



Northeastern University

Network Science Institute

Khoury College of Computer and Information Sciences

Just Machine Learning

Tina Eliassi-Rad

tina@eliassi.org

[@tinaeliassi](#)

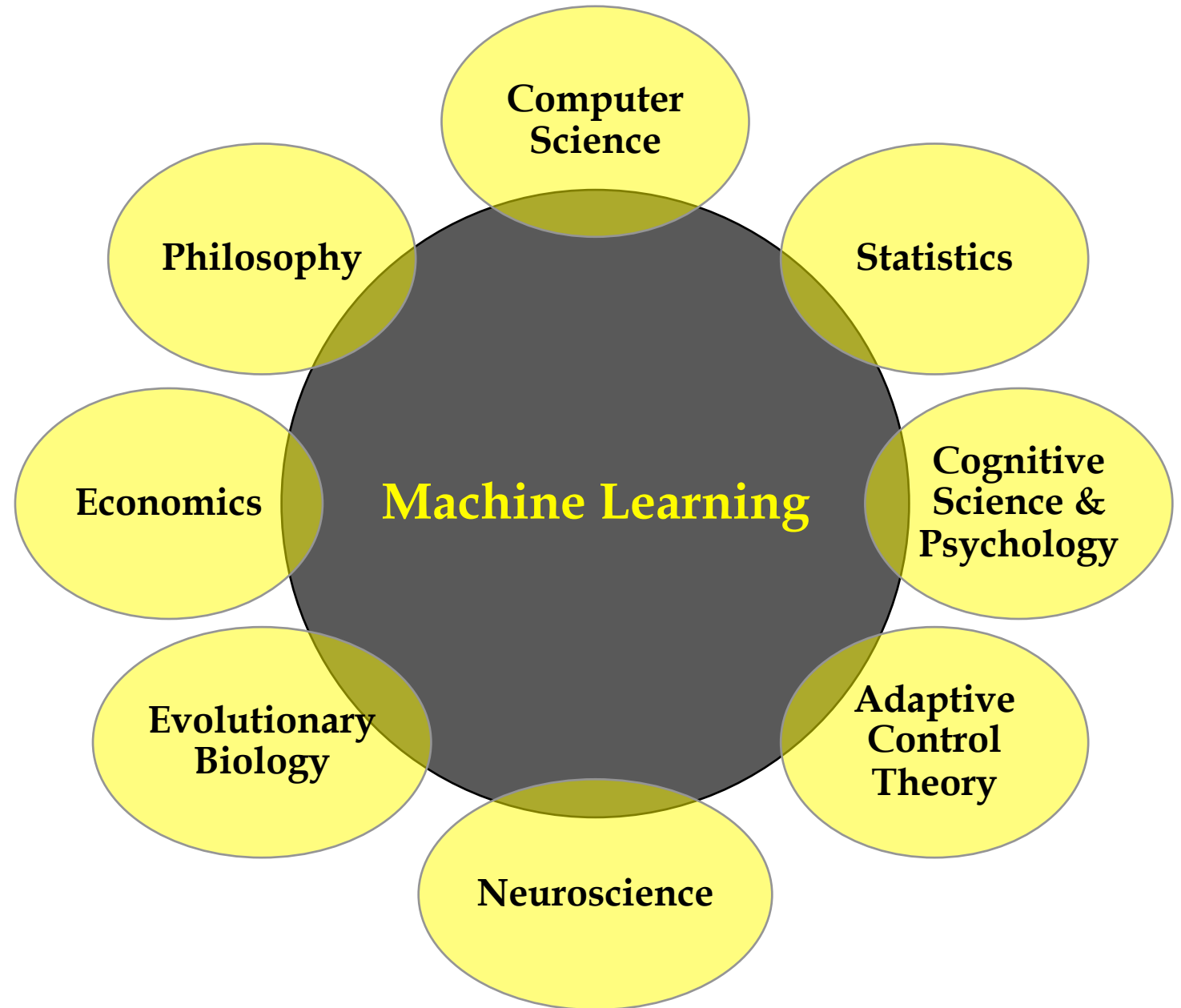
http://eliassi.org/tina_justML_2018.pdf

- Arthur Samuel coined the term machine learning (1959)
 - *Field of study that gives computers the ability to learn without being explicitly programmed*
 - The Samuel Checkers-playing Program



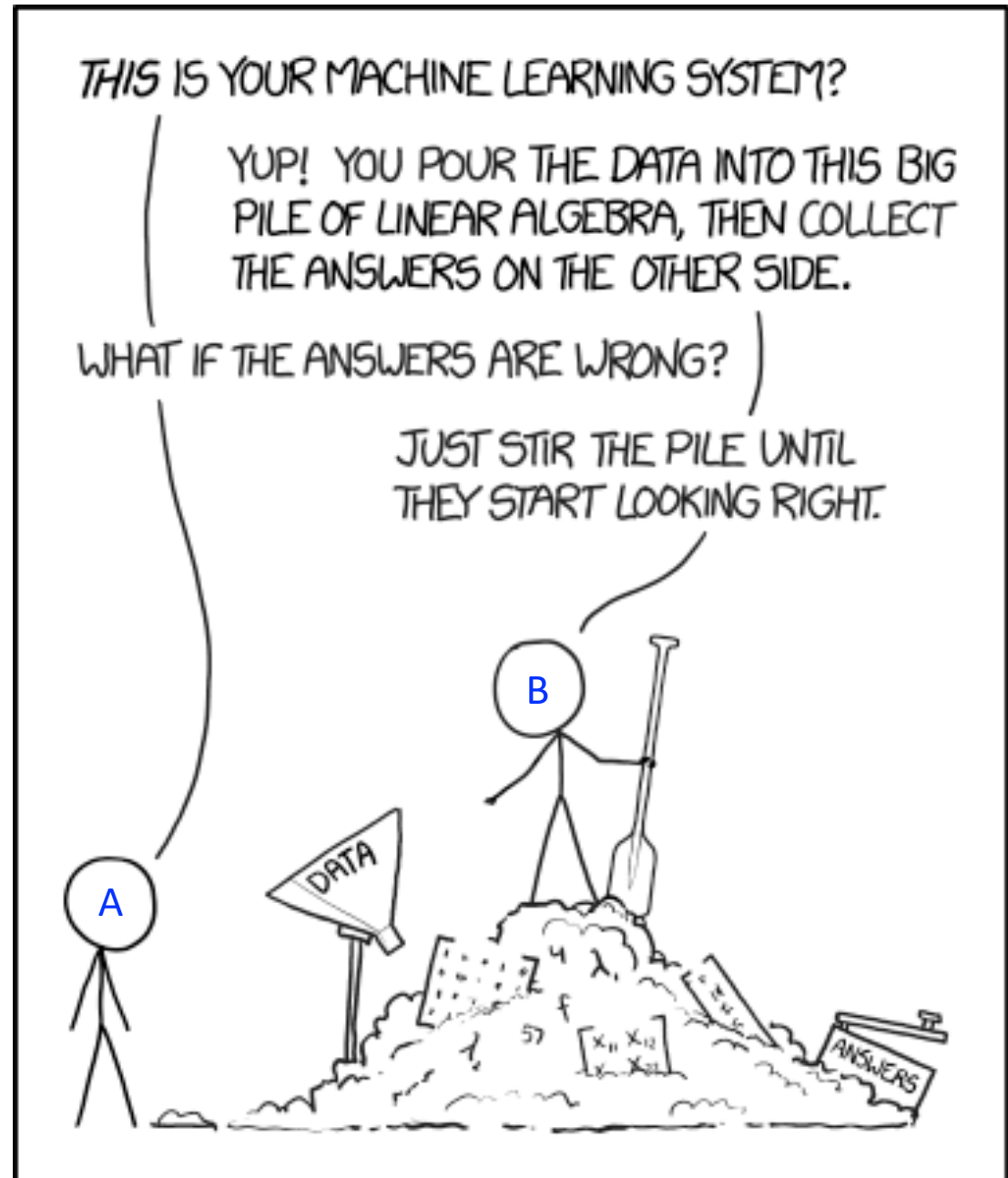
Machine Learning

in Theory



Machine Learning

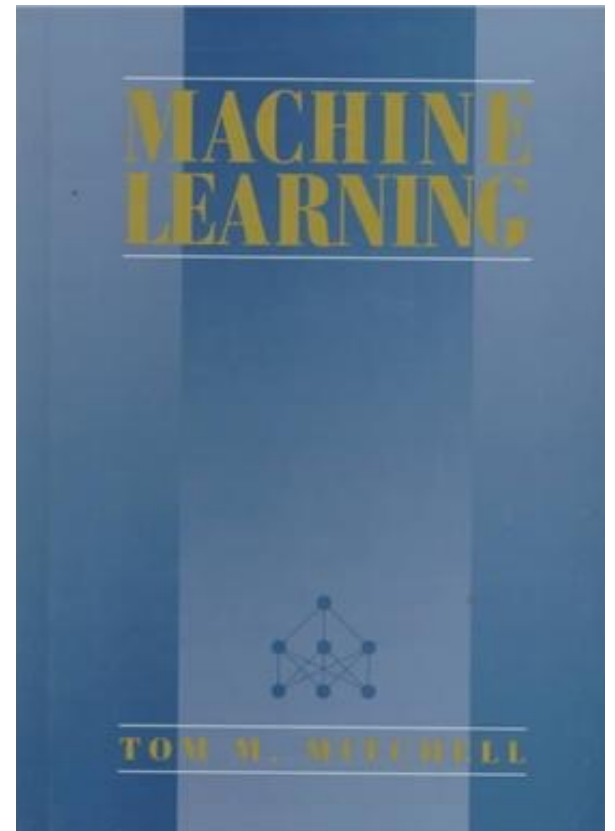
in Practice



The well-posed learning problem

- A computer program is said to **learn** from **experience E** w.r.t. some **task T** and some performance **measure P**, if its performance on T, as measured by P, **improves** with experience E.

-- Tom Mitchell (1997)



Some “success” stories

- IBM Watson defeats the best human competitors in Jeopardy!
- Google AlphaGo Model defeats Euro Go Campaign
- Speech recognition: Amazon Alexa, Apple Siri, Google Go, ...
- Image recognition
- Translation
- Fraud detection
- Self-driving cars
- Recommendation systems: Amazon, NetFlix, ...



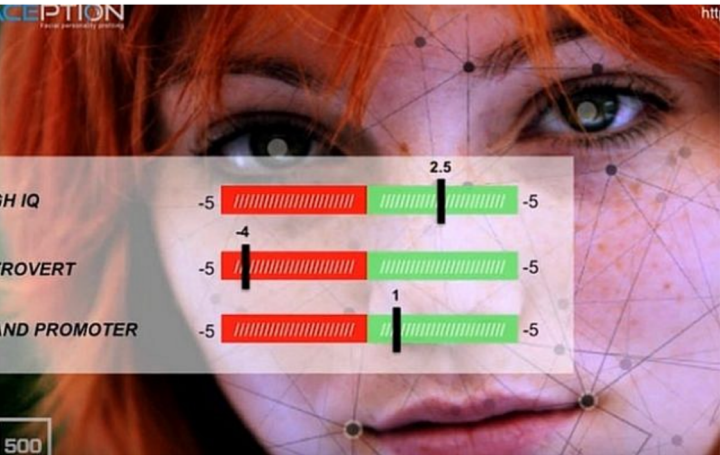
Racist Robots in the News

FACEPCION 'CAN MATCH AN INDIVIDUAL WITH VARIOUS PERSONALITY TRAITS AND TYPES WITH A HIGH LEVEL OF ACCURACY'

New Israeli facial imaging claims to identify terrorists and pedophiles

Tel Aviv start-up Facepion says its face 'classifiers' can spot criminals and even great poker players in a split second, but the experts are not convinced

By **SUE SURKES**
24 May 2016, 10:52 pm | 9



An image taken from a May 2016 presentation by Facepion co-founder Shai Gilboa (screen capture: YouTube)

A Tel-Aviv based start-up company says it has developed a program to identify personality types such as terrorists, pedophiles, white collar offenders and even great poker players from facial analysis that takes just a fraction of a second.

OPINION | **TECH**

'Gaydar' Shows How Creepy Algorithms Can Get

Imagine what an oppressive government could do with it.

By **Cathy O'Neil**
409 September 25, 2017, 6:30 AM EDT



Watch out. Photographer: Jin Lee/Bloomberg

Artificial intelligence keeps getting creepier. In one [controversial](#) study, researchers at Stanford University have [demonstrated](#) that facial recognition technology can identify gay people with surprising precision, although many caveats apply. Imagine how that could be used in the [many](#) countries where homosexuality is a criminal offense.

GOOGLE

Google Photos Mistakenly Labels Black People 'Gorillas'

BY CONOR DOUGHERTY JULY 1, 2015 7:01 PM 41

- Email
- Share
- Tweet
- Save
- More

Google continued to apologize Wednesday for a flaw in Google Photos, which was released to [great fanfare](#) in May, that led the new application to mistakenly label photos of black people as “gorillas.”

The company said it had fixed the problem and was working to figure out exactly how it happened.

“We’re appalled and genuinely sorry that this happened,” said a Google representative in an emailed statement. “We are taking immediate action to prevent this type of result from appearing.”

From self-driving cars to photos, Google, like every technology company, is constantly releasing cutting-edge technologies with the understanding that problems will arise and that it will have to fix them as it goes. The idea is that you never know what problems might arise until you get the technologies in the hands of real-world users.

In the case of the Google Photos app — which uses a combination of advanced computer vision and machine learning techniques to help users collect, search and categorize photos — errors are easy to spot. When the app was unveiled at the company’s annual developer show, executives went through carefully staged demonstrations to show how it can recognize landmarks like the Eiffel Tower and give users the ability to search their photos for people, places or things — even things as specific as a particular dog breed.

Facial Recognition Software Is Bad At Identifying Darker Skinned People

Computing
The Observer

Sun 28 May 2017 08.27 EDT



3,817 498

Interview

'A white mask worked better': why algorithms are not colour blind

By Ian Tucker

When Joy Buolamwini found that a robot recognised her face better when she wore a white mask, she knew a problem needed fixing



▲ Joy Buolamwini gives her TED talk on the bias of algorithms Photograph: TED

Joy Buolamwini is a graduate researcher at the MIT Media Lab and founder of the [Algorithmic Justice League](#) - an organisation that aims to challenge the biases in decision-making software. She grew up in Mississippi, gained a Rhodes scholarship, and she is also a Fulbright fellow, an Astronaut scholar and a Google Anita Borg scholar. Earlier this year [she won a \\$50,000 scholarship](#) funded by the makers of the film [Hidden Figures](#) for her work fighting coded discrimination.

Google's Speech Recognition Has a Gender Bias

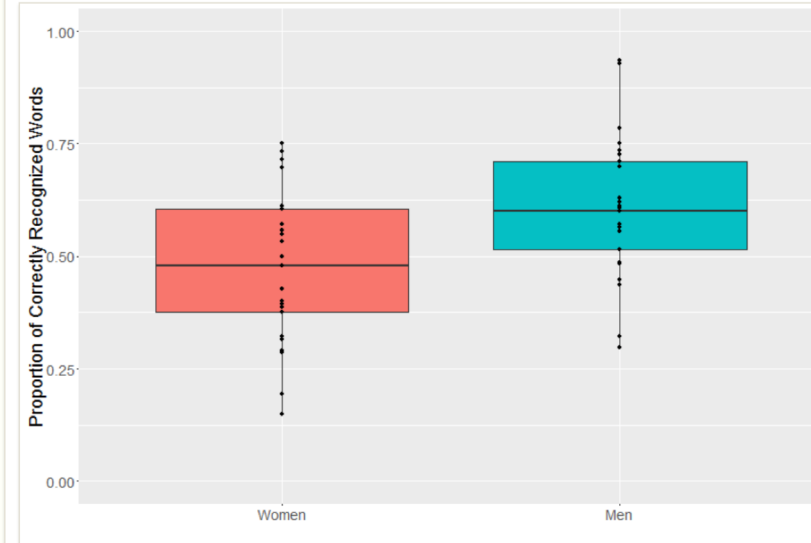
July 12, 2016

GOOGLE'S SPEECH RECOGNITION HAS A GENDER BIAS

Posted by **Rachael Tatman** in **Uncategorized** and tagged with **computational linguistics, gender, linguistics, sociolinguistics, speech recognition, speech signal, speech technology**

[In my last post](#), I looked at how Google's automatic speech recognition worked with different dialects. To get this data, I hand-checked annotations more than 1500 words from fifty different accent tag videos .

Now, because I'm a sociolinguist and I know that it's important to [stratify your samples](#), I made sure I had an equal number of male and female speakers for each dialect. And when I compared performance on male and female talkers, I found something deeply disturbing: YouTube's auto captions consistently performed better on male voices than female voice ($t(47) = -2.7$, $p < 0.01$) . (You can see my data and analysis [here](#).)



On average, for each female speaker less than half (47%) her words were captioned correctly. The average male speaker, on the other hand, was captioned correctly 60% of the time.

It's not that there's a consistent but small effect size, either, 13% is a pretty big effect. The Cohen's d was 0.7 which means, in non-math-speak, that if you pick a random

TayTweets: Microsoft's Twitter Bot

SECTIONS

HOME SEARCH

The New York Times

TECHNOLOGY

Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.

By DANIEL VICTOR MARCH 24, 2016



Tay's Twitter account. The bot was developed by Microsoft's technology and research and Bing teams.

REPORT

COGNITIVE SCIENCE

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan,^{1*} Joanna J. Bryson,^{1,2*} Arvind Narayanan^{1*}

Machine learning is a means to derive artificial intelligence by discovering patterns in existing data. Here, we show that applying machine learning to ordinary human language results in human-like semantic biases. We replicated a spectrum of known biases, as measured by the Implicit Association Test, using a widely used, purely statistical machine-learning model trained on a standard corpus of text from the World Wide Web. Our results indicate that text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names. Our methods hold promise for identifying and addressing sources of bias in culture, including technology.

We show that standard machine learning can acquire stereotyped biases from textual data that reflect everyday human culture. The general idea that text corpora capture semantics, including cultural stereotypes and empirical associations, has long been known in corpus linguistics (1, 2), but our findings add to this knowledge in three ways. First, we used word embeddings (3), a powerful tool to extract associations captured in text corpora; this method substantially amplifies the signal found in raw statistics. Second, our replication of documented human biases may yield tools and insights for studying prejudicial attitudes and behavior in humans. Third, since we performed our experiments on off-the-shelf machine learning components (primarily the Global Vectors for

response times when subjects are asked to pair two concepts they find similar, in contrast to two concepts they find different. We developed our first method, the Word-Embedding Association Test (WEAT), a statistical test analogous to the IAT, and applied it to a widely used semantic representation of words in AI, termed word embeddings. Word embeddings represent each word as a vector in a vector space of about 300 dimensions, based on the textual context in which the word is found. We used the distance between a pair of vectors (more precisely, their cosine similarity score, a measure of correlation) as analogous to reaction time in the IAT. The WEAT compares these vectors for the same set of words used by the IAT. We describe the WEAT in more detail below.

Most closely related to this paper is concurrent

the reaction latencies of four pairings (flowers + pleasant, insects + unpleasant, flowers + unpleasant, and insects + pleasant). Greenwald *et al.* measured effect size in terms of Cohen's *d*, which is the difference between two means of log-transformed latencies in milliseconds, divided by the standard deviation. Conventional small, medium, and large values of *d* are 0.2, 0.5, and 0.8, respectively. With 32 participants, the IAT comparing flowers and insects resulted in an effect size of 1.35 ($P < 10^{-8}$). Applying our method, we observed the same expected association with an effect size of 1.50 ($P < 10^{-7}$). Similarly, we replicated Greenwald *et al.*'s finding (5) that musical instruments are significantly more pleasant than weapons (see Table 1).

Notice that the word embeddings "know" these properties of flowers, insects, musical instruments, and weapons with no direct experience of the world and no representation of semantics other than the implicit metrics of words' co-occurrence statistics with other nearby words.

We then used the same technique to demonstrate that machine learning absorbs stereotyped biases as easily as any other. Greenwald *et al.* (5) found extreme effects of race as indicated simply by name. A bundle of names associated with being European American was found to be significantly more easily associated with pleasant than unpleasant terms, compared with a bundle of African-American names.

In replicating this result, we were forced to slightly alter the stimuli because some of the original African-American names did not occur in the corpus with sufficient frequency to be included. We therefore also deleted the same number of European-American names, chosen at random, to balance the number of elements in the sets of two concepts. Omissions and deletions are indicated in our list of keywords (see the supplementary materials).

In another widely publicized study, Bertrand and Mullainathan (7) sent nearly 5000 identical résumés in response to 1300 job advertisements, varying only the names of the candidates. They

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

Tolga Bolukbasi¹, Kai-Wei Chang², James Zou², Venkatesh Saligrama^{1,2}, Adam Kalai²

¹Boston University, 8 Saint Mary's Street, Boston, MA

²Microsoft 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 the words *receptionist* and *female*, while maintaining desired associations such as between the words *queen* and *female*. 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.

ProPublica's Study of NorthPointe Software

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

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of \$80.

Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was picked up for shoplifting \$86.35 worth of tools from a nearby Home Depot store.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)



Can an Algorithm Tell When Kids Are in Danger?

Child protective agencies are haunted when they fail to save kids. Pittsburgh officials believe a new data analysis program is helping them make better judgment calls.

By DAN HURLEY JAN. 2, 2018



For people of color, banks are shutting the door to homeownership

By Aaron Glantz and Emmanuel Martinez / February 15, 2018

Fifty years after the federal Fair Housing Act banned racial discrimination in lending, African Americans and Latinos continue to be routinely denied conventional mortgage loans at rates far higher than their white counterparts.

COMMUNICATIONS OF THE ACM

HOME | CURRENT ISSUE | **NEWS** | BLOGS | OPINION | RESEARCH | PRACTICE | CAREERS | ARCHIVE | VIDEOS

Home / News / Amazon Scraps Secret AI Recruiting Tool That Showed... / Full Text

ACM TECHNEWS

Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women

By Reuters
October 12, 2018
[Comments](#)

VIEW AS: SHARE:



Amazon had discontinued development of an artificial intelligence recruiting tool after determining it was biased against women.

Credit: V3.co.uk

recruiters to evaluate candidates.”

From Reuters
[View Full Article](#)

Amazon discontinued an artificial intelligence recruiting tool its machine learning specialists developed to automate the hiring process because they determined it was biased against women.

Starting in 2014, a group of Amazon researchers created 500 computer models focused on specific job functions and locations, training each to recognize about 50,000 terms that showed up on past Amazon job candidates' resumes.

However, because most resumes submitted to Amazon had come from men, the models tended to favor candidates who described themselves using verbs more commonly found on male engineers' resumes, such as "executed" and "captured."

In addition, the program penalized resumes that included the word "women's" and downgraded graduates of two all-women's colleges.

Although Amazon declined to comment on the technology's issues, the company said the tool was "never used by Amazon

SIGN IN for Full Access

User Name

Password

[» Forgot Password?](#)

[» Create an ACM Web Account](#)

SIGN IN

MORE NEWS & OPINIONS

An Electronic Rescue Dog

ETH Zurich

Want Facebook to Censor Speech? Be Careful What You Wish For

Wired

Moving Computing Education Past Argument from Authority: Stuart Reges and Women Who Code

Mark Guzdial

ACM RESOURCES

The Sun Solaris Operating System

Courses

Bias in computer systems

(Friedman & Nissenbaum, 1996)

- Identified three sources of bias
 1. Preexisting bias from social institutions, practices, and attitudes
 2. Technical bias from technical constraints or considerations
 3. Emergent bias from context of use
- “We conclude by suggesting that freedom from bias should be counted among the select set of criteria—including reliability, accuracy, and efficiency—according to which the quality of systems in use in society should be judged.”

Lots of activity recently

- “Autonomous Systems” by David Danks and Alex John London (IJCAI 2017)

- <http://bit.ly/2zrdbnX>

- UC Berkeley Course on Fairness in Machine Learning

- <https://fairmlclass.github.io>

- Fairness, accountability, and transparency

- FatML Conferences: <https://www.fatml.org>

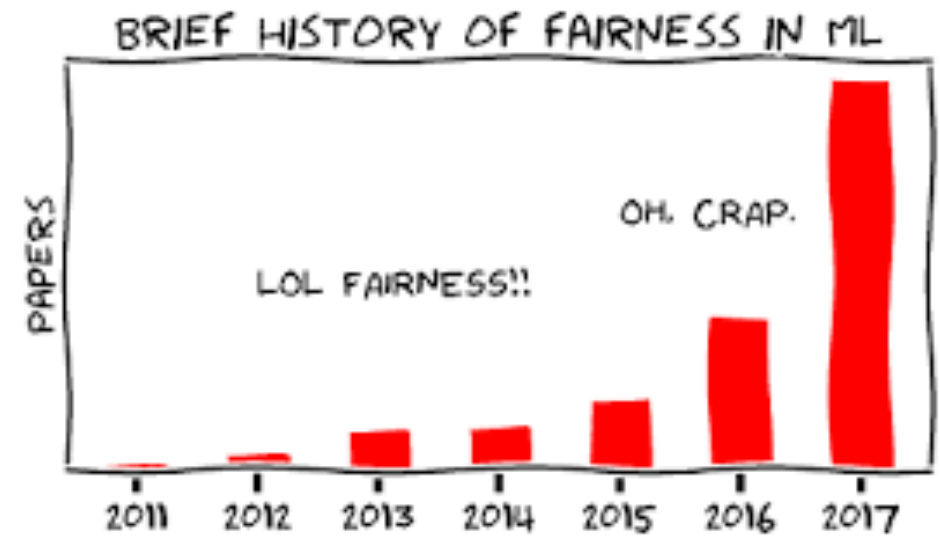


Figure from <https://fairmlclass.github.io>

Can we make ML algorithms “fair”?

- Should we change
 - the **task T**,
 - the **experience E**, or
 - the **performance measure P**?

How do computer scientists define fairness?

- Probabilistically
- Lots of parity (i.e., “fairness”) definitions
 - Decisions should be in some sense probabilistically independent of sensitive features values (such as gender, race)
- There are many possible senses

Lots of parity definitions

(Probabilistic definitions of different kinds of fairness)

- Demographic parity
- Accuracy parity
- True positive parity
- False positive parity
- Positive rate parity
- Precision parity
- Positive predictive value parity
- Negative predictive value parity
- Predictive value parity
- ...

See <https://fairmlclass.github.io> for definitions.

Lots of parity definitions

(Probabilistic definitions of different kinds of fairness)

- Demographic parity: The output of the classifier does not depend on the sensitive attribute (e.g., gender, race, education level, etc).
- Accuracy parity
- True positive parity
- False positive parity
- Positive rate parity
- Precision parity
- Predictive value parity
- ...

See <https://fairmlclass.github.io> for definitions.

Lots of parity definitions

(Probabilistic definitions of different kinds of fairness)

- Demographic parity
- Accuracy parity: The accuracy of the classifier does not depend on the sensitive attribute (e.g., gender, race, education level, etc).
- True positive parity
- False positive parity
- Positive rate parity
- Precision parity
- Predictive value parity
- ...

See <https://fairmlclass.github.io> for definitions.

Confusion matrix

	Predicted: NO	Predicted: YES
Actual: NO	TN	FP
Actual: YES	FN	TP

- **Accuracy:** How often is the classifier correct? $(TP+TN)/total$
- **Misclassification** (a.k.a. Error) **Rate:** How often is it wrong? $(FP+FN)/total$
- **True Positive Rate** (TPR, a.k.a. Sensitivity or Recall): When it's actually yes, how often does it predict yes? $TP/actual\ yes$
- **False Positive Rate** (FPR) : When it's actually no, how often does it predict yes? $FP/actual\ no$
- **Specificity** ($1 - FPR$) : When it's actually no, how often does it predict no? $TN/actual\ no$
- **Precision** (a.k.a. **Positive Predictive Value**): When it predicts yes, how often is it correct? $TP/predicted\ yes$
- **Negative Predictive Value:** When it predicts no, how often is it correct? $TN/predicted\ no$
- **Prevalence:** How often does the yes condition actually occur in our sample? $actual\ yes/total$

Confusion matrix

	Predicted: NO	Predicted: YES
Actual: NO	TN	FP
Actual: YES	FN	TP

- **Accuracy**: How often is the classifier correct? $(TP+TN)/total$
- **Misclassification** (a.k.a. Error) **Rate**: How often is it wrong? $(FP+FN)/total$
- **True Positive Rate** (TPR, a.k.a. Sensitivity or Recall): When it's actually yes, how often does it predict yes? $TP/actual\ yes$
- **False Positive Rate** (FPR) : When it's actually no, how often does it predict yes? $FP/actual\ no$
- **Specificity** ($1 - FPR$) : When it's actually no, how often does it predict no? $TN/actual\ no$
- **Precision** (a.k.a. **Positive Predictive Value**): When it predicts yes, how often is it correct? $TP/predicted\ yes$
- **Negative Predictive Value**: When it predicts no, how often is it correct? $TN/predicted\ no$
- **Prevalence**: How often does the yes condition actually occur in our sample? $actual\ yes/total$

What about the task T?

- The most popular task has been to **asses risk estimates**.
- Examples:
 - Jack's risk of defaulting on a loan is 8; Jill's is 2.
 - Ed's risk of recidivism is 9; Peter's is 1.
 - ...

Impossibility results ☹️

- Kleinberg, Mullainathan, Raghavan (2016)
 - <https://arxiv.org/abs/1609.05807>
- Chouldechova (2016)
 - <https://arxiv.org/abs/1610.07524>
- You can't have your cake and eat it too

Some definitions

- **X** contains features of an individual (e.g., medical records)
 - **X** often incorporates all sorts of measurement biases
- **A** is a sensitive attribute (e.g., race, gender, ...)
 - **A** is often unknown, ill-defined, misreported, or inferred
- **Y** is the true outcome (a.k.a. the ground truth; e.g., whether patient has cancer)
- **C** is the machine learning algorithm that uses **X** and **A** to predict the value of **Y** (e.g., predict whether the patient has cancer)

Some simplifying assumptions

- The sensitive attribute **A** divides the population into two groups **a** (e.g., whites) and **b** (e.g., non-whites)
- The machine learning algorithm **C** outputs 0 (e.g., predicts not cancer) or 1 (e.g., predicts cancer)
- The true outcome **Y** is 0 (e.g., not cancer) or 1 (e.g., cancer)

Impossibility results

- Kleinberg, Mullainathan, Raghavan (2016), Chouldechova (2016)
- Assume differing base rates – i.e., $\Pr_a(Y=1) \neq \Pr_b(Y=1)$ – **and** an imperfect machine learning algorithm ($C \neq Y$), then you can not simultaneously achieve
 - a) Precision parity: $\Pr_a(Y=1|C=1) = \Pr_b(Y=1|C=1)$.
 - b) True positive parity: $\Pr_a(C=1|Y=1) = \Pr_b(C=1|Y=1)$
 - c) False positive parity: $\Pr_a(C=1|Y=0) = \Pr_b(C=1|Y=0)$

Impossibility results

- Kleinberg, Mullainathan, Raghavan (2016), Chouldechova (2016)
- Assume differing base rates – i.e., $\Pr_a(Y=1) \neq \Pr_b(Y=1)$ – **and** an imperfect machine learning algorithm ($C \neq Y$), then you can not simultaneously achieve
 - a) Precision parity: $\Pr_a(Y=1|C=1) = \Pr_b(Y=1|C=1)$.
 - b) True positive parity: $\Pr_a(C=1|Y=1) = \Pr_b(C=1|Y=1)$
 - c) False positive parity: $\Pr_a(C=1|Y=0) = \Pr_b(C=1|Y=0)$

These impossibility results also hold for overlapping groups. -- Eliassi-Rad & Fitelson (2018)

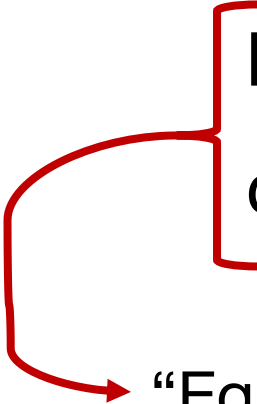
Impossibility results

- Kleinberg, Mullainathan, Raghavan (2016), Chouldechova (2016)
- Assume differing base rates – i.e., $\Pr_a(Y=1) \neq \Pr_b(Y=1)$ – **and** an imperfect machine learning algorithm ($C \neq Y$), then you can not simultaneously achieve

a) Precision parity: $\Pr_a(Y=1|C=1) = \Pr_b(Y=1|C=1)$

b) True positive parity: $\Pr_a(C=1|Y=1) = \Pr_b(C=1|Y=1)$

c) False positive parity: $\Pr_a(C=1|Y=0) = \Pr_b(C=1|Y=0)$



“Equalized odds” -- Hardt, Price, Srebro (2016)

Variations on the impossibility result

Condition	Chouldechova (2016)
Precision parity: $\Pr_a(Y = 1 \mid C = 1) = \Pr_b(Y = 1 \mid C = 1)$	YES
True positive parity: $\Pr_a(C = 1 \mid Y = 1) = \Pr_b(C = 1 \mid Y = 1)$	YES
False positive parity: $\Pr_a(C = 1 \mid Y = 0) = \Pr_b(C = 1 \mid Y = 0)$	YES
Unequal base rates: $\Pr_a(Y = 1) \neq \Pr_b(Y = 1)$	YES
Mutual exclusivity between groups a and b	YES
Statistical parity: $\Pr_a(C = 1) = \Pr_b(C = 1)$	YES
Imperfect classifier: $\Pr_a(C = 1 \mid Y = 0) \neq 0$ and $\Pr_b(C = 1 \mid Y = 0) \neq 0$ and $\Pr_a(C = 1 \mid Y = 1) \neq 1$ and $\Pr_b(C = 1 \mid Y = 1) \neq 1$	
Non-identical variables: $\Pr_a(Y = 1 \mid C = 1) \neq 0$ or $\Pr_b(Y = 1 \mid C = 1) \neq 0$	

Variations on the impossibility result

Condition	Chouldechova (2016)	Kleinberg et al. (2016)
Precision parity: $\Pr_a(Y = 1 \mid C = 1) = \Pr_b(Y = 1 \mid C = 1)$	YES	YES
True positive parity: $\Pr_a(C = 1 \mid Y = 1) = \Pr_b(C = 1 \mid Y = 1)$	YES	YES
False positive parity: $\Pr_a(C = 1 \mid Y = 0) = \Pr_b(C = 1 \mid Y = 0)$	YES	YES
Unequal base rates: $\Pr_a(Y = 1) \neq \Pr_b(Y = 1)$	YES	YES
Mutual exclusivity between groups a and b	YES	YES
Statistical parity: $\Pr_a(C = 1) = \Pr_b(C = 1)$	YES	
Imperfect classifier: $\Pr_a(C = 1 \mid Y = 0) \neq 0$ and $\Pr_b(C = 1 \mid Y = 0) \neq 0$ and $\Pr_a(C = 1 \mid Y = 1) \neq 1$ and $\Pr_b(C = 1 \mid Y = 1) \neq 1$		YES
Non-identical variables: $\Pr_a(Y = 1 \mid C = 1) \neq 0$ or $\Pr_b(Y = 1 \mid C = 1) \neq 0$		YES

Variations on the impossibility result

Condition	Chouldechova (2016)	Kleinberg et al. (2016)	Eliassi-Rad & Fitzelson (2018)
Precision parity: $\Pr_a(Y = 1 \mid C = 1) = \Pr_b(Y = 1 \mid C = 1)$	YES	YES	YES
True positive parity: $\Pr_a(C = 1 \mid Y = 1) = \Pr_b(C = 1 \mid Y = 1)$	YES	YES	YES
False positive parity: $\Pr_a(C = 1 \mid Y = 0) = \Pr_b(C = 1 \mid Y = 0)$	YES	YES	YES
Unequal base rates: $\Pr_a(Y = 1) \neq \Pr_b(Y = 1)$	YES	YES	YES
Mutual exclusivity between groups a and b	YES	YES	
Statistical parity: $\Pr_a(C = 1) = \Pr_b(C = 1)$	YES		
Imperfect classifier: $\Pr_a(C = 1 \mid Y = 0) \neq 0$ and $\Pr_b(C = 1 \mid Y = 0) \neq 0$ and $\Pr_a(C = 1 \mid Y = 1) \neq 1$ and $\Pr_b(C = 1 \mid Y = 1) \neq 1$		YES	YES
Non-identical variables: $\Pr_a(Y = 1 \mid C = 1) \neq 0$ or $\Pr_b(Y = 1 \mid C = 1) \neq 0$		YES	YES

Impossibility results

“Suppose we want to **determine the risk that a person is a carrier for a disease Y**, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test’s probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups.”

-- Kleinberg, Mullainathan, Raghavan (2016)

Impossibility results

“Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that **a higher fraction of women than men are carriers**. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test’s probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups.”

-- Kleinberg, Mullainathan, Raghavan (2016)

Impossibility results

“Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test’s probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups.”

-- Kleinberg, Mullainathan, Raghavan (2016)

Impossibility results

“Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test’s **probability estimates are systematically skewed upward or downward for at least one gender**; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups.”

-- Kleinberg, Mullainathan, Raghavan (2016)

Impossibility results



“Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test’s **probability estimates are systematically skewed upward or downward for at least one gender**; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups.”

-- Kleinberg, Mullainathan, Raghavan (2016)

Impossibility results

“Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test’s probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a **higher average risk estimate to healthy people (non-carriers) in one gender than the other**; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups.”

-- Kleinberg, Mullainathan, Raghavan (2016)

~~FALSE POSITIVE
PARITY~~

Impossibility results

“Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test’s probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a **higher average risk estimate to healthy people (non-carriers) in one gender than the other**; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups.”

-- Kleinberg, Mullainathan, Raghavan (2016)

Impossibility results

“Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test’s probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a **higher average risk estimate to carriers of the disease in one gender than the other**. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups.”

-- Kleinberg, Mullainathan, Raghavan (2016)

~~TRUE POSITIVE
PARITY~~

Impossibility results

“Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test’s probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a **higher average risk estimate to carriers of the disease in one gender than the other**. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups.”

-- Kleinberg, Mullainathan, Raghavan (2016)

Impossibility results

“Suppose we want to determine the risk that a person is a carrier for a disease Y, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of Y, at least one of the following undesirable properties must hold: (a) the test’s probability estimates are systematically skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that **this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups.**” -- Kleinberg, Mullainathan, Raghavan (2016)

ProPublica and NorthPointe

- ProPublica's main charge was that **black defendants experienced higher false positive rate**
- Northpointe's main defense was that **their risk assessment scores satisfy precision parity**: $\Pr_a(Y=1|C=1) = \Pr_b(Y=1|C=1)$
- Due to the impossibility results, Northpointe's algorithm **cannot satisfy “equalized odds”**
 - Disproportionately high false positive rate for blacks
 - Disproportionately high false negative rate for whites

Fallout from the impossibility theorems

- Get rid of one of the parities
- Put bounds on the parities
- Deborah Hellman (University of Virginia Law School)
 - Precision parity captures “what you ought to believe”
 - True positive and false positive parities capture “what you ought to do”
 - The algorithm ought not be thinking about the **right-making properties** when deliberating in many cases
 - ➔ If you are going to drop a parity, drop precision parity

a) Precision parity: $\Pr_a(Y=1|C=1) = \Pr_b(Y=1|C=1)$

b) True positive parity: $\Pr_a(C=1|Y=1) = \Pr_b(C=1|Y=1)$

c) False positive parity: $\Pr_a(C=1|Y=0) = \Pr_b(C=1|Y=0)$

Group vs. individual fairness

- “Fairness through awareness” by Dwork, Hardt, Pitassi, Reingold, Zemel (2012)
- “People who are similar w.r.t. a specific (classification) task should be treated similarly.”
- Does not get around the impossibility results
- Assuming you have equal base rates, treating everyone equally is a good move

Solutions considered from the machine learning side so far (1/2)

- Preprocessing or “massaging” the data to make it less biased
- Learning fair representations: encode data while obfuscating sensitive attributes
- Penalize the algorithm to encourage it to learn fairly
 - During training (e.g., through regularization or constraints) or as a post-processing step
- Allow the sensitive attributes to be used during training, but do not make them available to the model during inference time

Solutions considered from the machine learning side so far (2/2)

- Causal modeling
 - “Everything else being equal” cases
 - Findings depend strongly on model and assumptions
- Excellent tutorial at NIPS 2017 by Solon Barocas and Moritz Hardt
 - Slides: <http://mrtz.org/nips17/>
 - Video: <https://vimeo.com/248490141>



Directed graphical model with extra structure

Structural equation: $V \leftarrow f_V(U, W, N_V)$

<http://mrtz.org/nips17/#/84>

Humans vs. algorithms

- Julia Dressel and Hany Farid: The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), 17 Jan 2018.
 - <http://advances.sciencemag.org/content/4/1/eaao5580>
 - *“Algorithms for predicting recidivism are commonly used to assess a criminal defendant’s likelihood of committing a crime. These predictions are used in pretrial, parole, and sentencing decisions. Proponents of these systems argue that big data and advanced machine learning make these analyses more accurate and less biased than humans. We show, however, that the widely used commercial risk assessment software COMPAS is no more accurate or fair than predictions made by people with little or no criminal justice expertise. In addition, despite COMPAS’s collection of 137 features, the same accuracy can be achieved with a simple linear classifier with only two features.”*

Publicly available software

- University of Chicago's **Aequitas Toolkit**
 - <https://dsapp.uchicago.edu/aequitas/>
- Google's **What-If-Tool**
 - <https://pair-code.github.io/what-if-tool/>
 - Commercial at: <https://bit.ly/2xJYdqv>
- IBM Research's **AI Fairness 360 Interactive Experience**
 - <http://aif360.mybluemix.net/>
 - Code: <https://github.com/ibm/aif360>

Solutions considered from the policy side

- Regulations
- The EU has General Data Protections Regulation (GDPR) data laws which went into effect on May 25, 2018
- These laws grant users a “right to explanation” of any automated decision-making as applied to them
- Wikipedia entry: <http://bit.ly/1ImrNJz>

From fairness to justice: just machine learning in an unjust world?

- Racist/sexist humans – e.g., biased judges
- Unjust algorithms are already in use – e.g., three-strikes laws, mandatory minimum sentencing
 - They don't take enough empirical data into account
 - Machine learning can help here, but what are the suitable task, performance measure, and experience

The Just Machine Learning Project

- How should we represent implicit vs. explicit bias?
 - Is explicit bias represented as rules?
 - Is implicit bias a set of examples from which to draw conclusions?
- Gabby Johnson (UCLA): “The Structure of Bias”.
<https://tinyurl.com/y7k2te92>

The Just Machine Learning Project

- How should we represent implicit vs. explicit bias?
- How should we capture intent in machine learning?
 - The US anti-discrimination laws incentivize the framing of cases in terms of intent
 - Josh Simons (Harvard): “The Politics of Machine Learning: Discrimination, Fairness, and Equality”.

The Just Machine Learning Project

- How should we represent implicit vs. explicit bias?
- How should we capture intent in machine learning?
- Are purely data-driven approaches ideal in all scenarios?
 - Data are the results of cases meeting the laws/guidelines and subject matter experts.
 - **Laws** can be thought of as **constraints** and **policies** as **implementations** of those constraints.
 - Ideally, we'd like ML to change the biased policies and laws.

The Just Machine Learning Project

- What should the objective function be?
 - Sometimes there are multiple objective functions that are at odds with each other – e.g., child protective services
- Do we care about harm or do we care about benefit?
- Do we care about treatment or do we care about impact?
- Can we create a procedure that helps formulate objective functions?

Moving away from
assessing risk estimates

A different representation of the task

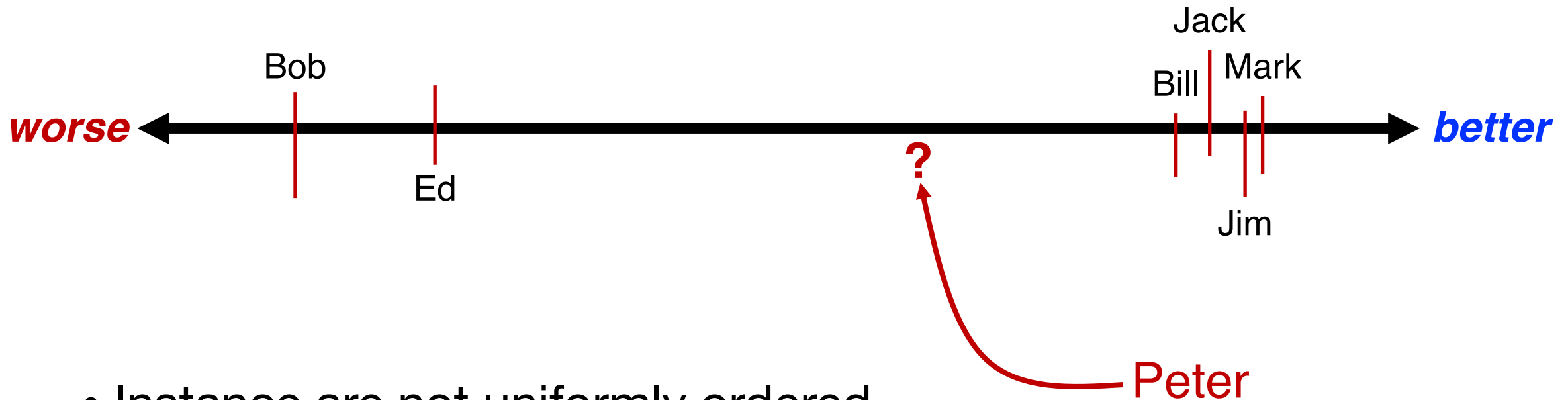


Xindi Wang



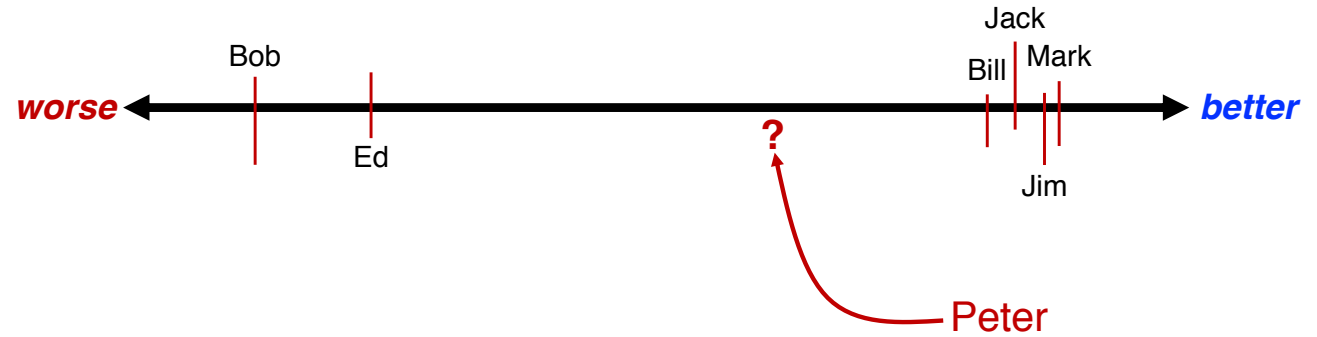
Onur Varol

- Given a sequence of ordered instances, where should we place a new instance?



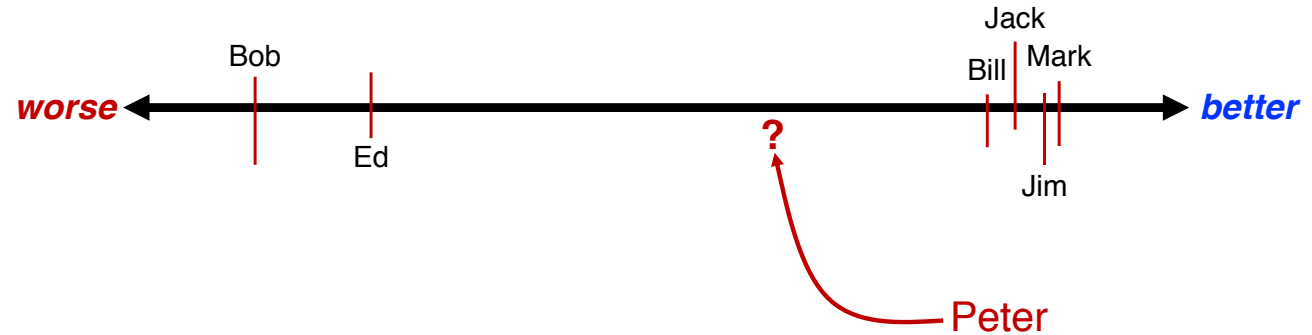
- Instances are not uniformly ordered

Learning to Place



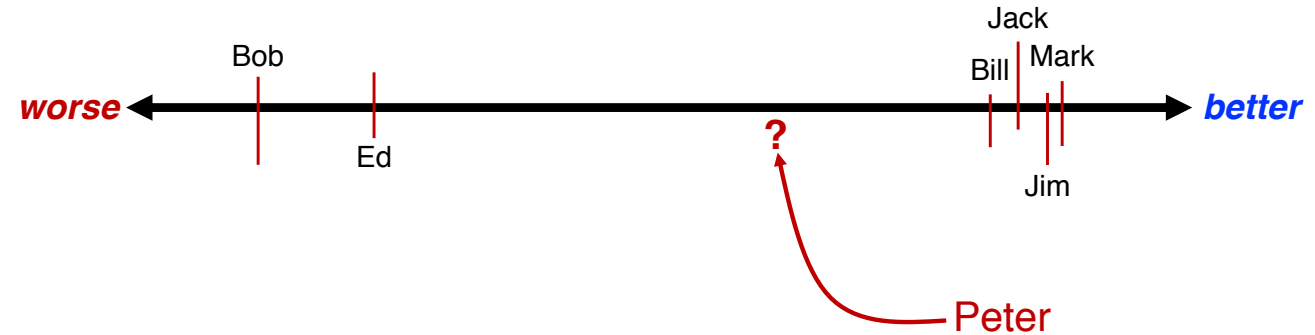
- A two-step approach:
 1. Learn pairwise preferences
 2. Generate a partial ordering from the pairwise preferences

Learning to Place



- A two-step approach:
 1. Learn pairwise preferences
 - Build a classifier by giving it a set of training pairs: $\langle \langle i, j \rangle, b \rangle$
 - i and j are instances in the training set
 - b is the binary response variable:
 - $b = 0$ if $f(i) < f(j)$
 - $b = 1$ if $f(i) > f(j)$
 - When a new instance arrives, the classifier is asked to predict b for each instance the train set $\langle \langle \text{train instance}, \text{new instance} \rangle, b \rangle$
 2. Generate a partial ordering from the pairwise preferences

Learning to Place



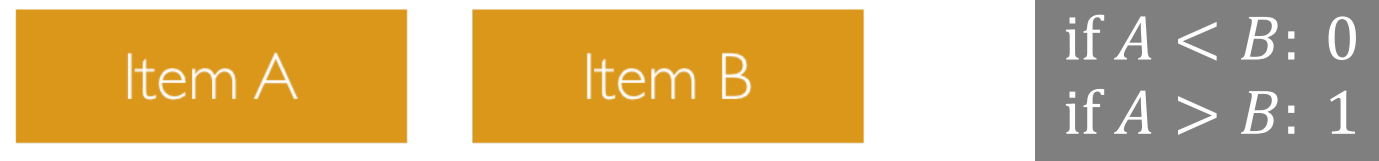
- A two-step approach:
 1. Learn pairwise preferences
 2. Generate a partial ordering from the pairwise preferences
 - Rank the instances. This can be based on
 - Voting
 - Hamilton path of a weighted tournament graph (WTG)
 - Learning to order things [Cohen, Schapire, Singer, JAIR 1999]
 - FAS-PIVOT [Ailon, Charikar, and Newman, STOC 2005]
 - WTG-Wave [Wang, et al., CompleNet 2018]
 - SpringRank [De Bacco, Larremore, and Moore, arXiv 2017]
 - ...

Why not regression?

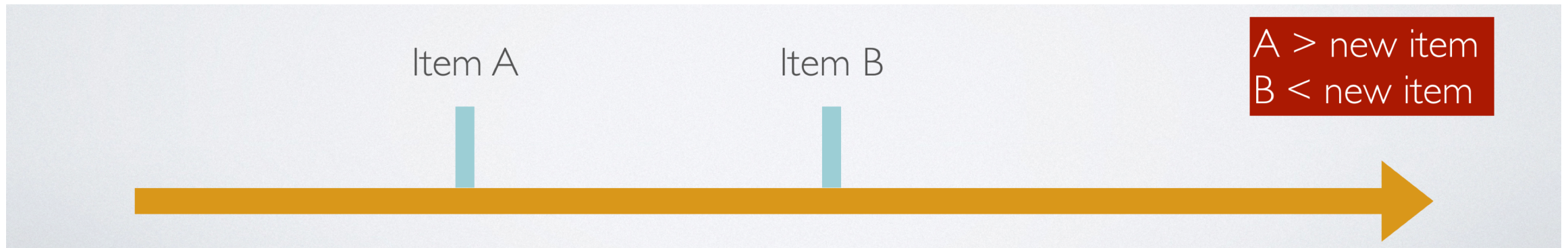
- Property of interest (i.e., target variable) is heavy-tailed
- This leads to a big class imbalance problem
- Methods, like linear regression, heavily under-predict the “big and rare” instances
 - Possible solution: Median-of-means [Hsu & Sabato, JMLR 2016]
 - Has free parameters that need to be tuned
 - No human in the loop

Learning to Place

- Phase I: Build a classifier for pairwise preferences (based on a given training set)



- Phase II: Find places for new items



Learning to Place

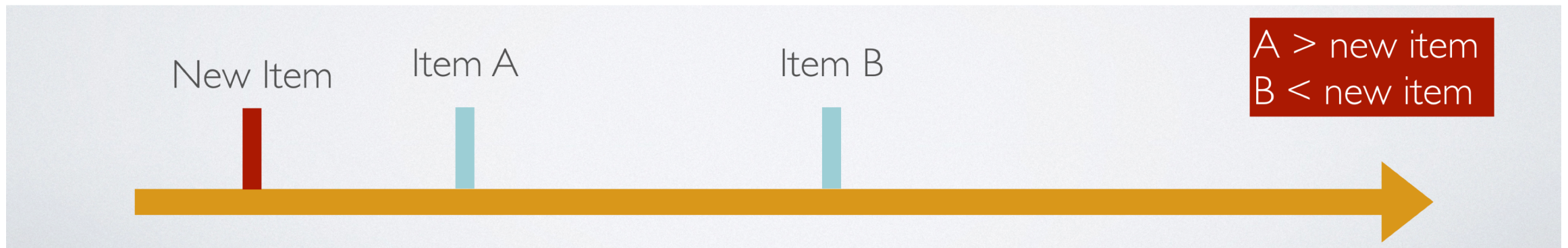
- Phase I: Build a classifier for pairwise preferences (based on a given training set)

Item A

Item B

if $A < B$: 0
if $A > B$: 1

- Phase II: Find places for new items



Learning to Place

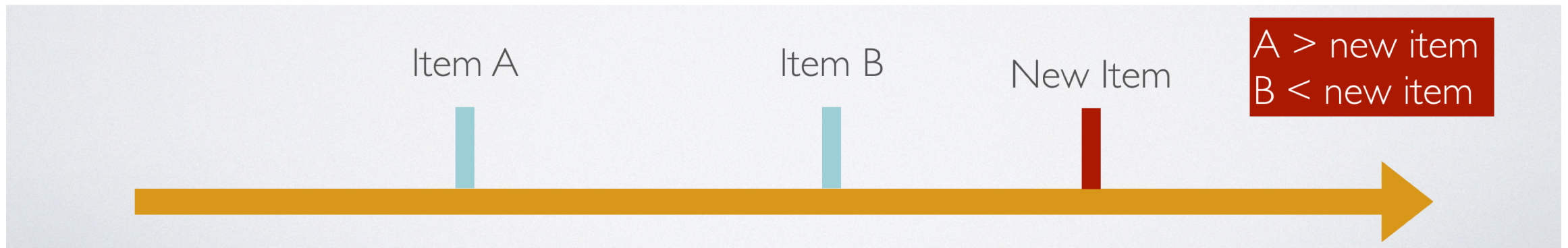
- Phase I: Build a classifier for pairwise preferences (based on a given training set)

Item A

Item B

if $A < B$: 0
if $A > B$: 1

- Phase II: Find places for new items



Learning to Place

- Phase I: Build a classifier for pairwise preferences (based on a given training set)

Item A

Item B

if $A < B$: 0
if $A > B$: 1

- Phase II: Find places for new items

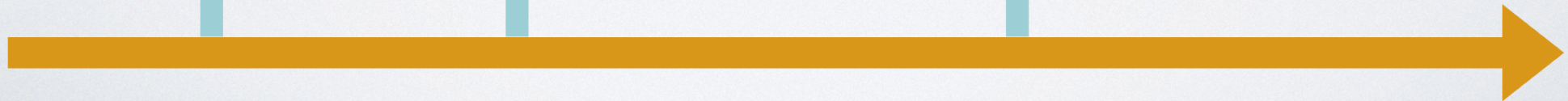
Item 4

?

Item 2

Item 1

Item 3



Learning to Place

- Phase I: Build a classifier for pairwise preferences (based on a given training set)

Item A

Item B

if $A < B$: 0
if $A > B$: 1

- Phase II: Find places for new items

Item 2

Item 1

Item 3

Item 4

?

Item 4 > Item 1
Item 4 > Item 2
Item 4 < Item 3

Learning to Place

- Phase I: Build a classifier for pairwise preferences (based on a given training set)

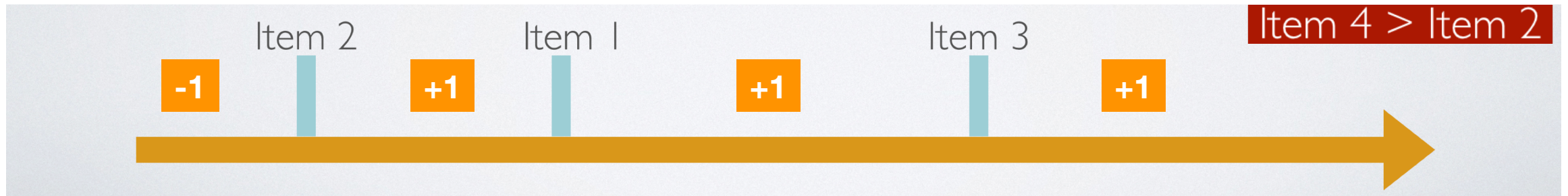
Item A

Item B

if $A < B$: 0
if $A > B$: 1

Item 4
?|

- Phase II: Find places for new items



Learning to Place

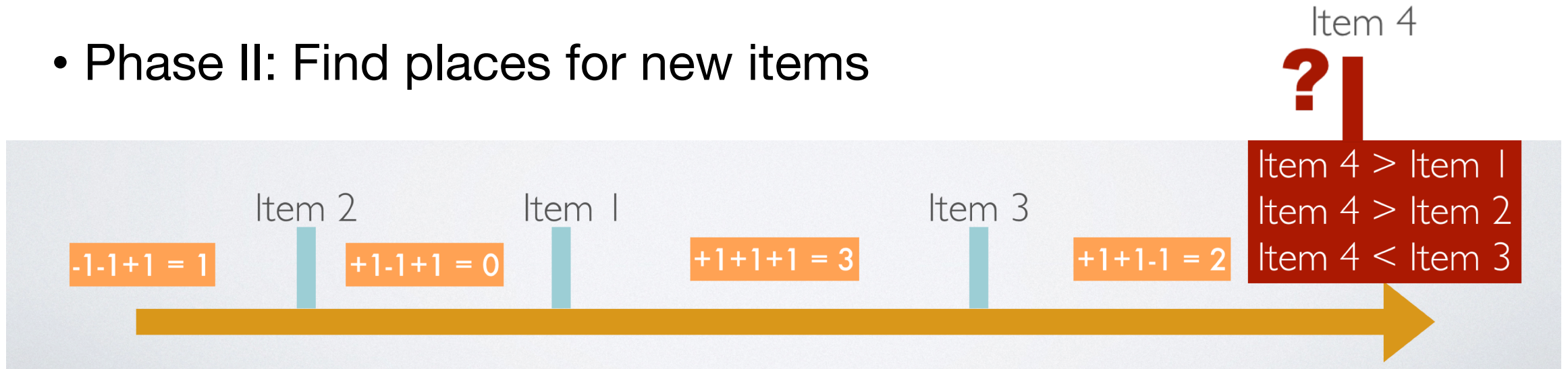
- Phase I: Build a classifier for pairwise preferences (based on a given training set)

Item A

Item B

if $A < B$: 0
if $A > B$: 1

- Phase II: Find places for new items



Learning to Place

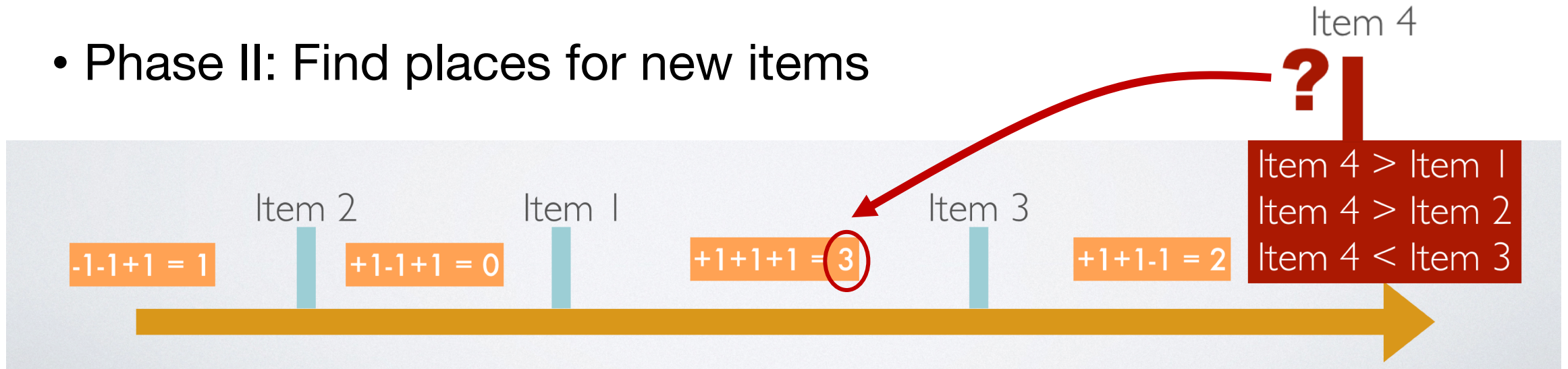
- Phase I: Build a classifier for pairwise preferences (based on a given training set)

Item A

Item B

if $A < B$: 0
if $A > B$: 1

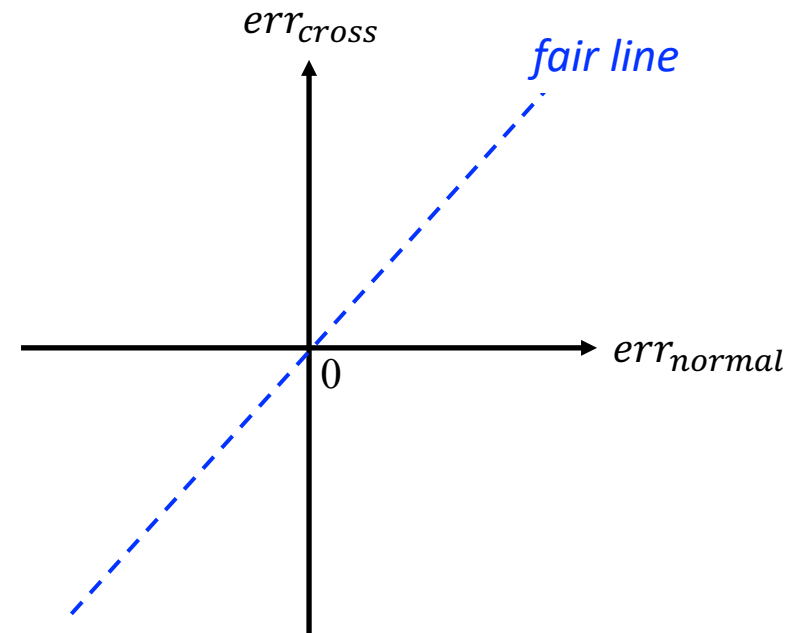
- Phase II: Find places for new items



What is fair?

What is fair?

- **Normal** test: the value(s) for the sensitive feature(s) is the same among the train and test sets
- **Cross** test: the value(s) of the sensitive feature(s) is **not** the same among the train and test sets
- **Error** = predicted value – actual value
 - If error > 0 , then overpredicting
 - If error = 0, then fair
 - If error < 0 , then underpredicting



What does Learning to Place
find on the COMPAS data?

COMPAS Data

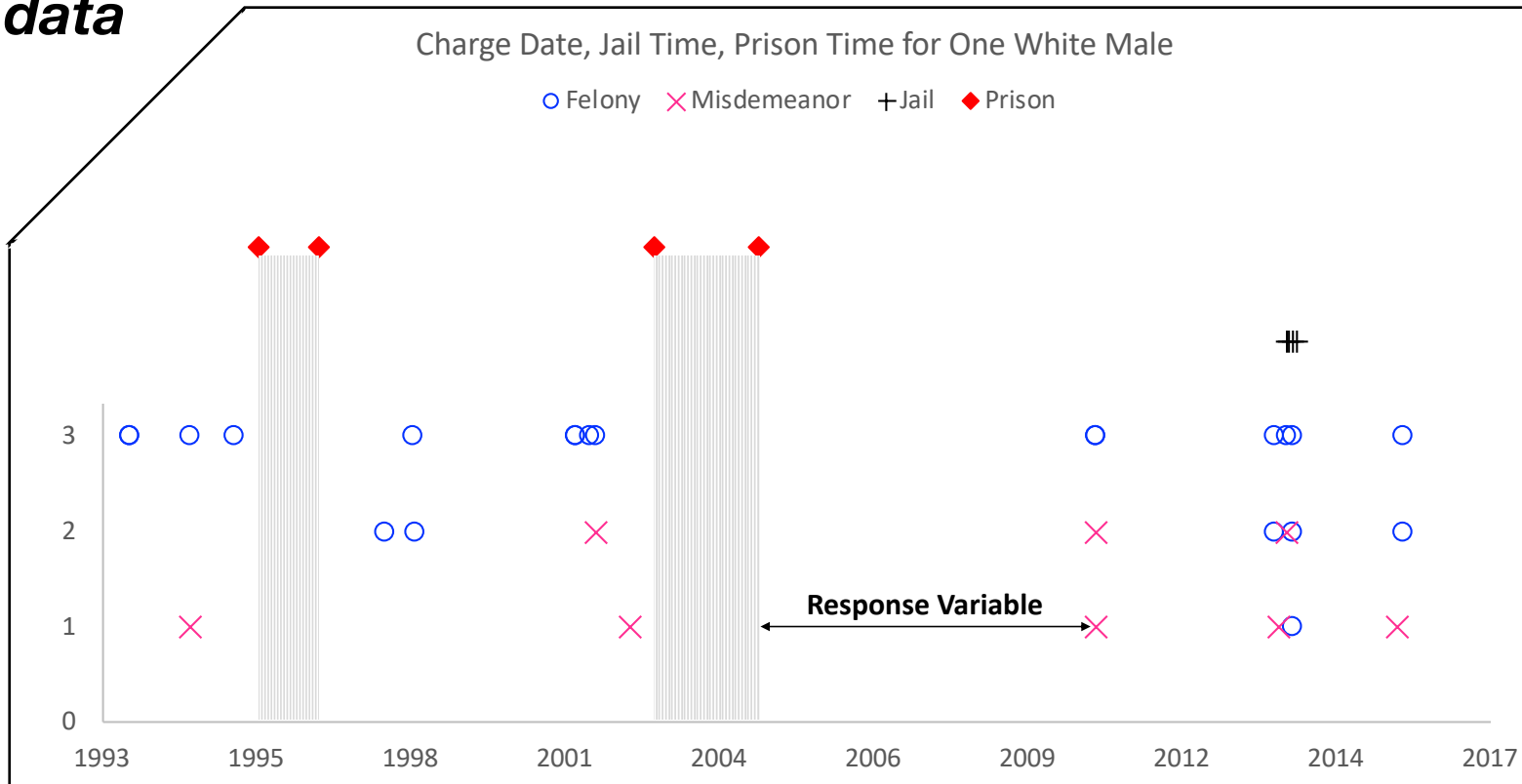
- The data is from the ProPublica story "Machine Bias",¹ where they analyze the COMPAS Recidivism Algorithm.
- There is a main dataset, **compas.db**, containing several tables including:
 - Case arrest: arrest records of criminals
 - Charge: charge records of criminals
 - Compas: COMPAS screening of criminals
 - Jail history: jail history of criminals
 - People: basic demographic data of criminals, together with some processed crime data²
 - Prison history: prison history of criminals
 - Summary: a summary table directly used for ProPublica analysis

¹ <https://github.com/propublica/compas-analysis>

² There is no source of data statement for the processed crime data.

The Just Machine Learning Project

- **Learning to place**
 - A two-step approach: pairwise preferences + ranking
- **Task on COMPAS recidivism data**
 - Order instances based on time interval (in days) between charges, or prison release and next charge
- **Our data:** Starting from compas.db, we curated crime history data for each individual



The Just Machine Learning Project

- ***Learning to place***

- A two-step approach: pairwise preferences + ranking

- ***Response variable***

- Time interval (in days) between charges

- ***Covariates***

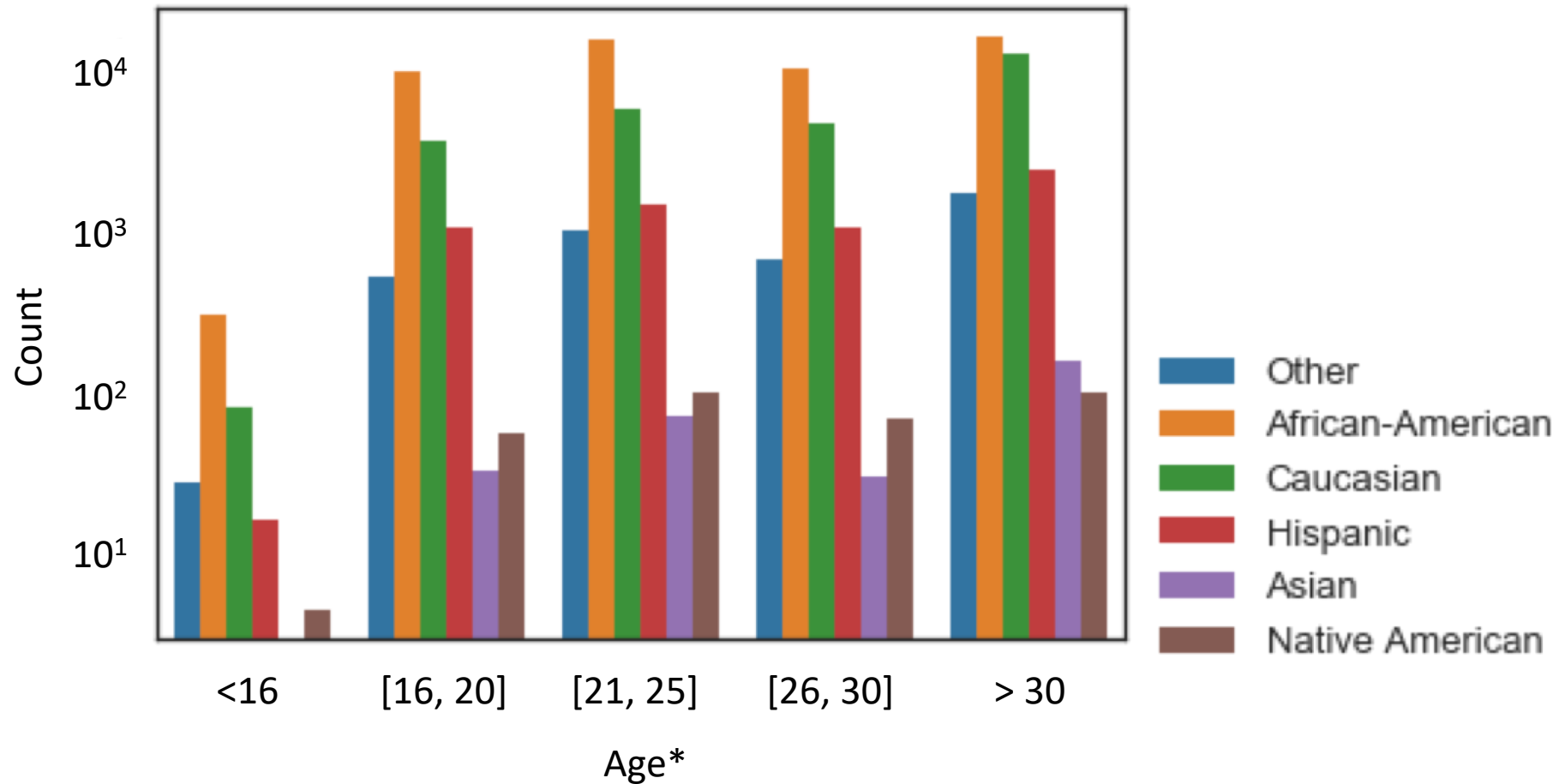
Gender & age at charge time	length of crime career	prior felony count	prior jail count	prior prison count
average & gradient of previous intervals	prior case & charge counts	prior misdemeanor count	prior jail length	prior prison length

Breakdown by charge type

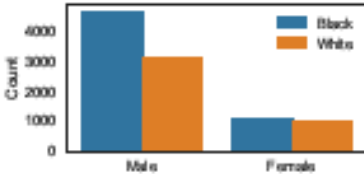
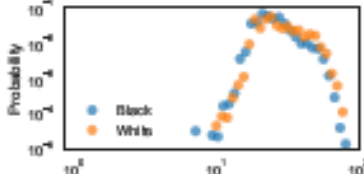
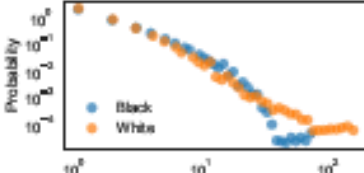
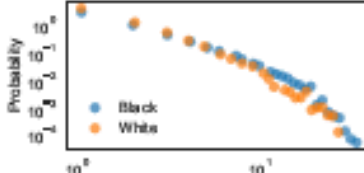
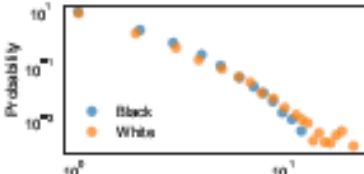
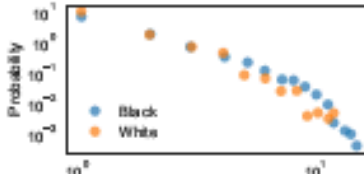
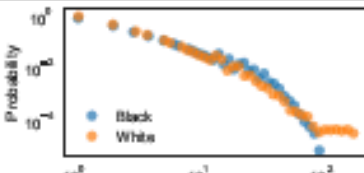
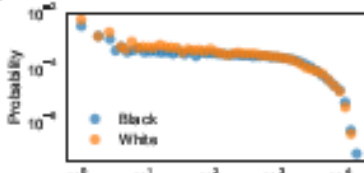
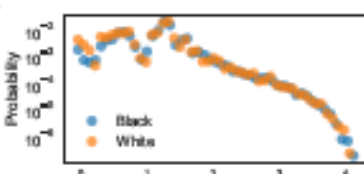
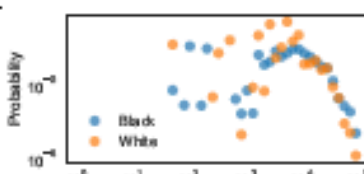
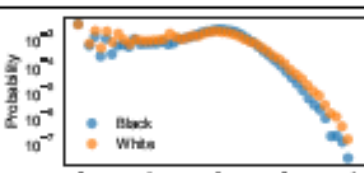
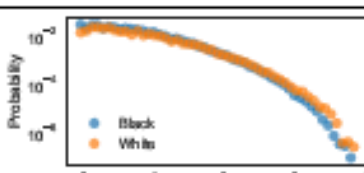
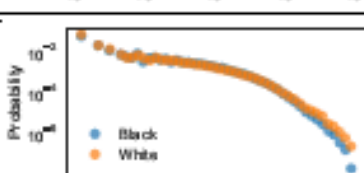
Charge Type	% Among Blacks	% Among Whites
Felony 1	2.37	1.47
Felony 2	10.70	6.48
Felony 3	57.68	53.28
Misdemeanor 1	21.84	30.13
Misdemeanor 2	7.40	8.64

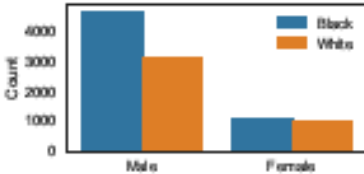
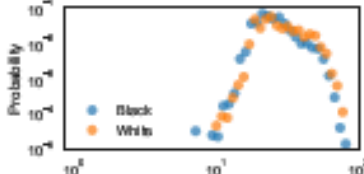
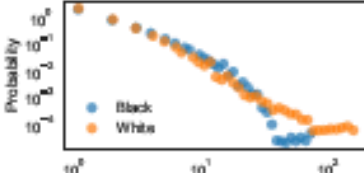
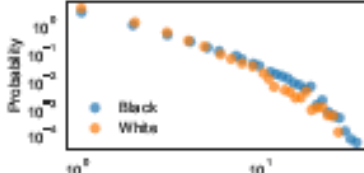
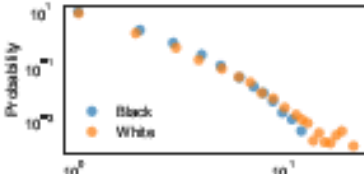
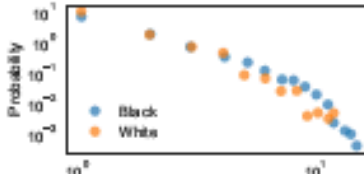
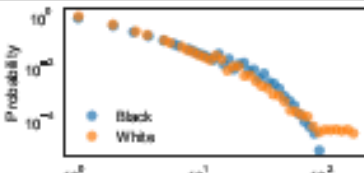
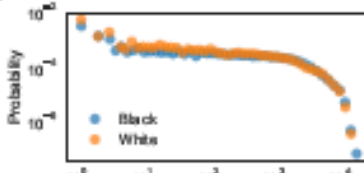
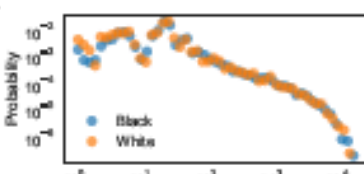
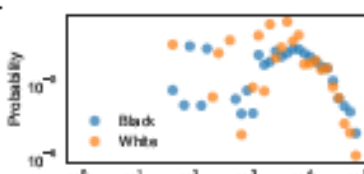
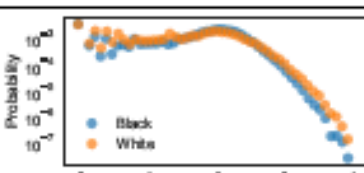
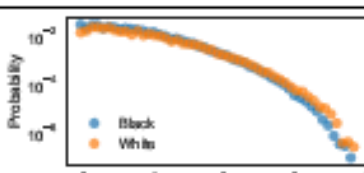
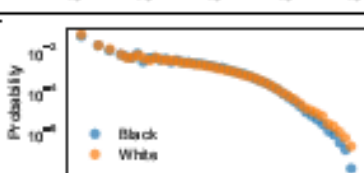
Is a black felony a white misdemeanor?

Age at the time of the charge



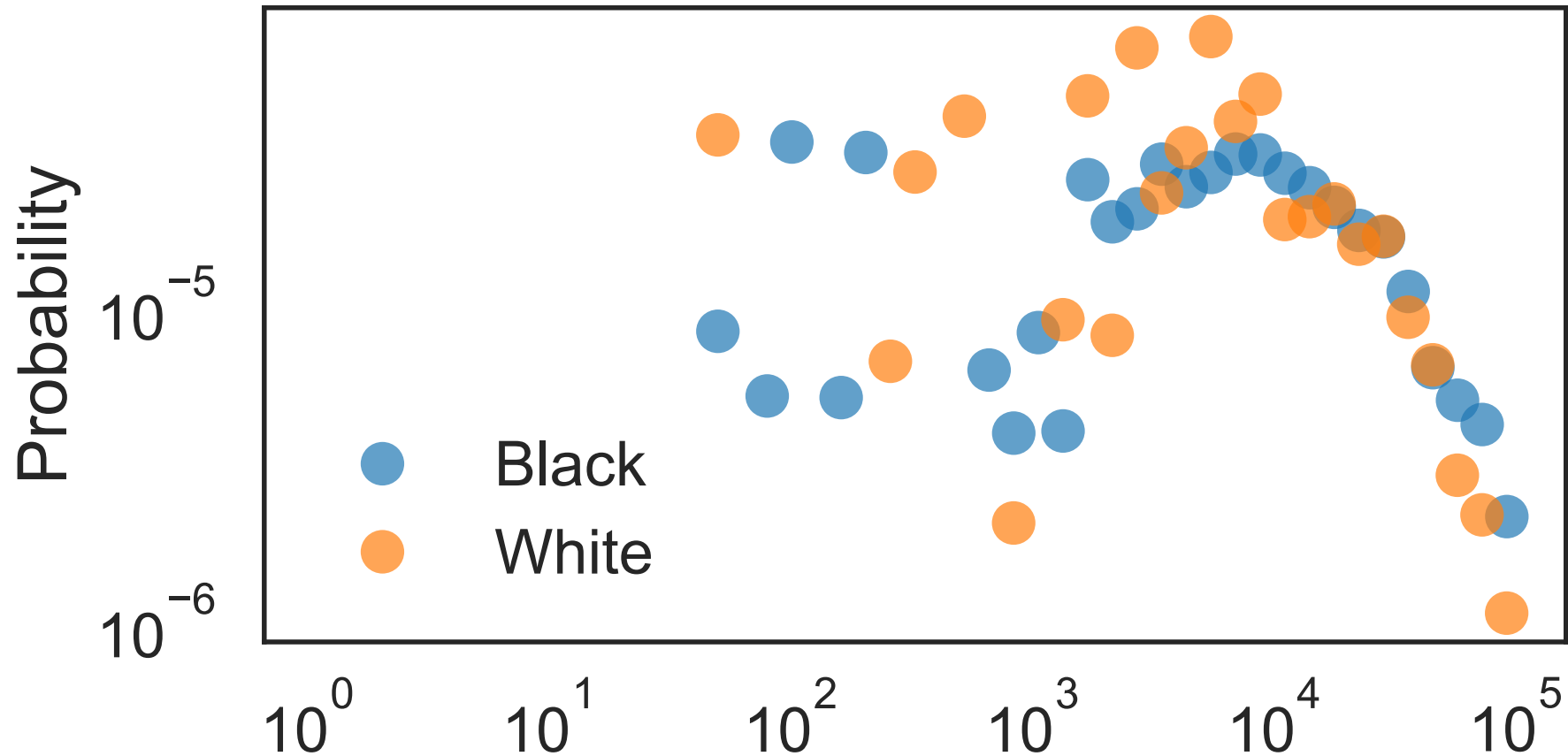
* We categorized age based on input from Jack McDevitt, a criminology professor at Northeastern University.

List of features	Distribution	List of features	Distribution
Gender		Age at the charge time: the age of the individual at the current case	
Prior misdemeanor count		Prior felony count	
Prior jail count		Prior prison count	
Prior charge counts: the number of charges the individual have before the current case.		Length of crime career: time since the first case till the current case, measure in days	
Prior jail length: time this individual spent in jail before the current case, measure in hour		Prior prison length: time this individual spent in prison before the current case, measure in hour	
Average interval between charges: the average of between charge intervals before this crime, measure in days		Trend of interval between charges: for the interval between charges intervals before this case, fit a linear line and get the slope of the fitted line.	
Target Variable	Current interval between charges: the interval between charges from the previous charge to the current charge. The distribution of current interval between charges is heavy-tailed.		

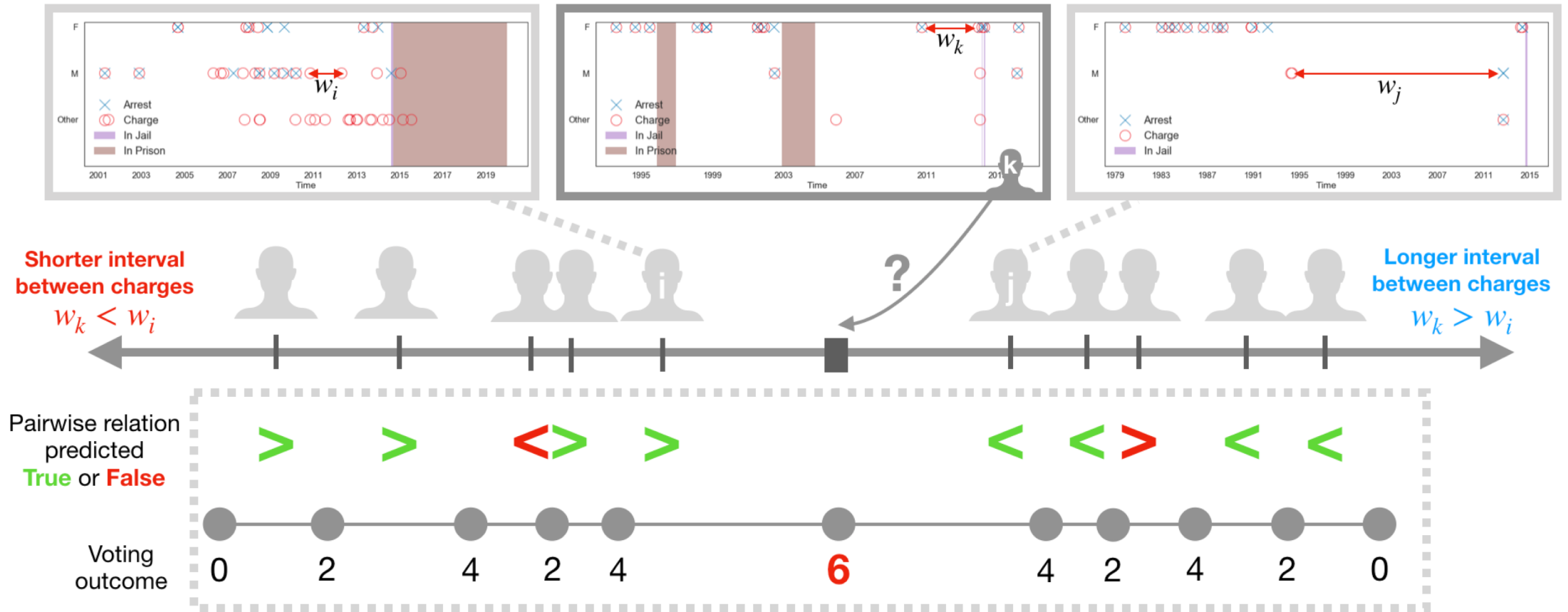
List of features	Distribution	List of features	Distribution
Gender		Age at the charge time: the age of the individual at the current case	
Prior misdemeanor count		Prior felony count	
Prior jail count		Prior prison count	
Prior charge counts: the number of charges the individual have before the current case.		Length of crime career: time since the first case till the current case, measure in days	
Prior jail length: time this individual spent in jail before the current case, measure in hour		Prior prison length: time this individual spent in prison before the current case, measure in hour	
Average interval between charges: the average of between charge intervals before this crime, measure in days		Trend of interval between charges: for the interval between charges intervals before this case, fit a linear line and get the slope of the fitted line.	
Target Variable	Current interval between charges: the interval between charges from the previous charge to the current charge. The distribution of current interval between charges is heavy-tailed.		

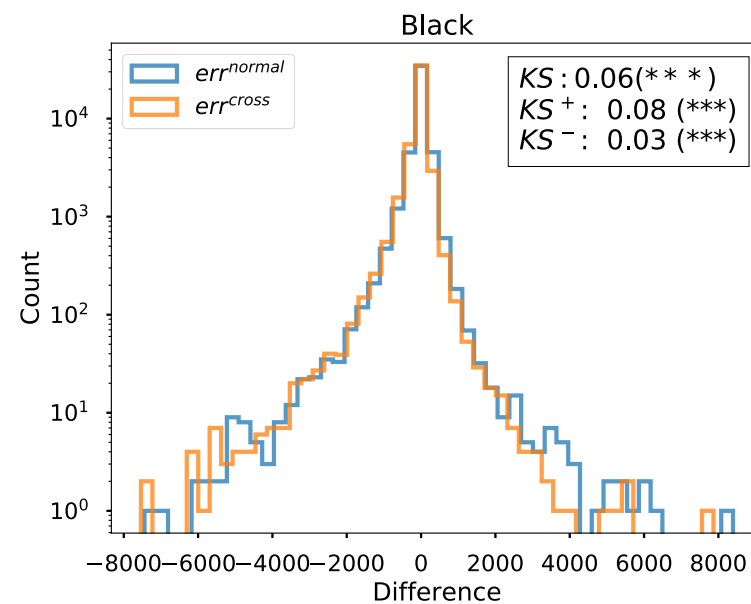
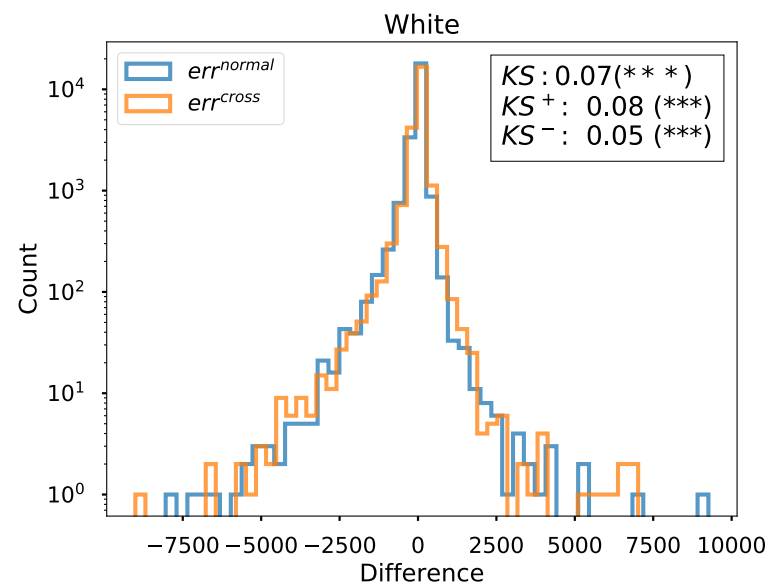
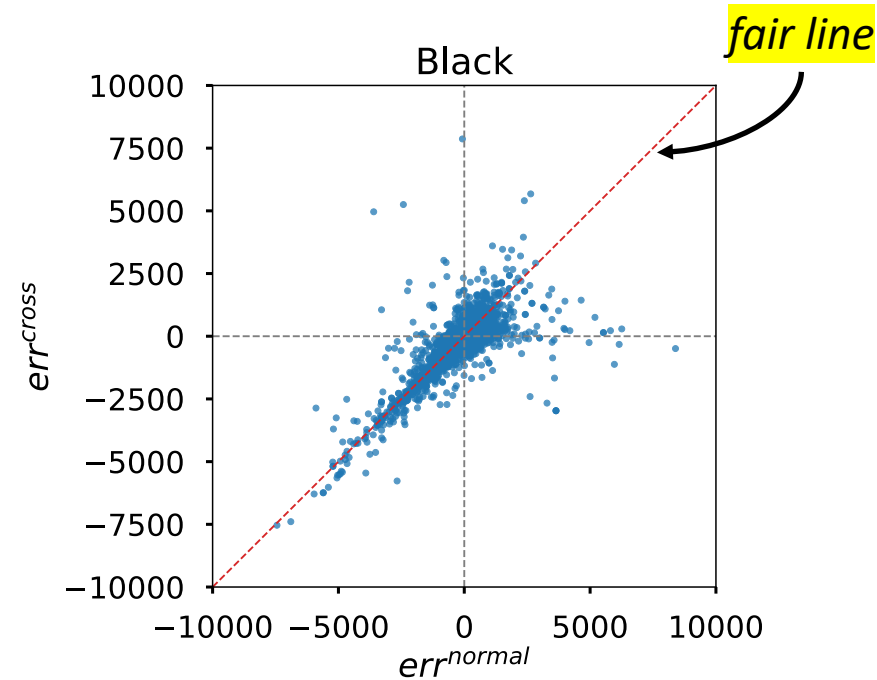
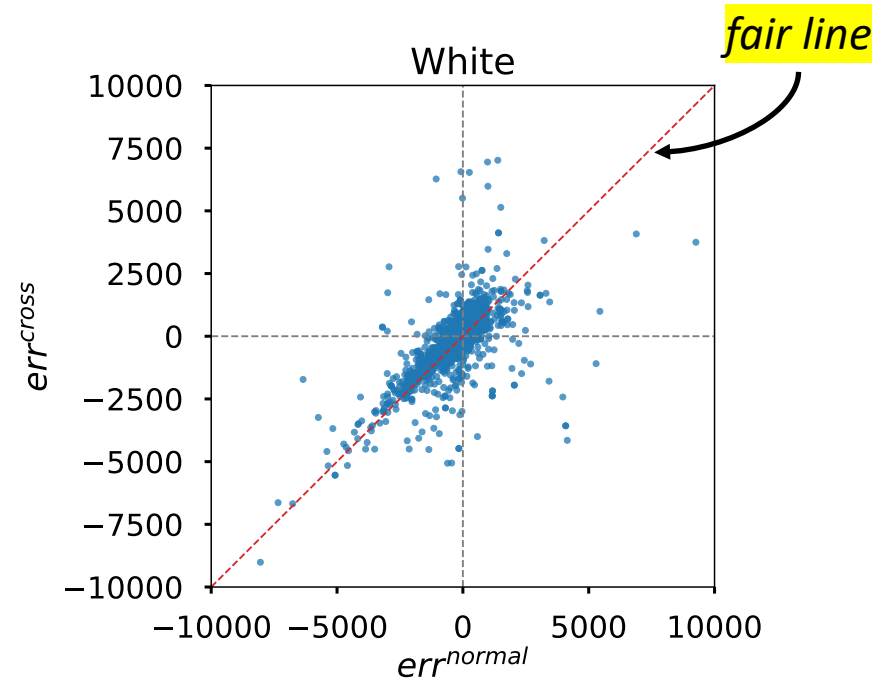


Prior prison length is different between blacks and whites

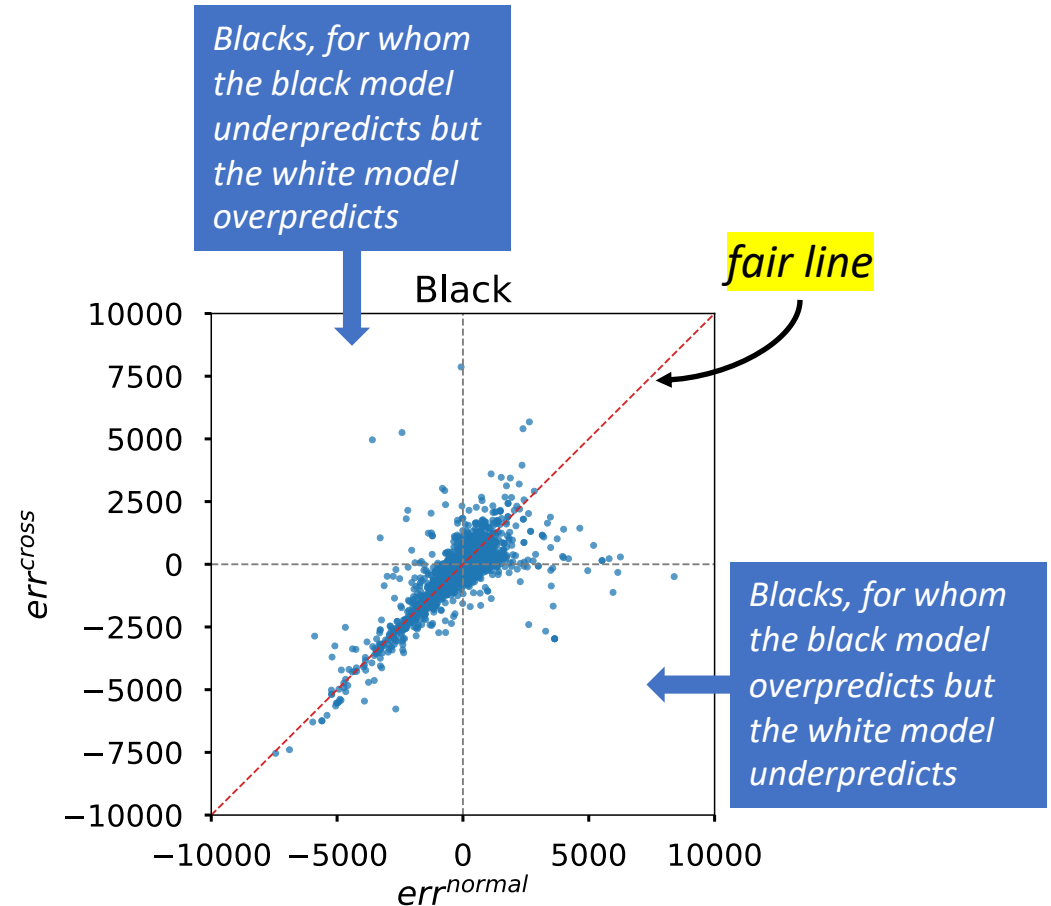
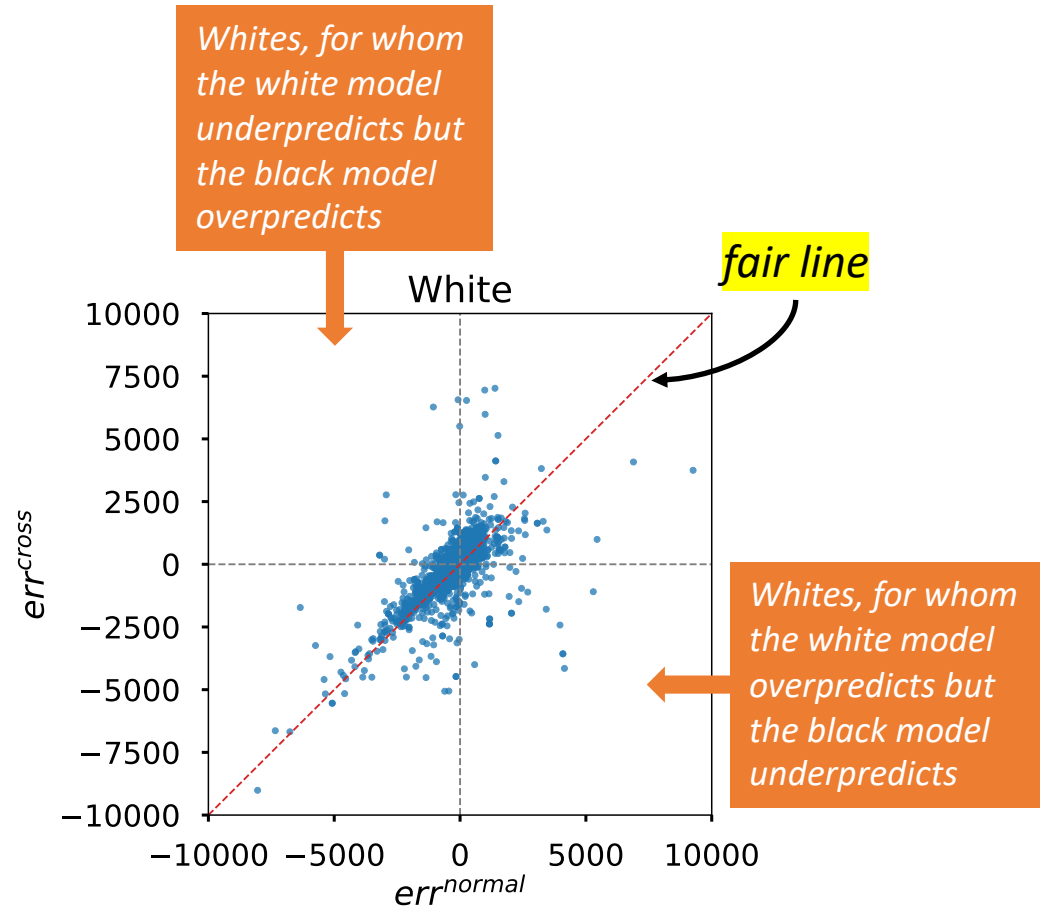


Learning to Place on COMPAS data



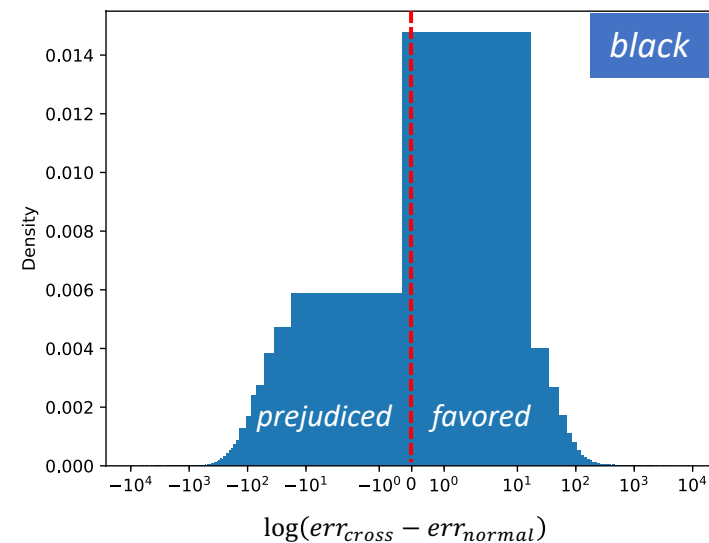
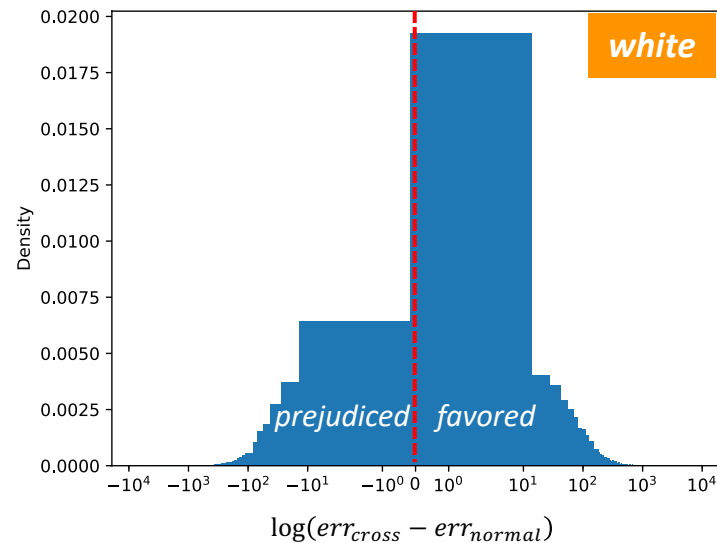


What is fair?



Favored vs. prejudiced rates

- Compute the kernel density estimation for $err_{cross} - err_{normal}$
- **Favored region:** area under curve for $err_{cross} - err_{normal} > 0$
- **Prejudiced region:** area under curve for $err_{cross} - err_{normal} < 0$

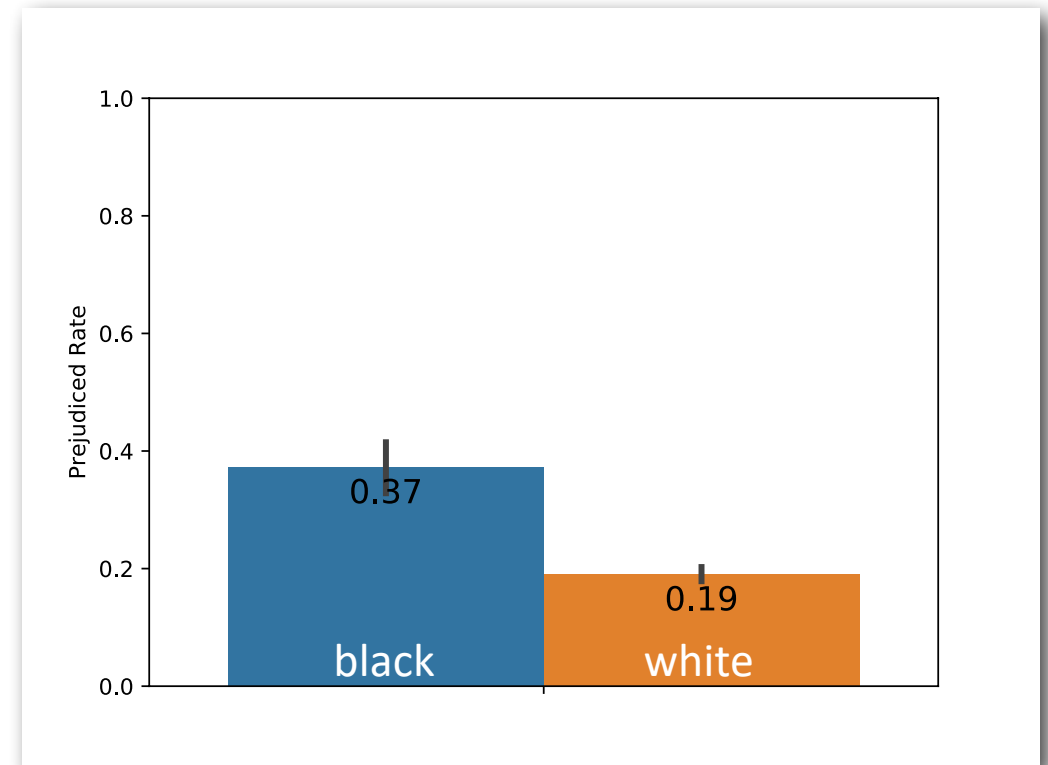
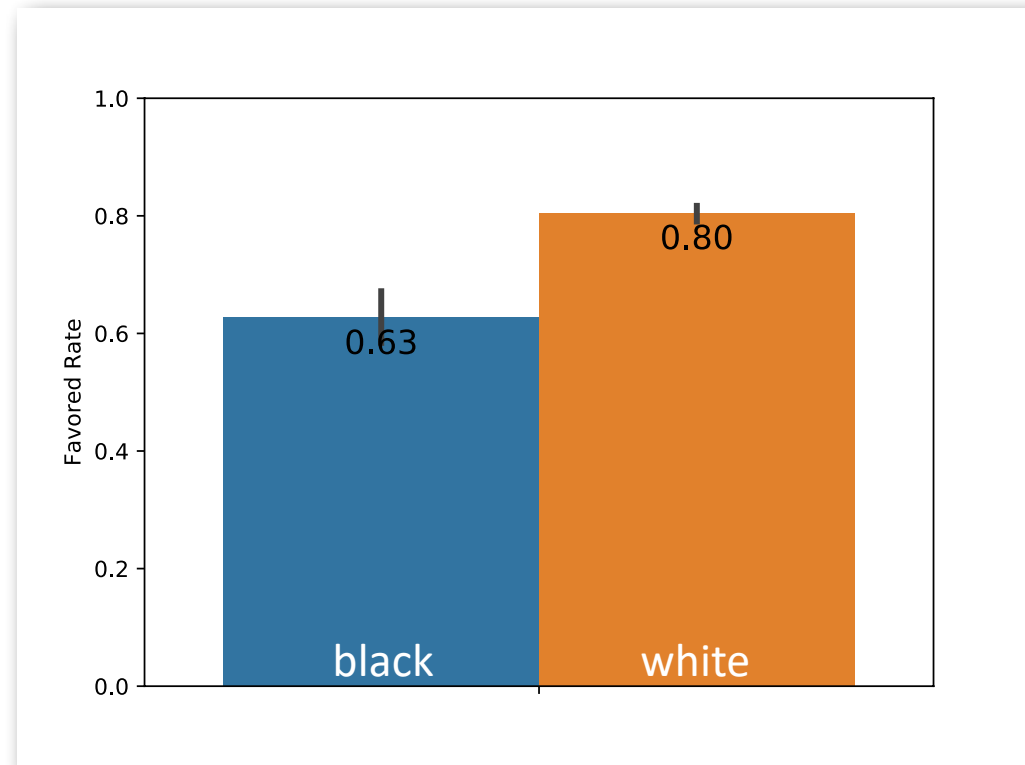


Experimental methodology

- 5×2 cross-validation
 - Five iterations of two fold cross-validation
 - Generates low type 1 error
 - Type 1 error: incorrectly detecting a difference when no difference exists
 - Is slightly more powerful than 10-fold cross-validation
 - Power = 1 – Type II error: ability to detect algorithm differences when they do exist

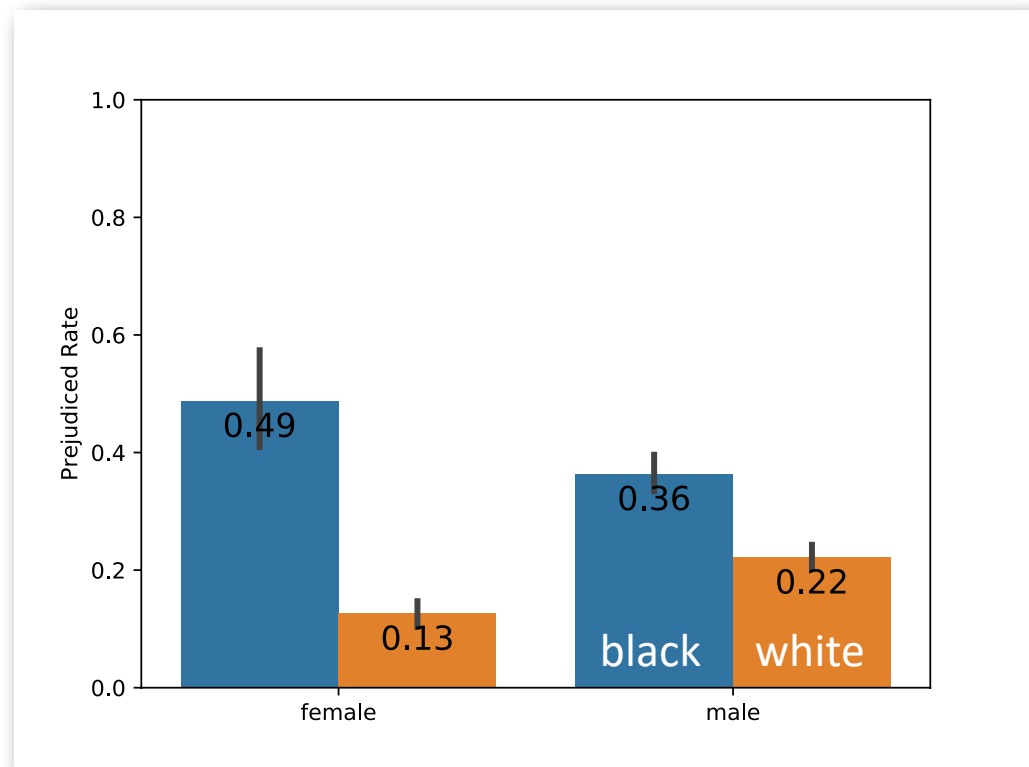
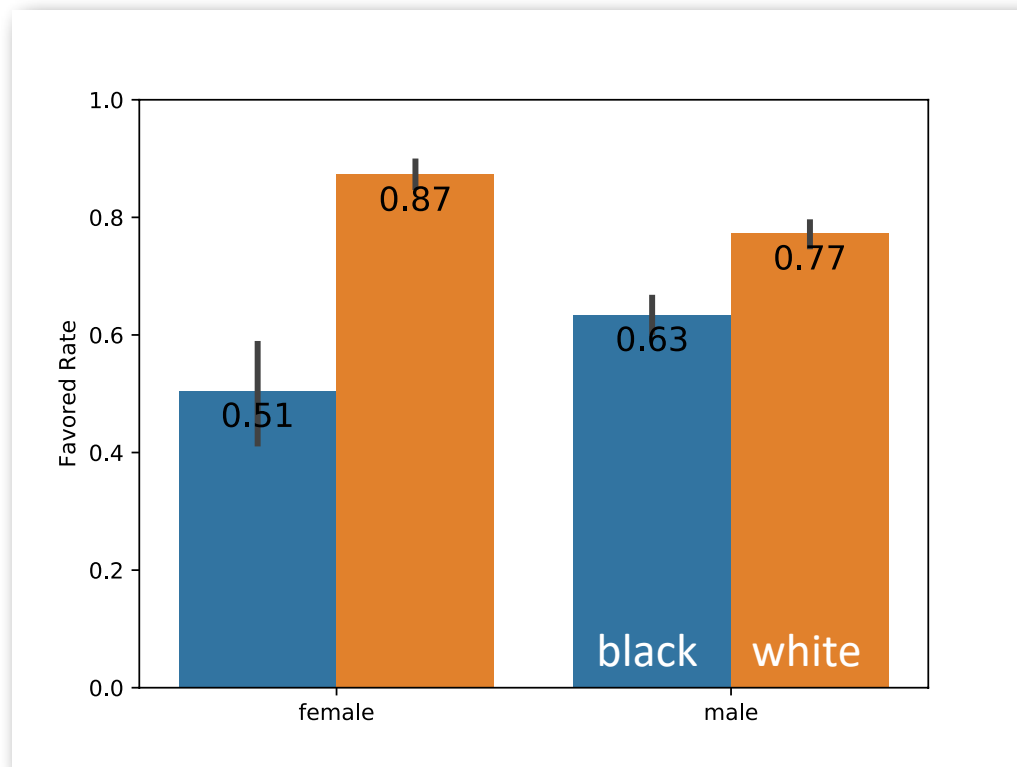
Favored vs. prejudiced by race

- Whites are favored; blacks are prejudiced against.
- Standard deviations on favoritism and prejudice are larger among blacks than among whites.



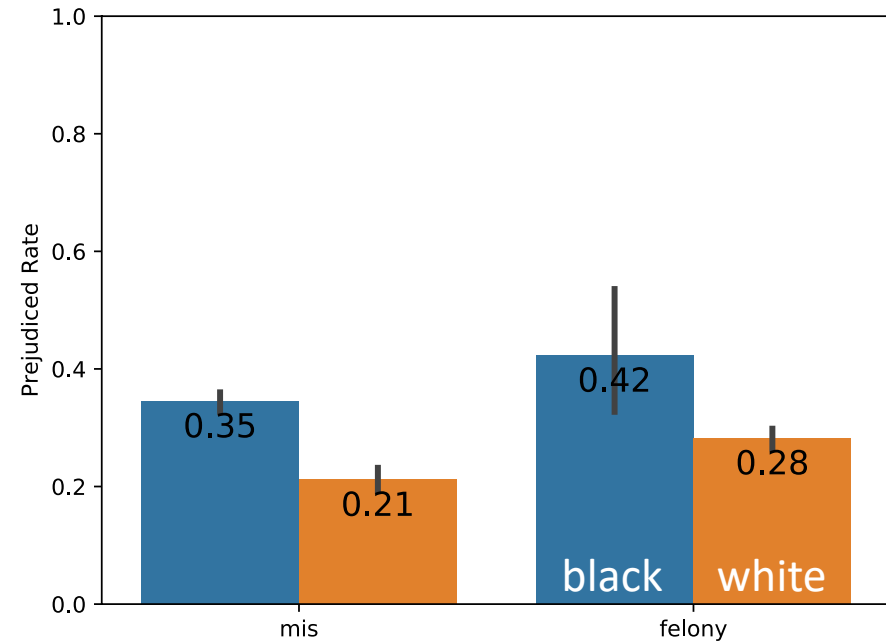
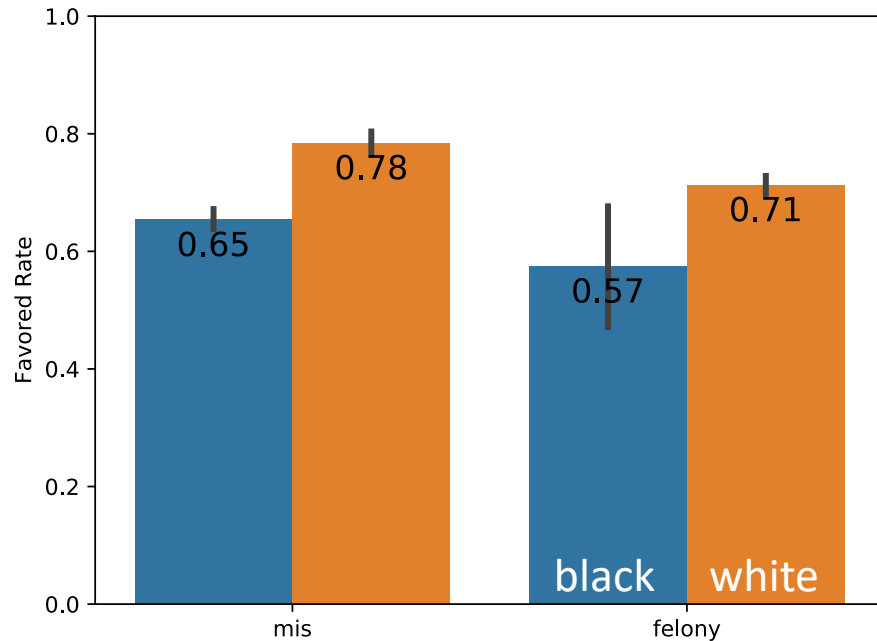
Favored vs. prejudiced by race & gender

- The gaps between white & black females are larger than among white & black males.
- Among whites, females are favored more and males are prejudiced against more.
- Standard deviations on favoritism and prejudice are larger among black females than on black males or whites of either sex.



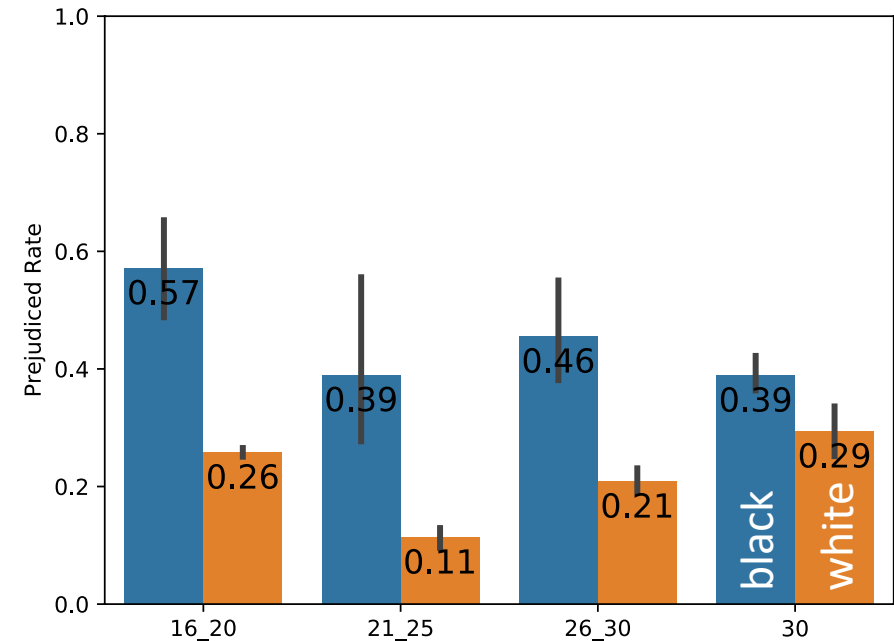
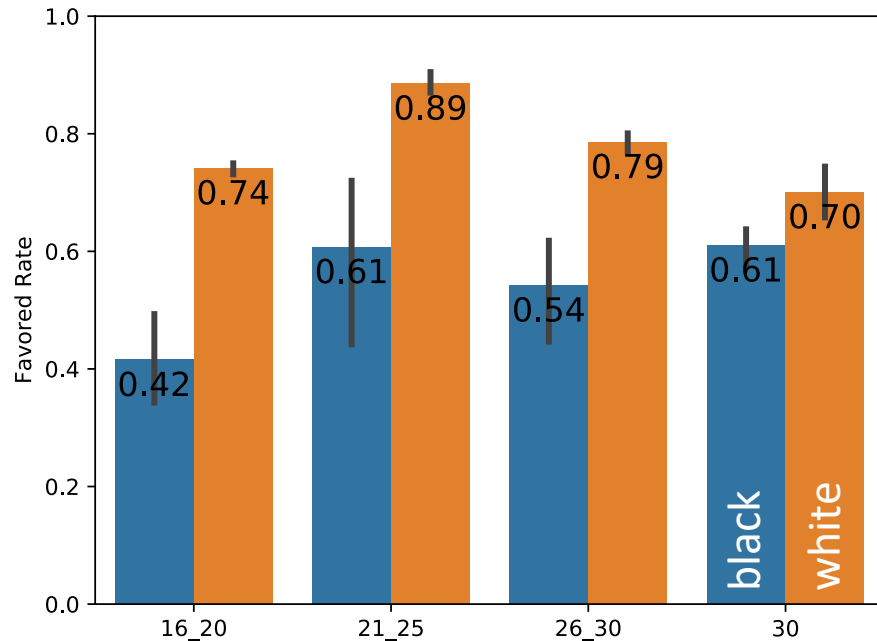
Favored vs. prejudiced by race & degree of charge

- Standard deviations on favoritism and prejudice are larger among black felonies than on white felonies or misdemeanors of either race .
- The favored rates for black and white felonies overlap.



Favored vs. prejudiced by race & age

- The gaps between whites and blacks between 16 & 20 is larger than any other age range.
- Between 21 & 25, whites are the most favored and the least prejudiced against.
- Between 21 & 25, standard deviations on favoritism and prejudice are larger among blacks than any other age for blacks or any age for whites.



So what should we do?

- Presented one way to quantify favoritism vs. prejudice
 - We have many others
 - Challenges:
 - Null models?
 - Connections to social and justice theories?

So what should we do?

- Presented one way to quantify in-group favoritism vs. out-group prejudice
 - We have many others
 - Challenges: Null models? Connections to social and justice theories?
- Perhaps we should employ a kind of affirmative action for machine learning algorithms when we observe favoritism for one group and prejudice for another
 - Example: In the COMPAS data, use the model trained on white individuals to analyze the cases of black individuals

Where do we go from here?

“Computers may be intelligent, but they are not wise. Everything they know, we taught them, and we taught them our biases. They are not going to un-learn them without transparency and corrective action by humans.”

-- Ellora Thadaney Israni

Education

- Lots of college courses on data/digital + ethics
- A list is publicly available in a Google spreadsheet at <https://tinyurl.com/yc74sdpe>

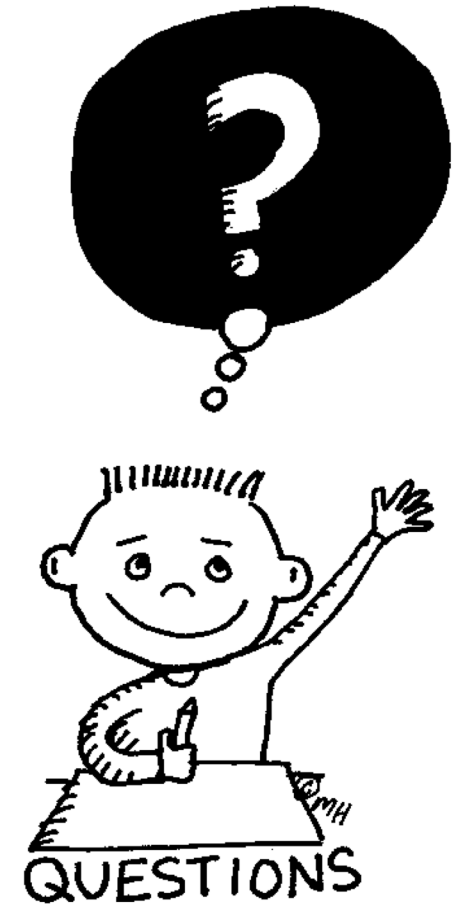
Embedded EthiCS: Integrating Ethics Broadly Across Computer Science Education

Barbara J. Grosz, David Gray Grant, Kate Vredenburg, Jeff Behrends, Lily Hu, Alison Simmons, Jim Waldo

(Submitted on 16 Aug 2018)

Computing technologies have become pervasive in daily life, sometimes bringing unintended but harmful consequences. For students to learn to think not only about what technology they could create, but also about what technology they should create, computer science curricula must expand to include ethical reasoning about the societal value and impact of these technologies. This paper presents Embedded EthiCS, a novel approach to integrating ethics into computer science education that incorporates ethical reasoning throughout courses in the standard computer science curriculum. It thus changes existing courses rather than requiring wholly new courses. The paper describes a pilot Embedded EthiCS program that embeds philosophers teaching ethical reasoning directly into computer science courses. It discusses lessons learned and challenges to implementing such a program across different types of academic institutions.

- Thanks to Danielle Allen, Jack McDevitt, Branden Fitelson, and Ron Sandler.
- Thank you for listening.
- COMPAS data
 - <https://github.com/propublica/compas-analysis/>
- Slides
 - http://eliassi.org/tina_justML_2018.pdf
- Contact info
 - tina@eliassi.org
 - [@tinaeliassi](#)



- Consider joining the RADLAB
- 3 postdoc positions at the intersection of machine learning, data mining, and network science
- Details at <http://eliassi.org/postdocs18.htm>