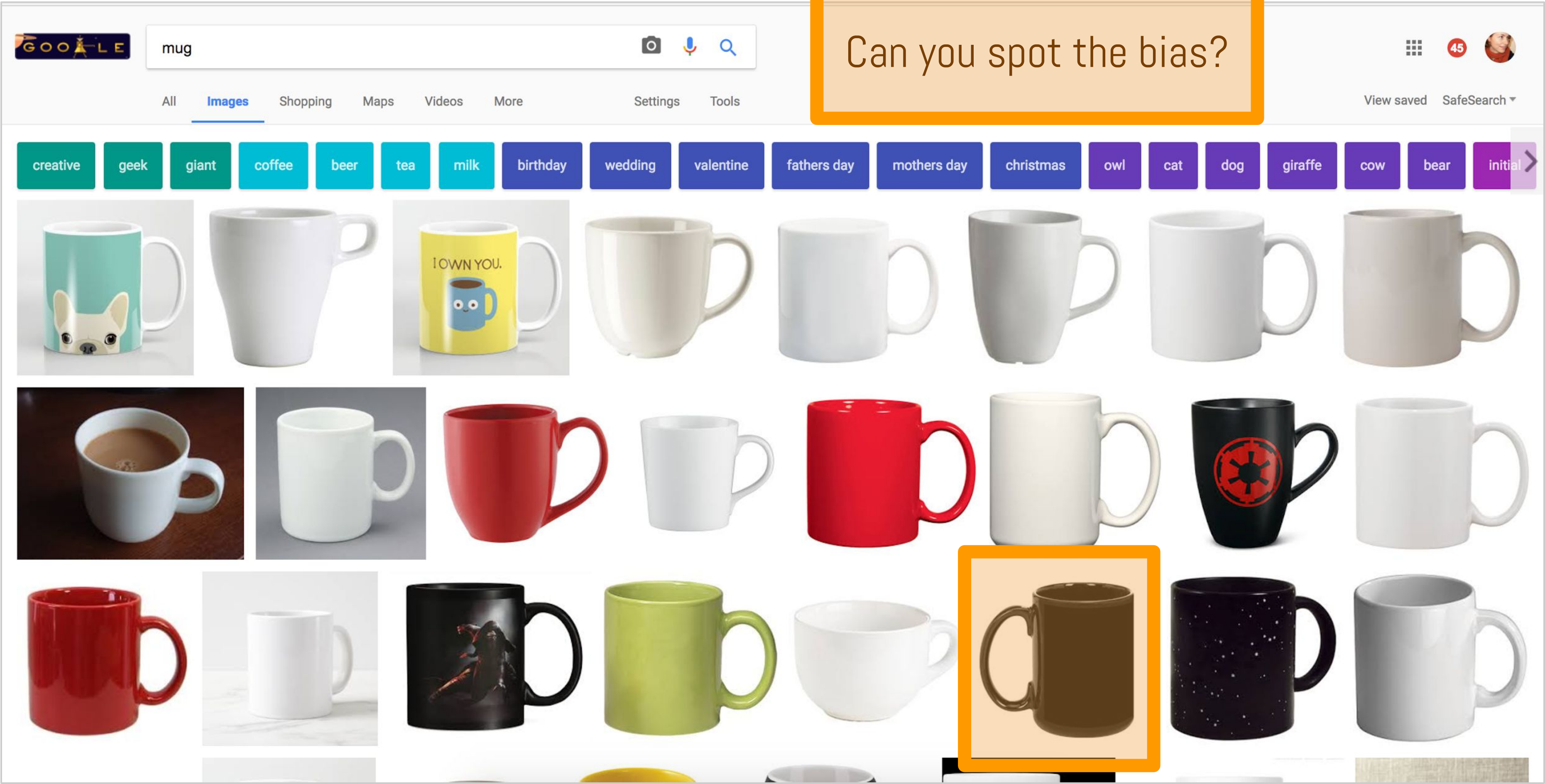
A real-world example

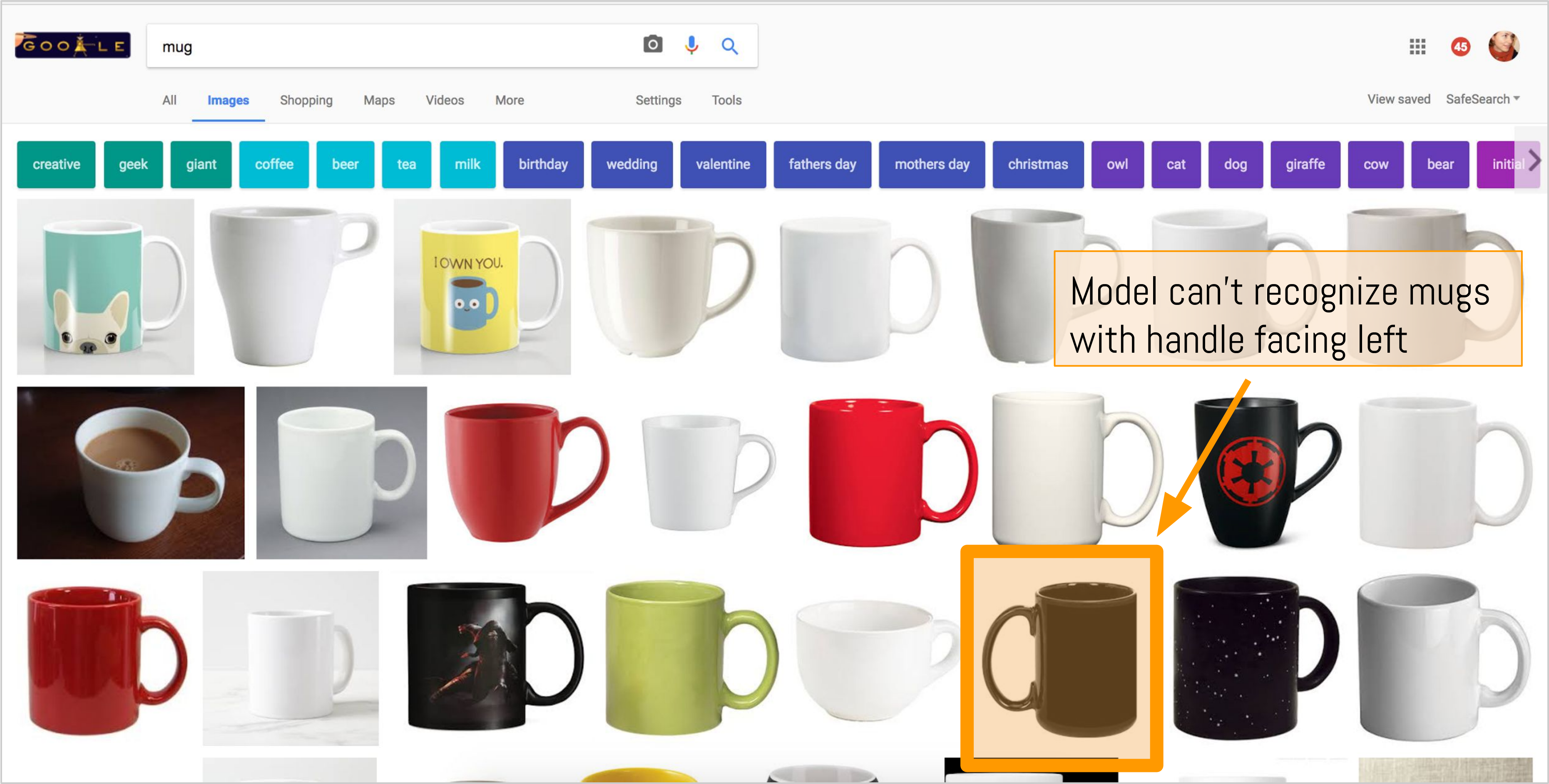

mug
45
All
Images
Shopping
Maps
Videos
More
Settings
Tools
View saved
SafeSearch
creative
geek
giant
coffee
beer
tea
milk
birthday
wedding
valentine
fathers day
mothers day
christmas
owl
cat
dog
giraffe
cow
bear
initial
I OWN YOU.

Can you spot the bias?



Can you spot the bias?

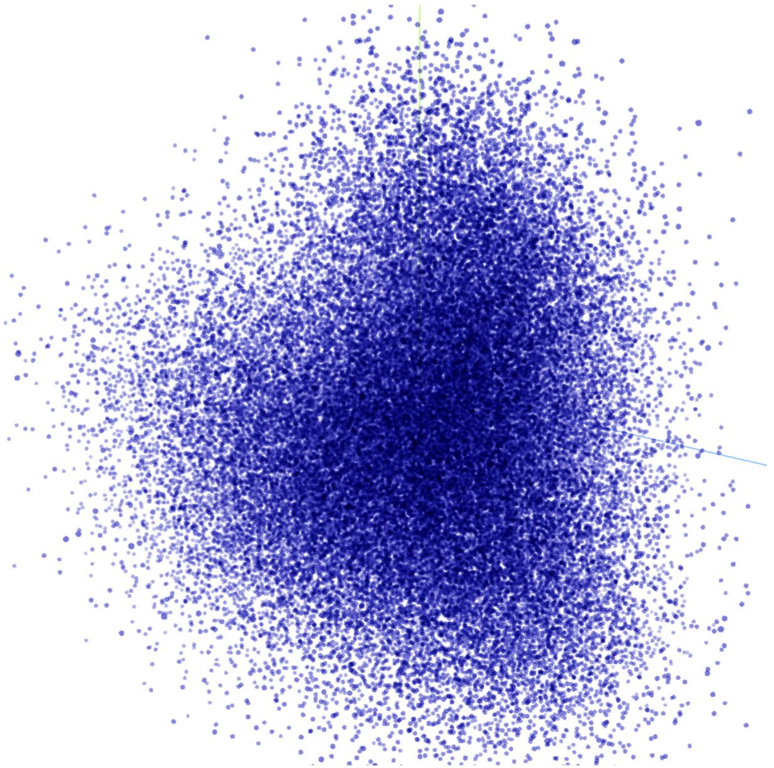



mug
All
Images
Shopping
Maps
Videos
More
Settings
Tools
View saved
SafeSearch
creative
geek
giant
coffee
beer
tea
milk
birthday
wedding
valentine
fathers day
mothers day
christmas
owl
cat
dog
giraffe
cow
bear
initial
I OWN YOU.
Model can't recognize mugs with handle facing left

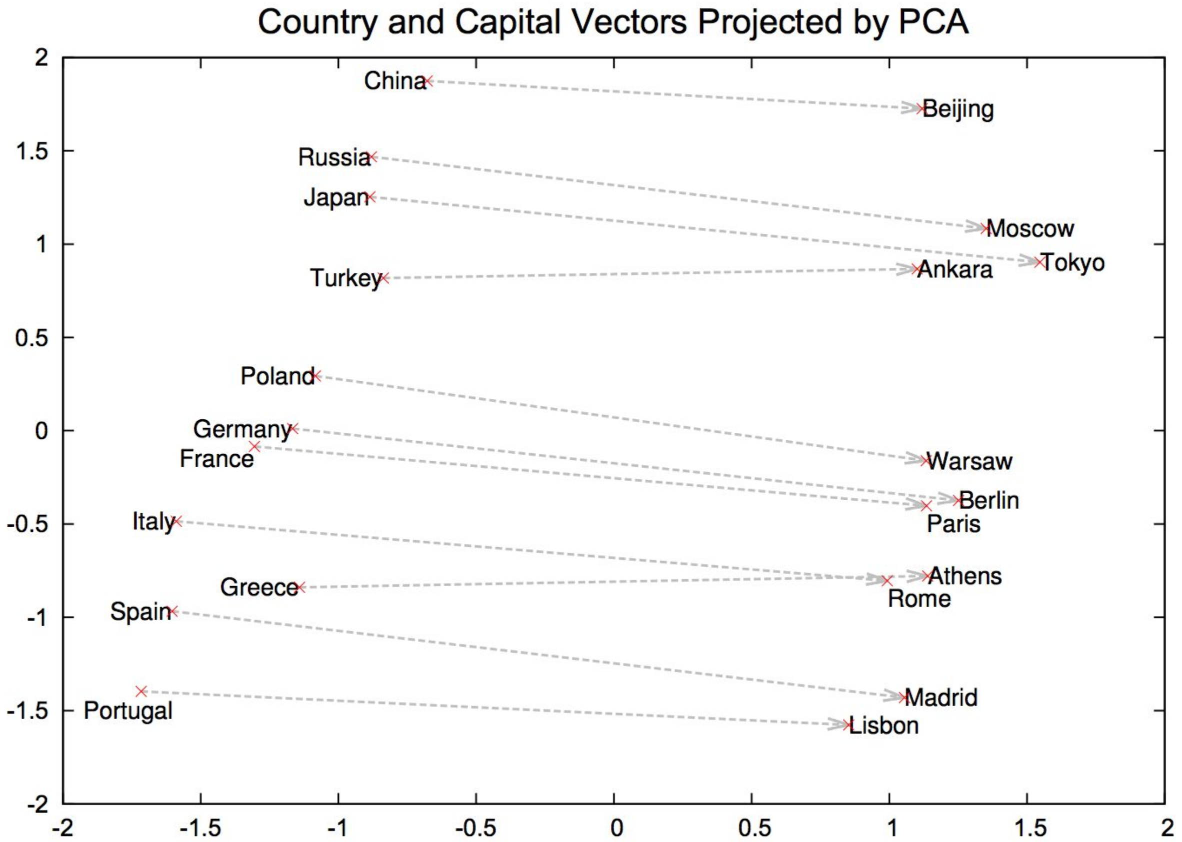
How can this lead to unfairness?


THOMPSON

word embeddings

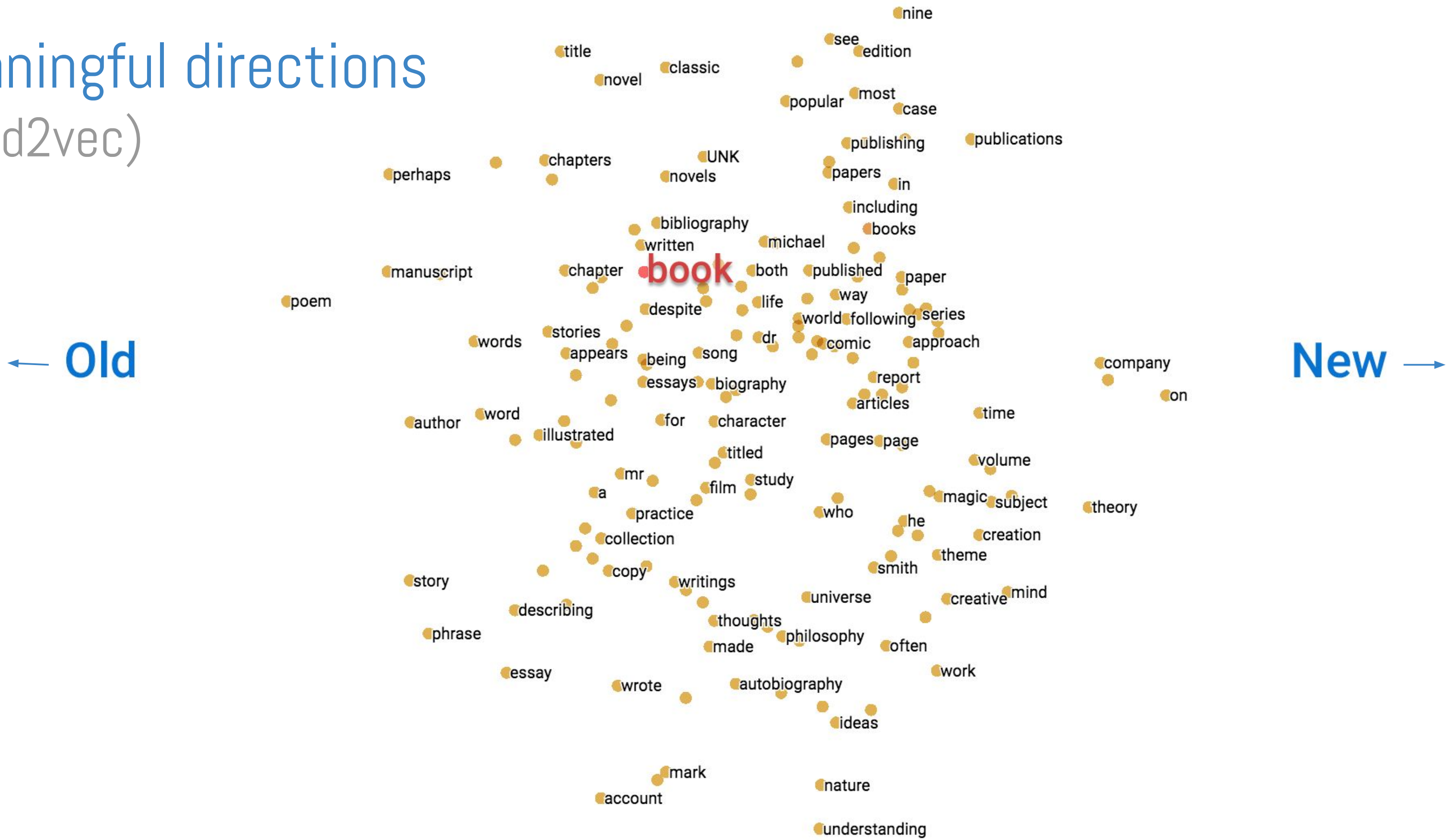


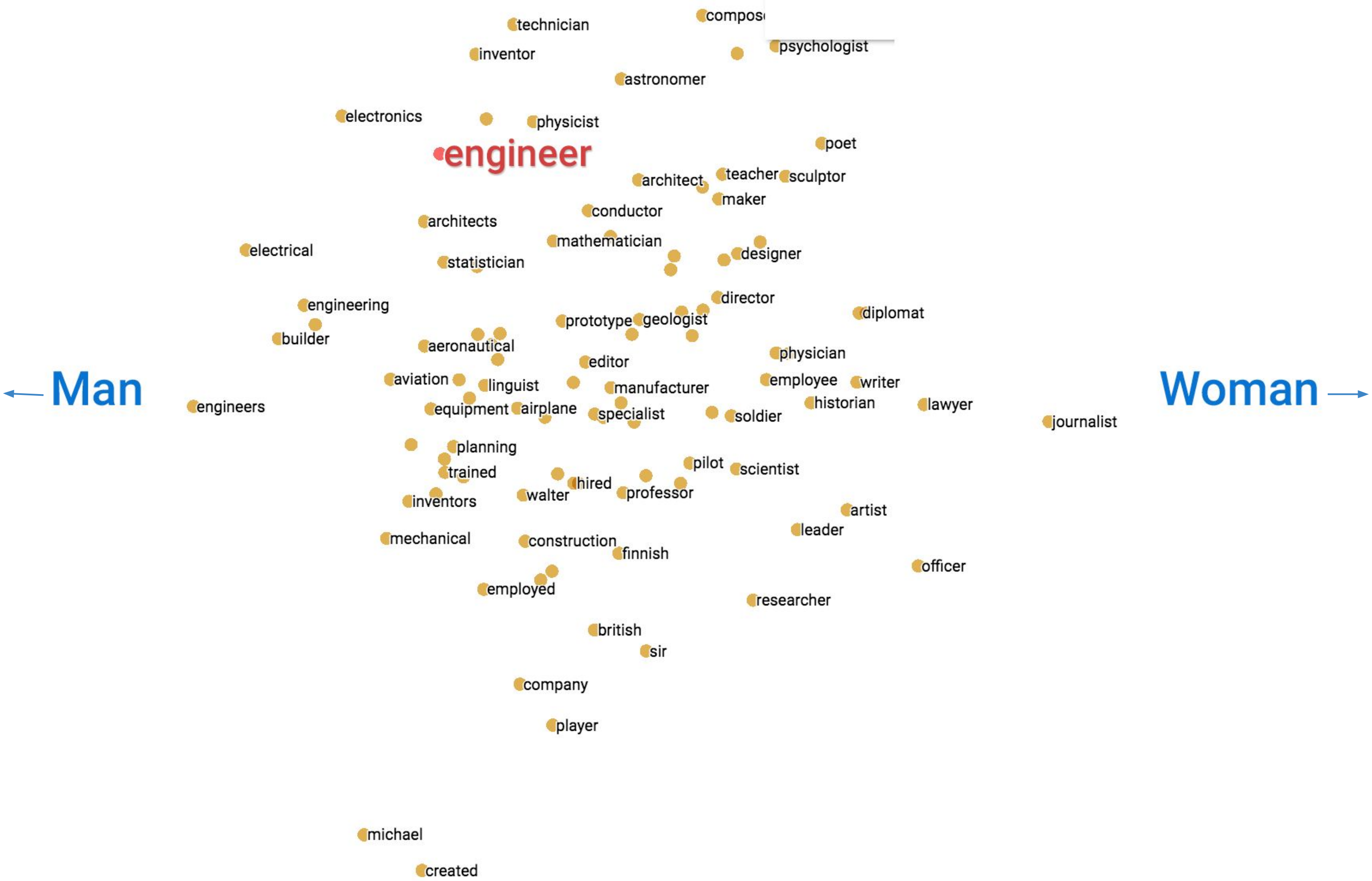
Word embeddings

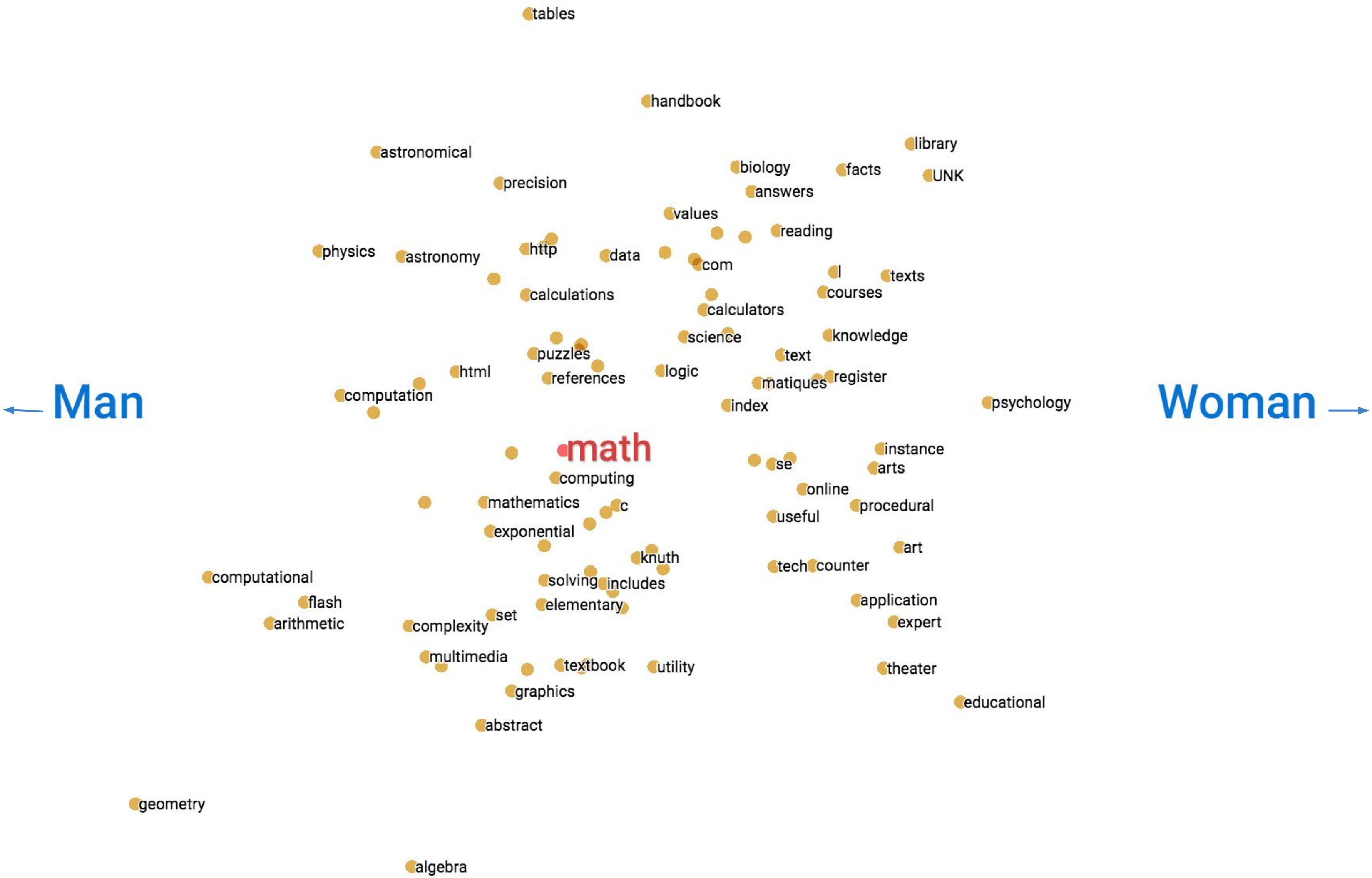


Distributed Representations of Words and Phrases and their Compositionality
Mikolov et al. 2013

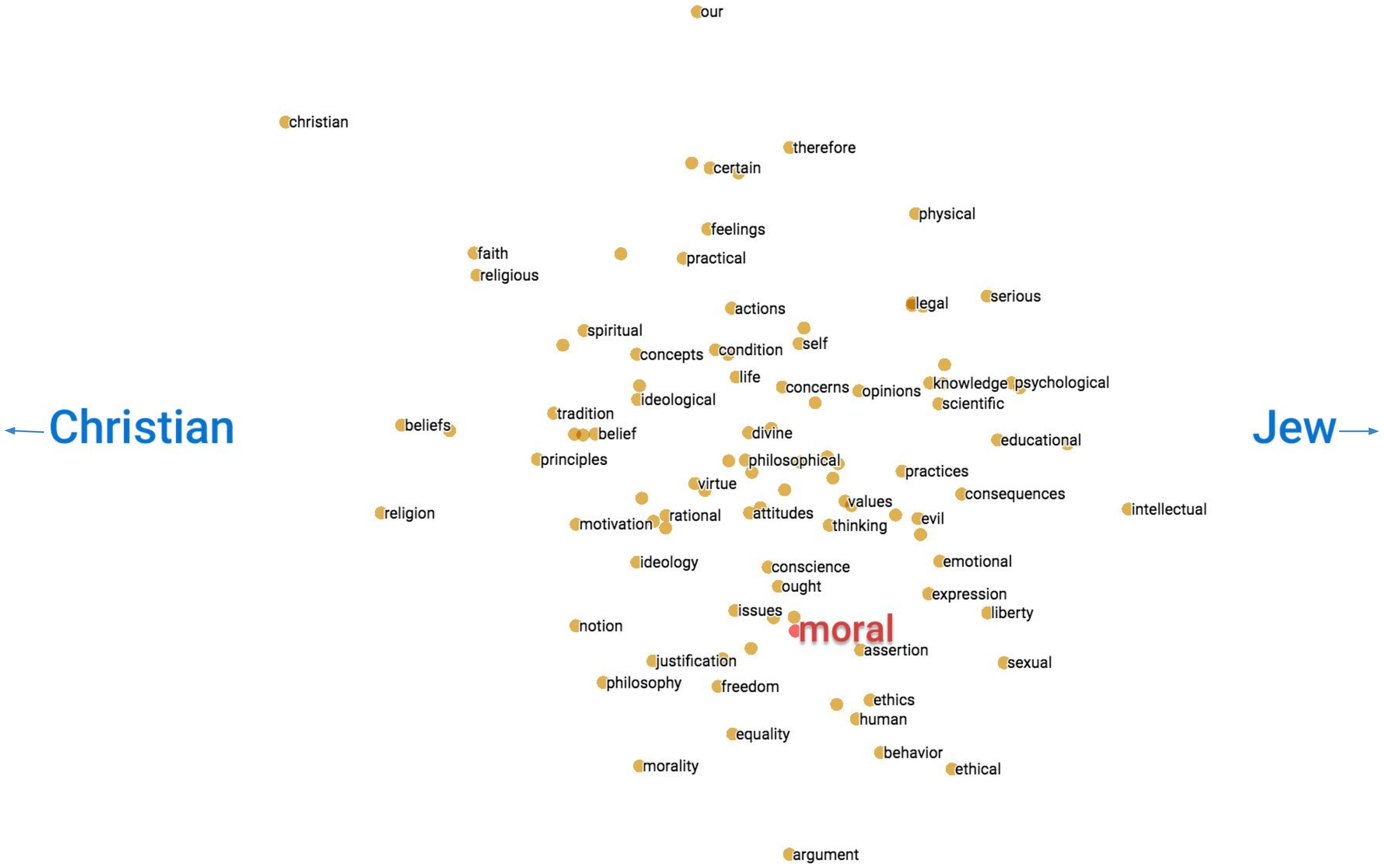
Meaningful directions
(word2vec)




Man
Woman
engineer
technician
compos
psychologist
inventor
astronomer
electronics
physicist
poet
architect
teacher
sculptor
maker
conductor
architects
mathematician
designer
electrical
statistician
director
engineering
prototype
geologist
diplomat
builder
aeronautical
physician
editor
aviation
linguist
manufacturer
employee
writer
engineers
equipment
airplane
specialist
soldier
historian
lawyer
journalist
planning
pilot
scientist
trained
hired
walter
professor
inventors
artist
leader
mechanical
construction
finnish
officer
employed
researcher
british
sir
company
player
michael
created



tables
handbook
library
astronomical
biology
facts
UNK
precision
answers
values
reading
physics
astronomy
http
data
com
l
texts
courses
calculations
calculators
science
knowledge
puzzles
text
html
references
logic
register
matiques
Man
computation
index
psychology
Woman
math
instance
se
arts
computing
online
mathematics
c
procedural
useful
exponential
art
knuth
tech
counter
computational
solving
includes
application
flash
elementary
arithmetic
set
expert
complexity
multimedia
textbook
utility
theater
graphics
educational
abstract
geometry
algebra


our
christian
therefore
certain
physical
feelings
faith
practical
religious
legal
serious
actions
spiritual
self
concepts
condition
life
concerns
opinions
knowledge
psychological
ideological
scientific
Christian
tradition
Jew
beliefs
belief
divine
educational
principles
philosophical
practices
virtue
consequences
values
religion
rational
attitudes
intellectual
evil
motivation
thinking
ideology
emotional
conscience
ought
expression
issues
liberty
notion
moral
assertion
justification
sexual
philosophy
freedom
ethics
human
equality
behavior
ethical
morality
argument

Can we "de-bias" embeddings?

Can we "de-bias" embeddings?

Bolukbasi *et al.*: this may be possible.

Idea: "collapse" dimensions corresponding to key attributes, such as gender.

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between between the words *receptionist* and *female*, while maintaining desired associations such as between the words *queen* and *female*. We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

1 Introduction

There have been hundreds or thousands of papers written about word embeddings and their applications, from Web search [27] to parsing Curriculum Vitae [16]. However, none of these papers have recognized how blatantly sexist the embeddings are and hence risk introducing biases of various types into real-world systems.

A word embedding that represent each word (or common phrase) w as a d-dimensional *word vector* $\vec{w} \in \mathbb{R}^d$. Word embeddings, trained only on word co-occurrence in text corpora, serve as a dictionary of sorts for computer programs that would like to use word meaning. First, words with similar semantic meanings tend to have vectors that are close together. Second, the vector differences between words in embeddings have been shown to represent relationships between words [32, 26]. For example given an analogy puzzle, "man is to king as woman is to x" (denoted as *man*:*king* :: *woman*:x), simple arithmetic of the embedding vectors finds that x=*queen* is the best answer because:

$$\overrightarrow{\mathrm{man}} - \overrightarrow{\mathrm{woman}} \approx \overrightarrow{\mathrm{king}} - \overrightarrow{\mathrm{queen}}$$

Similarly, x=*Japan* is returned for *Paris*:*France* :: *Tokyo*:x. It is surprising that a simple vector arithmetic can simultaneously capture a variety of relationships. It has also excited practitioners because such a tool could be useful across applications involving natural language. Indeed, they are being studied and used in a variety of downstream applications (e.g., document ranking [27], sentiment analysis [18], and question retrieval [22]).

However, the embeddings also pinpoint sexism implicit in text. For instance, it is also the case that:

$$\overrightarrow{\mathrm{man}} - \overrightarrow{\mathrm{woman}} \approx \overrightarrow{\mathrm{computer\ programmer}} - \overrightarrow{\mathrm{homemaker}}.$$

How can we build systems that are fair?

First, we need to decide what we mean by “fair”…

Interesting fact:
You can't always get what you want in terms of “fairness”!

Fairness: you can't always get what you want!

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

COMPAS (from company called Northpointe)

- Estimates chances a defendant will be re-arrested
  - Issue: "rearrest" != "committed crime"
- Meant to be used for bail decisions
  - Issue: also used for sentencing

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

These contingency tables reveal that the algorithm is more likely to misclassify a black defendant as higher risk than a white defendant. Black defendants who do not recidivate were nearly twice as likely to be classified by COMPAS as higher risk compared to their white counterparts (45 percent vs. 23 percent). However, black defendants who scored higher did recidivate slightly more often than white defendants (63 percent vs. 59 percent).

This conclusion came from applying COMPAS to historical arrest records.

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Enter the computer scientists...

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

Alexandra Chouldechova

Heinz College, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA, USA
achould@cmu.edu

Abstract

Recidivism prediction instruments (RPI's) provide decision makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time. While such instruments are gaining increasing popularity across the country, their use is attracting tremendous controversy. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. This paper discusses a fairness criterion originating in the field of educational and psychological testing that has recently been applied to assess the fairness of recidivism prediction instruments. We demonstrate how adherence to the criterion may lead to considerable disparate impact when recidivism prevalence differs across groups.

1 Introduction

Risk assessment instruments are gaining increasing popularity within the criminal justice system, with versions of such instruments being used or considered for use in pre-trial decision-making, parole decisions, and in some states even sentencing [1, 2]. In each of these cases, a high-risk classification—particularly a high-risk misclassification—may have a direct adverse impact on a criminal defendant's outcome. If RPI's are to continue to be used, it is important to ensure that they do not result in unethical practices that disparately affect different groups.

Within the psychometrics literature, there exist widely accepted and adopted standards for assessing whether an instrument is fair in the sense of being free of *predictive bias*. These standards have recently been applied to the COMPAS [3] and PCRA [4] instruments, with initial findings suggesting that there is evidence of predictive bias when it comes to gender, but not when it comes to race [5, 6, 7].

In a recent widely popularized investigation of the COMPAS RPI conducted by a team at ProPublica, a different approach to assessing instrument bias told what appears to be a contradictory story [8]. The authors found that the likelihood of a non-recidivating Black defendant being assessed as high-risk is nearly twice that of White defendants. While this analysis has met with much criticism, it has also made headlines. There is no doubt that it is now embedded in the national conversation on the use of RPI's.

In this paper we show that the differences in false positive and false negative rates cited as evidence of racial bias in the ProPublica article are a direct consequence of applying an instrument that is free from predictive bias[1] to a population in which recidivism prevalence differs across groups. Our main contribution is twofold. (1) First, we make precise the connection between the psychometric notion of test fairness and error rates in classification. (2) Next, we demonstrate how using an RPI that has different false postive and false negative rates between groups can lead to disparate impact when individuals assessed as high risk receive stricter penalties. Throughout our discussion we use the term *disparate impact* to refer to settings where a penalty policy has unintended disproportionate adverse impact on a particular group.

It is important to bear in mind that fairness itself—along with the notion of disparate impact—is a social and ethical concept, not a statistical one. An instrument that is free from predictive bias may nevertheless result in disparate impact depending on how and where it is used. In this paper we consider

[1] in the psychometric sense

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg * Sendhil Mullainathan † Manish Raghavan ‡

Abstract

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

1 Introduction

There are many settings in which a sequence of people comes before a decision-maker, who must make a judgment about each based on some observable set of features. Across a range of applications, these judgments are being carried out by an increasingly wide spectrum of approaches ranging from human expertise to algorithmic and statistical frameworks, as well as various combinations of these approaches.

Along with these developments, a growing line of work has asked how we should reason about issues of bias and discrimination in settings where these algorithmic and statistical techniques, trained on large datasets of past instances, play a significant role in the outcome. Let us consider three examples where such issues arise, both to illustrate the range of relevant contexts, and to surface some of the challenges.

A set of example domains. First, at various points in the criminal justice system, including decisions about bail, sentencing, or parole, an officer of the court may use quantitative *risk tools* to assess a defendant's probability of recidivism — future arrest — based on their past history and other attributes. Several recent analyses have asked whether such tools are mitigating or exacerbating the sources of bias in the criminal justice system; in one widely-publicized report, Angwin et al. analyzed a commonly used statistical method for assigning risk scores in the criminal justice system — the COMPAS risk tool — and argued that it was biased against African-American defendants [2, 23]. One of their main contentions was that the tool's errors were asymmetric: African-American defendants were more likely to be incorrectly labeled as higher-risk than they actually were, while white defendants were more likely to be incorrectly labeled as lower-risk than they actually were. Subsequent analyses raised methodological objections to this report, and also observed that despite the COMPAS risk tool's errors, its estimates of the probability of recidivism are equally well calibrated to the true outcomes for both African-American and white defendants [1, 10, 13, 17].

*Cornell University
†Harvard University
‡Cornell University

Equality of Opportunity in Supervised Learning

Moritz Hardt Eric Price Nathan Srebro

October 11, 2016

Abstract

We propose a criterion for discrimination against a specified sensitive attribute in supervised learning, where the goal is to predict some target based on available features. Assuming data about the predictor, target, and membership in the protected group are available, we show how to optimally *adjust* any learned predictor so as to remove discrimination according to our definition. Our framework also improves incentives by shifting the cost of poor classification from disadvantaged groups to the decision maker, who can respond by improving the classification accuracy.

In line with other studies, our notion is *oblivious*: it depends only on the joint statistics of the predictor, the target and the protected attribute, but not on interpretation of individual features. We study the inherent limits of defining and identifying biases based on such oblivious measures, outlining what can and cannot be inferred from different oblivious tests.

We illustrate our notion using a case study of FICO credit scores.

1 Introduction

As machine learning increasingly affects decisions in domains protected by anti-discrimination law, there is much interest in algorithmically measuring and ensuring fairness in machine learning. In domains such as advertising, credit, employment, education, and criminal justice, machine learning could help obtain more accurate predictions, but its effect on existing biases is not well understood. Although reliance on data and quantitative measures can help quantify and eliminate existing biases, some scholars caution that algorithms can also introduce new biases or perpetuate existing ones [BS16]. In May 2014, the Obama Administration's Big Data Working Group released a report [PPM+14] arguing that discrimination can sometimes "be the inadvertent outcome of the way big data technologies are structured and used" and pointed toward "the potential of encoding discrimination in automated decisions". A subsequent White House report [Whi16] calls for "equal opportunity by design" as a guiding principle in domains such as credit scoring.

Despite the demand, a vetted methodology for avoiding discrimination against *protected attributes* in machine learning is lacking. A naïve approach might require that the algorithm should ignore all protected attributes such as race, color, religion, gender, disability, or family status. However, this idea of "fairness through unawareness" is ineffective due to the existence of *redundant encodings*, ways of predicting protected attributes from other features [PRT08].

Another common conception of non-discrimination is *demographic parity*. Demographic parity requires that a decision—such as accepting or denying a loan application—be independent of the protected attribute. In the case of a binary decision $\hat{Y} \in \{0,1\}$ and a binary protected attribute $A \in \{0,1\}$, this constraint can be formalized by asking that $\Pr\{\hat{Y} = 1 \mid A = 0\} = \Pr\{\hat{Y} =$

Algorithmic decision making and the cost of fairness

Sam Corbett-Davies
Stanford University
scorbett@stanford.edu

Emma Pierson
Stanford University
emmap1@stanford.edu

Avi Feller
Univ. of California, Berkeley
afeller@berkeley.edu

Sharad Goel
Stanford University
scgoel@stanford.edu

Aziz Huq
University of Chicago
huq@uchicago.edu

ABSTRACT

Algorithms are now regularly used to decide whether defendants awaiting trial are too dangerous to be released back into the community. In some cases, black defendants are substantially more likely than white defendants to be incorrectly classified as high risk. To mitigate such disparities, several techniques recently have been proposed to achieve *algorithmic fairness*. Here we reformulate algorithmic fairness as constrained optimization: the objective is to maximize public safety while satisfying formal fairness constraints designed to reduce racial disparities. We show that for several past definitions of fairness, the optimal algorithms that result require detaining defendants above race-specific risk thresholds. We further show that the optimal *unconstrained* algorithm requires applying a single, uniform threshold to all defendants. The unconstrained algorithm thus maximizes public safety while also satisfying one important understanding of equality: that all individuals are held to the same standard, irrespective of race. Because the optimal constrained and unconstrained algorithms generally differ, there is tension between improving public safety and satisfying prevailing notions of algorithmic fairness. By examining data from Broward County, Florida, we show that this trade-off can be large in practice. We focus on algorithms for pretrial release decisions, but the principles we discuss apply to other domains, and also to human decision makers carrying out structured decision rules.

1 INTRODUCTION

Judges nationwide use algorithms to help decide whether defendants should be detained or released while awaiting trial [13, 31]. One such algorithm, called COMPAS, assigns defendants risk scores between 1 and 10 that indicate how likely they are to commit a violent crime based on more than 100 factors, including age, sex and criminal history. For example, defendants with scores of 7 reoffend at twice the rate as those with scores of 3. Accordingly, defendants classified as high risk are much more likely to be detained while awaiting trial than those classified as low risk.

These algorithms do not explicitly use race as an input. Nevertheless, an analysis of defendants in Broward County, Florida [2] revealed that black defendants are substantially more likely to be classified as high risk. Further, among defendants who ultimately did not reoffend, blacks were more than twice as likely as whites to be labeled as risky. Even though these defendants did not go on to commit a crime, being classified as high risk meant they were subjected to harsher treatment by the courts. To reduce racial disparities of this kind, several authors recently have proposed a variety of *fair* decision algorithms [16, 21, 23, 24, 26].[1]

Here we reformulate algorithmic fairness as constrained optimization: the objective is to maximize public safety while satisfying formal fairness constraints. We show that for several past definitions of fairness, the optimal algorithms that result require applying multiple, race-specific thresholds to individuals' risk scores. One might, for example, detain white defendants who score above 4, but detain black defendants only if they score above 6. We further show that the optimal *unconstrained* algorithm requires applying a single, uniform threshold to all defendants. This safety-maximizing rule thus satisfies one important understanding of equality: that all individuals are held to the same standard, irrespective of race. Since the optimal constrained and unconstrained algorithms in general differ, there is tension between reducing racial disparities and improving public safety. By examining data from Broward County, we demonstrate that this tension is more than theoretical. Adhering to past fairness definitions can substantially decrease public safety; conversely, optimizing for public safety alone can produce stark racial disparities.

We focus here on the problem of designing algorithms for pretrial release decisions, but the principles we discuss apply to other domains, and also to human decision makers carrying out structured decision rules. We emphasize at the outset that algorithmic decision making does not preclude additional, or alternative, policy interventions. For example, one might provide released defendants with robust social services aimed at reducing recidivism, or conclude that it is more effective and equitable to replace pretrial detention with non-custodial supervision. Moreover, regardless of the algorithm used, human discretion may be warranted in individual cases.

2 BACKGROUND

2.1 Defining algorithmic fairness

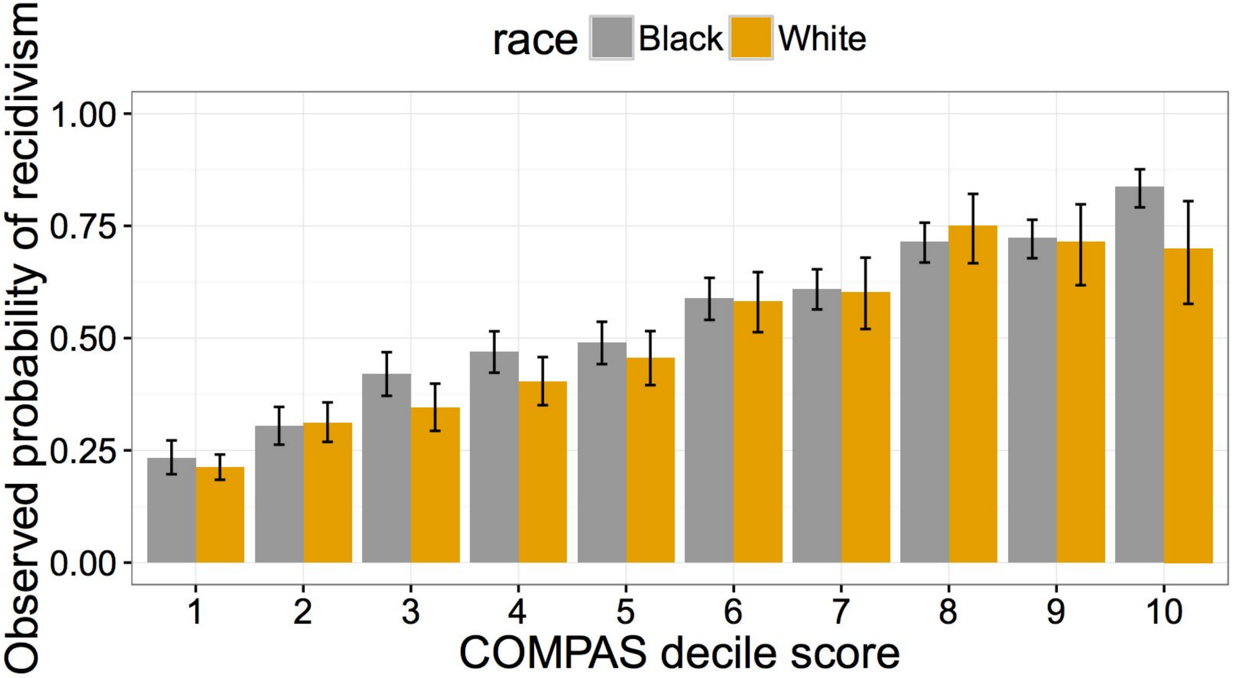
Existing approaches to algorithmic fairness typically proceed in two steps. First, a formal criterion of fairness is defined; then, a decision rule is developed to satisfy that measure, either exactly or approximately. To formally define past fairness measures, we introduce a general notion of (randomized) decision rules. Suppose we have a vector $x_i \in \mathbb{R}^p$ that we interpret as the visible attributes of individual i. For example, x might represent a defendant's age, gender, race, and criminal history. We consider binary decisions

Working paper, Stanford University
2017. 978-x-xxxx-xxxx-x/YY/MM.
DOI: 10.1145/nnnnnnn.nnnnnnn

[1] We consider racial disparities because they have been at the center of many recent debates in criminal justice, but the same logic applies across a range of possible attributes, including gender.

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

Alexandra Chouldechova

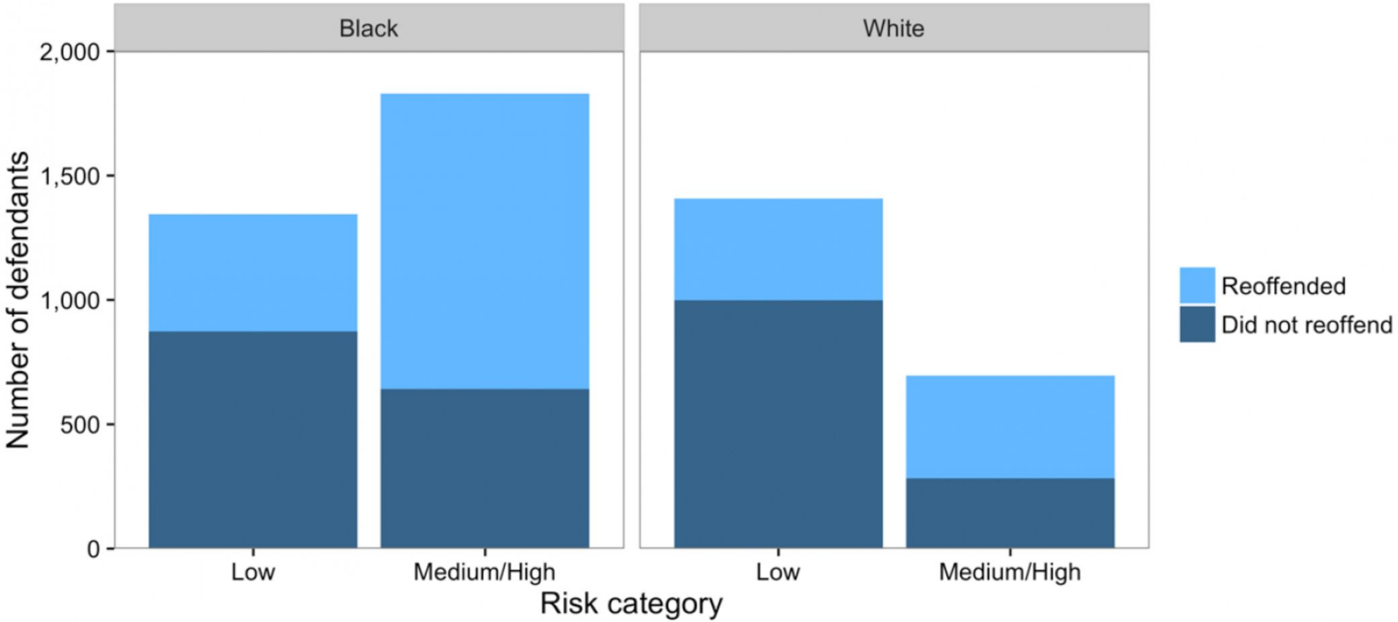


A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel October 17, 2016

The Washington Post

Democracy Dies in Darkness

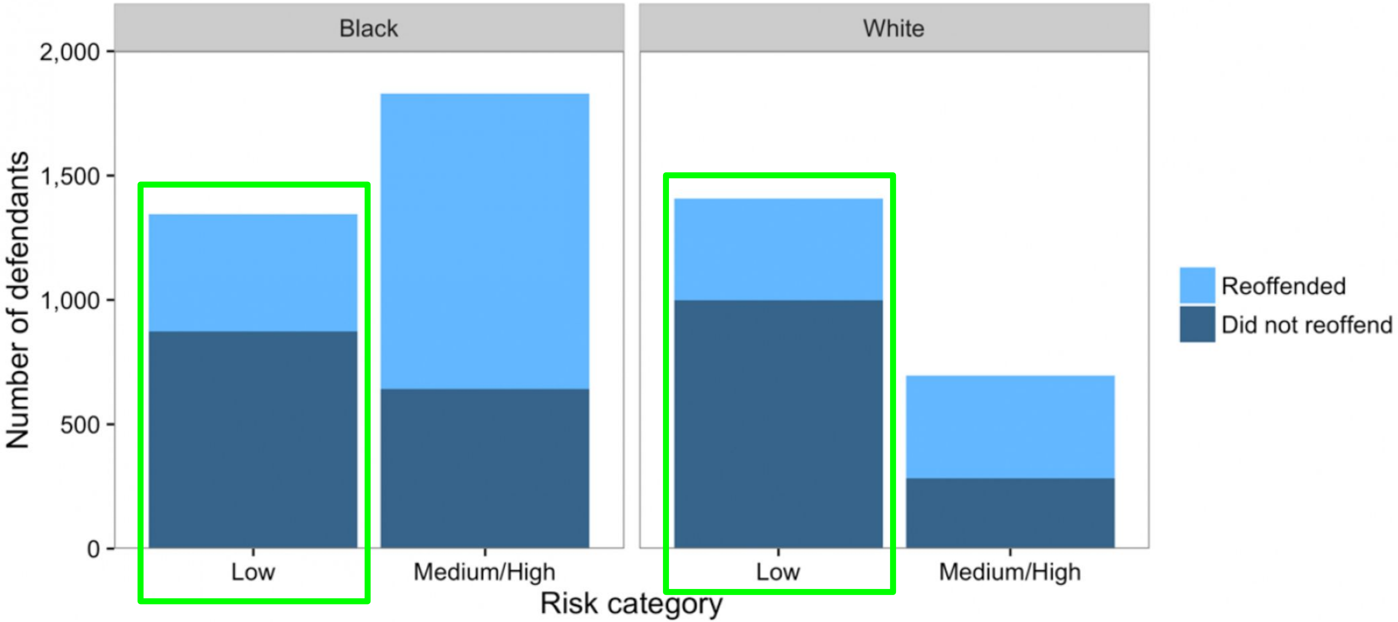


A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel October 17, 2016

The Washington Post

Democracy Dies in Darkness



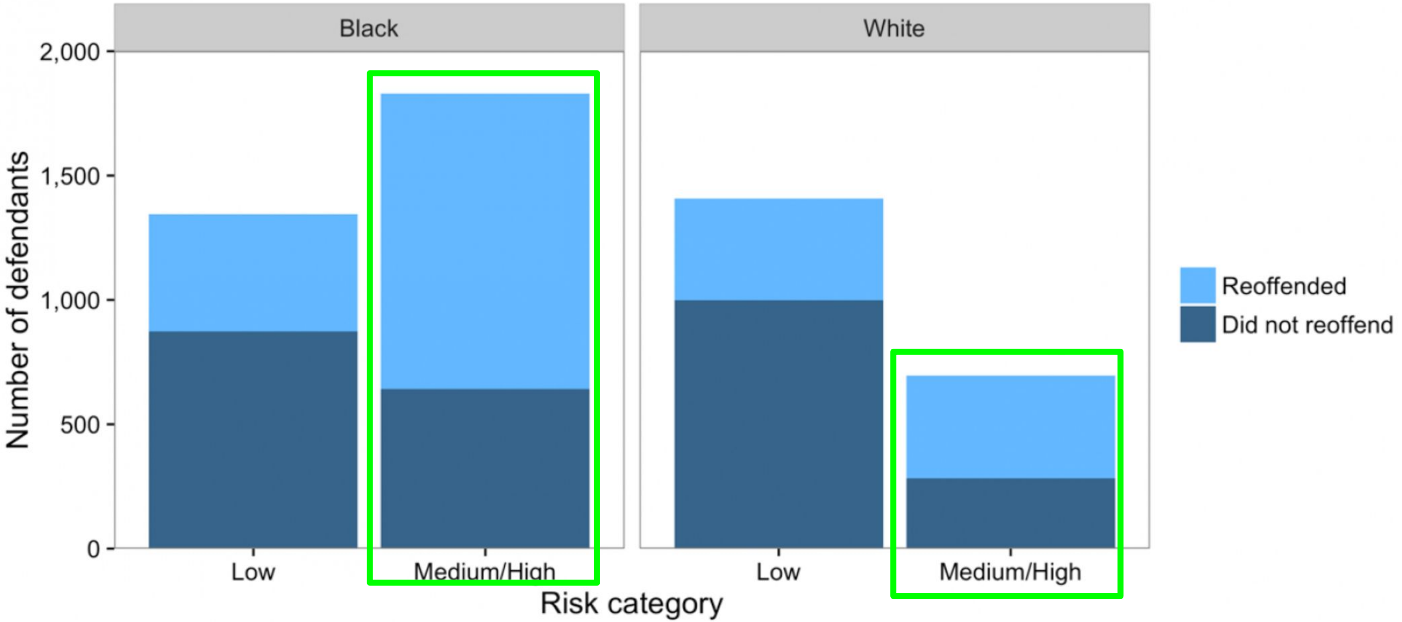
Low-risk

FAIR

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

The Washington Post
Democracy Dies in Darkness



Low-risk
FAIR

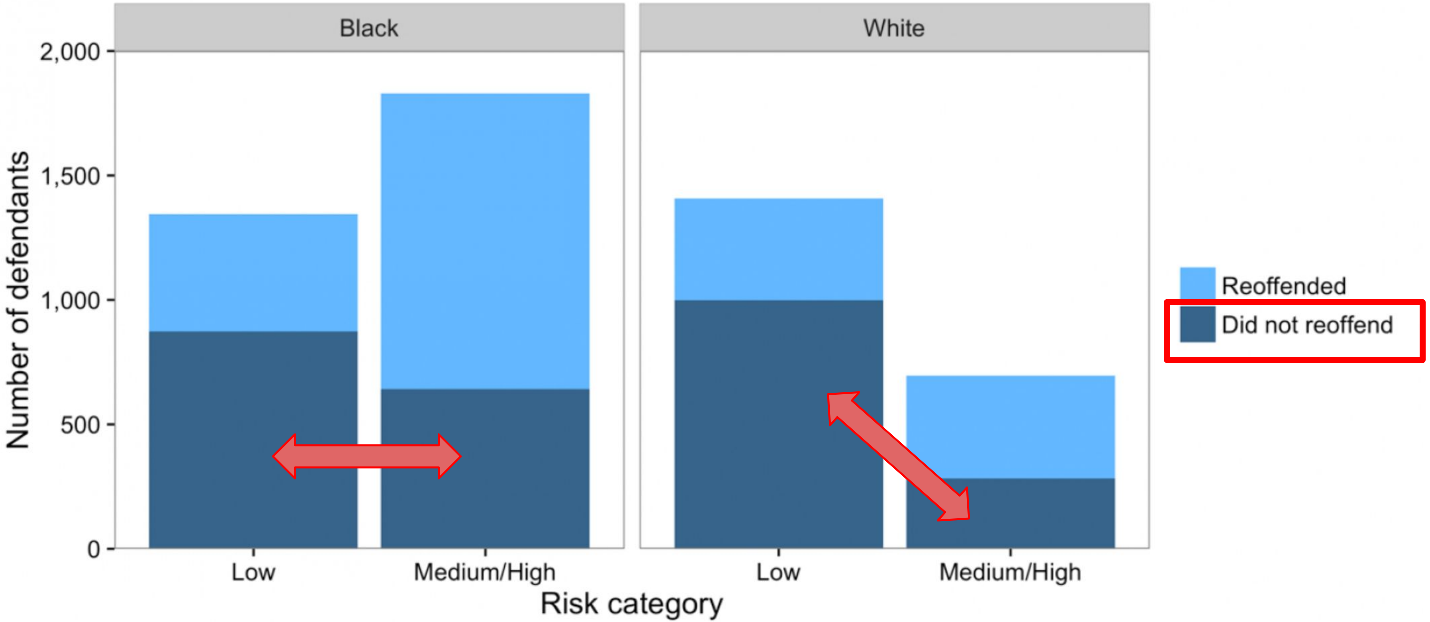
Medium-high
FAIR

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

The Washington Post

Democracy Dies in Darkness



Low-risk
FAIR

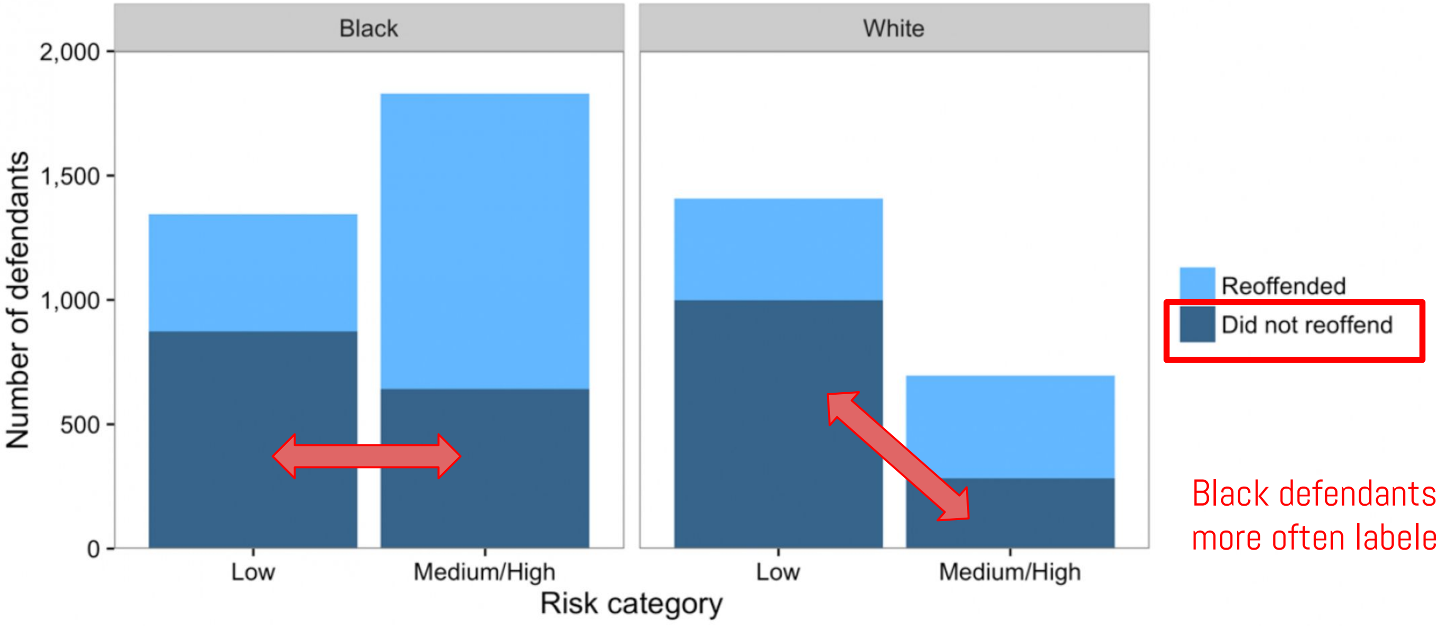
Medium-high
FAIR

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

The Washington Post

Democracy Dies in Darkness



Low-risk
FAIR

Medium-high
FAIR

Did not reoffend
UNFAIR

Black defendants who did not reoffend were more often labeled "high risk"

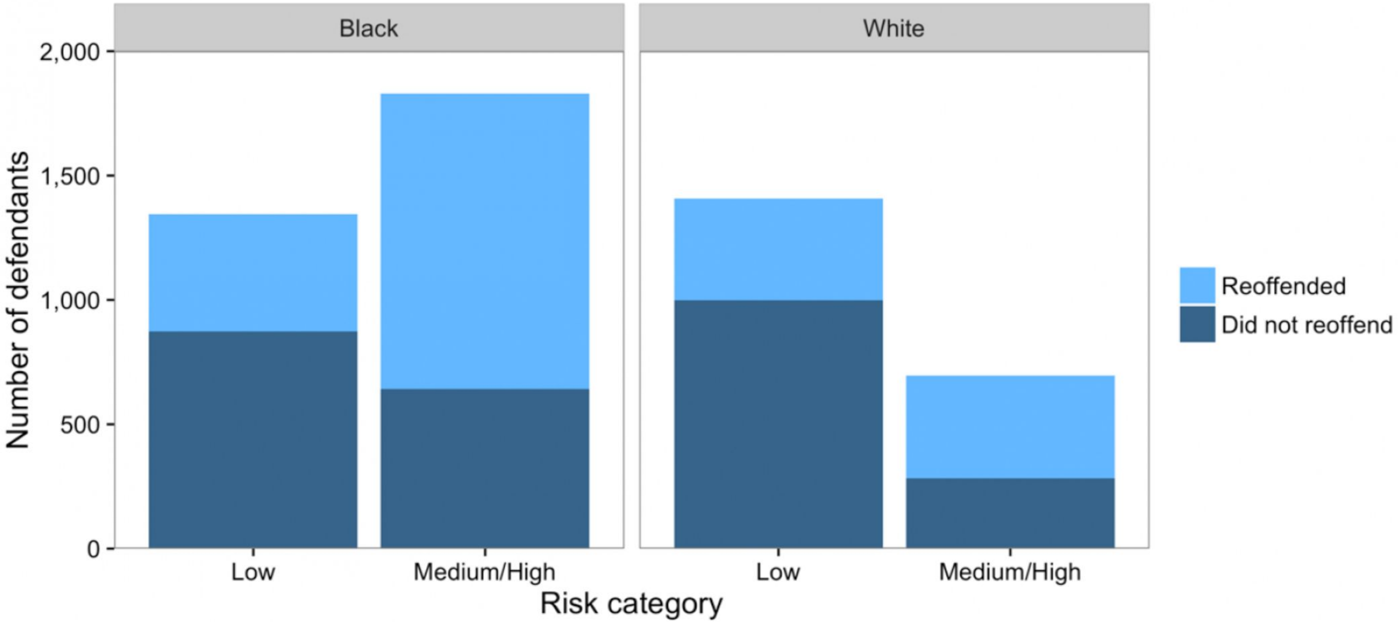
A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

The Washington Post

Democracy Dies in Darkness

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel October 17, 2016

Unless classifier is perfect, can't **all** be fair due to different base rates



Low-risk
FAIR

Medium-high
FAIR

Did not reoffend
UNFAIR

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg * Sendhil Mullainathan † Manish Raghavan ‡

When the two groups have equal base rates, then the risk assignment that gives the same score to everyone in the population achieves statistical parity along with conditions (A), (B), and (C). But when the two groups do not have equal base rates, it is immediate to show that statistical parity is inconsistent with both the calibration condition (A) and with the conjunction of the two balance conditions (B) and (C). To see the inconsistency of statistical parity with the calibration condition, we take Equation (1) from the proof above, sum the coordinates of the vectors on both sides, and divide by N_t, the number of people in group t. Statistical parity requires that the right-hand sides of the resulting equation be the same for $t = 1, 2$, while the assumption that the two groups have unequal base rates implies that the left-hand sides of the equation must be different for $t = 1, 2$. To see the inconsistency of statistical parity with the two balance conditions (B) and (C), we simply observe that if the average score assigned to the positive class and to the negative class are the same in the two groups, then the average of the scores over all members of the two groups cannot be the same provided they do not contain the same proportion of positive-class and negative-class members.

3 The Approximate Theorem

In this section we prove Theorem 1.2. First, we must first give a precise specification of the approximate fairness conditions:

$$(1-\varepsilon)[n_t^\top XV]_b \leq [n_t^\top PX]_b \leq (1-\varepsilon)[n_t^\top XV]_b \tag{A'}$$

$$(1-\varepsilon)\left(\frac{1}{N_2-\mu_2}\right) n_t^\top (I-P)Xv \leq \left(\frac{1}{N_1-\mu_1}\right) n_t^\top (I-P)Xv \leq (1+\varepsilon)\left(\frac{1}{N_2-\mu_2}\right) n_t^\top (I-P)Xv \tag{B'}$$

$$(1-\varepsilon)\left(\frac{1}{\mu_2}\right) n_t^\top PXv \leq \left(\frac{1}{\mu_1}\right) n_t^\top PXv \leq (1+\varepsilon)\left(\frac{1}{\mu_2}\right) n_t^\top PXv \tag{C'}$$

For (B') and (C'), we also require that these hold when μ_1 and μ_2 are interchanged.

We also specify the approximate versions of perfect prediction and equal base rates in terms of $f(\varepsilon)$, which is a function that goes to 0 as ε goes to 0.

- *Approximate perfect prediction.* $\gamma_1 \geq 1 - f(\varepsilon)$ and $\gamma_2 \geq 1 - f(\varepsilon)$
- *Approximately equal base rates.* $|\mu_1/N_1 - \mu_2/N_2| \leq f(\varepsilon)$

A brief overview of the proof of Theorem 1.2 is as follows. It proceeds by first establishing an approximate form of Equation (1) above, which implies that the total expected score assigned in each group is approximately equal to the total size of the positive class. This in turn makes it possible to formulate approximate forms of Equations (3) and (4). When the base rates are close together, the approximation is too loose to derive bounds on the predictive power; but this is okay since in this case we have approximately equal base rates. Otherwise, when the base rates differ significantly, we show that most of the expected score must be assigned to the positive class, giving us approximately perfect prediction.

The remainder of this section provides the full details of the proof.

Total scores and the number of people in the positive class. First, we will show that the total score for each group is approximately μ_t, the number of people in the positive class. Define $\hat{\mu}_t = n_t^\top Xv$. Using (A'),

we have

$$\begin{aligned}
\hat{\mu}_t &= n_t^\top Xv \\
&= n_t^\top XVe \\
&= \sum_{b=1}^{B} [n_t^\top PX]_b \\
&\leq (1+\varepsilon)\sum_{b=1}^{B} [n_t^\top PX]_b \\
&= (1+\varepsilon) n_t^\top PXe \\
&= (1+\varepsilon)\mu_t
\end{aligned}$$

Similarly, we can lower bound $\hat{\mu}_t$ as

$$\begin{aligned}
\hat{\mu}_t &= \sum_{b=1}^{B} [n_t^\top PX]_b \\
&\geq (1-\varepsilon)\sum_{b=1}^{B} [n_t^\top PX]_b \\
&= (1-\varepsilon)\mu_t
\end{aligned}$$

Combining these, we have

$$(1-\varepsilon)\mu_t \leq \hat{\mu}_t \leq (1+\varepsilon)\mu_t. \tag{7}$$

The portion of the score received by the positive class. We can use (C') to show that $\gamma_1 \approx \gamma_2$. Recall that γ_t, the average of the expected scores assigned to members of the positive class in group t, is defined as $\gamma_t = \frac{1}{\mu_t} n_t PXv$. Then, it follows trivially from (C') that

$$(1-\varepsilon)\gamma_2 \leq \gamma_1 \leq (1+\varepsilon)\gamma_2. \tag{8}$$

The relationship between the base rates. We can apply this to (B') to relate μ_1 and μ_2, using the observation that the score not received by people of the positive class must fall instead to people of the negative class. Examining the left inequality of (B'), we have

$$\begin{aligned}
(1-\varepsilon)\left(\frac{1}{N_2-\mu_2}\right) n_t^\top (I-P)Xv &= (1-\varepsilon)\left(\frac{1}{N_2-\mu_2}\right)(n_t^\top Xv - n_t^\top PXv) \\
&= (1-\varepsilon)\left(\frac{1}{N_2-\mu_2}\right)(\hat{\mu}_2 - \gamma_2\mu_2) \\
&\geq (1-\varepsilon)\left(\frac{1}{N_2-\mu_2}\right)((1-\varepsilon)\mu_2 - \gamma_2\mu_t) \\
&= (1-\varepsilon)\left(\frac{\mu_2}{N_2-\mu_2}\right)(1-\varepsilon-\gamma_2) \\
&\geq (1-\varepsilon)\left(\frac{\mu_2}{N_2-\mu_2}\right)\left(1-\varepsilon-\frac{\gamma_1}{1-\varepsilon}\right) \\
&= (1-2\varepsilon+\varepsilon^2-\gamma_1)\left(\frac{\mu_2}{N_2-\mu_2}\right)
\end{aligned}$$

Thus, the left inequality of (B') becomes

$$(1-2\varepsilon+\varepsilon^2-\gamma_1)\left(\frac{\mu_2}{N_2-\mu_2}\right) \leq \left(\frac{1}{N_1-\mu_1}\right) n_t^\top (I-P)Xv \tag{9}$$

By definition, $\hat{\mu}_1 = n_t^\top Xv$ and $\gamma_t\mu_t = n_t^\top PXv$, so this becomes

$$(1-2\varepsilon+\varepsilon^2-\gamma_1)\left(\frac{\mu_2}{N_2-\mu_2}\right) \leq \left(\frac{1}{N_1-\mu_1}\right)(\hat{\mu}_1 - \gamma_1\mu_1) \tag{10}$$

If the base rates differ. Let ρ_1 and ρ_2 be the respective base rates, i.e. $\rho_1 = \mu_1/N_1$ and $\rho_2 = \mu_2/N_2$. Assume that $\rho_1 \leq \rho_2$ (otherwise we can switch μ_1 and μ_2 in the above analysis), and assume towards contradiction that the base rates differ by at least $\sqrt{\varepsilon}$, meaning $\rho_1 + \sqrt{\varepsilon} < \rho_2$. Using (10),

$$\begin{aligned}
\frac{\rho_1+\sqrt{\varepsilon}}{1-\rho_1=\sqrt{\varepsilon}} &\leq \frac{\rho_2}{1-\rho_2} \\
&\leq \left(\frac{1+\varepsilon-\gamma_1}{1-2\varepsilon+\varepsilon^2-\gamma_1}\right)\left(\frac{\rho_1}{1-\rho_1}\right) \\
(\rho_1+\sqrt{\varepsilon})(1-\rho_1)(1-2\varepsilon+\varepsilon^2-\gamma_1) &\leq \rho_1(1-\rho_1-\sqrt{\varepsilon})(1+\varepsilon-\gamma_1) \\
(\rho_1+\sqrt{\varepsilon})(1-\rho_1)(1-2\varepsilon) - \rho_1(1-\rho_1-\sqrt{\varepsilon})(1+\varepsilon) &\leq \gamma_1\left[(\rho_1+\sqrt{\varepsilon})(1-\rho_1) - \rho_1(1-\rho_1-\sqrt{\varepsilon})\right] \\
\rho_1[(1-\rho_1)(1-2\varepsilon) - (1-\rho_1-\sqrt{\varepsilon})(1+\varepsilon)] + \sqrt{\varepsilon}(1-\rho_1)(1-2\varepsilon) &\leq \gamma_1[\sqrt{\varepsilon}(1-\rho_1) + \sqrt{\varepsilon}\rho_1] \\
\rho_1(-2\varepsilon+2\varepsilon\rho_1-\varepsilon+\varepsilon\rho_1+\sqrt{\varepsilon}+\varepsilon\sqrt{\varepsilon}) + \sqrt{\varepsilon}(1-2\varepsilon-\rho_1+2\varepsilon\rho_1) &\leq \gamma_1\sqrt{\varepsilon} \\
\rho_1(-3\varepsilon+3\varepsilon\rho_1+\sqrt{\varepsilon}+\varepsilon\sqrt{\varepsilon}-\sqrt{\varepsilon}+2\varepsilon\sqrt{\varepsilon}) + \sqrt{\varepsilon}(1-2\varepsilon) &\leq \gamma_1\sqrt{\varepsilon} \\
\varepsilon\rho_1(-3+3\rho_1+3\sqrt{\varepsilon}) + \sqrt{\varepsilon}(1-2\varepsilon) &\leq \gamma_1\sqrt{\varepsilon} \\
3\varepsilon\rho_1(-1+\rho_1) + \sqrt{\varepsilon}(1-2\varepsilon) &\leq \gamma_1\sqrt{\varepsilon} \\
1-2\varepsilon-3\sqrt{\varepsilon}\rho_1(1-\rho_1) &\leq \gamma_1 \\
1-\sqrt{\varepsilon}\left(2\sqrt{\varepsilon}+\frac{3}{4}\right) &\leq \gamma_1
\end{aligned}$$

Recall that $\gamma_2 \geq \gamma_1(1-\varepsilon)$, so

$$\begin{aligned}
\gamma_2 &\geq (1-\varepsilon)\gamma_1 \\
&\geq (1-\varepsilon)\left(1-\sqrt{\varepsilon}\left(2\sqrt{\varepsilon}+\frac{3}{4}\right)\right) \\
&\geq 1-\varepsilon-\sqrt{\varepsilon}\left(2\sqrt{\varepsilon}+\frac{3}{4}\right) \\
&= 1-\sqrt{\varepsilon}\left(3\sqrt{\varepsilon}+\frac{3}{4}\right)
\end{aligned}$$

Let $f(\varepsilon) = \sqrt{\varepsilon}\max(1, 3\sqrt{\varepsilon}+3/4)$. Note that we assumed that ρ_1 and ρ_2 differ by an additive $\sqrt{\varepsilon} \leq f(\varepsilon)$. Therefore if the ε-fairness conditions are met and the base rates are not within an additive $f(\varepsilon)$, then $\gamma_1 \geq 1-f(\varepsilon)$ and $\gamma_2 \geq 1-f(\varepsilon)$. This completes the proof of Theorem 1.2.

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
ProPublica, Dec. 30, 2016, 4:44 p.m.

https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say

Lessons learned

- Fairness of an algorithm depends in part on how it's used
- In fairness, you can't always get (everything) you want
  - Must make a careful choice of quantitative metrics
  - This involves case-by-case policy decisions
  - These tradeoffs affect human decisions too!
- Improvements to fairness may come with their own costs to other values we want to retain (privacy, performance, etc.)

Can computer scientists do anything besides depress us?

Equality of Opportunity in Supervised Learning

Moritz Hardt Eric Price Nathan Srebro

October 11, 2016

Abstract

We propose a criterion for discrimination against a specified sensitive attribute in supervised learning, where the goal is to predict some target based on available features. Assuming data about the predictor, target, and membership in the protected group are available, we show how to optimally *adjust* any learned predictor so as to remove discrimination according to our definition. Our framework also improves incentives by shifting the cost of poor classification from disadvantaged groups to the decision maker, who can respond by improving the classification accuracy.

In line with other studies, our notion is *oblivious*: it depends only on the joint statistics of the predictor, the target and the protected attribute, but not on interpretation of individual features. We study the inherent limits of defining and identifying biases based on such oblivious measures, outlining what can and cannot be inferred from different oblivious tests.

We illustrate our notion using a case study of FICO credit scores.

1 Introduction

As machine learning increasingly affects decisions in domains protected by anti-discrimination law, there is much interest in algorithmically measuring and ensuring fairness in machine learning. In domains such as advertising, credit, employment, education, and criminal justice, machine learning could help obtain more accurate predictions, but its effect on existing biases is not well understood. Although reliance on data and quantitative measures can help quantify and eliminate existing biases, some scholars caution that algorithms can also introduce new biases or perpetuate existing ones [BS16]. In May 2014, the Obama Administration's Big Data Working Group released a report [PPM+14] arguing that discrimination can sometimes "be the inadvertent outcome of the way big data technologies are structured and used" and pointed toward "the potential of encoding discrimination in automated decisions". A subsequent White House report [Whi16] calls for "equal opportunity by design" as a guiding principle in domains such as credit scoring.

Despite the demand, a vetted methodology for avoiding discrimination against *protected attributes* in machine learning is lacking. A naïve approach might require that the algorithm should ignore all protected attributes such as race, color, religion, gender, disability, or family status. However, this idea of "fairness through unawareness" is ineffective due to the existence of *redundant encodings*, ways of predicting protected attributes from other features [PRT08].

Another common conception of non-discrimination is *demographic parity*. Demographic parity requires that a decision—such as accepting or denying a loan application—be independent of the protected attribute. In the case of a binary decision $\widehat{Y} \in \{0,1\}$ and a binary protected attribute $A \in \{0,1\}$, this constraint can be formalized by asking that $\Pr\{\widehat{Y} = 1 \mid A = 0\} = \Pr\{\widehat{Y} =$

Hardt, Price, Srebro (2016)
On forcing a threshold classifier to be "fair" by various definitions:

- Group-unaware
  Same threshold for each group
- Demographic Parity
  Same proportion of **positive** classifications
- "Equal opportunity"
  Same proportion of **true** positives

Can computer scientists do anything besides depress us?

Equality of Opportunity in Supervised Learning

Moritz Hardt Eric Price Nathan Srebro

October 11, 2016

Abstract

We propose a criterion for discrimination against a specified sensitive attribute in supervised learning, where the goal is to predict some target based on available features. Assuming data about the predictor, target, and membership in the protected group are available, we show how to optimally *adjust* any learned predictor so as to remove discrimination according to our definition. Our framework also improves incentives by shifting the cost of poor classification from disadvantaged groups to the decision maker, who can respond by improving the classification accuracy.

In line with other studies, our notion is *oblivious*: it depends only on the joint statistics of the predictor, the target and the protected attribute, but not on interpretation of individual features. We study the inherent limits of defining and identifying biases based on such oblivious measures, outlining what can and cannot be inferred from different oblivious tests.

We illustrate our notion using a case study of FICO credit scores.

1 Introduction

As machine learning increasingly affects decisions in domains protected by anti-discrimination law, there is much interest in algorithmically measuring and ensuring fairness in machine learning. In domains such as advertising, credit, employment, education, and criminal justice, machine learning could help obtain more accurate predictions, but its effect on existing biases is not well understood. Although reliance on data and quantitative measures can help quantify and eliminate existing biases, some scholars caution that algorithms can also introduce new biases or perpetuate existing ones [BS16]. In May 2014, the Obama Administration's Big Data Working Group released a report [PPM+14] arguing that discrimination can sometimes "be the inadvertent outcome of the way big data technologies are structured and used" and pointed toward "the potential of encoding discrimination in automated decisions". A subsequent White House report [Whi16] calls for "equal opportunity by design" as a guiding principle in domains such as credit scoring.

Despite the demand, a vetted methodology for avoiding discrimination against *protected attributes* in machine learning is lacking. A naïve approach might require that the algorithm should ignore all protected attributes such as race, color, religion, gender, disability, or family status. However, this idea of "fairness through unawareness" is ineffective due to the existence of *redundant encodings*, ways of predicting protected attributes from other features [PRT08].

Another common conception of non-discrimination is *demographic parity*. Demographic parity requires that a decision—such as accepting or denying a loan application—be independent of the protected attribute. In the case of a binary decision $\widehat{Y} \in \{0,1\}$ and a binary protected attribute $A \in \{0,1\}$, this constraint can be formalized by asking that $\Pr\{\widehat{Y} = 1 \mid A = 0\} = \Pr\{\widehat{Y} =$

A visual explanation...

Attacking discrimination in ML mathematically

Would default on loan

Would pay back loan

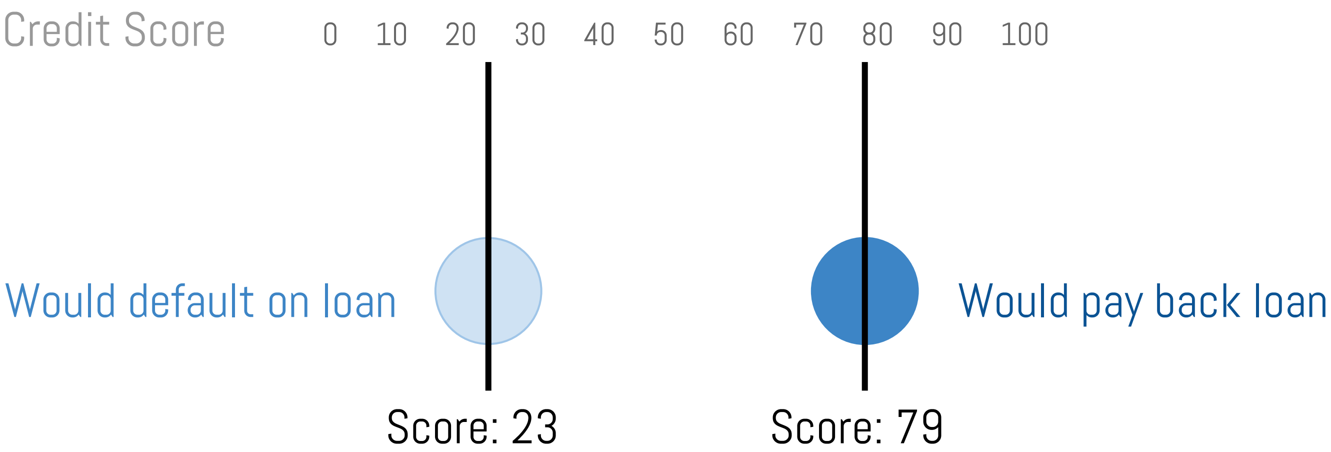
Attacking discrimination in ML mathematically

Credit Score 0 10 20 30 40 50 60 70 80 90 100

Would default on loan

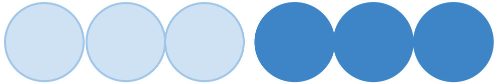
Would pay back loan

Attacking discrimination in ML mathematically

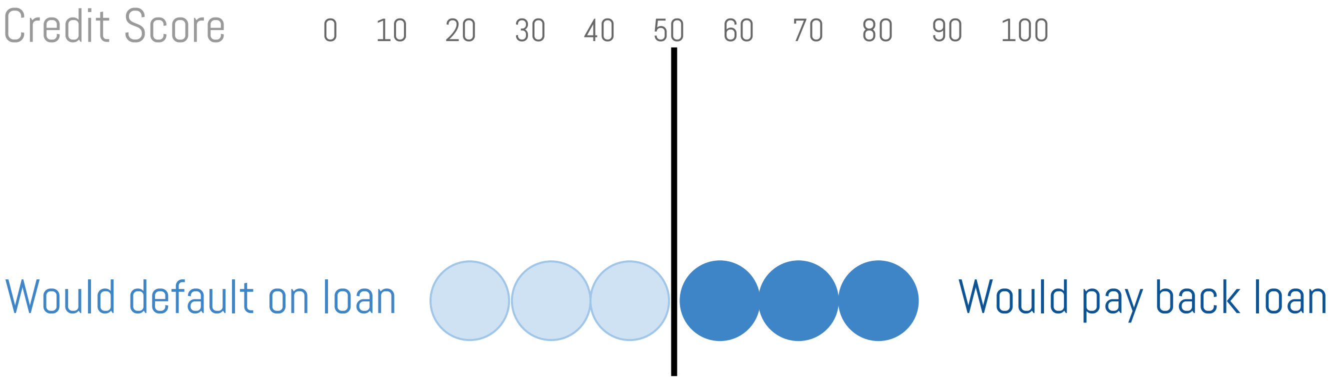


Attacking discrimination in ML mathematically

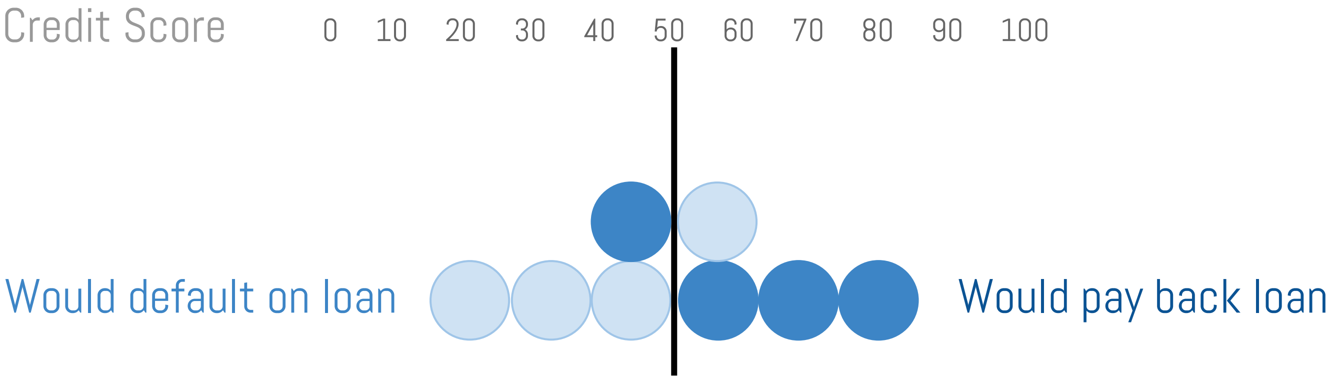
Credit Score 0 10 20 30 40 50 60 70 80 90 100

Would default on loan  Would pay back loan

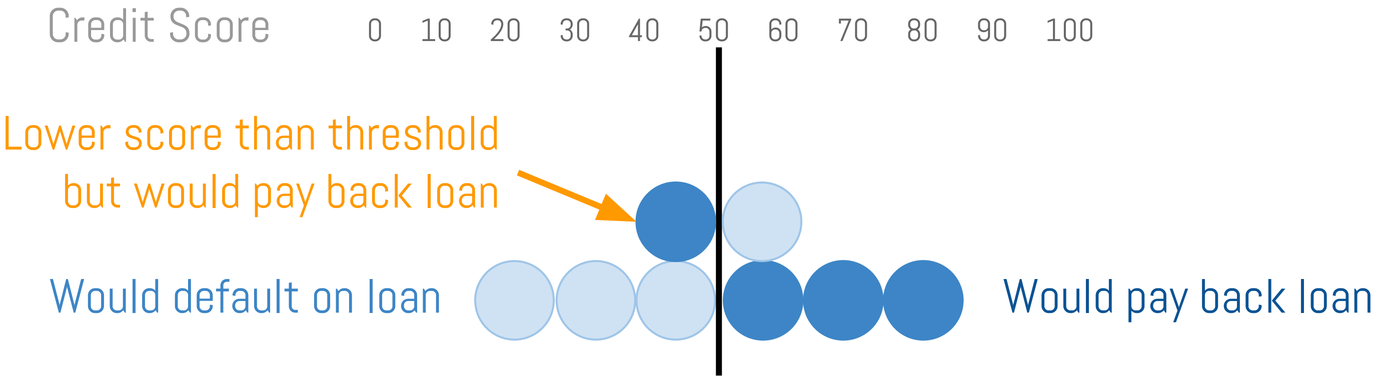
Attacking discrimination in ML mathematically



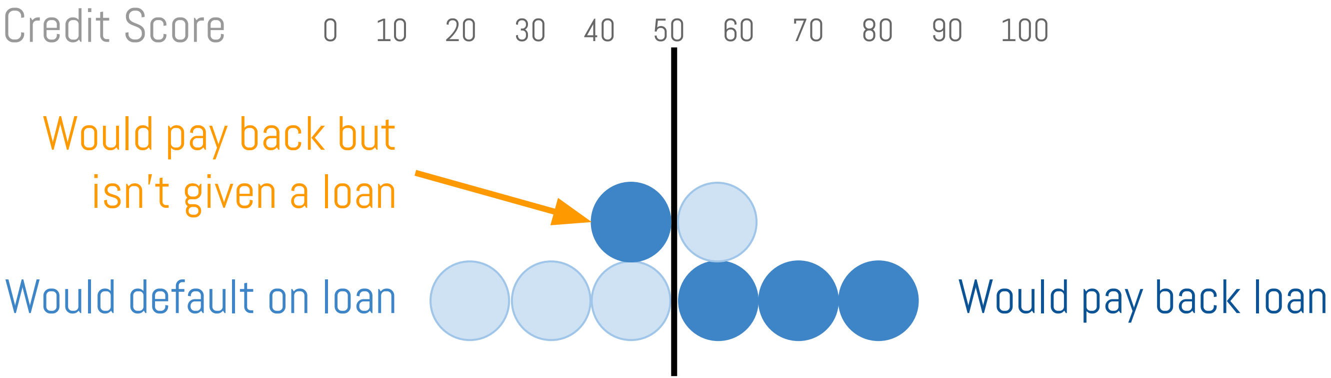
Attacking discrimination in ML mathematically



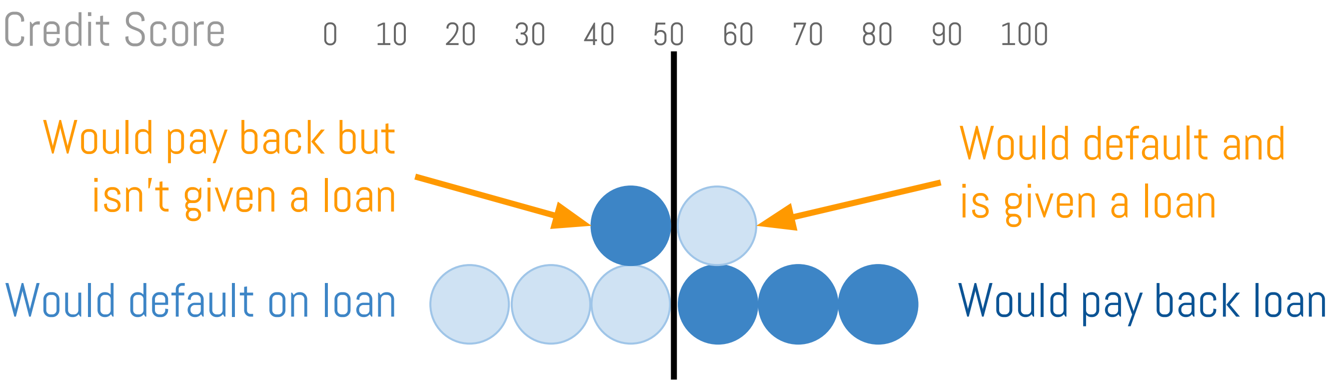
Attacking discrimination in ML mathematically



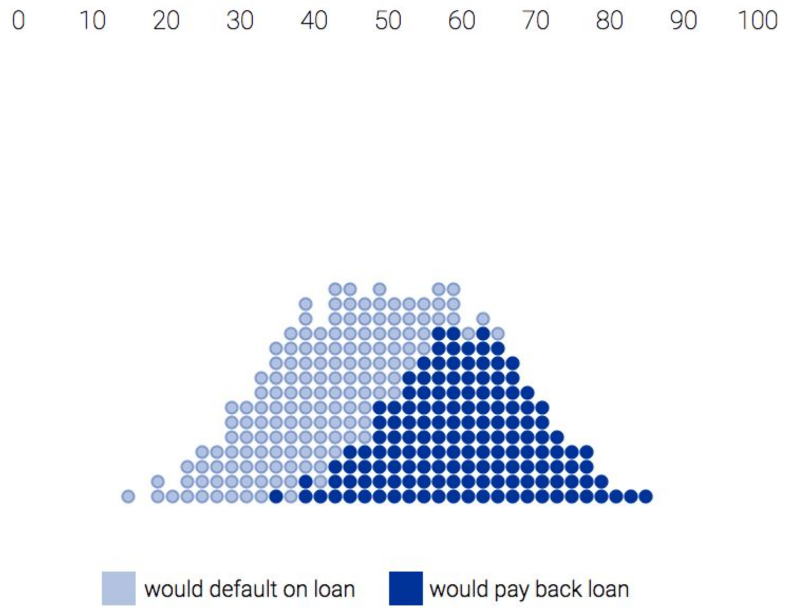
Attacking discrimination in ML mathematically



Attacking discrimination in ML mathematically



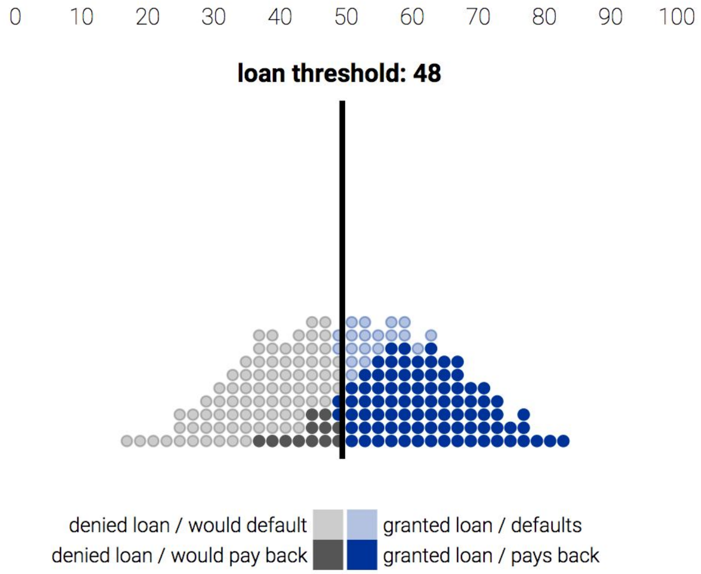
Attacking discrimination in ML mathematically



Attacking discrimination in ML mathematically

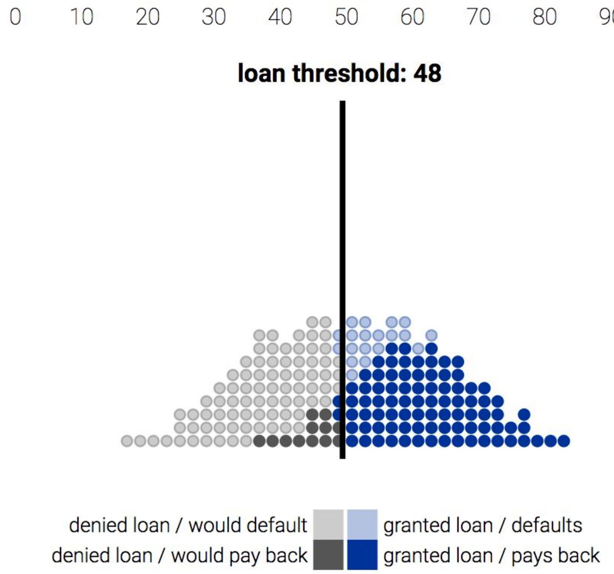
Threshold Decision

Profit: 1.2800



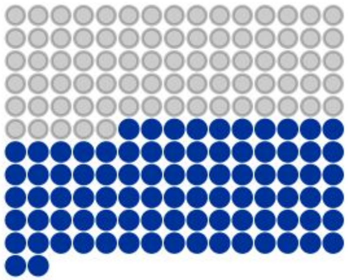
Attacking discrimination in ML mathematically

Threshold Decision



Profit: 1.2800

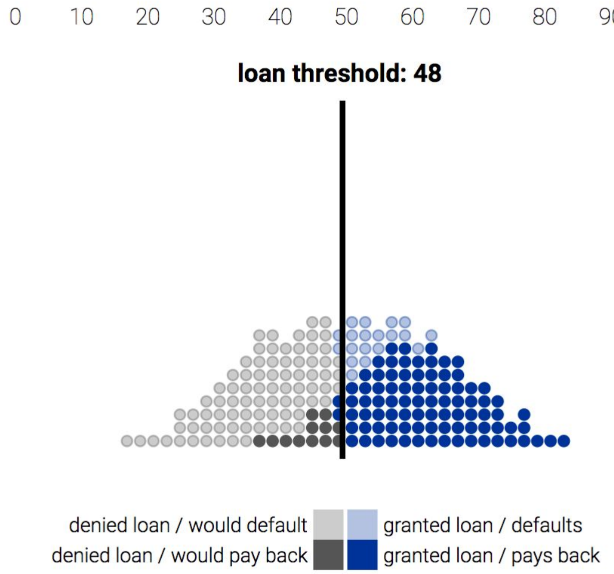
Correct 84%
loans granted to paying applicants and denied to defaulters



Incorrect 16%
loans denied to paying applicants and granted to defaulters

Attacking discrimination in ML mathematically

Threshold Decision



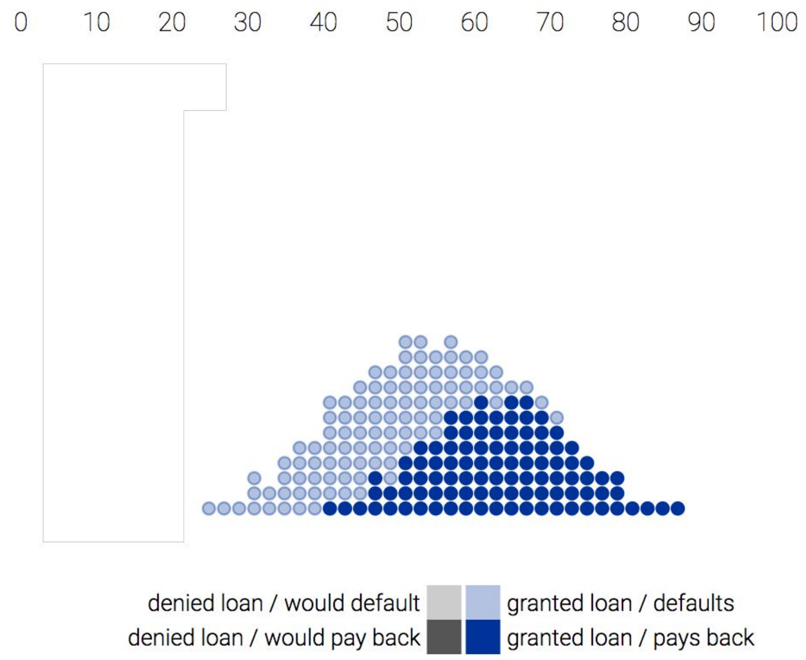
Profit: 1.2800

True Positive Rate 88%
percentage of paying applications getting loans

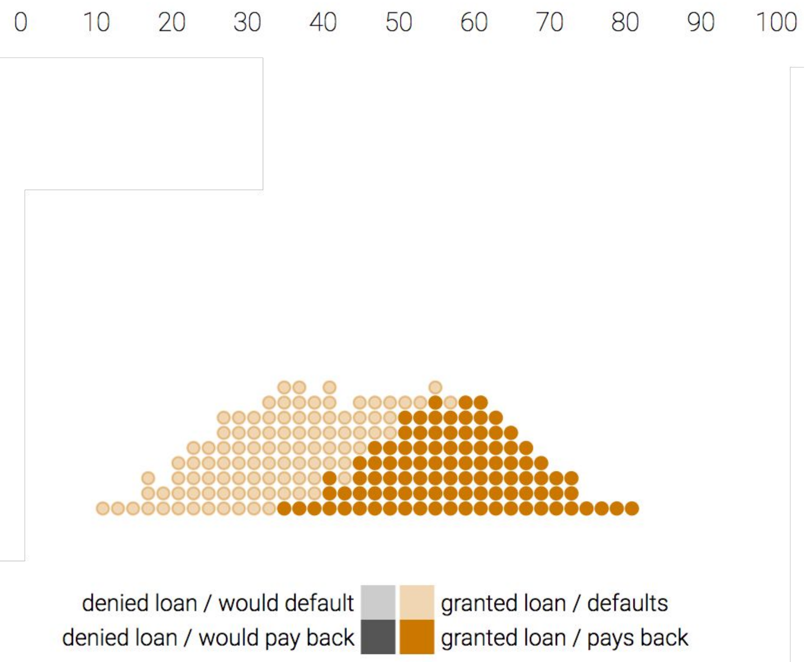
Positive Rate 54%
percentage of all applications getting loans

Multiple groups and multiple distributions

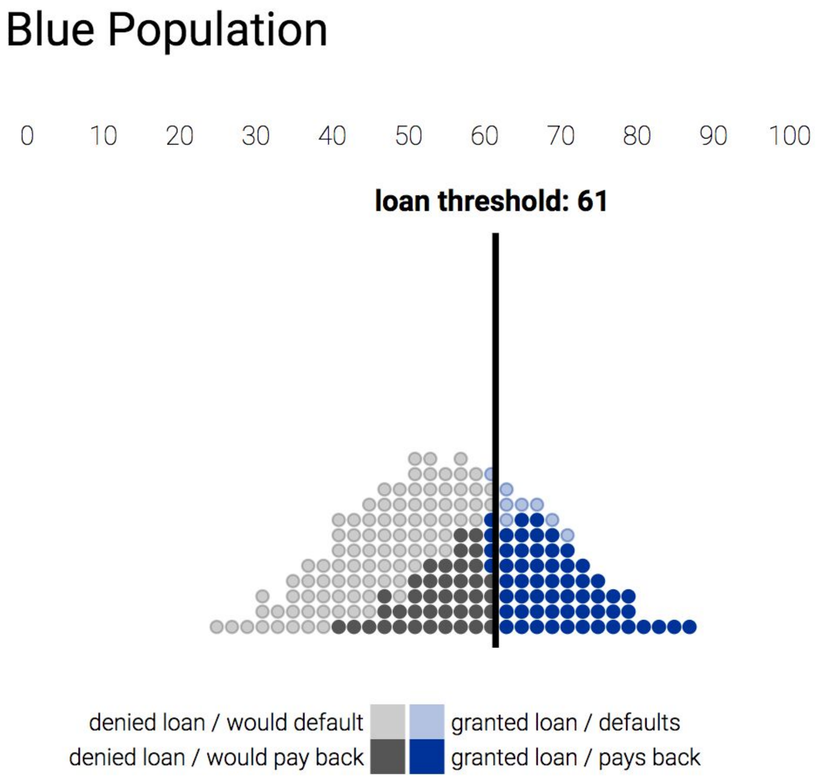
Blue Population

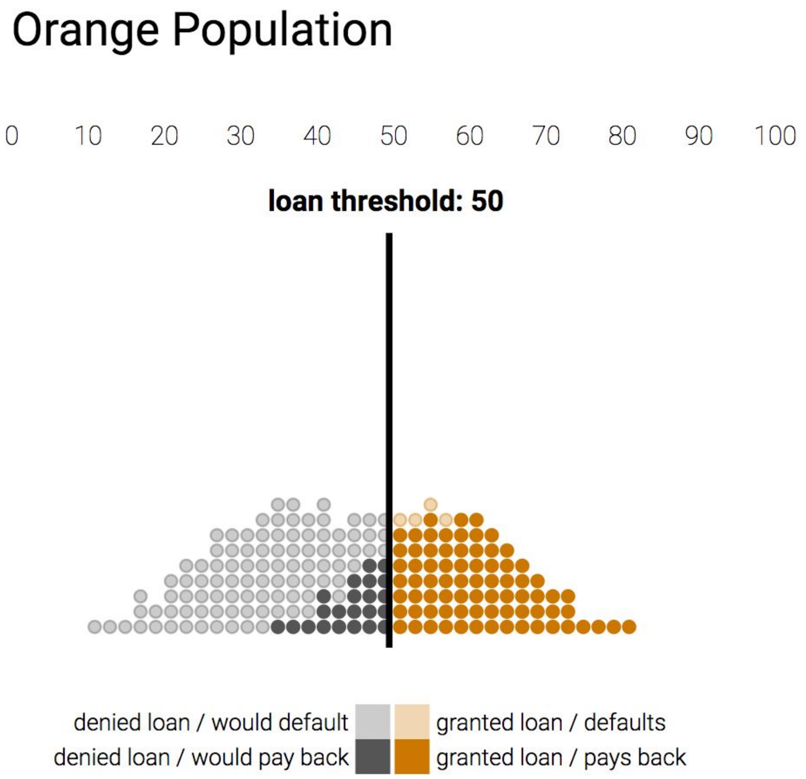


Orange Population



Multiple groups and multiple distributions

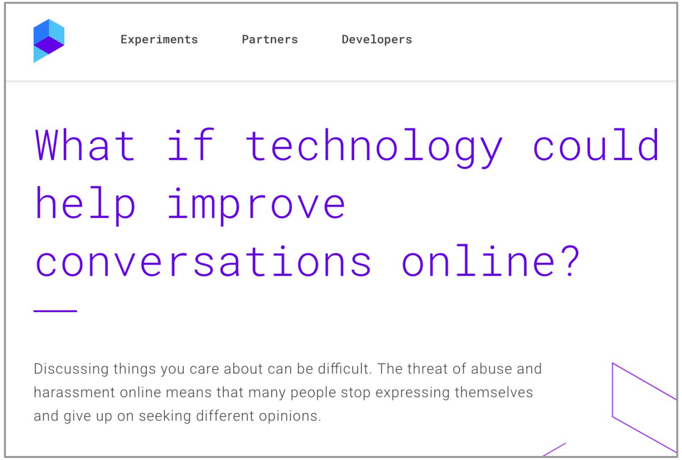




Demo

Case Study

Conversation AI / Perspective API (Jigsaw / CAT / others)



Jigsaw / CAT / others

Conversation AI / Perspective API

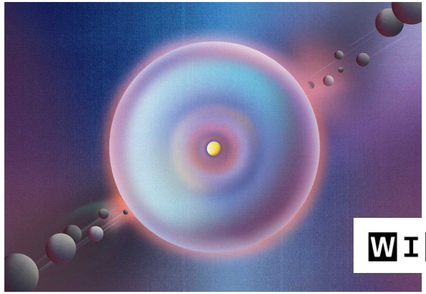
Experiments Partners Developers

What if technology could help improve conversations online?

Discussing things you care about can be diffic
harassment online means that many people st
and give up on seeking different opinions.

ANDY GREENBERG SECURITY 02.23.17 7:00 AM

NOW ANYONE CAN DEPLOY GOOGLE'S TROLL-FIGHTING AI



WIRED

Conversation AI / Perspective API

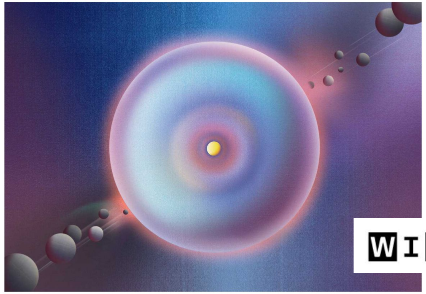
Experiments Partners Developers

What if technology could help improve conversations online?

Discussing things you care about can be diffic
harassment online means that many people st
and give up on seeking different opinions.

ANDY GREENBERG SECURITY 02.23.17 7:00 AM

NOW ANYONE CAN DEPLOY GOOGLE'S TROLL-FIGHTING AI



WIRED

lynn cyrin
@lynncyrin

Follow

smh, I quite enjoyed the pears #actually

61% similar to comments people said were "toxic"

SEEM WRONG?

Black Trans Woman Eats Can of Pears, Really Enjoys It

RETWEETS 7

LIKES 20

7:53 PM - 23 Feb 2017

3 7 20

False "toxic" positives

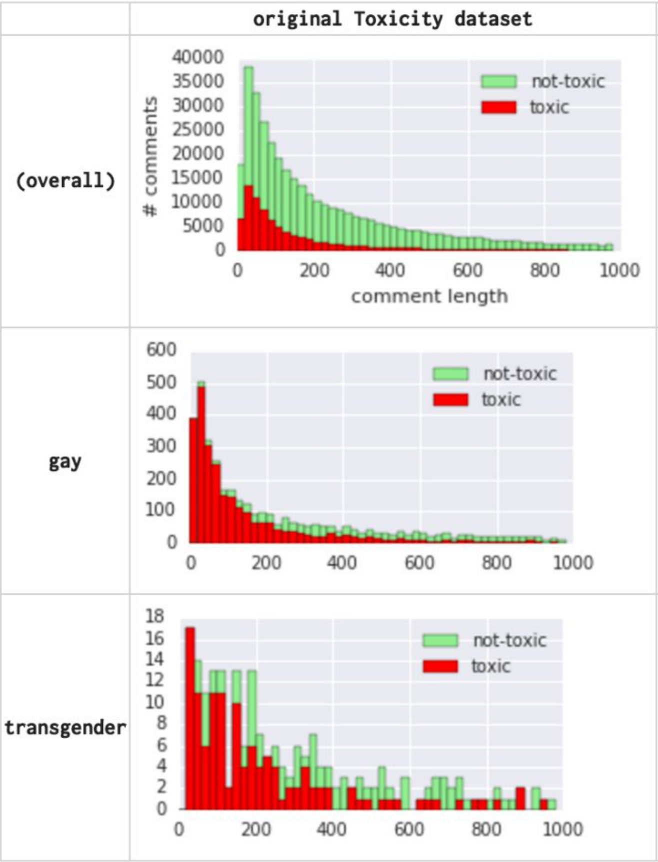
Comment	Toxicity score
The Gay and Lesbian Film Festival starts today.	82%
Being transgender is independent of sexual orientation.	52%
A Muslim is someone who follows or practices Islam.	46%

How did this happen?

term	**fraction labeled toxic**
(overall)	**22%**
"queer"	70%
"gay"	67%
"transgender"	55%
"lesbian"	54%
"homosexual"	51%
"feminist"	39%
"black"	34%
"white"	29%
"heterosexual"	24%

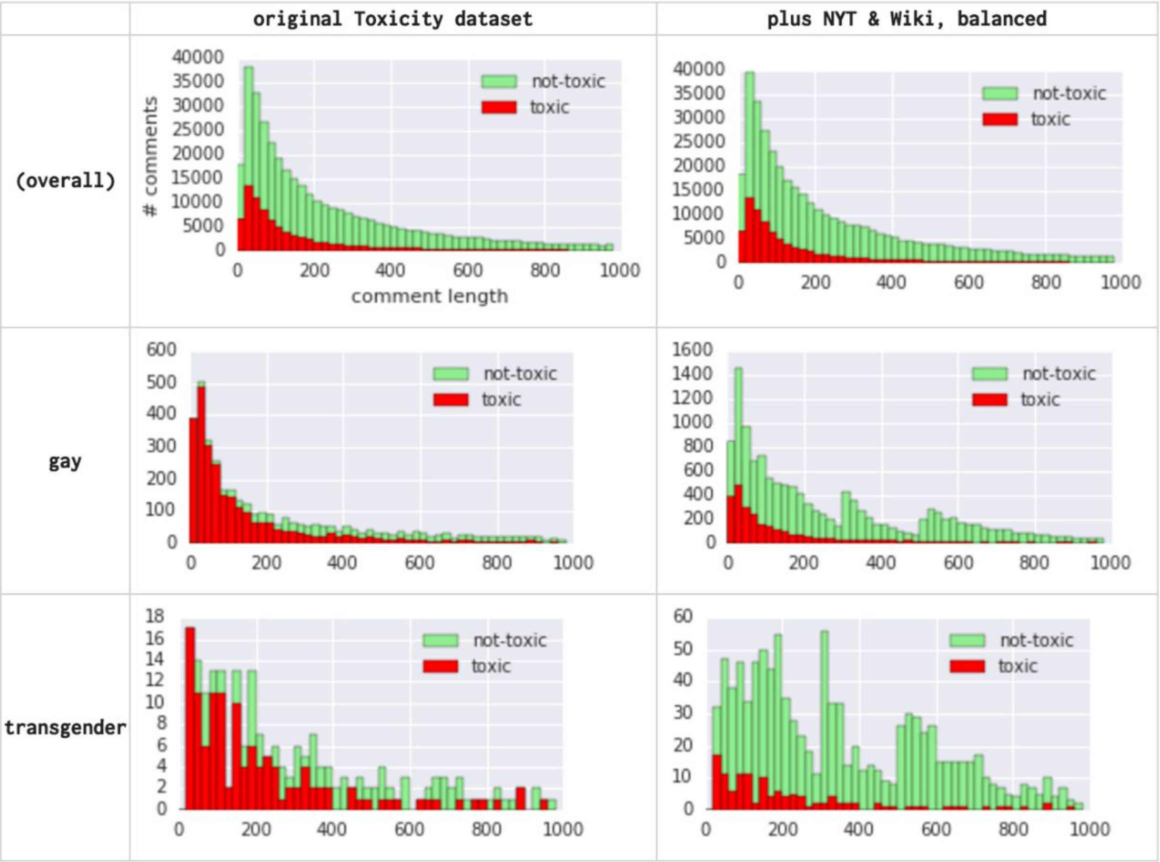
How did this happen?

term	fraction labeled toxic
(overall)	**22%**
"queer"	70%
"gay"	67%
"transgender"	55%
"lesbian"	54%
"homosexual"	51%
"feminist"	39%
"black"	34%
"white"	29%
"heterosexual"	24%



One possible fix

term	fraction labeled toxic
(overall)	**22%**
"queer"	70%
"gay"	67%
"transgender"	55%
"lesbian"	54%
"homosexual"	51%
"feminist"	39%
"black"	34%
"white"	29%
"heterosexual"	24%



False positives - some improvement

Comment	Old	**New**
The Gay and Lesbian Film Festival starts today.	82%	**1%**
Being transgender is independent of sexual orientation.	52%	**5%**
A Muslim is someone who follows or practices Islam.	46%	**13%**

Overall AUC for old and new classifiers was very close.

A common objection...

- Our algorithms are just mirrors of the world. Not our fault if they reflect bias!

A common objection...

- Our algorithms are just mirrors of the world. Not our fault if they reflect bias!

Some replies:

- If the effect is unjust, why shouldn't we fix it?
- Would you apply this same standard to raising a child?

Another objection

- Objection: People are biased and opaque.
- Why should ML systems be any different?
    - True: this won't be easy
    - We have a chance to do better with ML

Another objection

- Objection: People are biased and opaque.
- Why should ML systems be any different?
  - True: this won't be easy
  - We have a chance to do better with ML

[–] **geoffhinton** Google Brain [S] 20 points 2 years ago

I suspect that in the end, understanding how big artificial neural networks work after they have learned will be quite like trying to understand how the brain works but with some very important differences:

1. We know exactly what each neuron computes.
2. We know the learning algorithm they are using.
3. We know exactly how they are connected.
4. We can control the input and observe the behaviour of any subset of the neurons for as long as we like.
5. We can interfere in all sorts of ways without filling in forms.

What can you do?

1. Include diverse perspectives in design and development
2. **Train** ML models on comprehensive data sets
3. **Test** products with diverse users
4. Periodically re-evaluate and be alert to errors

Fairness in Machine Learning

Fernanda Viégas @viegasf
Martin Wattenberg @wattenberg
Google Brain