

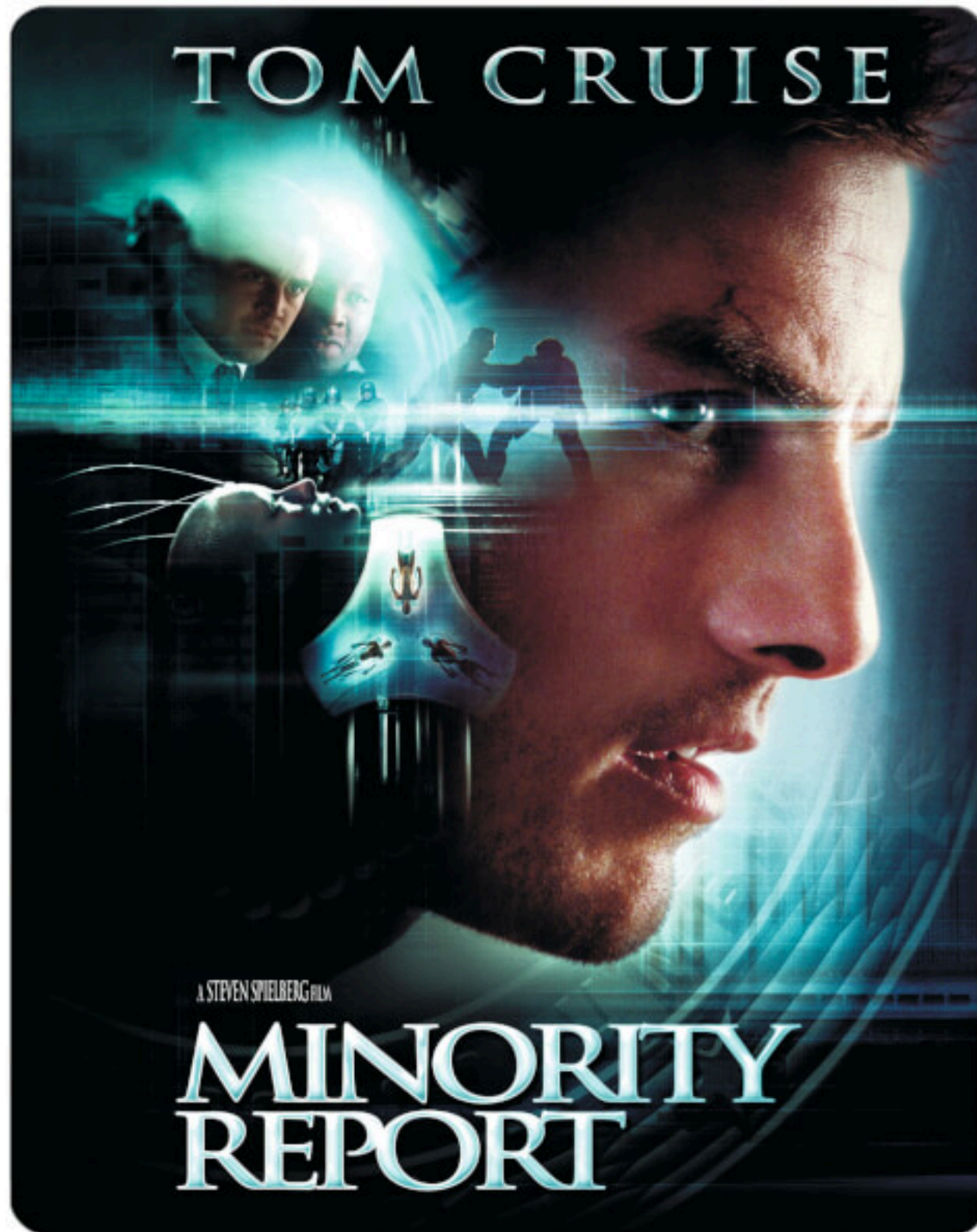
CS 520

Theory and Practice of Software Engineering
Fall 2019

Software Fairness

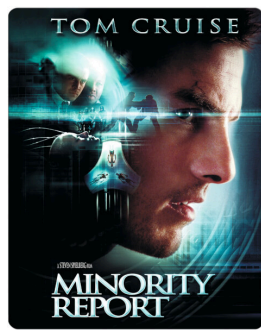
September 5, 2019

TOM CRUISE



A STEVEN SPIELBERG FILM

MINORITY REPORT



Resilient cities Cities

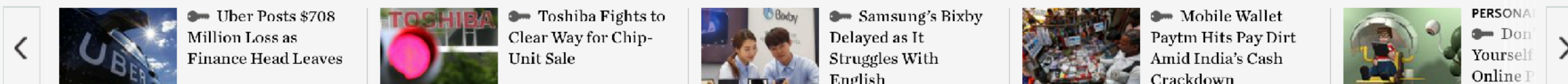
Predicting crime, LAPD-style

Cutting edge data-driven analysis directs Los Angeles patrol officers to likely future crime scenes - but critics worry that decision-making by machine will bring 'tyranny of the algorithm'

- [Join our live Q&A with Homicide Watch this Friday](#)



▲ PredPol co-developer P Jeffrey Brantingham at the Unified Command Post in Los Angeles. 'This is not Minority Report,' he said. Photograph: Damian Dovarganes/AP



TECH

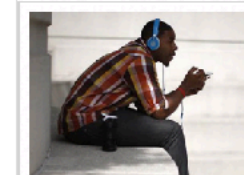
On Orbitz, Mac Users Steered to Pricier Hotels



Orbitz has found that Apple users spend as much as 30% more a night on hotels, so the online travel site is starting to show them different, and sometimes costlier, options than Windows visitors see. Dana Mattioli has details on The News Hub. Photo: Bloomberg.

By Dana Mattioli

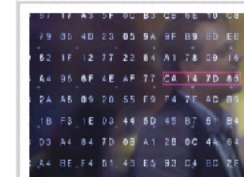
ADVERTISEMENT



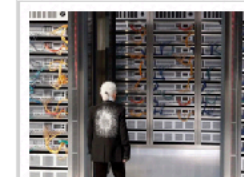
Techstars to Launch Accelerator for Music-Industry Tech Startups

THE CLOUD ALONE ISN'T "SMART."

rackspace.



The Skills Gap Is No Laughing Matter



Cloud IT Infrastructure Spending Up





The Algorithm That Beats Your Bank Manager

[HAAS NEWS](#) > [NEWS CATEGORIES](#) > [RESEARCH NEWS](#)

Minority homebuyers face widespread statistical lending discrimination, study finds

By [Laura Counts](#) | NOVEMBER 13, 2018

Face-to-face meetings between mortgage officers and homebuyers have been rapidly replaced by online applications and algorithms, but lending discrimination hasn't gone away.

A [new University of California, Berkeley study](#) has found that both online and face-to-face lenders charge higher interest rates to African American and Latino borrowers, earning 11 to 17 percent higher profits on such loans. All told, those homebuyers pay up to half a billion dollars more in interest every year than white borrowers with comparable credit scores do, researchers found.

The findings raise legal questions about the rise of statistical discrimination in the fintech era, and point to potentially widespread violations of U.S. fair lending laws, the researchers say. While lending discrimination has historically been caused by human prejudice, pricing disparities are increasingly the result of algorithms that use machine learning to target applicants who might shop around less for higher-priced loans.

"The mode of lending discrimination has shifted from human bias to algorithmic bias," said study co-author [Adair Morse](#), a finance professor at UC Berkeley's Haas School of Business. "Even if the people writing the

EDITOR'S PICK

Wisconsin Supreme Court allows state to continue using computer program to assist in sentencing

KATELYN FERRAL | The Capital Times | kferral@madison.com | [@katelynferral](https://twitter.com/katelynferral) Jul 13, 2016



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

**Software can make bad decisions.
Software can discriminate!**

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

Subscribe to the Series

English Spanish Turkish Detect language ▼



English Spanish Turkish ▼

Translate

He is a nurse.
She is a doctor.



31/5000

O bir hemşire.
O bir doktor.



English Spanish Turkish Detect language ▼



English Spanish Turkish ▼

Translate

O bir hemşire.
O bir doktor.



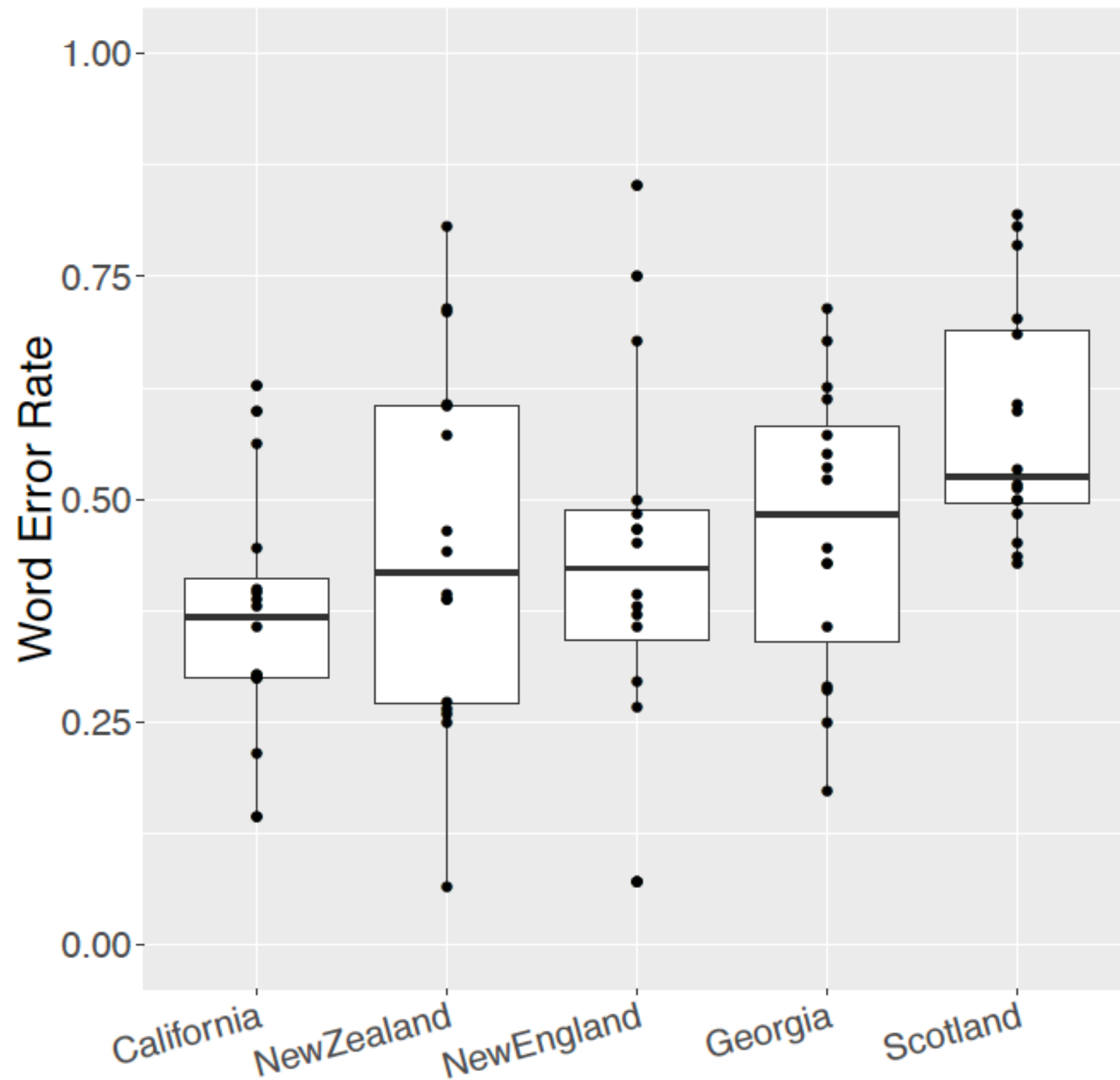
28/5000

She is a nurse.
He is a doctor.

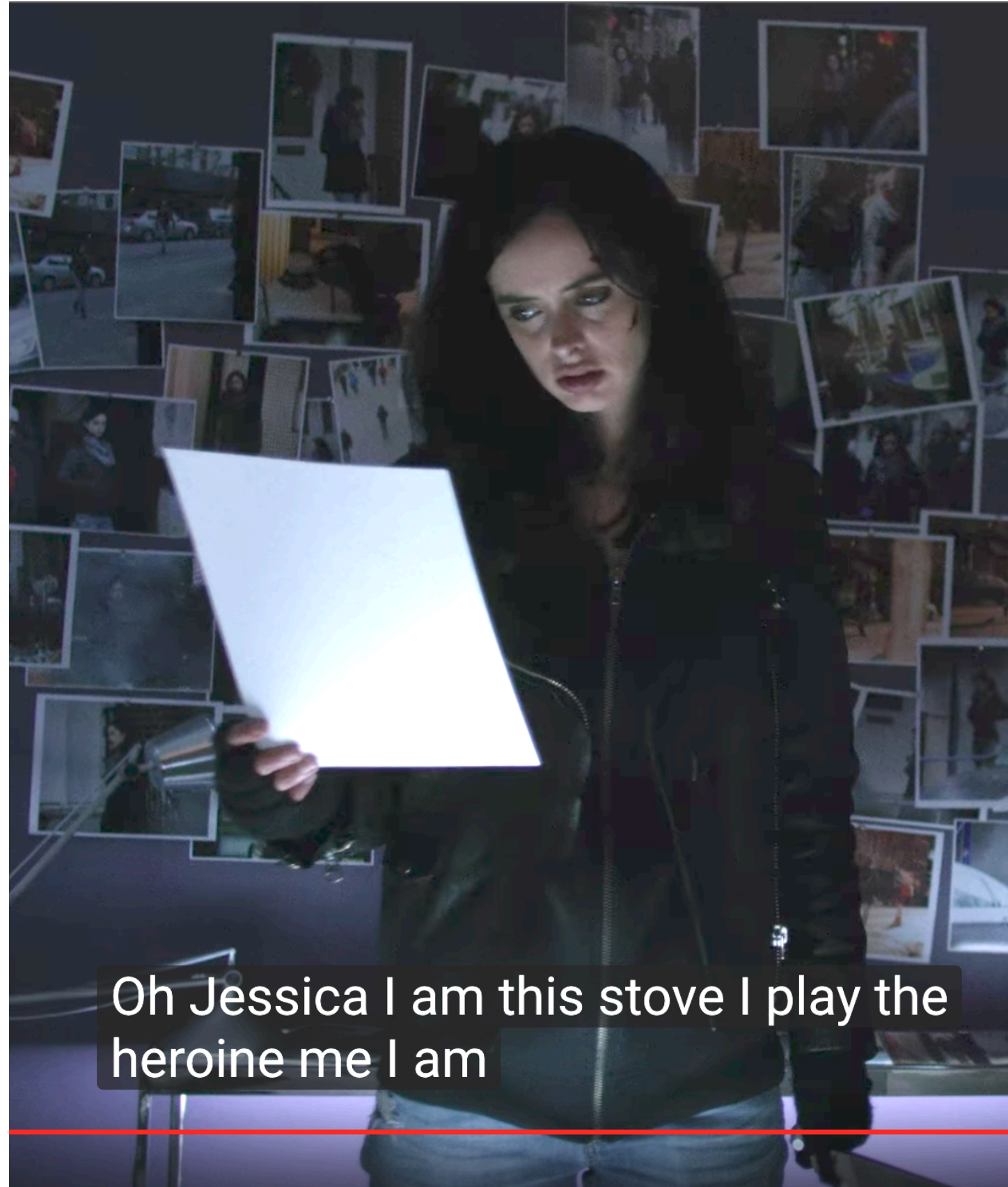




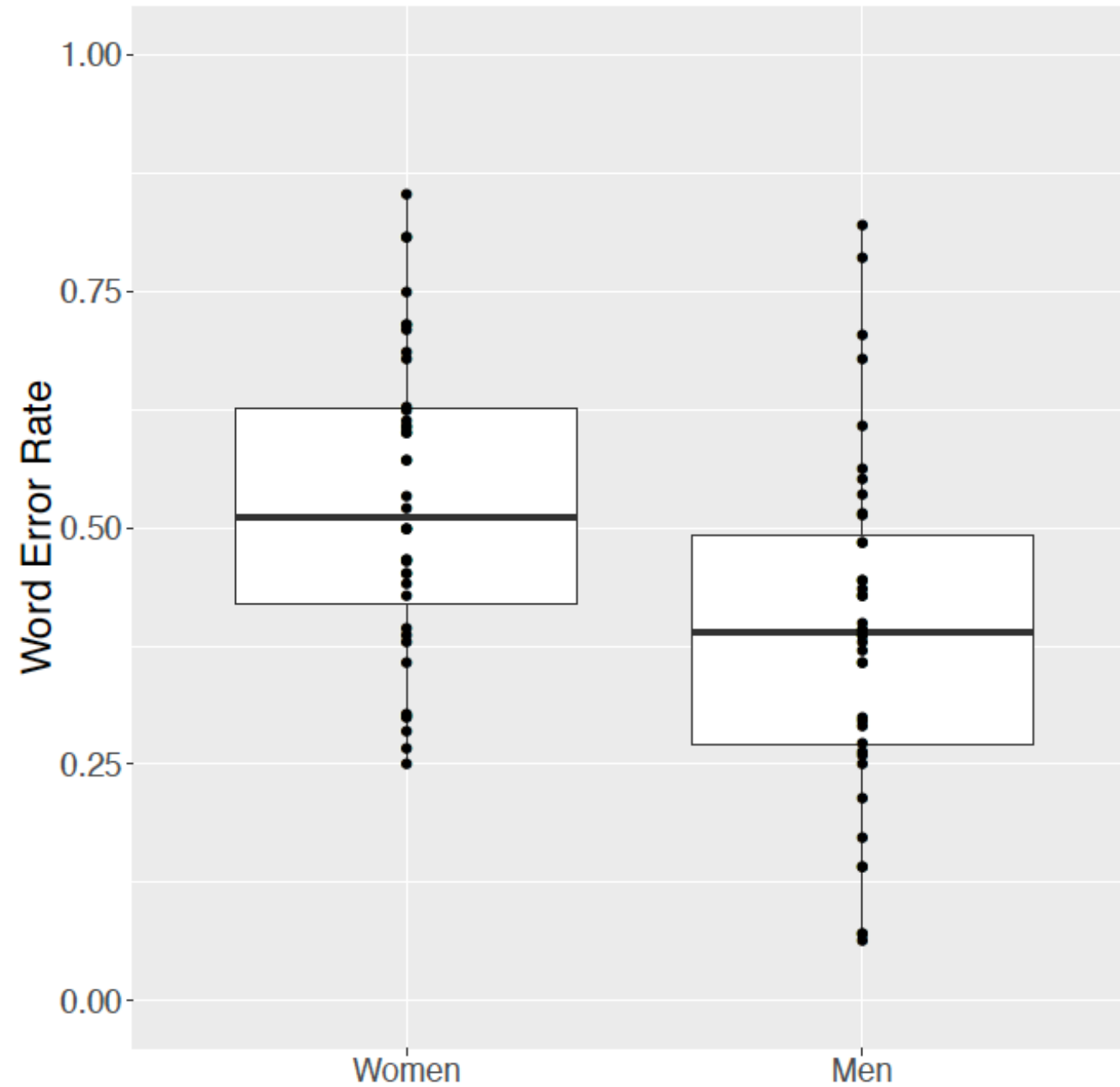
YouTube automatic captions



YouTube automatic captions



YouTube automatic captions





BUOLAMWINI

ow I'm fighting bias in algorithms

Joy Buolamwini

https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

how people want to use vision software

how people want to use vision software

HireVue

SEE A HIREVUE ASSESSMENTS DEMO

WANT TO SEE HIREVUE IN ACTION?

SEE A DEMO

The predictive power of a traditional assessment in a video interview.

- Predictive assessments tied to higher quality hires
- Assess and interview in one step and hire faster
- Better candidate experience
- Validated technology expertise

JOIN 700+ CUSTOMERS THAT USE HIREVUE

 **vodafone**

TJX[®]

TIFFANY & CO.

GEICO[®]

 **DR PEPPER
SNAPPLE**^{GROUP}

 **intel**

today's goals

Define software discrimination.

Operationalize measuring discrimination through causal software testing.

Provide provable fairness guarantees.

Design software to be fair

2011 11th IEEE International Conference on Data Mining

Handling Conditional Discrimination

Indre Žliobaitė
Bournemouth University, UK
izliobait@bournemouth.ac.uk

Faisal Kamiran
TU Eindhoven, the Netherlands
f.kamiran@tue.nl

Toon Calders
TU Eindhoven, the Netherlands
t.calders@tue.nl

Discrimination Aware Decision Tree Learning

Faisal Kamiran, Toon Calders and Mykola Pechenizkiy
Email: {f.kamiran,t.calders,m.pechenizkiy}@tue.nl
Eindhoven University of Technology, The Netherlands

Abstract—Recently, the following discrimination aware classification problem was introduced: given a labeled dataset and an attribute B , find a classifier with high predictive accuracy

It can be argued that in many real-life cases discrimination can be explained; e.g., it may very well be that females in an employment dataset overall have less years of working experience, justifying a correlation between the gender and the label. Nevertheless, in this paper we assume this not to be the case. We assume that the data is already divided up based on acceptable explanatory attributes. Within this framework, gender discrimination can no longer be justified. Within previous works [7], [3], simply removing a sensitive attribute from the training data does not work, as attributes may be correlated with the suppressed attribute. It was observed that classifiers tend to pick up

Building Classifiers with Independency Constraints

Toon Calders¹, Faisal Kamiran²
Eindhoven University of Technology
{t.calders, f.kamiran}@tue.nl

Abstract. 150 word abstract.

Fairness Constraints: Mechanisms for Fair Classification

Muhammad Bilal Zafar, Isabel Valera
Max Planck Institute for Intelligent Systems

Abstract

Algorithmic decision making systems are becoming ubiquitous across a wide variety of online and offline services. These systems rely on complex learning methods and vast amounts of data to optimize the service functionality and satisfaction of the end user and profitability. However, there is a growing concern that these automated decisions can lead, even in the absence of intent, to a lack of fairness, i.e., their outcomes can disproportionately hurt (or, benefit) particular groups of people sharing one or more sensitive attributes (e.g., race, sex). In this paper, we propose a flexible mechanism to control fairness by leveraging a novel information-theoretic decision boundary (un)fairness constraint. This mechanism works with a wide range of classifiers, logistic regression, and machine learning models, and shows that our mechanism allows for fine-grained control on the degree of fairness. A Python implementation is available at [fate-dev](https://github.com/fate-dev/fate-dev).

1 INTRODUCTION

Algorithmic decision making is becoming automated and (e.g., spam filtering, product recommendations, pretrial risk assessments) settings. However, as the scale of the analyzed data grows, growing concerns from citizens [2016], governments [Podese, 2016], and researchers [Swamy, 2016] about the loss of transparency, accountability, and fairness.

Anti-discrimination laws

Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, Florida, USA. JMLR, 2017. Copyright 2017 by the author(s).

Learning Fair Representations

Richard Zemel
Yu (Ledell) Wu
Kevin Swersky
Toniann Pitassi
University of Toronto, 10 King's College Rd., Toronto, Canada

Cynthia Dwork
Microsoft Research, 1065 La Avenida Mountain View, CA

Abstract

We propose a learning algorithm for fair classification that achieves both group fairness (the proportion of members in a protected group receiving positive classification is identical to the proportion in the overall population)

2012 IEEE 12th International Conference on Data Mining

Decision Theory for Discrimination-aware Classification

Faisal Kamiran*, Asim Karim[†], and Xiangliang Zhang*
*King Abdullah University of Science and Technology (KAUST), The Kingdom of Saudi Arabia
Email: faisal.kamiran, xiangliang.zhang@kaust.edu.sa
[†]Lahore University of Management Sciences, Pakistan
Email: akarim@lums.edu.pk

- Typically machine learning systems:
- Balance training sets
 - Introduce training noise
 - Constrain regression's loss function
 - Split criteria on sensitive inputs

needs to be processed again. Being restricted to a discrimination-aware classifier (e.g., naive Bayes or decision tree [2]) is also an issue because that classifier is not the best performing classifier for a given dataset. In this paper, we propose two flexible and easy-to-use discrimination-aware classification methods based on the hypothesis: discriminatory decisions are often made on the decision boundary because of decision confidence. We implement this hypothesis via decision boundary confidence and ensemble methods. Our first solution, called Reject Option based on Confidence (ROC), exploits the low confidence region of an ensemble of probabilistic classifiers for rejection. More specifically, ROC invokes the confidence of labels instances belonging to deprived groups in a manner that reduces discrimination. Our second solution, called Discrimination-Aware Ensemble (DAE), exploits the disagreement region of a classifier to reject both deprived and favored group instances. Our proposed solutions have favorable results over existing discrimination-aware classification methods.

Our methods are not restricted to a particular classifier. Our first solution works with any probabilistic classifier while our second solution works with general ensemble methods.

Our methods require neither modification of learning algorithms nor preprocessing of historical data – preprocessing can be made discrimination-aware in advance. Thus, the change in the sensitive attributes can be handled easily by decision makers. Our methods give better control and interpretability of discrimination-aware classification to decision makers. We provide an extensive experimental evaluation of our methods on real-world datasets. The results demonstrate that our methods achieve a better trade-off of discrimination and superior accuracy compared to existing discrimination-aware classification methods.

II. RELATED WORK

Recent work on social discrimination-aware data mining was done by Pedreschi et al. [3], [4], focusing on discovery of discriminatory classification rules from biased datasets.

Design alone is not enough

possible causes



biased data



**implementation
bugs**



**unintended interactions and
mismatched components**



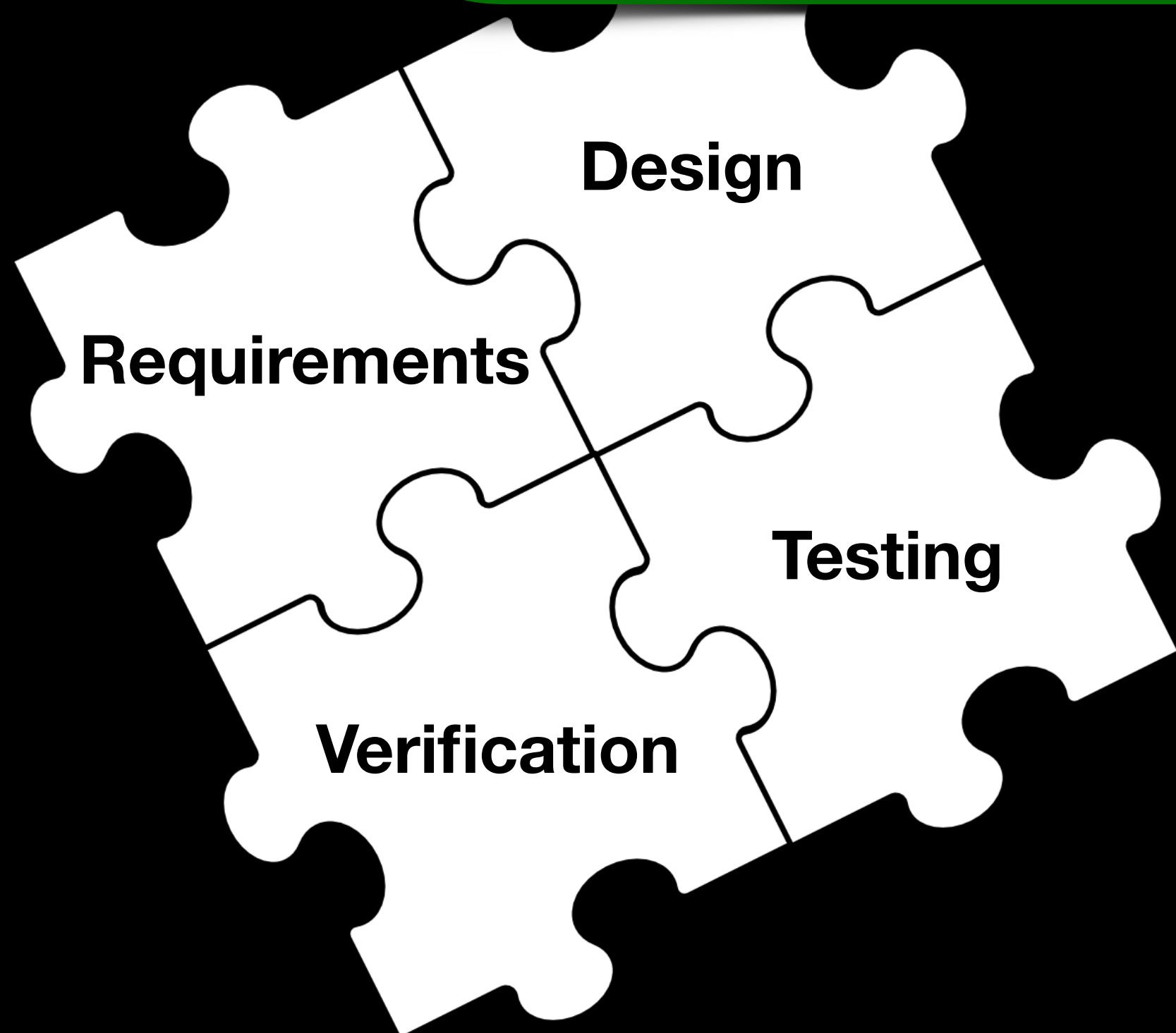
poor design

Fairness is just like quality and security

Fairness must be part of the
software engineering lifecycle

Call to Action!

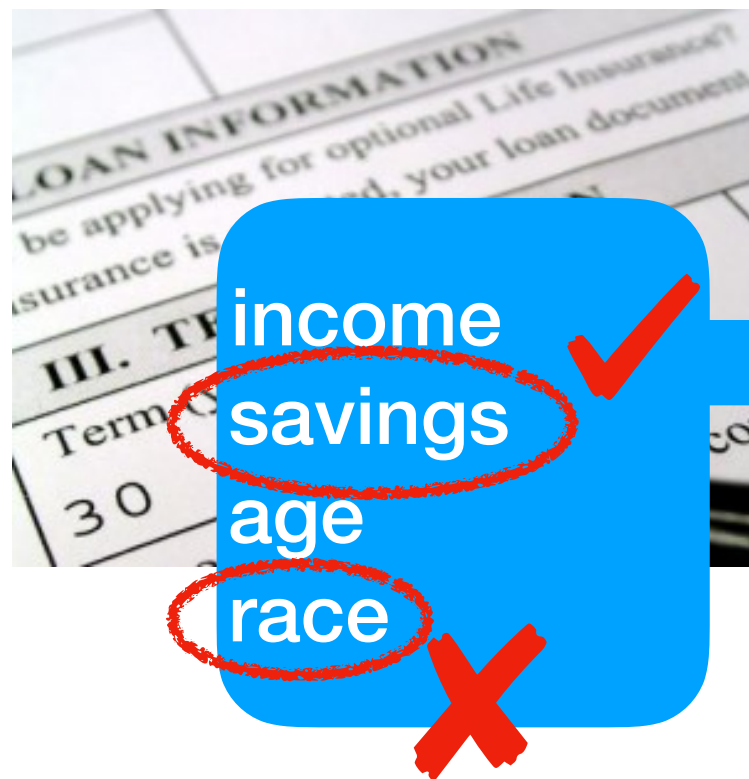
Fairness must be part of the software engineering lifecycle



Let's talk about requirements.

**What does it mean for
software to discriminate?**

LOAN program



LOAN

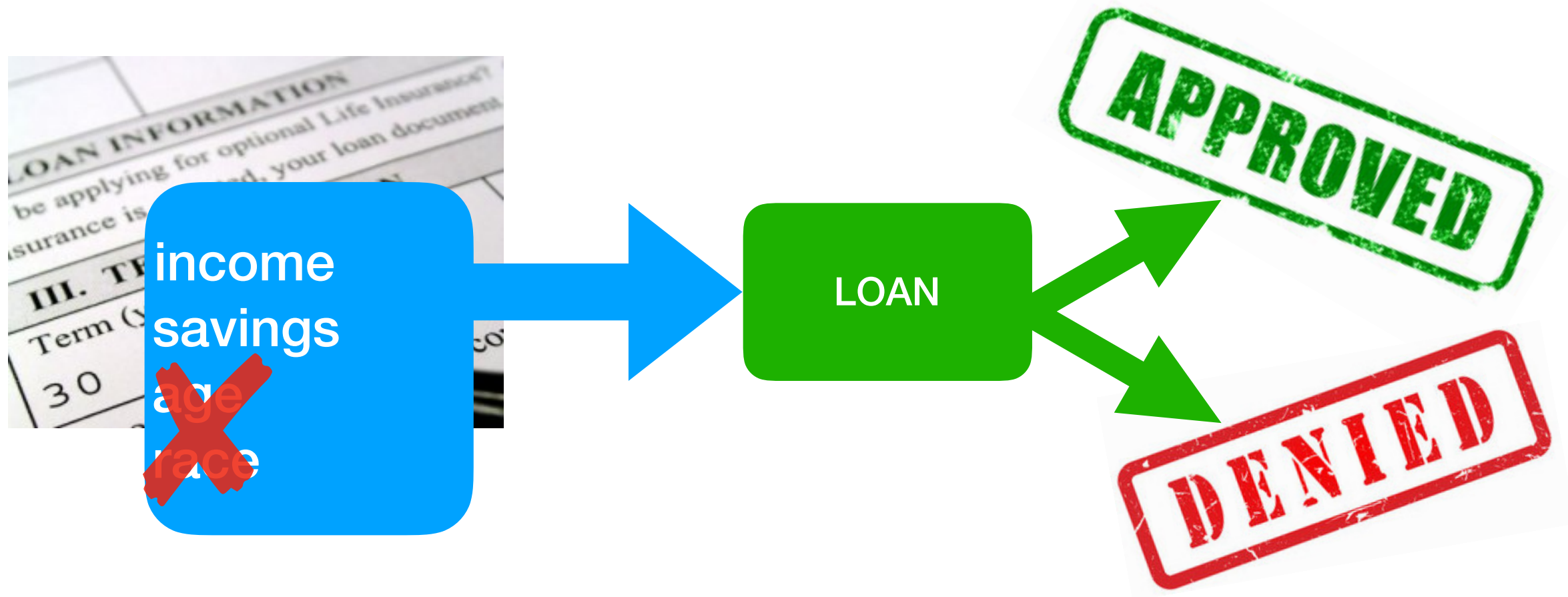
APPROVED

DENIED

This talk is not about policy.

Fairness: Disparate Treatment

Hide the data



Fairness: Disparate Treatment

Hide the data

Ads by Google

[Latanya Sweeney, Arrested?](#)

1) Enter Name and State. 2) Access Full Background Checks Instantly.

www.instantcheckmate.com/

Ineffective because of data correlation.

[Latanya Sweeney. Discrimination in online ad delivery. CACM 2013]

Amazon just showed us that 'unbiased' algorithms can be inadvertently racist

Rafi Letzter

Apr. 21, 2016, 4:50 PM 1,259



FACEBOOK



LINKEDIN



TWITTER

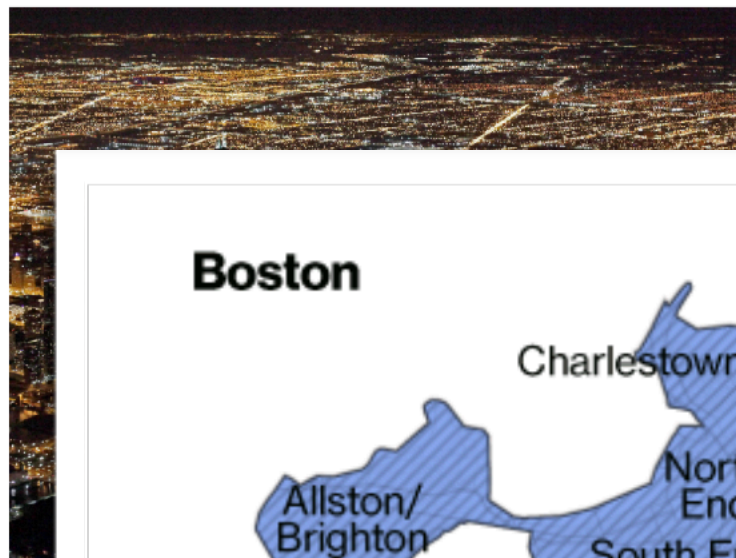


A Bloomberg report Thursday revealed that Amazon's same-day delivery service offered to Prime users around major US cities seems to routinely, if unintentionally, exclude black neighborhoods.

The maps, which you should check out on Bloomberg's site, show that in cities like Chicago, New York, and Atlanta, same-day delivery is not available at this point — except the north side of the city.

But the thing is that Amazon's director of PR Scott Stanzel wrote in an email to Bloomberg:

There are a number of factors that affect whether we can deliver same-day. Those include population density, proximity to a distribution center, local demand in an area, as well as the ability of carriers to deliver packages to 9:00 pm every single day, even Sunday.



Chicago, Illinois



Recommended For You

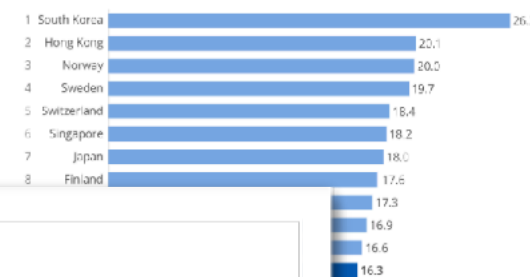


'None of it makes much sense': Experts are baffled by Comey's use of a fake Russian document to skirt the DOJ

Tech Chart of the Day

The Countries With The Fastest Internet

Average internet connection speed in Q3 2016 (in Mbps)



Source: Statista

Lenovo

BI Enterprise
on Twitter, Steve
investing
Zj1STQJ
Hm9NbE

2 hours ago

BI Enterprise
Levie is taking a
Bezos' playbook as
company for th...

Vs9FYN
2 hours ago

Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women

Isobel Asher Hamilton 4h



- Amazon tried building an artificial-intelligence tool to help with recruiting, but it showed a bias against women, [Reuters reports](#).
- Engineers reportedly found the AI was unfavorable toward female candidates because it had combed through male-dominated résumés to accrue its data.
- Amazon reportedly abandoned the project at the beginning of 2017.



disparate treatment: still not fair

Amazon wo
with hiring

Fairness: Demographic Parity

Compare subpopulation proportions



often called group discrimination

Fails to identify discrimination against individuals.

How group discrimination can fail

Europe

Asia



approve loans to all **green** deny
loans to all **purple** applicants

approve loans to all **purple** deny
loans to all **green** applicants

European and Asian discriminations cancel each other out,
and the group discrimination measure can be 0.

Fairness: Disparate Impact

Prohibits using a facially neutral practice that has an unjustified adverse impact on members of a protected class.

80% rule: Employer's hiring rates for protected groups may not differ by more than 80%.

Fairness: Delayed Impact

**Making seemingly fair decisions can
(but shouldn't), in the long term,
produce unfair consequences**

Liu et al., Delayed impact of fair machine learning. ICML 2018

Fairness: Predictive Equality

False positive rates should not differ

Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. FATML 2016
Corbett-Davies. Algorithmic decision making and the cost of fairness. KDD 2017

Fairness: Equal Opportunity

False negative rates should not differ

Hardt et al. Equality of Opportunity in Supervised Learning. NIPS 2016
Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments FATML 2016

Fairness: Equality of Odds

Proportionality of Quality of Opportunity

Learning. NIPS 2016

Fairness: Equality of Opportunity

Condition of Equality of Opportunity

Berk et al. Fairness in criminal justice

rt. Sociol. 2018

HOW STANDARDS PROLIFERATE:

(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)

SITUATION:
THERE ARE
14 COMPETING
STANDARDS.

14?! RIDICULOUS!
WE NEED TO DEVELOP
ONE UNIVERSAL STANDARD
THAT COVERS EVERYONE'S
USE CASES.



SOON:

SITUATION:
THERE ARE
15 COMPETING
STANDARDS.

Fairness: Correlation

correlation(race, **APPROVED**) = 0.8

mutual information(race, **APPROVED**) = 0.6

Correlation does not measure causation

What is fairness?

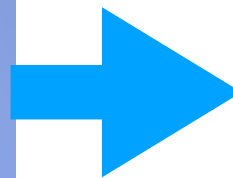
Sensitive inputs should not affect software behavior.

We want to measure causality!

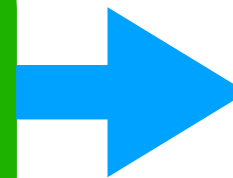
causal testing

Sensitive inputs should not affect software behavior.

hypothesis
testing:



LOAN



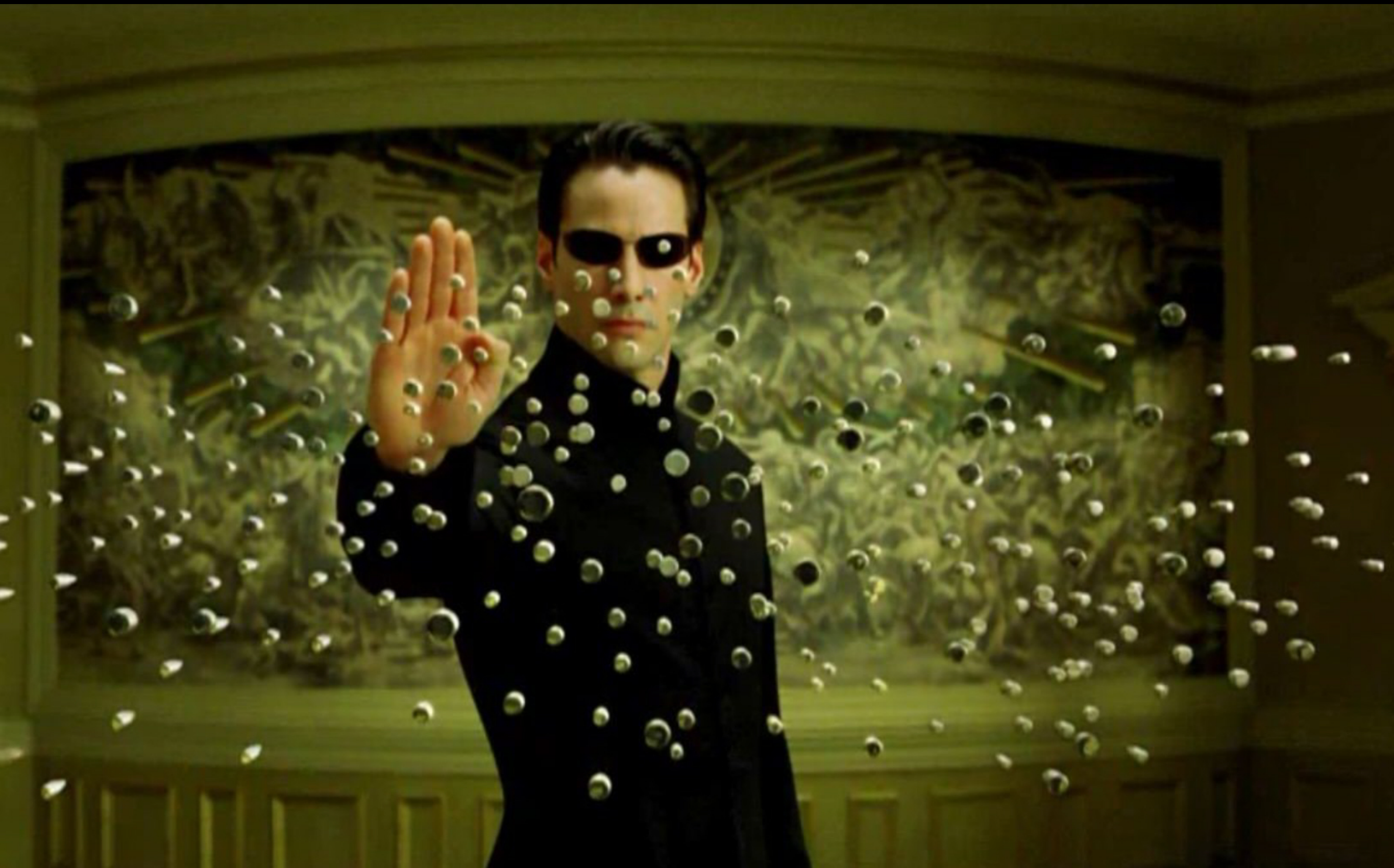
PROVED

causal testing



No need for an oracle!

causal testing



Themis

automated test-suite generator



How much does my software
discriminate with respect to ...?

Does my software discriminate more
than 10% of the time, and against what?

Themis generates a test suite or can use a manually written one

<http://fairness.cs.umass.edu>

discrimination measures

causal discrimination

$$\text{LOAN}(\text{img1}) \stackrel{?}{=} \text{LOAN}(\text{img2})$$

group discrimination



How does Themis work?

adaptive, confidence-driven sampling

input schema

confidence

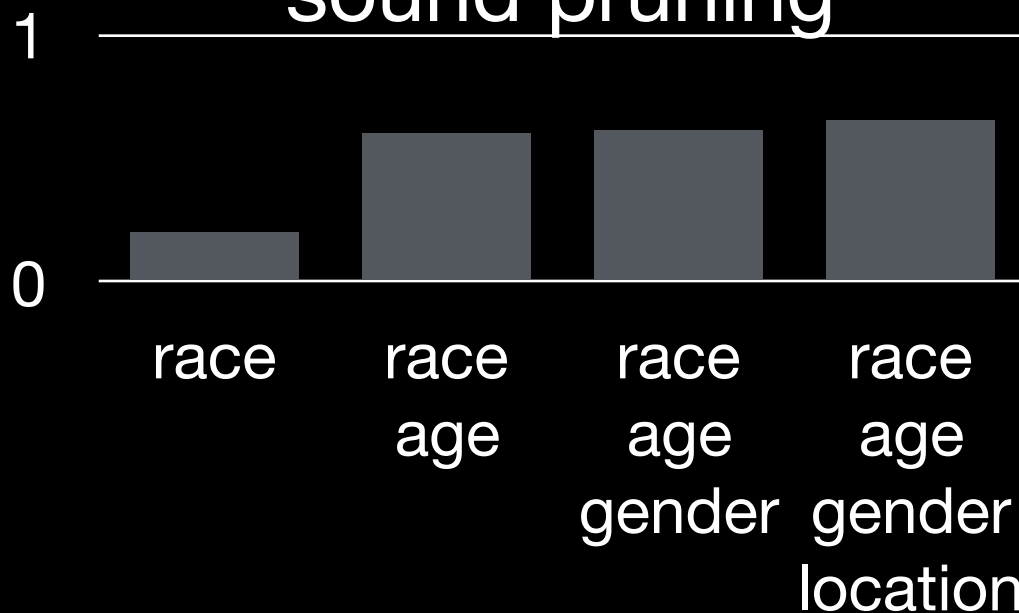
error bound



Themis

$$error = z^* \sqrt{\frac{p(1-p)}{r}}$$

sound pruning



Evaluation

Eight open-source decision systems trained on two public data sets

discrimination-aware logistic regression	[88]
discrimination-aware decision tree	[40]
discrimination-aware naive Bayes	[18]
discrimination-aware decision tree	[91]
naive Bayes	scikit-learn
decision tree	
logistic regression	
SVM	

- Census income dataset:
financial data
45K people
income > \$50K?
- Statlog German credit dataset:
credit data
1K people
“good” or “bad” credit?

findings

Group discrimination is not enough.

More than 11% of the individuals had the output flipped just by altering the individual's gender.

Decision tree trained not to group discriminate against gender causal discriminated against gender: 0.11.

findings

Trying to avoid group discrimination may introduce other discrimination.

Training a decision tree not to discriminate against gender made it discriminate against race 38.4% of the time.

findings

Pruning is highly effective.

- The more a system discriminates, the more efficient Themis is.
- On average, pruning reduced test suites by **148x** for causal and **2,849x** for group discrimination. Best improvement was **13,000x**.

What are we doing now?



Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots



By [Jacob Snow](#), Technology & Civil Liberties Attorney, ACLU of Northern California

JULY 26, 2018 | 8:00 AM

TAGS: [Face Recognition Technology](#), [Surveillance Technologies](#), [Privacy & Technology](#)



“The false matches were disproportionately of people of color, including six members of the Congressional Black Caucus, among them civil rights legend Rep. John Lewis (D-Ga.).”

nationwide, and today, there are 28 more causes for concern. In a test the ACLU recently conducted of the facial recognition tool, called “Rekognition,” the software incorrectly matched 28 members of Congress, identifying them as other people who have been arrested for a crime.

The members of Congress who were falsely matched with the mugshot



What are we doing now?

ACLU

GET UPDATES / DONATE

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots

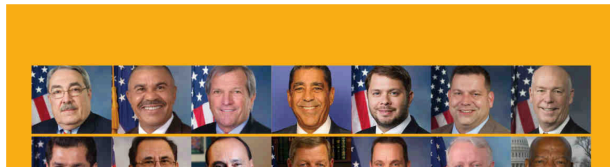


By [Jacob Snow](#), Technology & Civil Liberties Attorney, ACLU of Northern California
JULY 26, 2018 | 8:00 AM

TAGS: [Face Recognition Technology](#), [Surveillance Technologies](#), [Privacy & Technology](#)



Amazon's face surveillance technology is the target of growing opposition nationwide, and today, there are 28 more causes for concern. In a test the ACLU recently conducted of the facial recognition tool, called "Rekognition."



Fair computer vision



<https://thispersondoesnotexist.com/>

What are we doing now?

ACLU GET UPDATES / DONATE

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots

By Jacob Snow, Technology & Civil Liberties Attorney, ACLU of Northern California
JULY 26, 2018 | 8:00 AM

TAGS: [Face Recognition Technology](#), [Surveillance Technologies](#), [Privacy & Technology](#)

Amazon's face surveillance technology is the target of growing opposition nationwide, and today, there are 28 more causes for concern. In a test the ACLU recently conducted of the facial recognition tool, called "Rekognition," the software incorrectly matched 28 members of Congress, identifying them as other people who have been arrested for a crime.



The members of Congress who were falsely matched with the mugshot

Fair computer vision

Fair natural
language processing



English Spanish Turkish Detect language Translate

He is a nurse.
She is a doctor.

O bir hemşire.
O bir doktor.

English Spanish Turkish Detect language Translate

O bir hemşire.
O bir doktor.

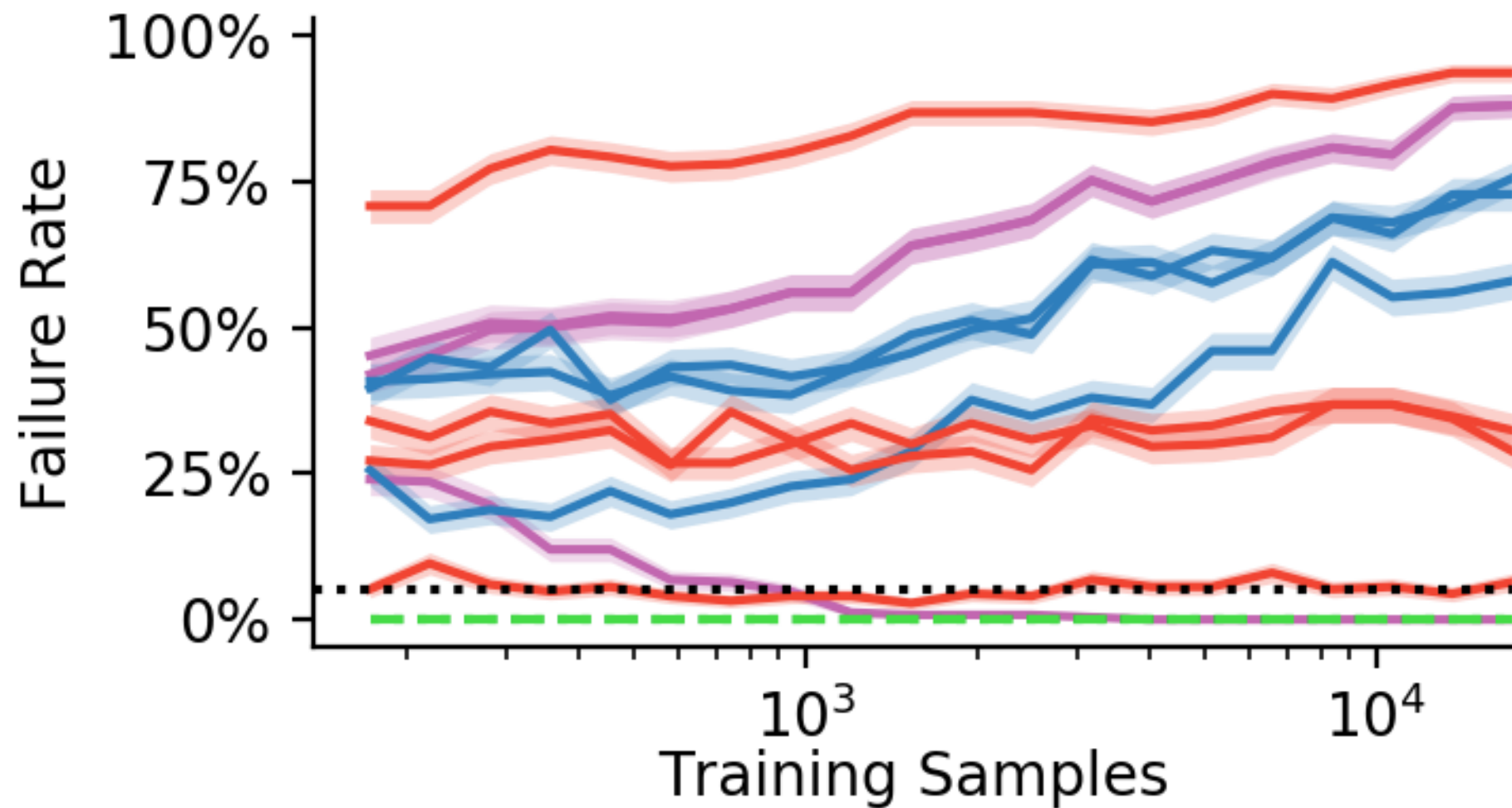
She is a nurse.
He is a doctor.

But what's the holy grail?

Provably fair machine learning:

Provide (high-probability)
guarantees that the classifier
is fair on unseen data.

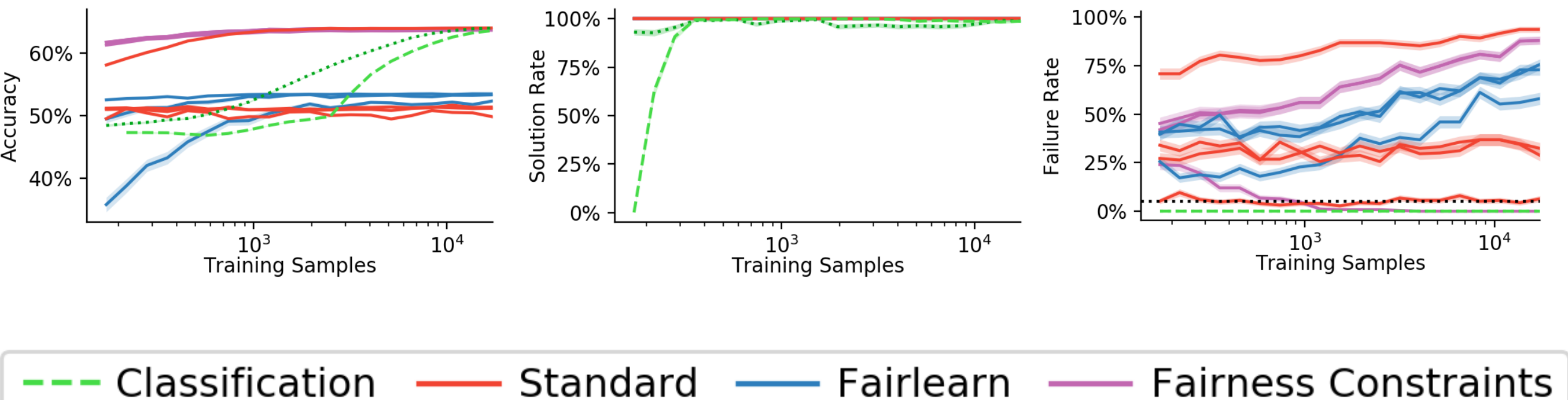
Disparate Impact



--- Classification — Standard — Fairlearn — Fairness Constraints

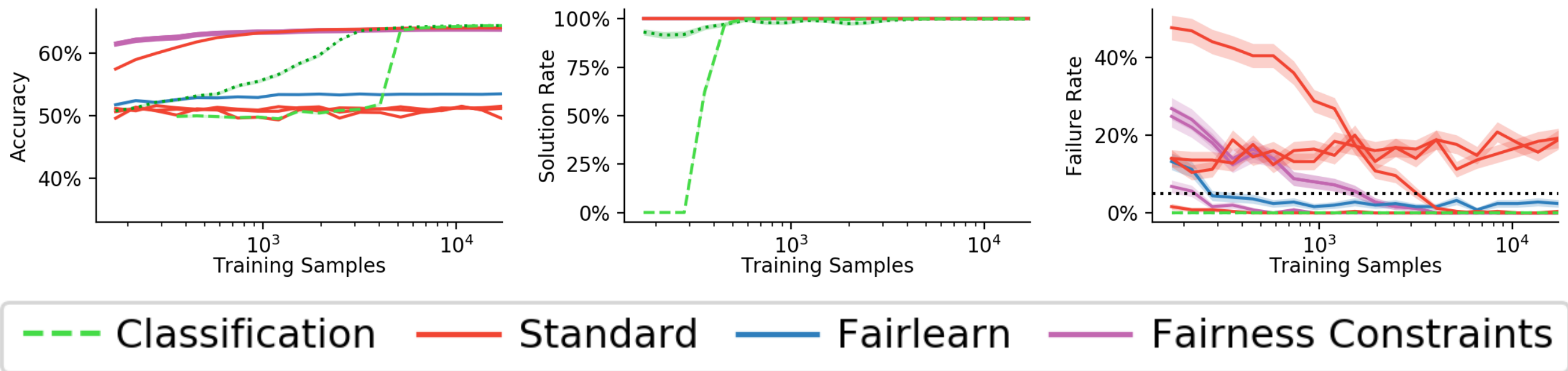
Fairlearn: Agarwal et al. A reductions approach to fair classification. ICML 2018.
Fairness Constraints: Zafar et al., Fairness Constraints: A Mechanism for Fair Classification. FATML 2015.

Disparate Impact

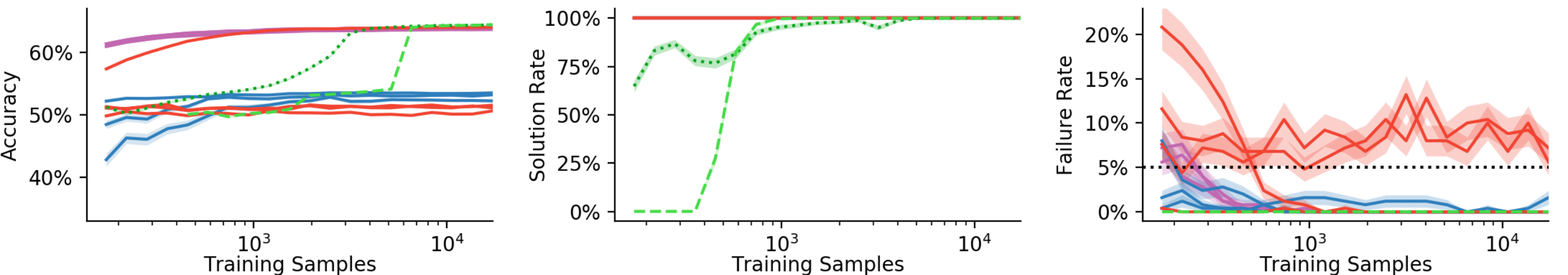


Fairlearn: Agarwal et al. A reductions approach to fair classification. ICML 2018.
Fairness Constraints: Zafar et al., Fairness Constraints: A Mechanism for Fair Classification. FATML 2015.

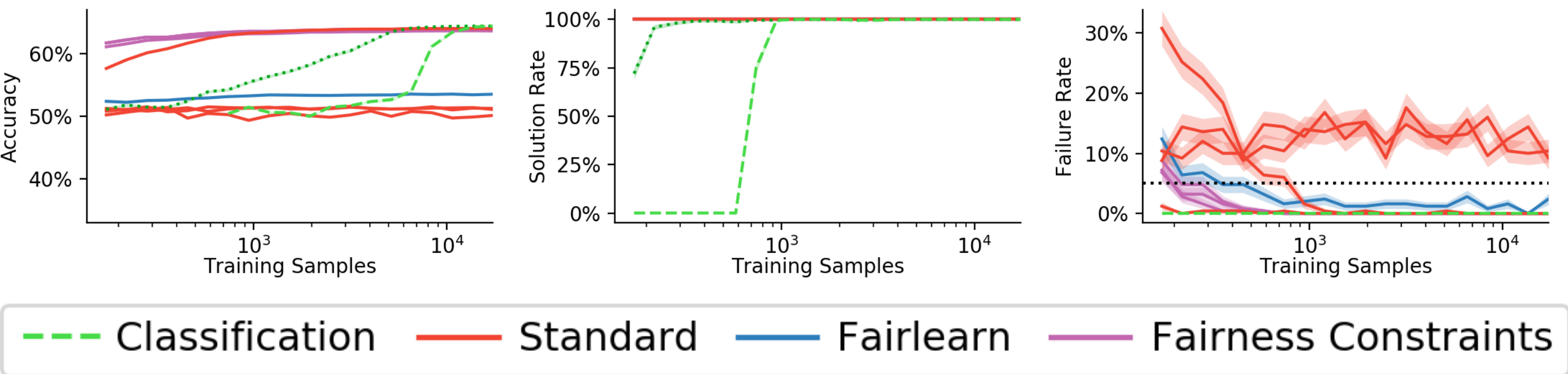
Demographic Parity



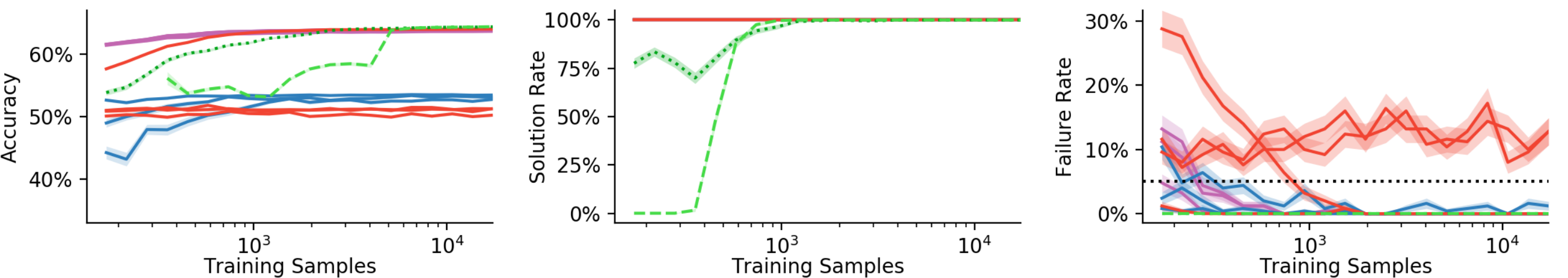
Equal Opportunity



Equalized Odds



Predictive Equality



Contributions

<http://fairness.cs.umass.edu>



- **Causality-based** definition and method for measuring software fairness
- Themis, an **automated test-suite generator** for fairness testing
- Evaluation on real-world software, demonstrating software is biased and **our methods can catch it**
- **Provable guarantees** on fairness in machine learning



Rico Angell



Brittany Johnson



Stephen Giguere



Sarah Brockman



Blossom Metevier



Sainyam Galhotra



Alexandra Meliou



Andy Barto



Bruno Castro
da Silva



Emma Brunskill



Philip Thomas



Yuriy Brun

<http://fairness.cs.umass.edu>

<https://tinyurl.com/FairnessPaper>



UMassAmherst

ORACLE®

Contributions

<http://fairness.cs.umass.edu>



- **Causality-based** definition and method for measuring software fairness
- Themis, an **automated test-suite generator** for fairness testing
- Evaluation on real-world software, demonstrating software is biased and **our methods can catch it**
- **Provable guarantees** on fairness in machine learning

Homework 1

- Due September 17, 9AM
- Will be posted shortly (you'll get an email)
- Learn some machine learning! Learn to use tools that help evaluate and mitigate bias in machine learning.
- Requires downloading a 5GB file, so do that early.