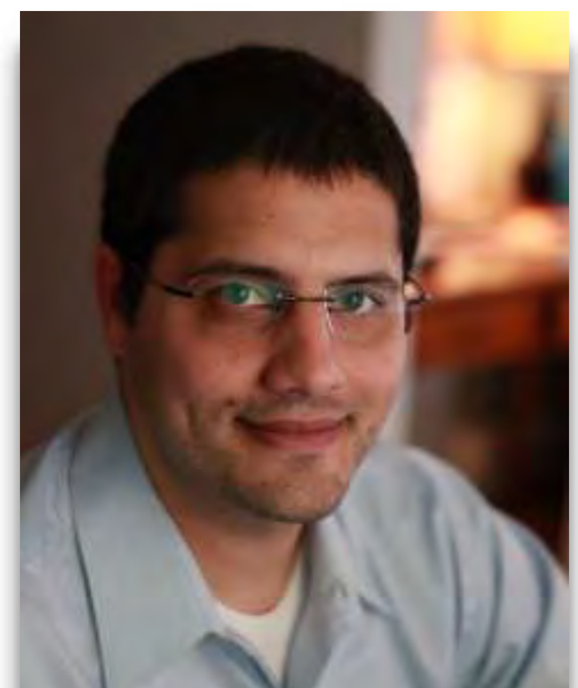
# Engineering Software to Prevent Undesirable Behavior of Intelligent Machines

Alexandra Meliou



Philip Thomas

Yuriy Brun

UMass Amherst

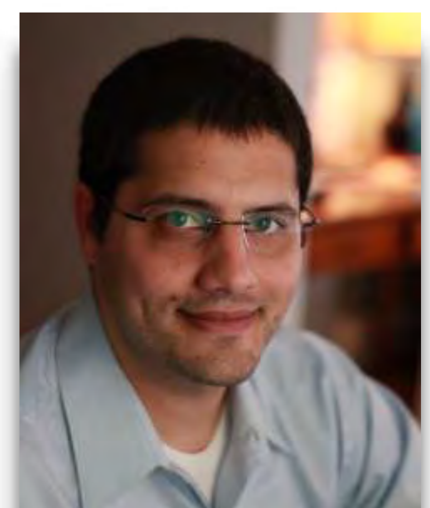
http://fairness.cs.umass.edu

http://aisafety.cs.umass.edu

# Engineering Software to Prevent Undesirable Behavior of Intelligent Machines

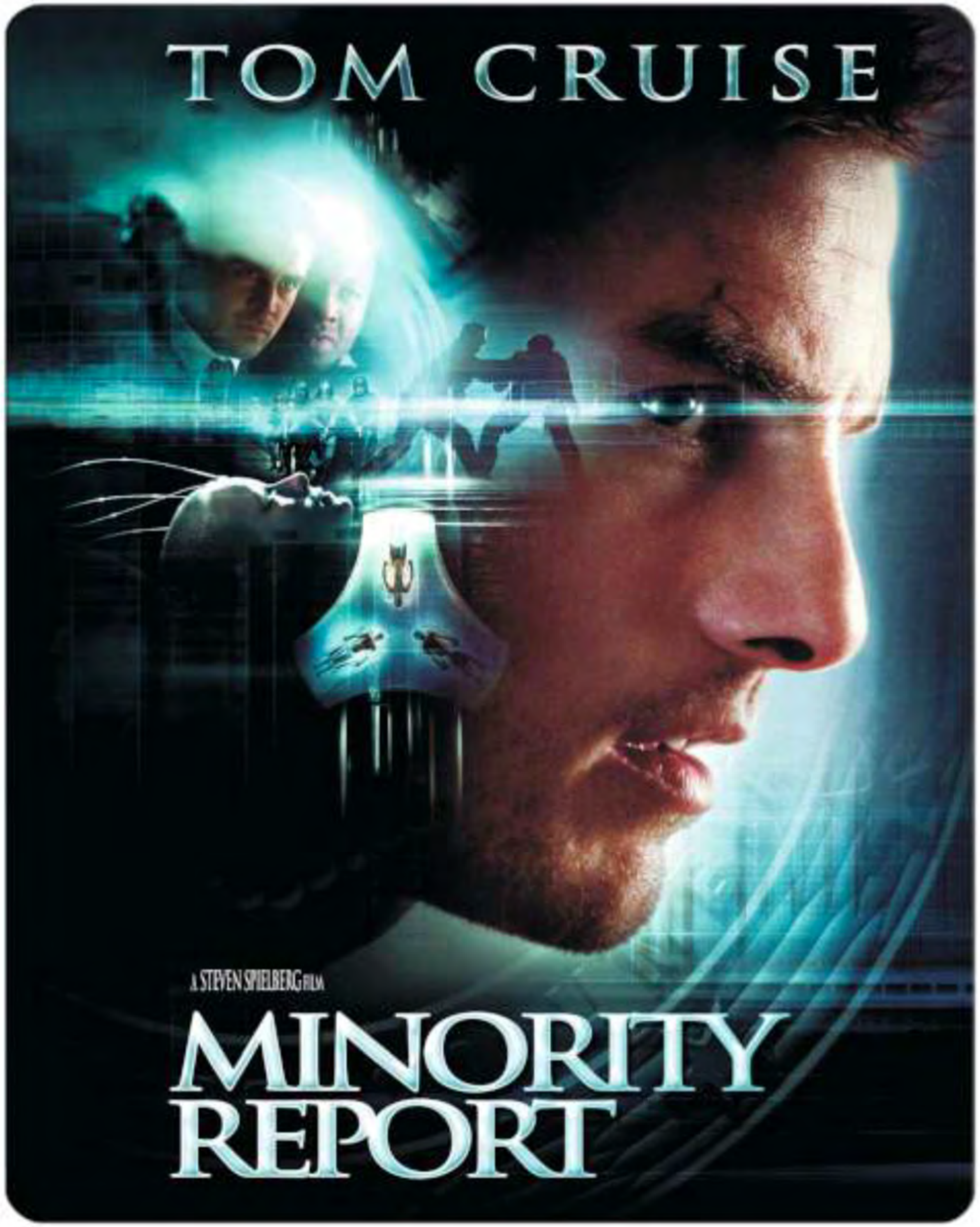

Yuriy


Alexandra


Philip

CCF-1763423
CCF-1744471
CCF-1453474

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.


TOM CRUISE
A STEVEN SPIELBERG FILM
MINORITY
REPORT


**Resilient cities** Cities

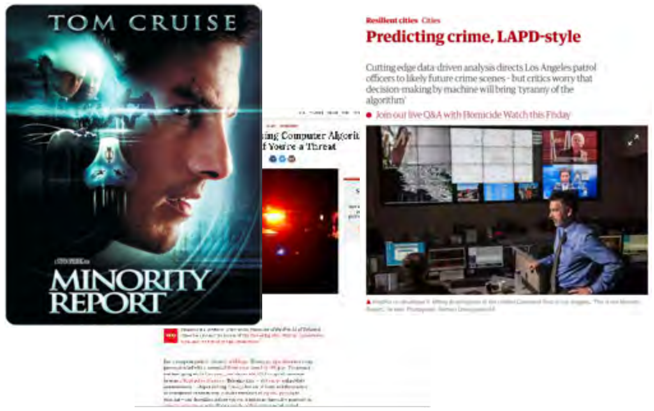
# Predicting crime, LAPD-style

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

https://www.theguardian.com/cities/2014/jun/25/predicting-crime-lapd-los-angeles-police-data-analysis-algorithm-minority-report




ACLU GET UPDATES / DONATE

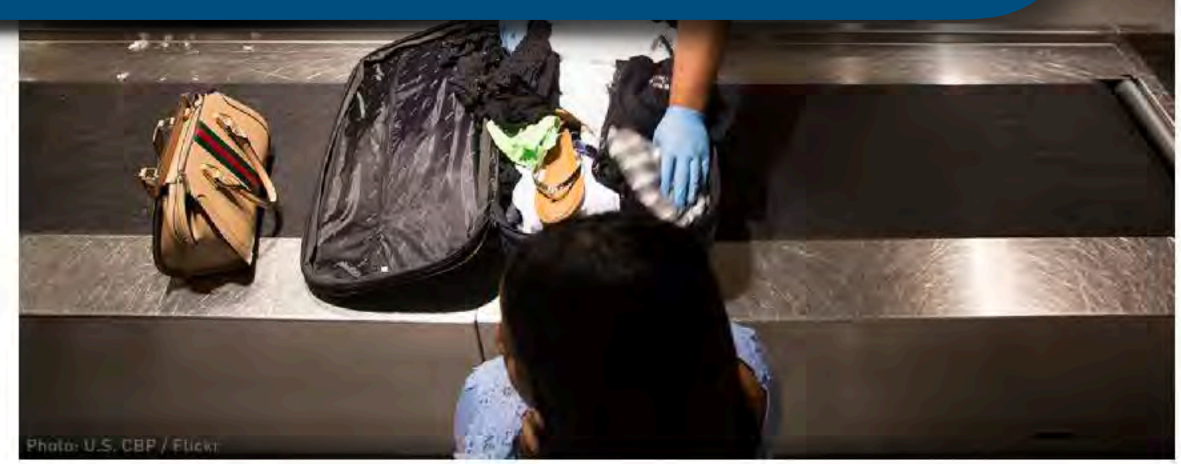
# The Government Is Blacklisting People Based on Predictions of Future Crimes

By Hina Shamsi, Director, ACLU National Security Project

TAG

Ima
with
blac
and
You
memb
funerals or religious obligations, and lose jobs because you can't travel or your employer finds out you're blacklisted.

You know what the government has done violates your constitutionally protected ability to travel and to be free from false stigma. You have

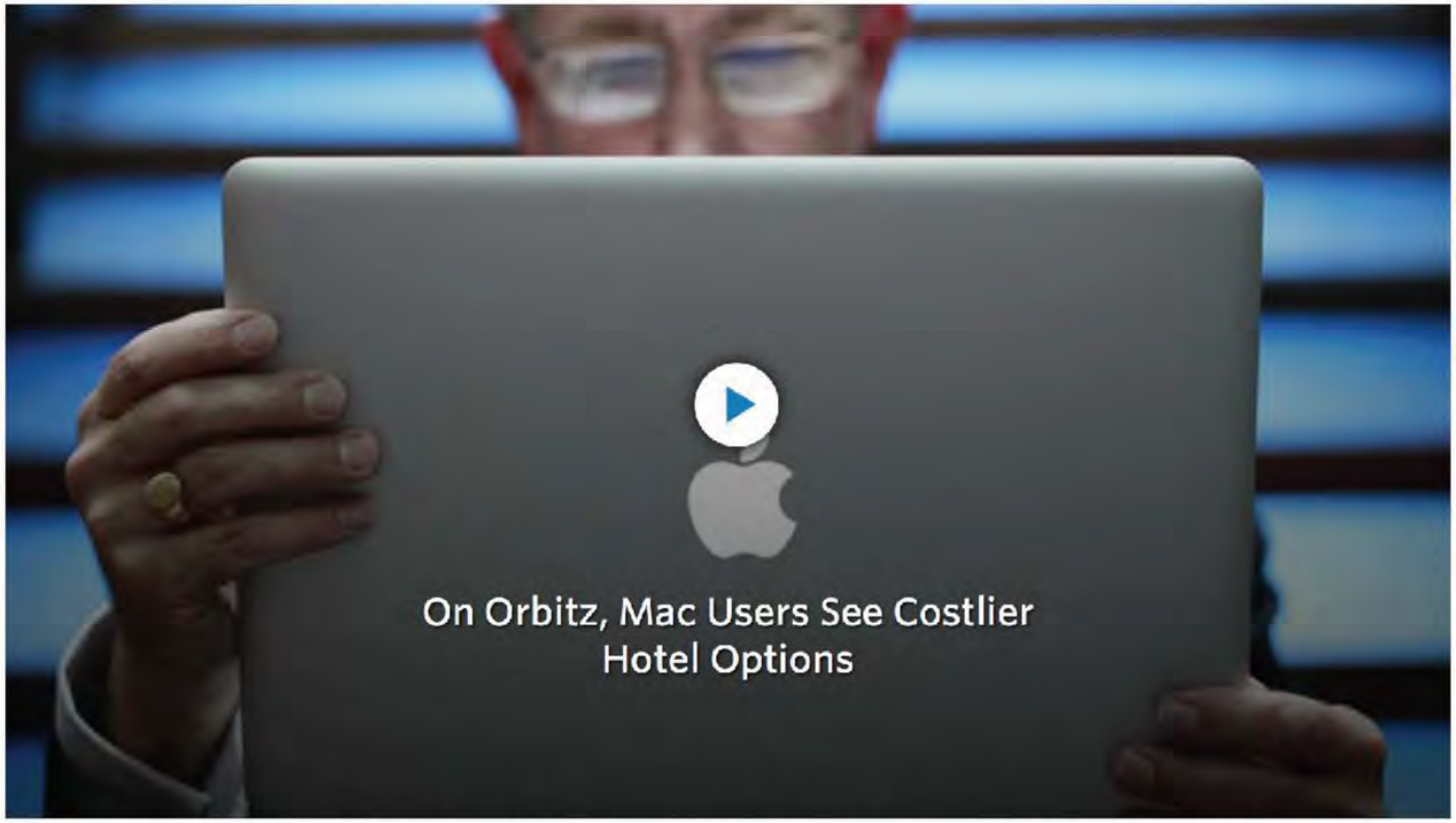



## Modern software influences critical decisions

https://www.aclu.org/blog/national-security/discriminatory-profiling/government-blacklisting-people-based-predictions

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

forbes.com

Forbes LOG IN

MAR 15, 2011 @ 12:55 PM 16,947

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

# MIT News

ON CAMPUS AND AROUND THE WORLD

SUBSCRIBE

## Using AI to predict breast cancer and personalize care

MIT/MGH's image-based deep learning model can predict breast cancer up to five years in advance.

Ada

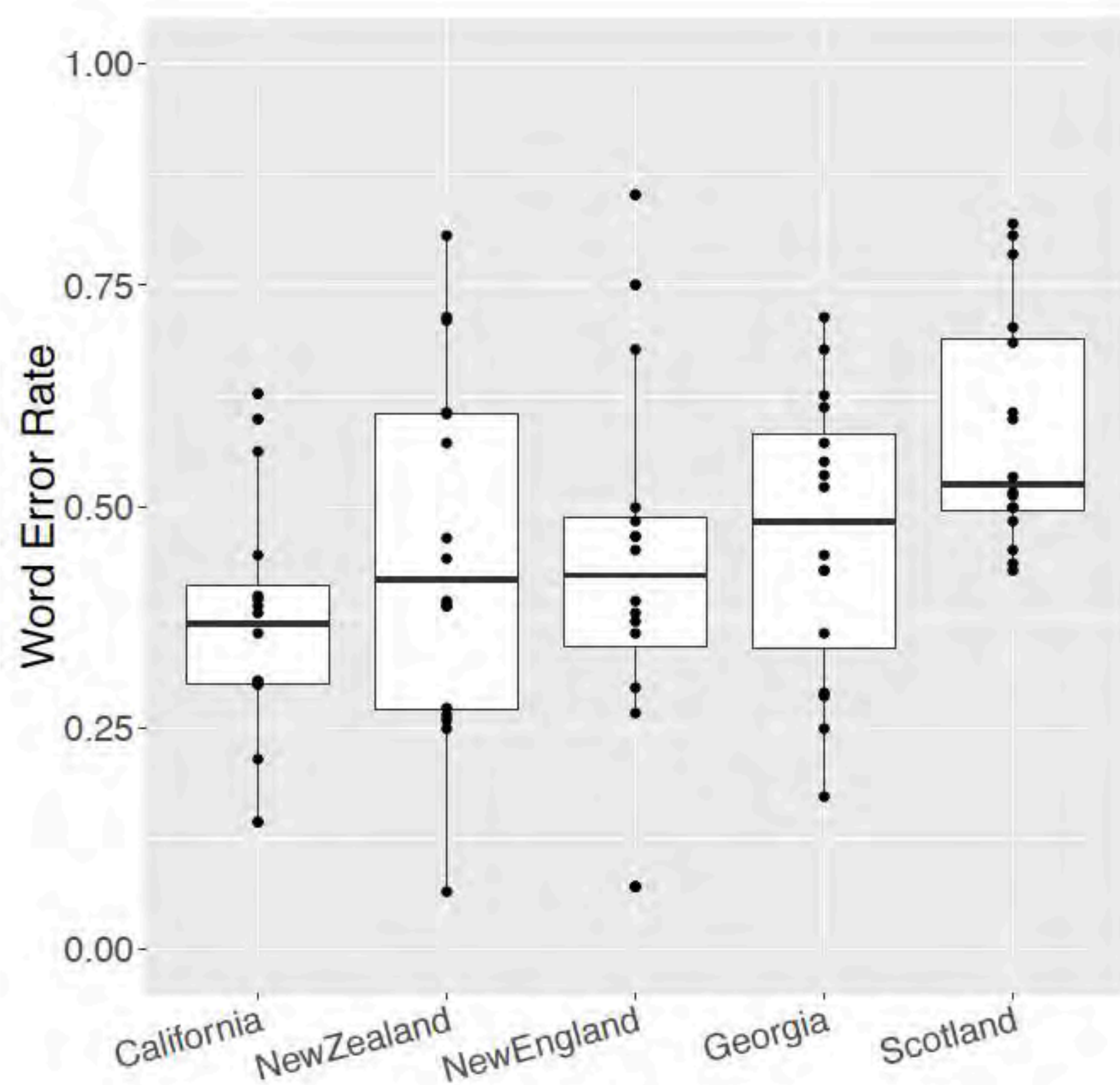
M

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

Google Translate

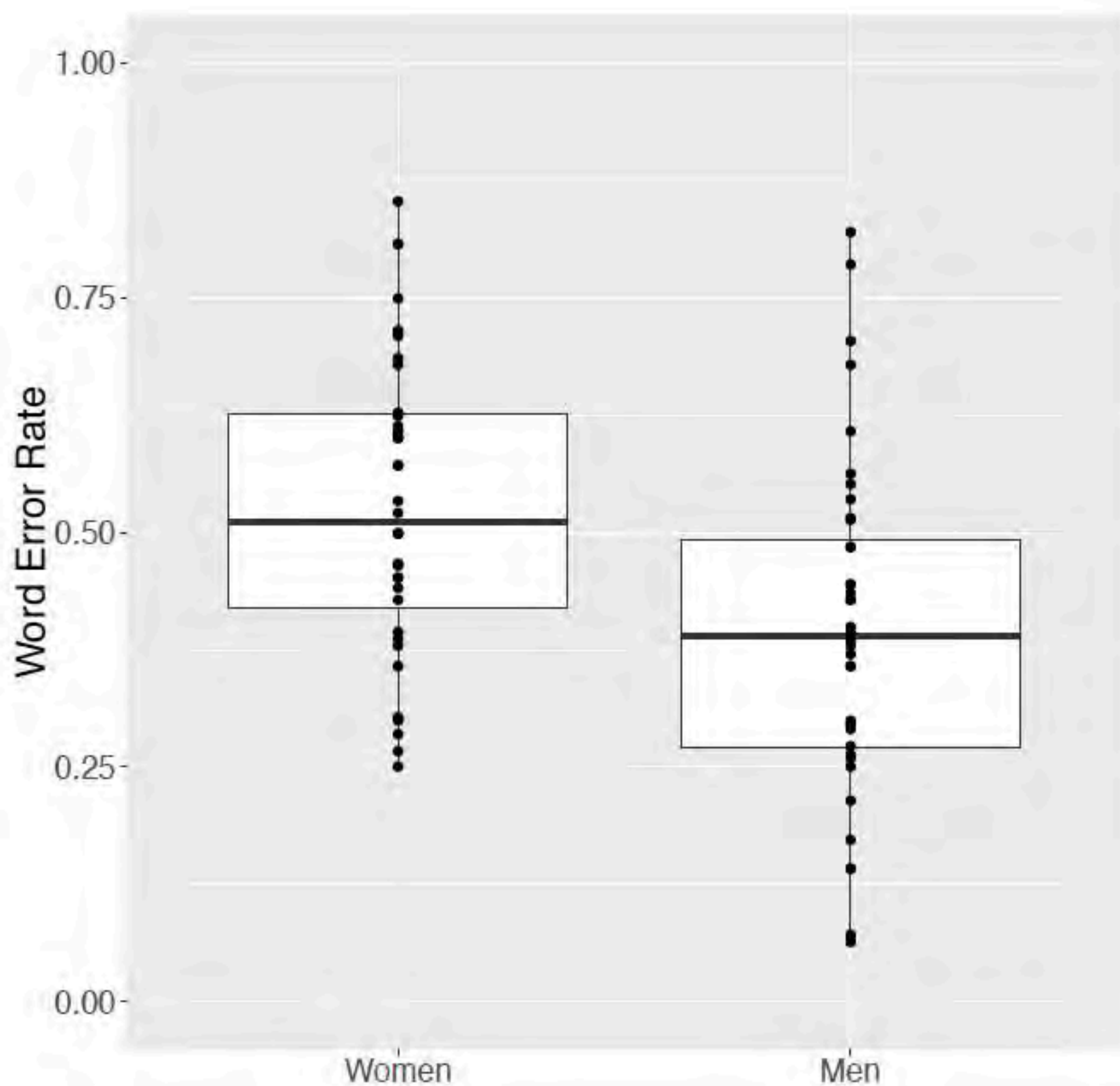
YouTube

# YouTube automatic captions




Rachael Tatman, "Gender and Dialect Bias in YouTube's Automatic Captions" in 2017 Workshop on Ethics in Natural Language Processing

# YouTube automatic captions




Rachael Tatman, "Gender and Dialect Bias in YouTube's Automatic Captions" in 2017 Workshop on Ethics in Natural Language Processing

Artificial intelligence Dec 20

# A US government study confirms most face recognition systems are racist


Almost 200 face recognition algorithms—a majority in the industry—had worse performance on nonwhite faces, according to a landmark study.

**What they tested:** The US National Institute of Standards and Technology (NIST) tested every algorithm on two of the most common tasks for face recognition. The first, known as "one-to-one" matching, involves matching a photo of someone to

**Joy Buolamwini**
https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

# today's goals

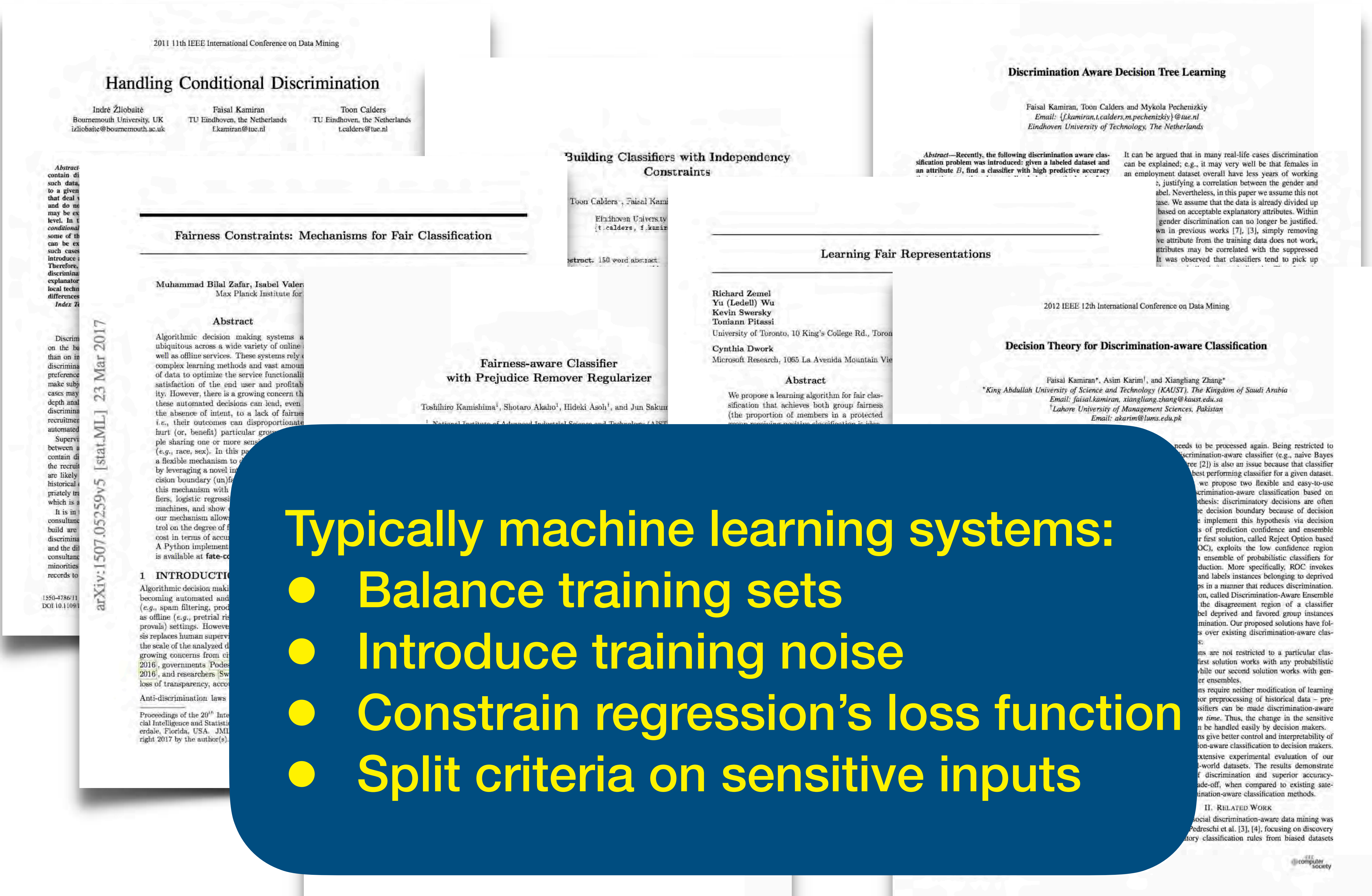
Define software discrimination.

Operationalize measuring discrimination through causal software testing.

Provide provable fairness guarantees.

Discuss fairness research landscape.

# Design software to be fair




Typically machine learning systems:

- Balance training sets
- Introduce training noise
- Constrain regression’s loss function
- Split criteria on sensitive inputs

Design alone is not enough

# possible causes

biased data


poor design

implementation bugs

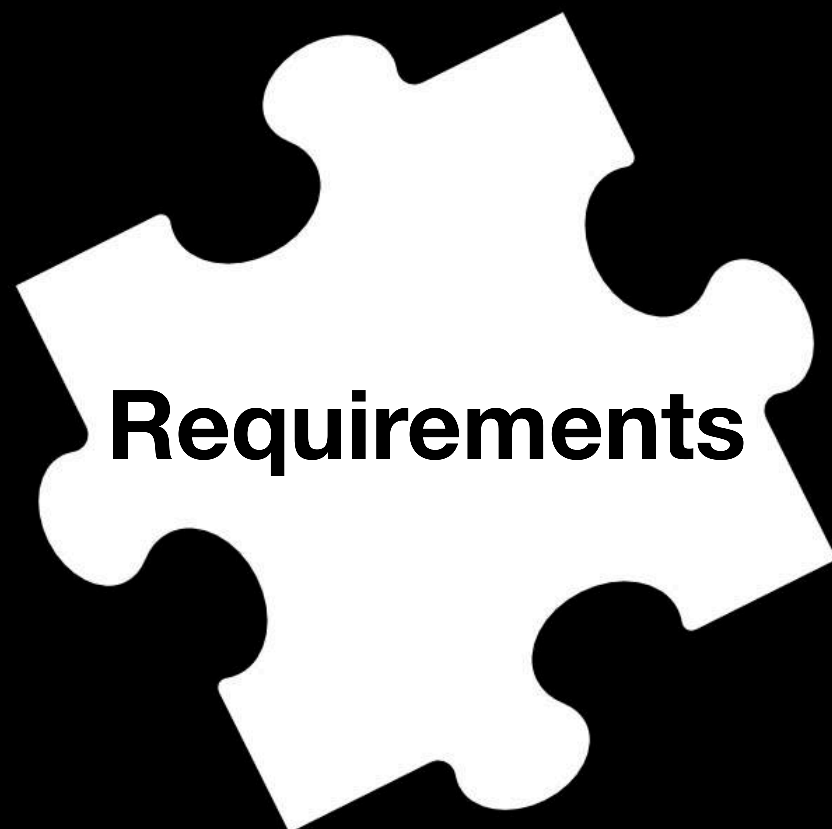
unintended interactions and mismatched components

# Fairness is just like quality and security

Fairness must be part of the software engineering lifecycle

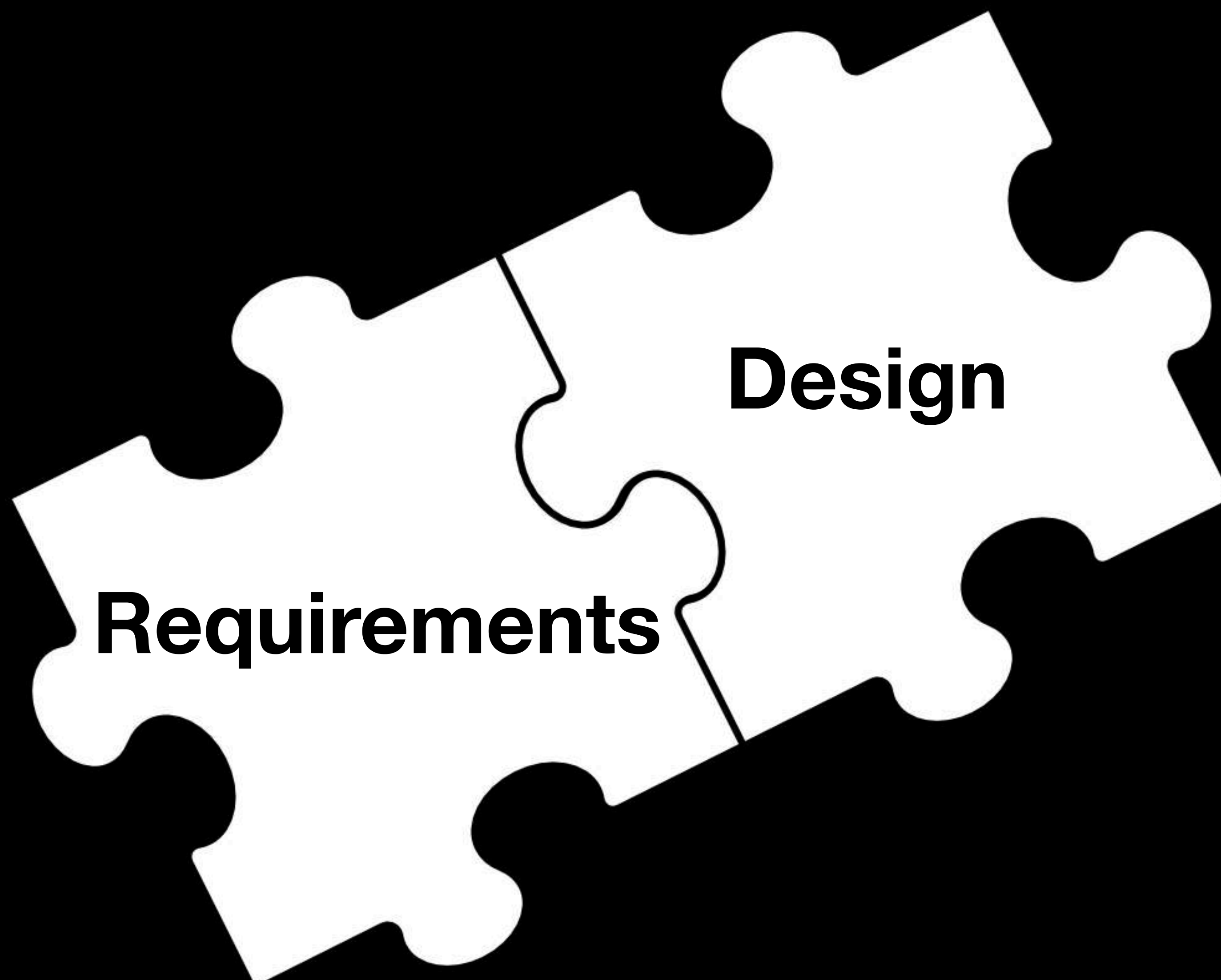
# Call to Action!

Fairness must be part of the software engineering lifecycle




**We need methods for specifying fairness requirements**

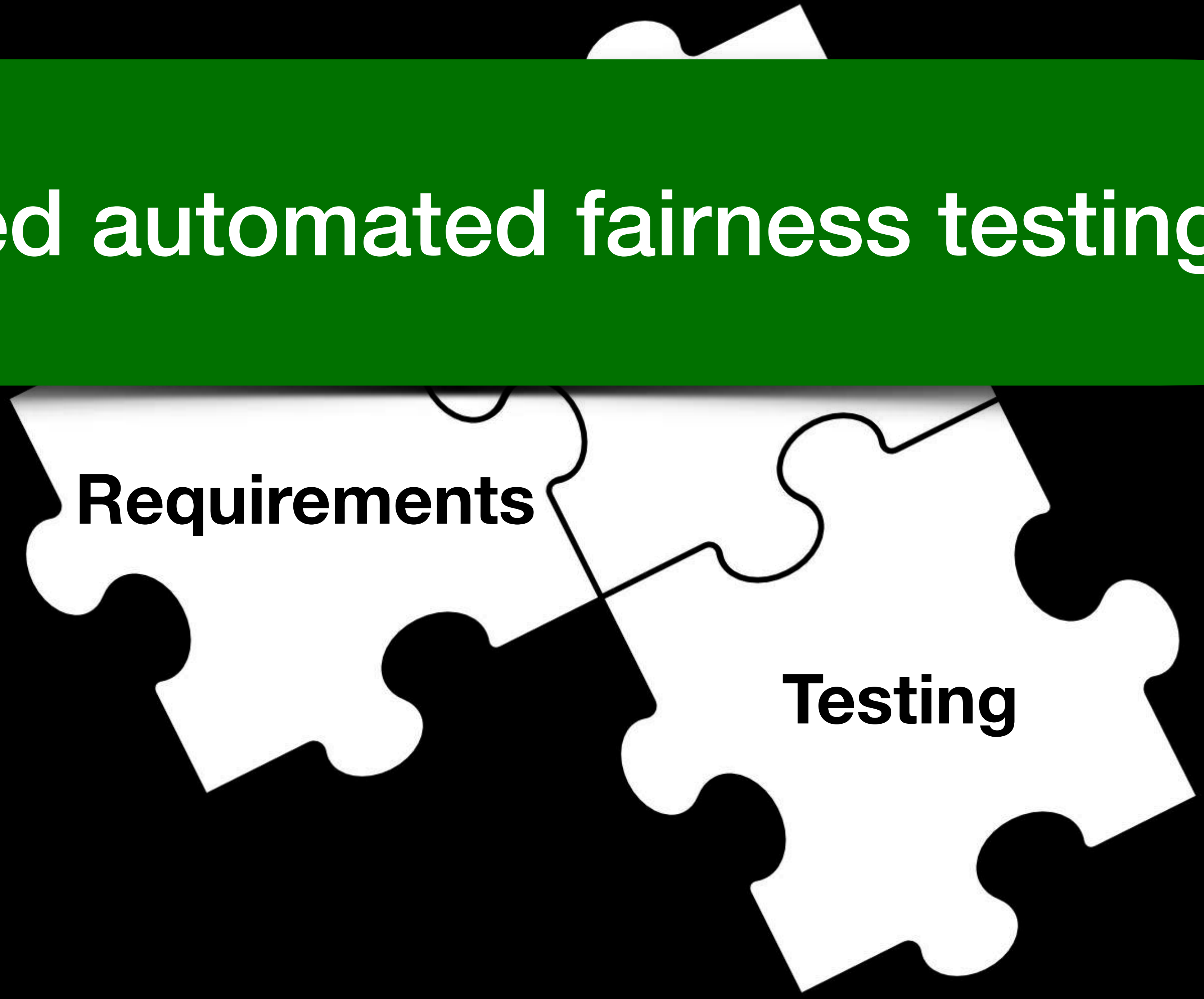
# Call to Action!




**We need fairness design principles**

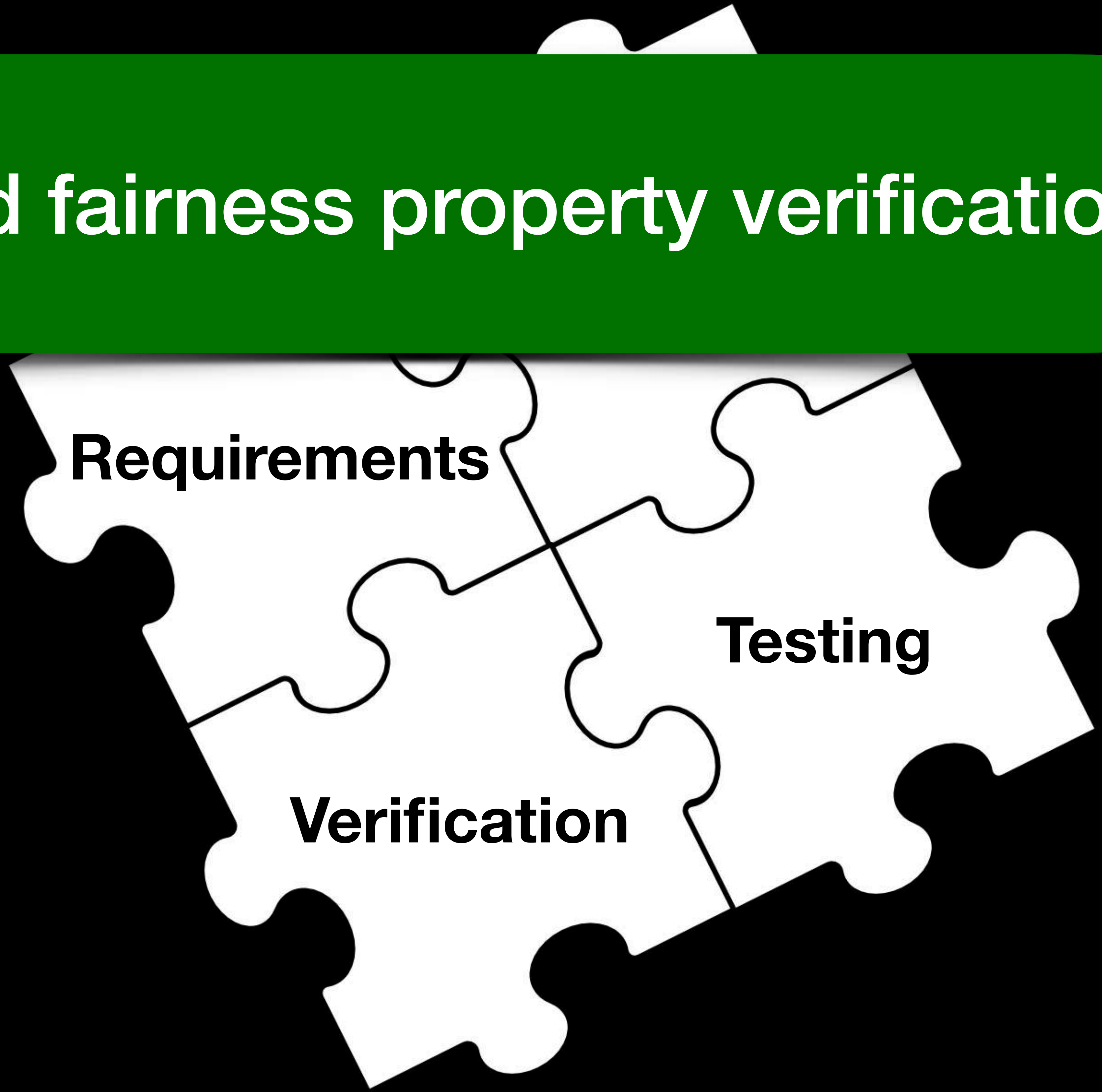
# Call to Action!

**We need automated fairness testing**



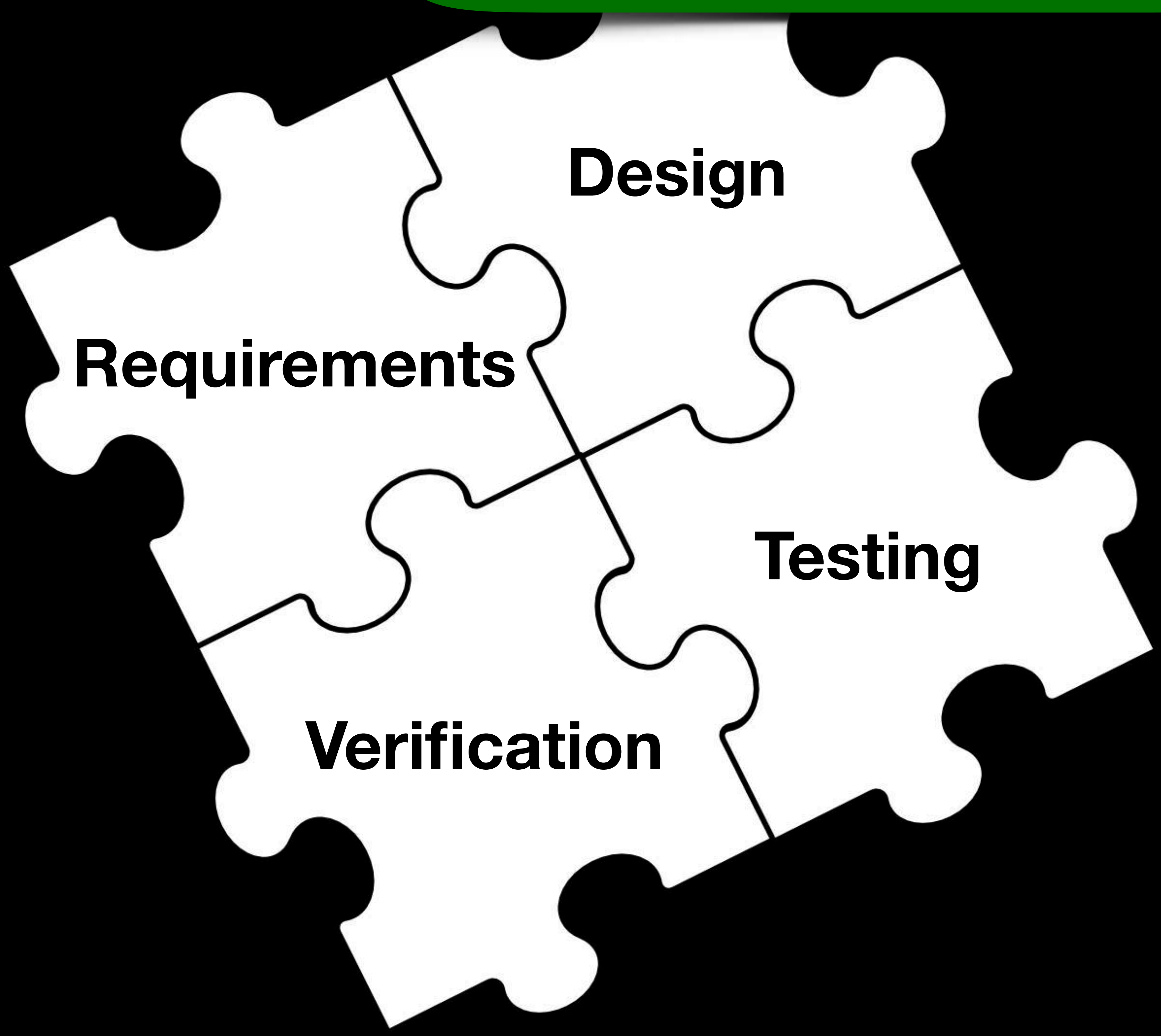
# Call to Action!

We need fairness property verification



# Call to Action!

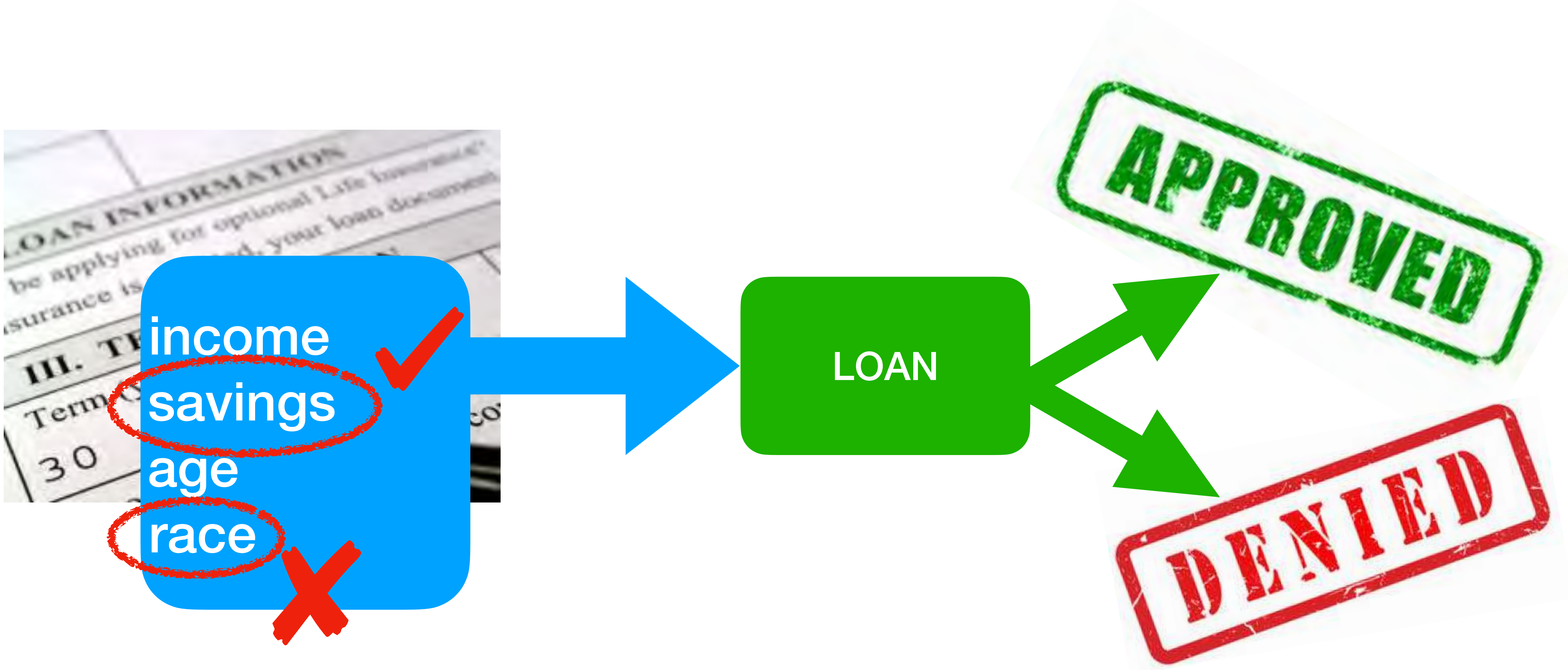
Fairness must be part of the software engineering lifecycle



Let’s talk about requirements.

What does it mean for software to discriminate?

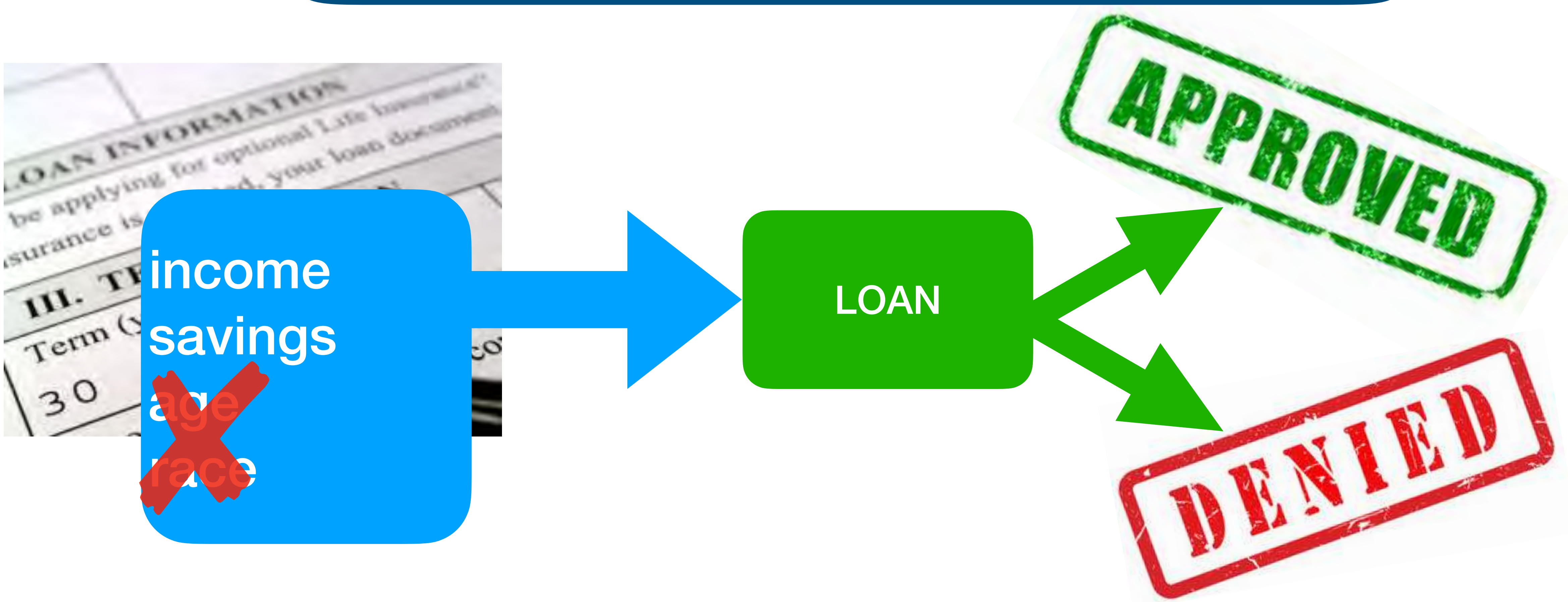
# LOAN program




This talk is not about policy.

# Fairness: Disparate Treatment

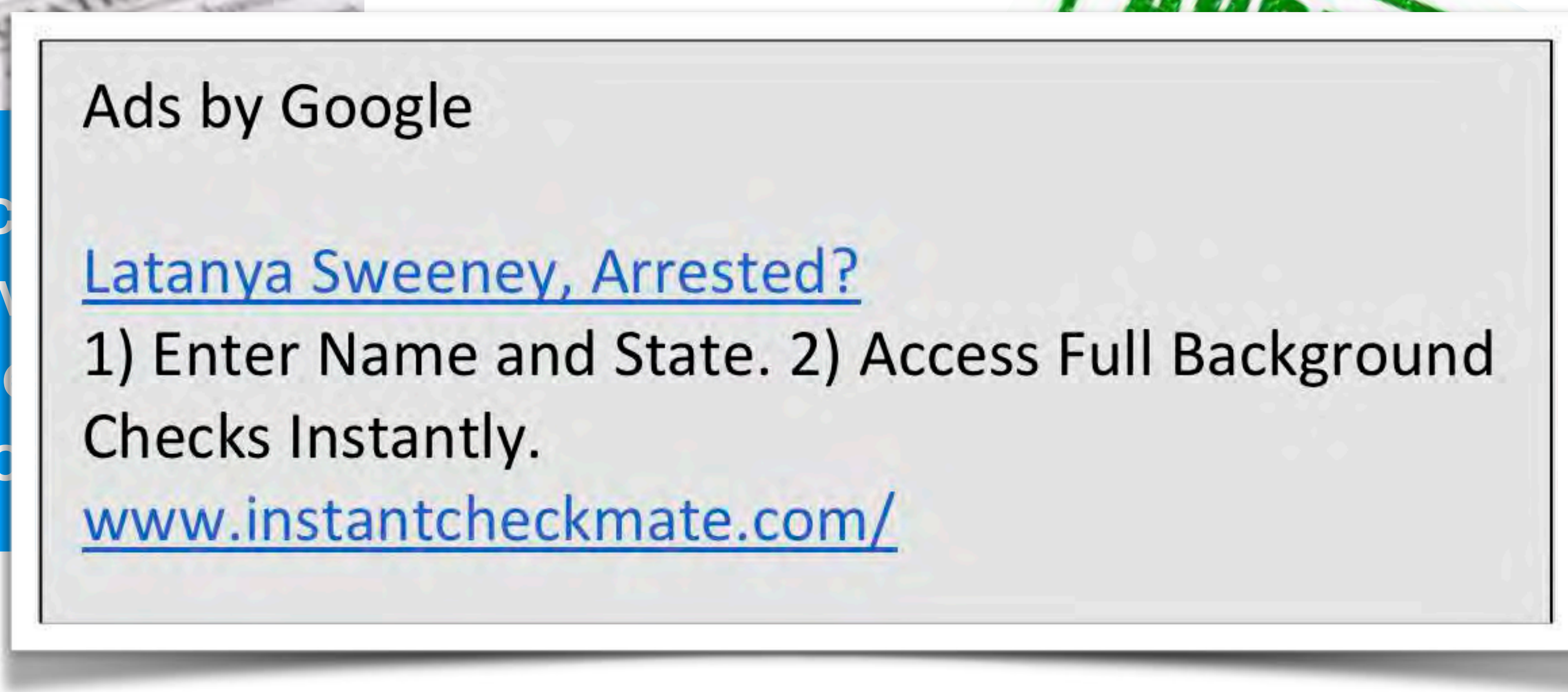
## Hide the data




Zafar et al. Fairness constraints: Mechanisms for fair classification. AISTATS 2017.

# Fairness: Disparate Treatment

**Hide the data**




**Ineffective because of data correlation.**
[Latanya Sweeney. Discrimination in online ad delivery. CACM 2013]

businessinsider.com

BUSINESS INSIDER TECH NEWS

# 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-da[...] code at this point — except the m[...]

Chi[...]
Blo[...]

But the thing is that Amazon's d[...] of PR Scott Stanzel wrote in an e[...]

> There are a number of factor[...] deliver same-day. Those incl[...] center, local demand in an ar[...] area, as well as the ability of [...] to 9:00 pm every single day, even Sunday".


## 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


[...]gers: [...]en [...]is here. Lenovo

[...]BI_Enterprise [...]n Twitter, Steve [...] investing [...]Zj1STQJ [...]Hm9NbE 2 hours ago

[...]BI_Enterprise [...] Levie is taking a [...]ezos' playbook as [...]mpany for th... [...]Vs9FYN 2 hours ago

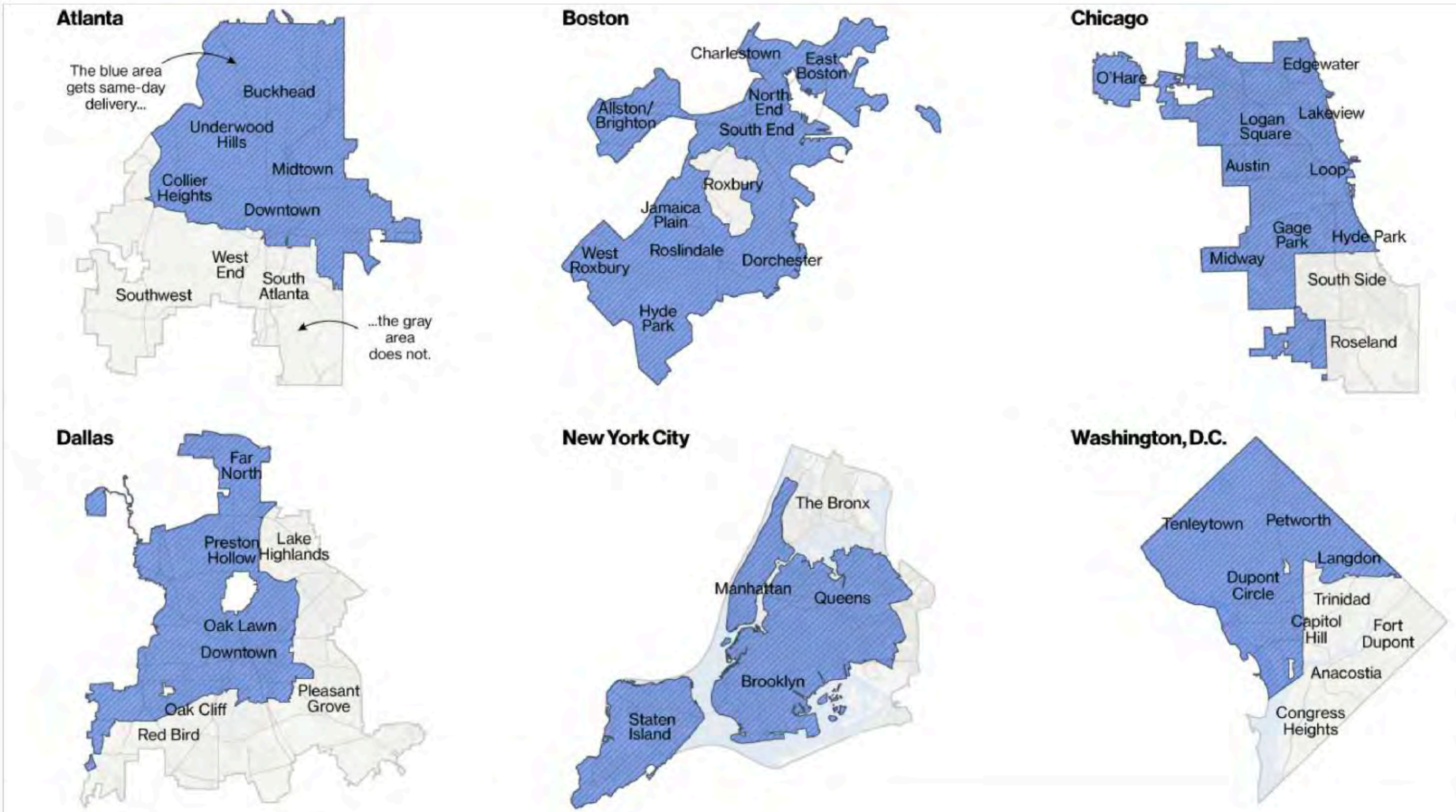
Real-time market data. Get the latest on stocks, commodities, currencies, funds, rates, ETFs, and businessinsider.com

BUSINESS INSIDER TECH NEWS

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

## Recommended For You

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




https://www.bloomberg.com/graphics/2016-amazon-same-day/

But the thing is that Amazon's defense here is entirely plausible. Director of PR Scott Stanzel wrote in an email to Tech Insider:

> There are a number of factors that go into determining where we can deliver same-day. Those include distance to the nearest fulfillment center, local demand in an area, numbers of Prime members in an area, as well as the ability of our various carrier partners to deliver up to 9:00 pm every single day, even Sunday .

2 hours ago


MARKETS INSIDER

Real-time market data. Get the latest on stocks, commodities, currencies, funds, rates, ETFs, and

# Fairness: Demographic Parity

often called group discrimination

## Compare subpopulation proportions


Fails to identify discrimination against individuals.


Dwork et al. Fairness through awareness. ITCS 2012.
Calders and Verwer. Three naive Bayes approaches for discrimination-free classification. DMKD 2010.

# How demographic parity can fail

B

## Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness

Michael Kearns [1] Seth Neel [1] Aaron Roth [1] Zhiwei Steven Wu [2]

**Abstract**

We introduce a new family of fairness definitions that interpolate between statistical and individual notions of fairness, obtaining some of the best properties of each. We show that checking whether these notions are satisfied is computationally hard in the worst case, but give practical oracle-efficient algorithms for learning subject to these constraints, and confirm our findings with experiments.

### 1. Introduction

As machine learning is being deployed in increasingly consequential domains (including policing (Rudin, 2013), criminal sentencing (Barry-Jester et al., 2015), and lending (Koren, 2016)), the problem of ensuring that learned models are *fair* has become urgent.

Approaches to fairness in machine learning can coarsely be divided into two kinds: *statistical* and *individual* notions of fairness. Statistical notions typically fix a small number of protected demographic groups $\mathcal{G}$ (such as racial groups), and then ask for (approximate) parity of some statistical measure across all of these groups. One popular statistical measure asks for equality of false positive or negative rates across all

One main attraction of statistical definitions of fairness is that they can in principle be obtained and checked without making any assumptions about the underlying population, and hence lead to more immediately actionable algorithmic approaches. On the other hand, individual notions of fairness ask for the algorithm to satisfy some guarantee which binds at the individual, rather than group, level. Individual notions of fairness have attractively strong semantics, but their main drawback is that achieving them seemingly requires more assumptions to be made about the setting under consideration.

The semantics of statistical notions of fairness would be significantly stronger if they were defined over a large number of *subgroups*, thus permitting a rich middle ground between fairness only for a small number of coarse pre-defined groups, and the strong assumptions needed for fairness at the individual level. Consider the kind of *fairness gerrymandering* that can occur when we only look for unfairness over a small number of pre-defined groups:

**Example 1.1.** *Imagine a setting with two binary features, corresponding to race (say black and white) and gender (say male and female), both of which are distributed independently and uniformly at random in a population. Consider a classifier that labels an example positive if and only if it corresponds to a black man, or a white woman. Then the classifier will appear to be equitable when one considers*

approve lo… **rple** deny
loans to al… icants

Europea… ther out,
and the demographic parity 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%.**

Zafar et al. Fairness constraints: Mechanisms for fair classification. AISTATS 2017.

# 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

# Fai rness: Equal odds

lictive eq

earning. NIPS 2016

# Fair Equality

to

eg

Berk et al. Fairness in criminal justice

rt. Sociol

18

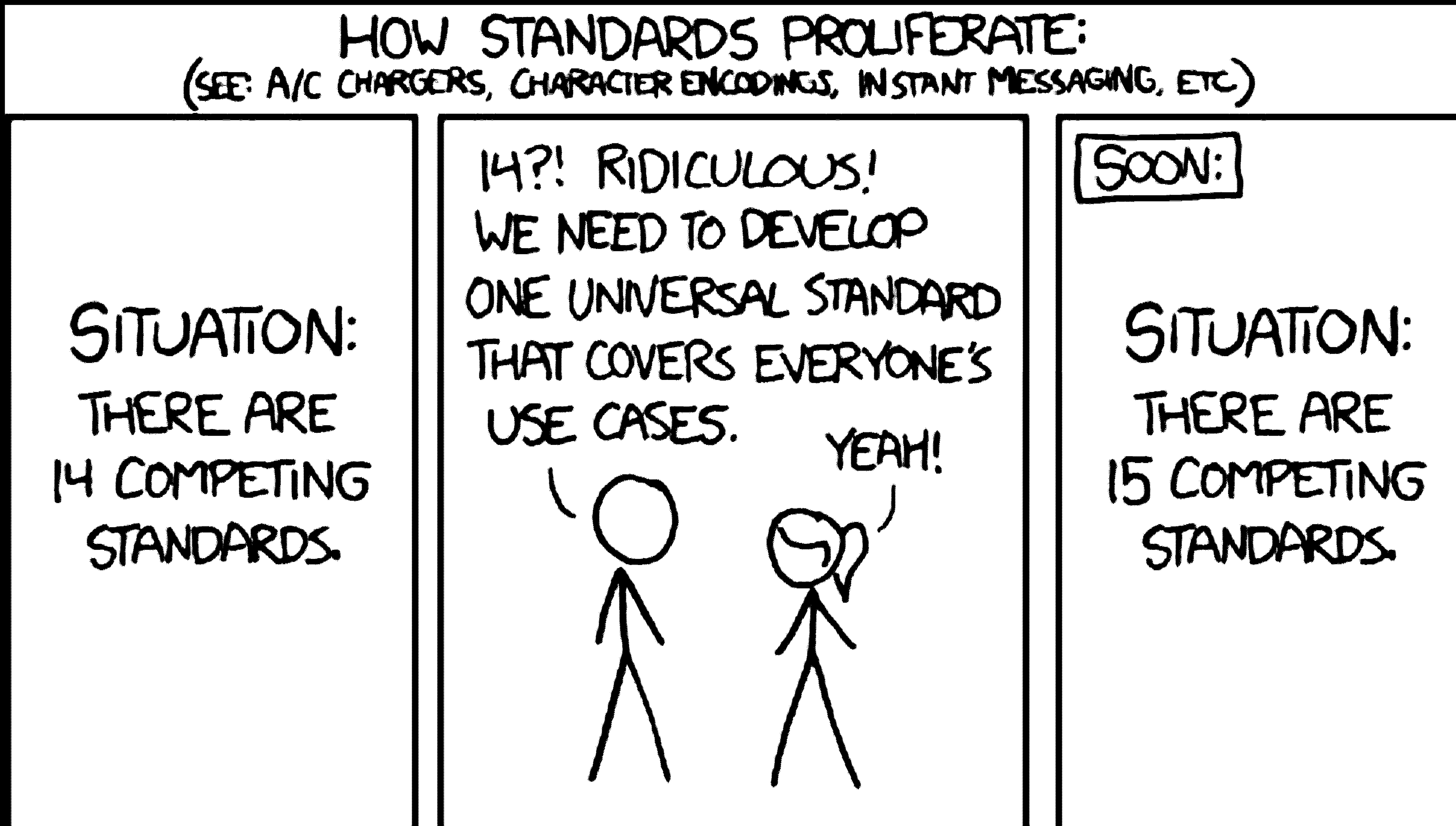
Use Accuracy

rness

ation Disparity

Accuracy Equality

Conditi


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.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

# Fairness: Correlation

correlation(race, 

) = 0.8

mutual information(race, 

) = 0.6

Correlation does not measure causation

Atlidakis et al. FairTest: Discovering unwarranted associations in data-driven applications. EuroS&P 2017

# What is fairness?

Sensitive inputs should not affect software behavior.

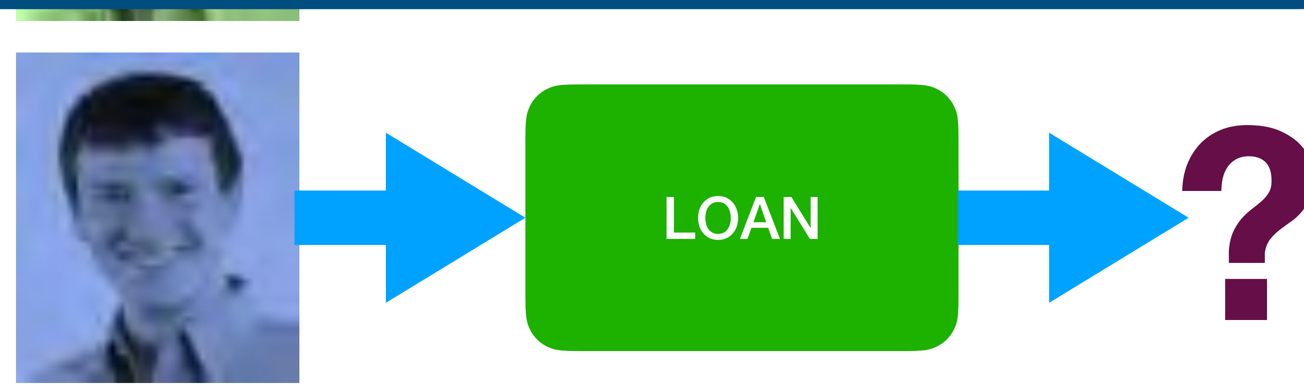
We want to measure causality!

Woodward. Making things happen: A theory of causal explanation. 2005

# causal testing

**Sensitive inputs should not affect software behavior.**

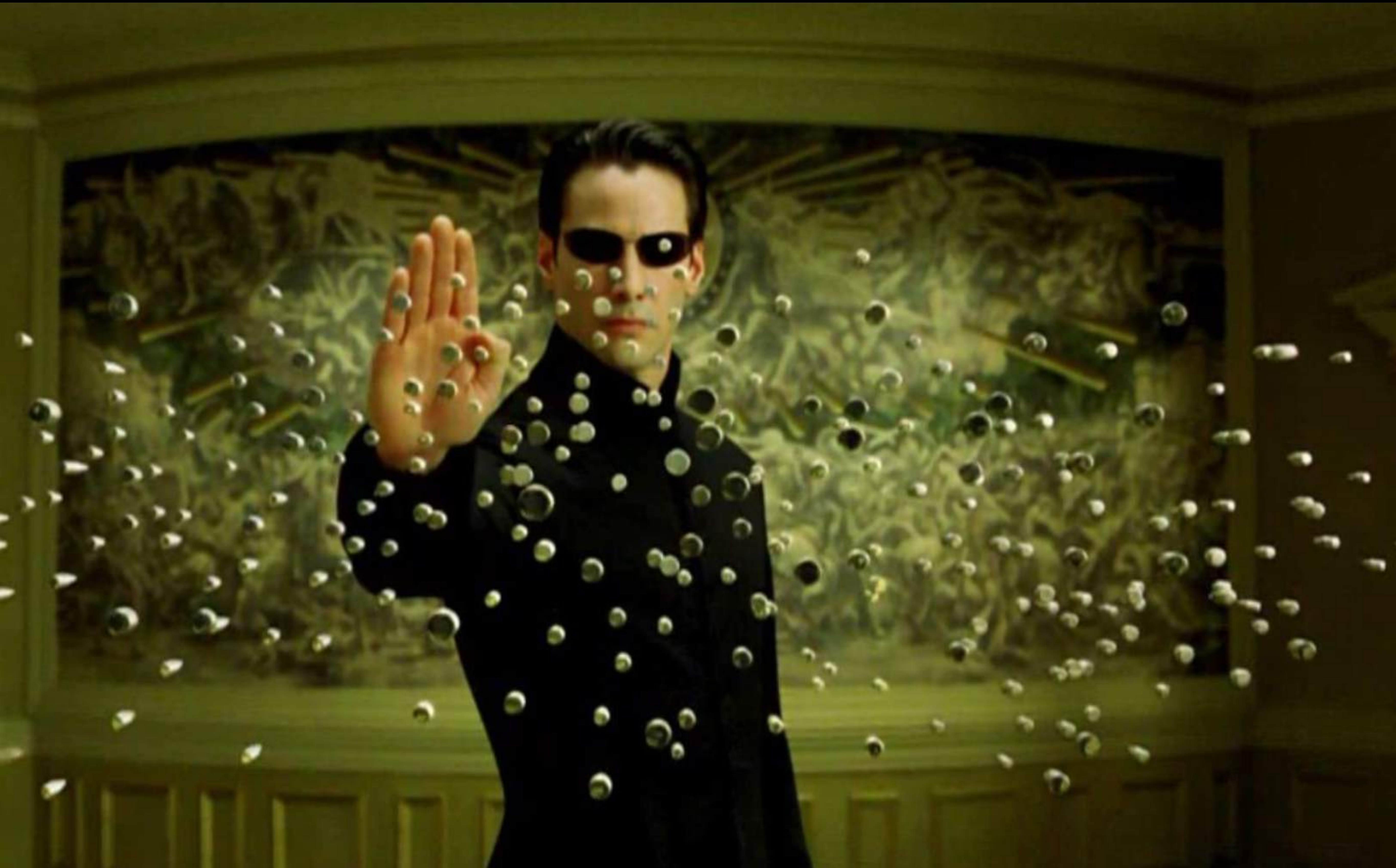
hypo
testing:




Galhotra, Brun, and Meliou, Fairness Testing: Testing Software for Discrimination. ESEC/FSE 2017

# causal testing

# 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

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

http://fairness.cs.umass.edu

Angell, Johnson, Brun, and Meliou, Themis: Automatically Testing Software for Discrimination. ESEC/FSE 2018 Demo

# How does Themis work?

## adaptive, confidence-driven sampling

input schema

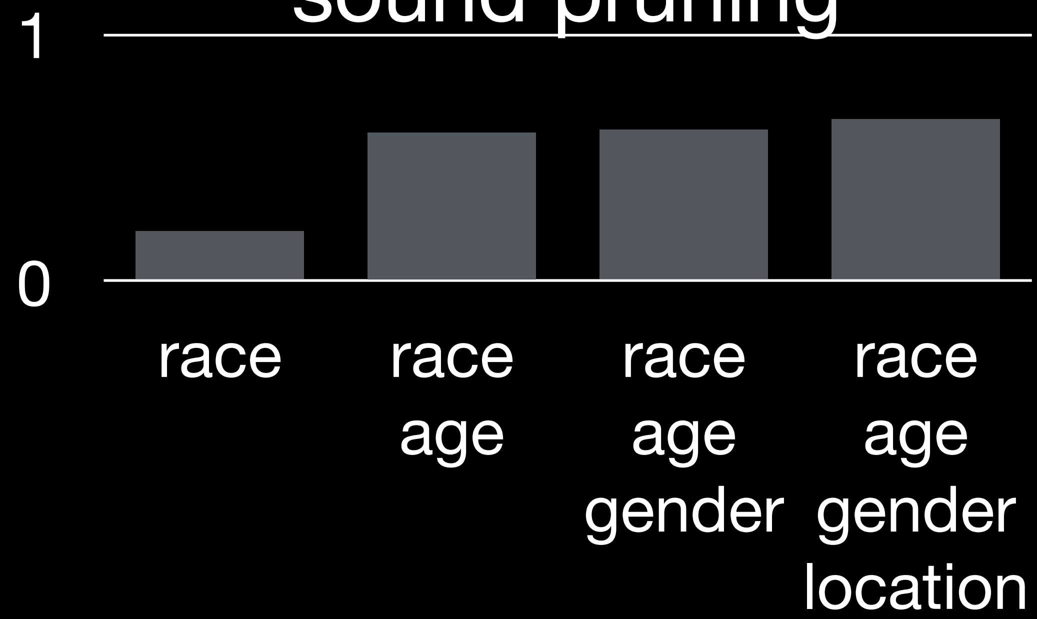
confidence

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

**Trained a bunch of systems.**
**Some are supposed to enforce fairness.**

- Census income dataset:
  financial data
  45K people
  income > $50K?

- Statlog German credit dataset:
  credit data
  1K people
  “good” or “bad” credit?

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

# findings

Demographic parity 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

Optimizing demographic parity 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 148× for causal and 2,849× for group discrimination. Best improvement was 13,000×.

# Causal Testing: more than bias detection

## Causal Testing: Understanding Defects' Root Causes

Brittany Johnson
University of Massachusetts Amherst
Amherst, MA, USA
bjohnson@cs.umass.edu

Yuriy Brun
University of Massachusetts Amherst
Amherst, MA, USA
brun@cs.umass.edu

Alexandra Meliou
University of Massachusetts Amherst
Amherst, MA, USA
ameli@cs.umass.edu

### ABSTRACT

Understanding the root cause of a defect is critical to isolating and repairing buggy behavior. We present Causal Testing, a new method of root-cause analysis that relies on the theory of counterfactual causality to identify a set of executions that likely hold key causal information necessary to understand and repair buggy behavior. Using the Defects4J benchmark, we find that Causal Testing could be applied to 71% of real-world defects, and for 77% of those, it can help developers identify the root cause of the defect. A controlled experiment with 37 developers shows that Causal Testing improves participants' ability to identify the cause of the defect from 80% of the time with standard testing tools to 86% of the time with Causal Testing. The participants report that Causal Testing provides useful information they cannot get using tools such as JUnit. Holmes, our prototype, open-source Eclipse plugin implementation of Causal Testing, is available at http://holmes.cs.umass.edu/.

### CCS CONCEPTS

• **Software and its engineering → Software testing and debugging.**

### KEYWORDS

Causal Testing, causality, theory of counterfactual causality, software debugging, test fuzzing, automated test generation, Holmes

**ACM Reference Format:**
Brittany Johnson, Yuriy Brun, and Alexandra Meliou. 2020. Causal Testing: Understanding Defects' Root Causes. In *42nd International Conference on Software Engineering (ICSE '20), May 23–29, 2020, Seoul, Republic of Korea.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3377811.3380377

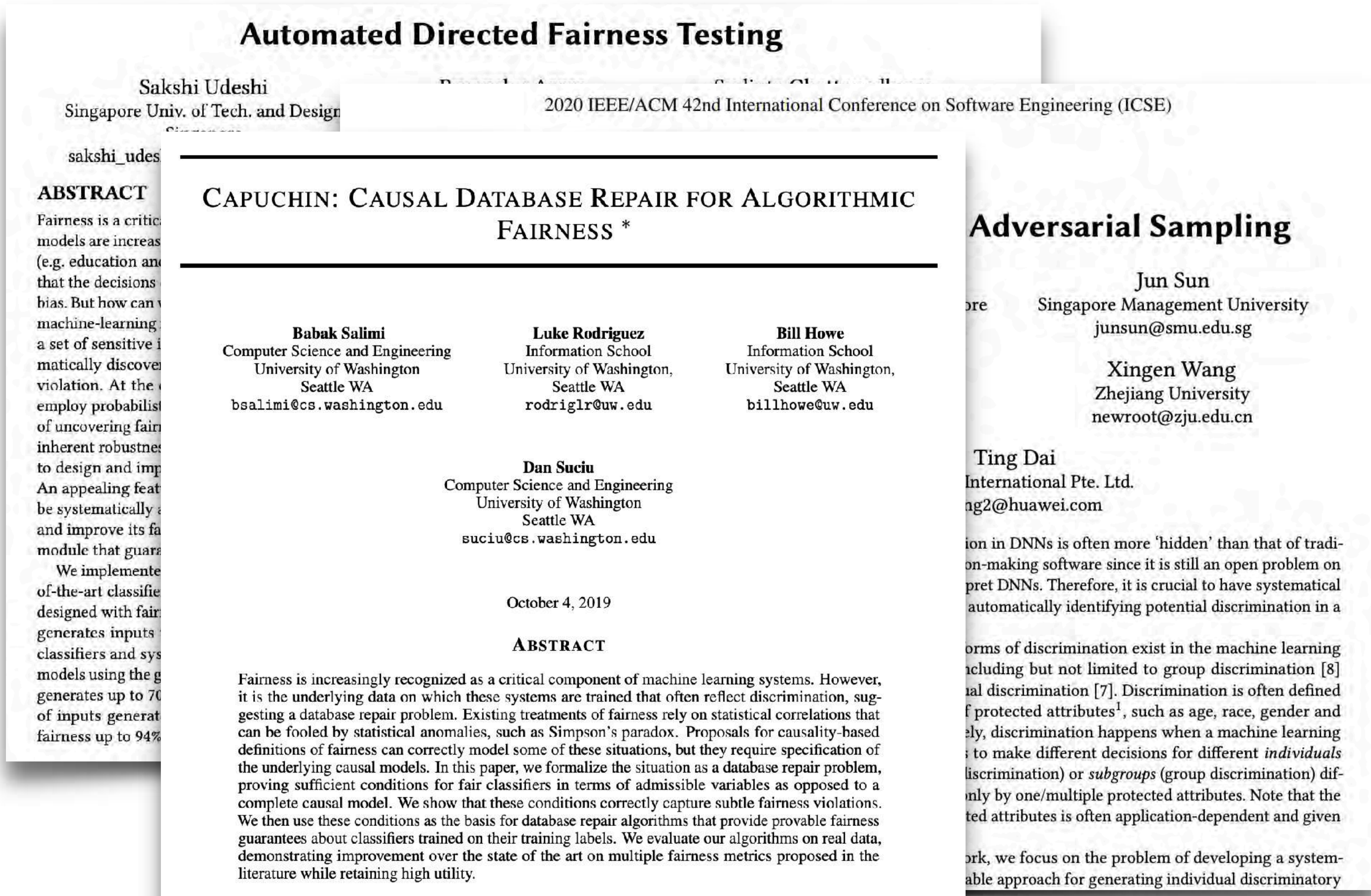
test input [74] and a set of test-breaking changes [73], they do not help explain *why* the code is faulty [40].

To address this shortcoming of modern debugging tools, this paper presents *Causal Testing*, a novel technique for identifying root causes of failing executions based on the theory of counterfactual causality. Causal Testing takes a manipulationist approach to causal inference [71], modifying and executing tests to observe causal relationships and derive causal claims about the defects' root causes.

Given one or more failing executions, Causal Testing conducts *causal experiments* by modifying the existing tests to produce a small set of executions that differ minimally from the failing ones but do not exhibit the faulty behavior. By observing a behavior and then purposefully changing the input to observe the behavioral changes, Causal Testing infers causal relationships [71]: The change in the input *causes* the behavioral change. Causal Testing looks for two kinds of minimally-different executions, ones whose inputs are similar and ones whose execution paths are similar. When the differences between executions, either in the inputs or in the execution paths, are small, but exhibit different test behavior, these small, causal differences can help developers understand what is causing the faulty behavior.

Consider a developer working on a web-based geo-mapping service (such as Google Maps or MapQuest) receiving a bug report that the directions between "New York, NY, USA" and "900 René Lévesque Blvd. W Montreal, QC, Canada" are wrong. The developer replicates the faulty behavior and hypothesizes potential causes. Maybe the special characters in "René Lévesque" caused a problem. Maybe the first address being a city and the second a specific building caused a mismatch in internal data types. Maybe the route is too long and the service's precomputing of some routes is causing the

# Debugging



Automated Directed Fairness Testing

Sakshi Udeshi
Singapore Univ. of Tech. and Design

2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)

Adversarial Sampling

Jun Sun
Singapore Management University
junsun@smu.edu.sg

Xingen Wang
Zhejiang University
newroot@zju.edu.cn

Ting Dai

CAPUCHIN: CAUSAL DATABASE REPAIR FOR ALGORITHMIC FAIRNESS *

**Babak Salimi**
Computer Science and Engineering
University of Washington
Seattle WA
bsalimi@cs.washington.edu

**Luke Rodriguez**
Information School
University of Washington,
Seattle WA
rodriglr@uw.edu

**Bill Howe**
Information School
University of Washington,
Seattle WA
billhowe@uw.edu

**Dan Suciu**
Computer Science and Engineering
University of Washington
Seattle WA
suciu@cs.washington.edu

October 4, 2019

ABSTRACT

Fairness is increasingly recognized as a critical component of machine learning systems. However, it is the underlying data on which these systems are trained that often reflect discrimination, suggesting a database repair problem. Existing treatments of fairness rely on statistical correlations that can be fooled by statistical anomalies, such as Simpson's paradox. Proposals for causality-based definitions of fairness can correctly model some of these situations, but they require specification of the underlying causal models. In this paper, we formalize the situation as a database repair problem, proving sufficient conditions for fair classifiers in terms of admissible variables as opposed to a complete causal model. We show that these conditions correctly capture subtle fairness violations. We then use these conditions as the basis for database repair algorithms that provide provable fairness guarantees about classifiers trained on their training labels. We evaluate our algorithms on real data, demonstrating improvement over the state of the art on multiple fairness metrics proposed in the literature while retaining high utility.

# More Complex Inputs

## Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models

**Julius Adebayo** JULIUSAD@MIT.EDU
**Lalana Kagal** LKAGAL@CSAIL.MIT.EDU
CSAIL, MIT, 32 Vassar Street Cambridge, MA 02139 USA.

### Abstract

Predictive models are increasingly deployed for the purpose of determining access to services such as credit, insurance, and employment. Despite potential gains in productivity and efficiency, several potential problems have yet to be addressed, particularly the potential for unintentional discrimination. We present an iterative procedure, based on orthogonal projection of input attributes, for enabling interpretability of black-box predictive models. Through our iterative procedure, one can quantify the relative dependence of a black-box model on its input attributes.The relative significance of the inputs to a predictive model can then be used to assess the fairness (or discriminatory extent) of such a model.

sexual orientation. A predictive model that significantly weights these protected attributes would tend to exhibit disparate outcomes for these groups of individuals. Hence, the focus of this paper is on auditing predictive models to determine the relative significance of a models inputs in determining outcomes. Given the relative significance of a model to its inputs, judgement can be more easily made about the model's fairness.

The potential increased efficiency and societal gains from leveraging predictive modeling seem limitless, and have rightly necessitated the widespread adoption of these models. In particular, use of predictive modeling for decision making in determining access to services is starting to become the defacto standard in industries such as banking, insurance, housing, and employment. As the need for more accurate forecasts or predictions has heightened, there has been an increase in the use of complicated, often uninterpretable predictive models in making forecasts from data. Increasingly, these predictive models tend to have millions

# 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

https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

# What are we doing now?

ACLU GET UPDATES / DONATE

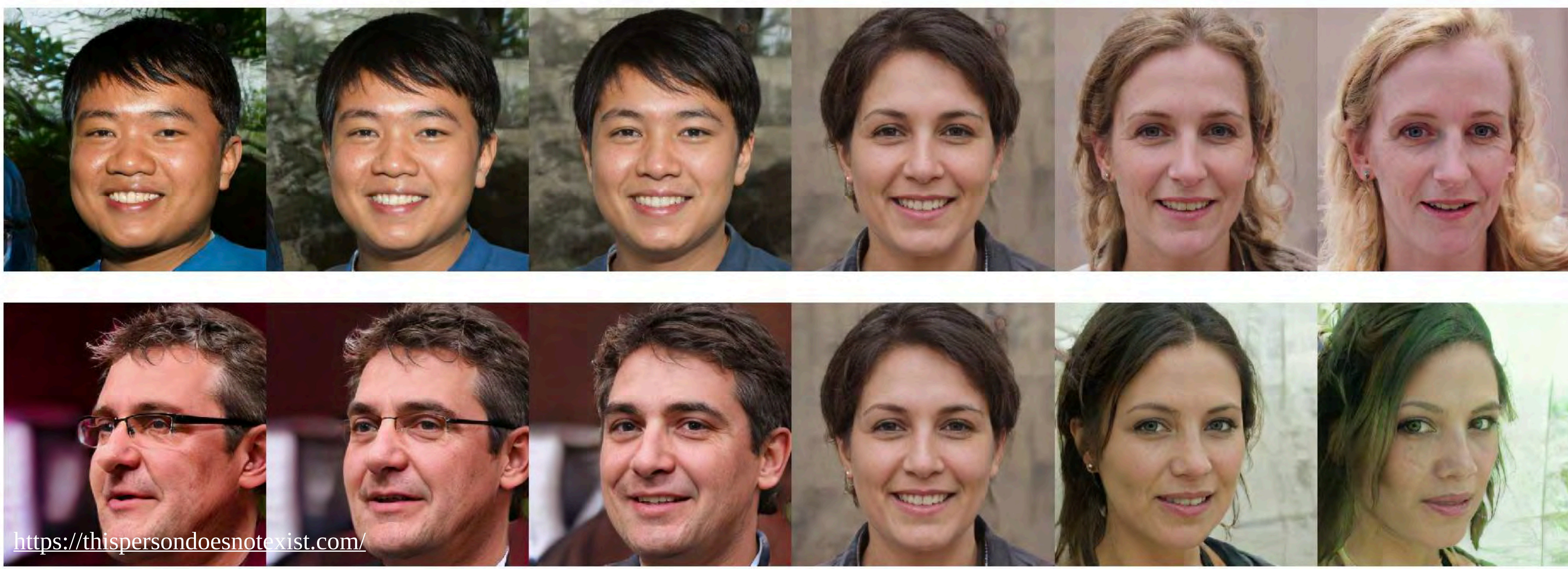
**Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots**

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

Fair computer vision

Fair natural language processing

# Testing versus Verifying

Provably fair machine learning:

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

# How would that work?

User specifies a definition of fairness.

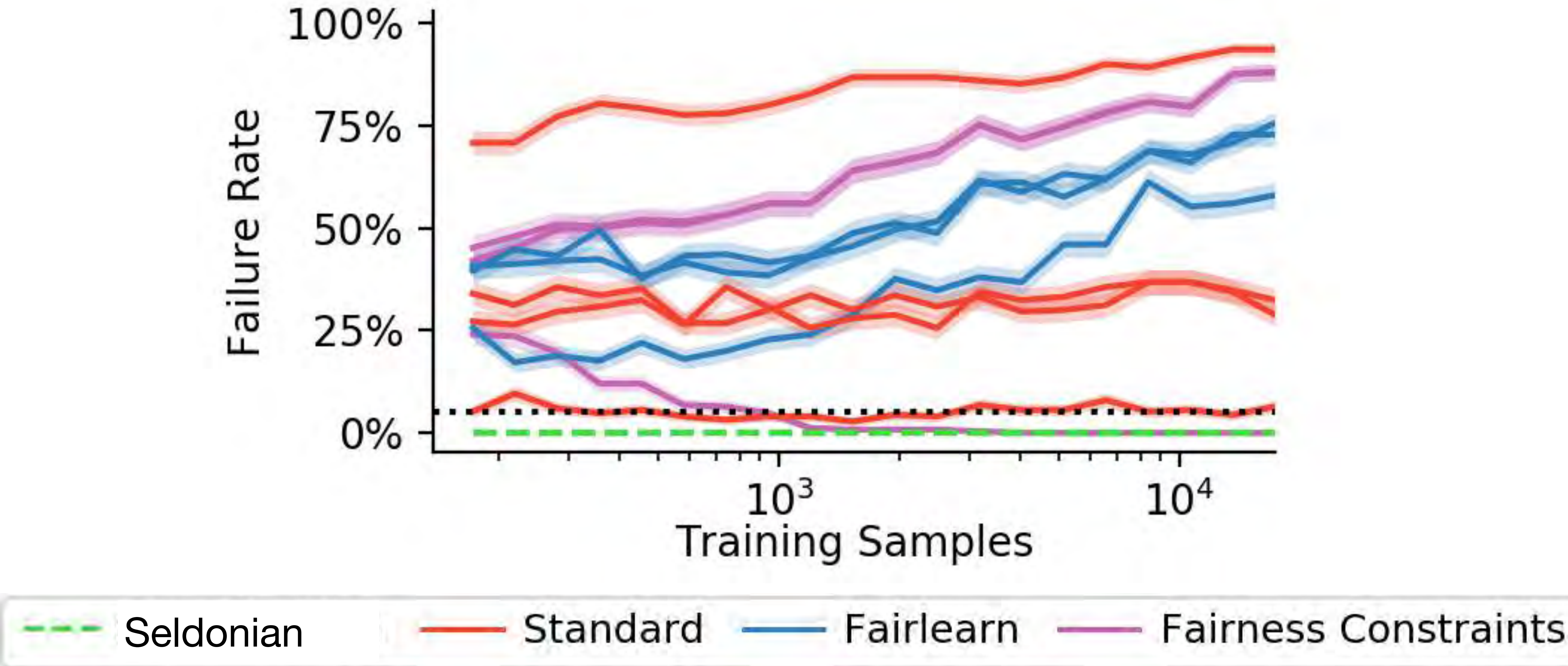

Train classifiers,
selects one to satisfy fairness,
verify safety on held-out suite.

# How would that work?

**Limitation: The algorithm has to be able to return "No Solution Found"**

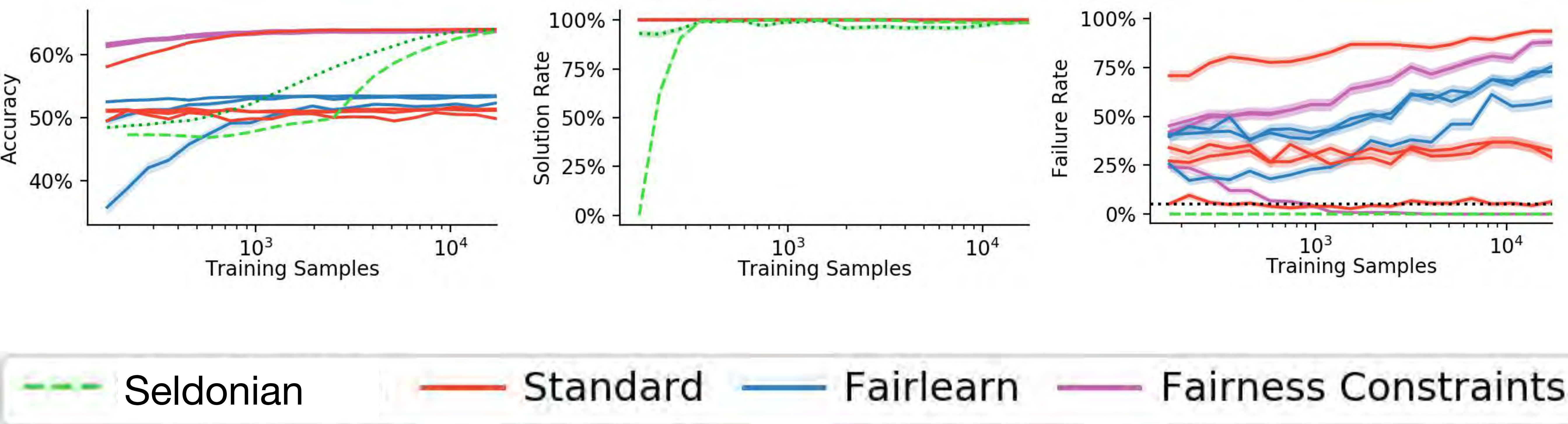
Train classifiers,
selects one to satisfy fairness,
verify safety on held-out suite.

# Disparate Impact



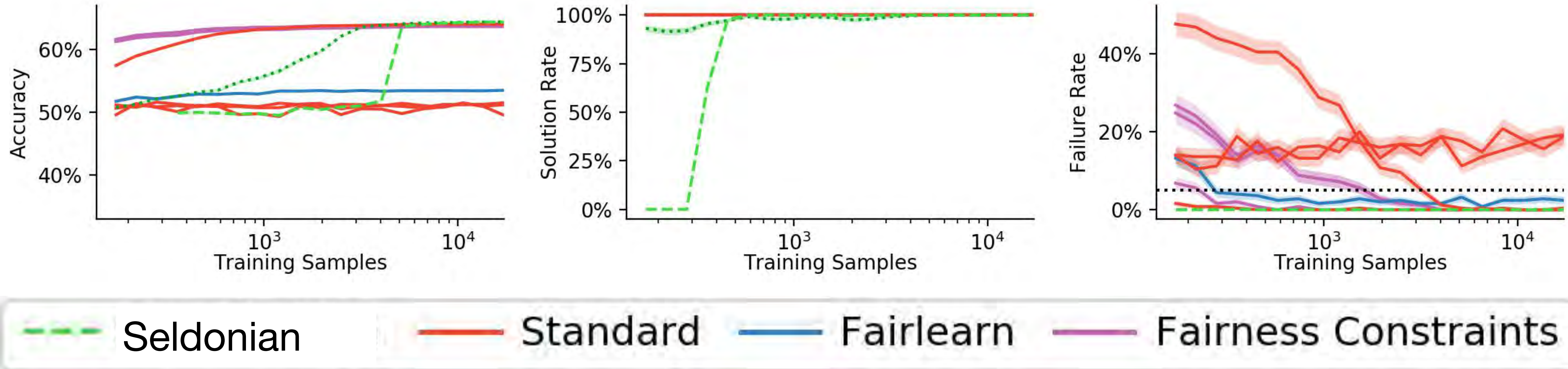

Fairlearn: Agarwal et al. A reductions approach to fair classification. ICML 2018.
Fairness Constraints: Zafar et al., Fairness Constraints: A Mechanism for Fair Classification. FATML 2015.

# Disparate Impact

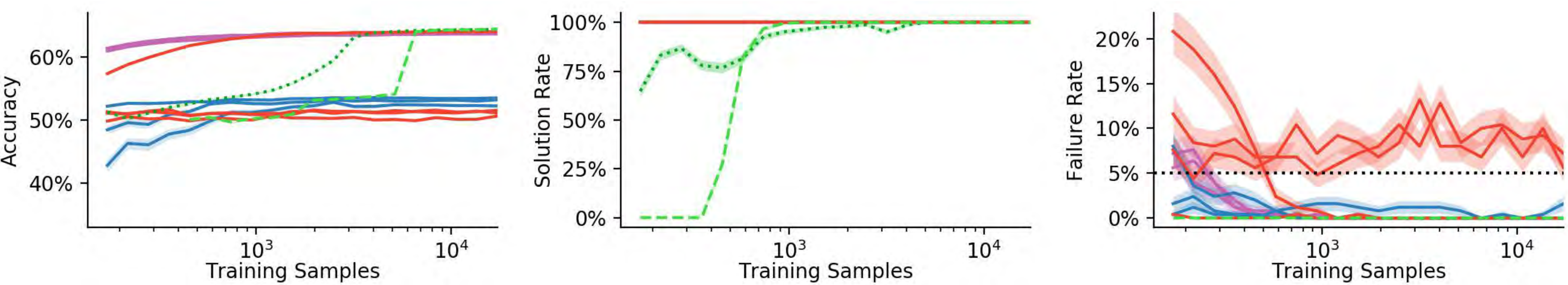



Fairlearn: Agarwal et al. A reductions approach to fair classification. ICML 2018.
Fairness Constraints: Zafar et al., Fairness Constraints: A Mechanism for Fair Classification. FATML 2015.

# Demographic Parity




# Equal Opportunity



# Equalized Odds

RESEARCH

COMPUTER SCIENCE

## Preventing undesirable behavior of intelligent machines

Philip S. Thomas[1*], Bruno Castro da Silva[2], Andrew G. Barto[1], Stephen Giguere[1], Yuriy Brun[1], Emma Brunskill[3]

Intelligent machines using machine learning algorithms are ubiquitous, ranging from simple data analysis and pattern recognition tools to complex systems that achieve superhuman performance on various tasks. Ensuring that they do not exhibit undesirable behavior—that they do not, for example, cause harm to humans—is therefore a pressing problem. We propose a general and flexible framework for designing machine learning algorithms. This framework simplifies the problem of specifying and regulating undesirable behavior. To show the viability of this framework, we used it to create machine learning algorithms that precluded the dangerous behavior caused by standard machine learning algorithms in our experiments. Our framework for designing machine learning algorithms simplifies the safe and responsible application of machine learning.

Machine learning (ML) algorithms are having an increasing impact on modern society. They are used by geologists to predict landslides (*1*) and by biologists working to create a vaccine for HIV (*2*); they also influence criminal sen-

Note that the algorithm does not know $f(\theta)$ for any $\theta \in \Theta$ (e.g., the true mean squared error); it can only reason about it from data (e.g., by using the sample mean squared error).

One problem with the standard ML approach is that the user of an ML algorithm

algorithm could output. Our framework mathematically defines what an algorithm should do in a way that allows the user to directly place probabilistic constraints on the solution, $a(D)$, returned by the algorithm. This differs from the standard ML approach wherein the user can only indirectly constrain $a(D)$ by restricting or modifying the feasible set $\Theta$ or objective function $f$. Concretely, algorithms constructed using our framework are designed to satisfy constraints of the form $\Pr(g(a(D)) \leq 0) \geq 1 - \delta$, where $g: \Theta \to \mathbb{R}$ defines a measure of undesirable behavior (as illustrated later by example) and $\delta \in [0, 1]$ limits the admissible probability of undesirable behavior.

Note that in these constraints, $D$ is the only source of randomness; we denote random variables by capital noncalligraphic letters to make clear which terms are random in statements of probability and expectation. Because these constraints define which algorithms $a$ are acceptable (rather than which solutions $\theta$ are acceptable), they must be satisfied during the design of the algorithm rather than when the algorithm is applied. This shifts the burden of ensuring that the algorithm is well-behaved

Thomas, Castro da Silva, Barto, Giguere, Brun, and Brunskill.
"Preventing Undesirable Behavior of Intelligent Machines", Science 366 (6468), Nov 22, 2019

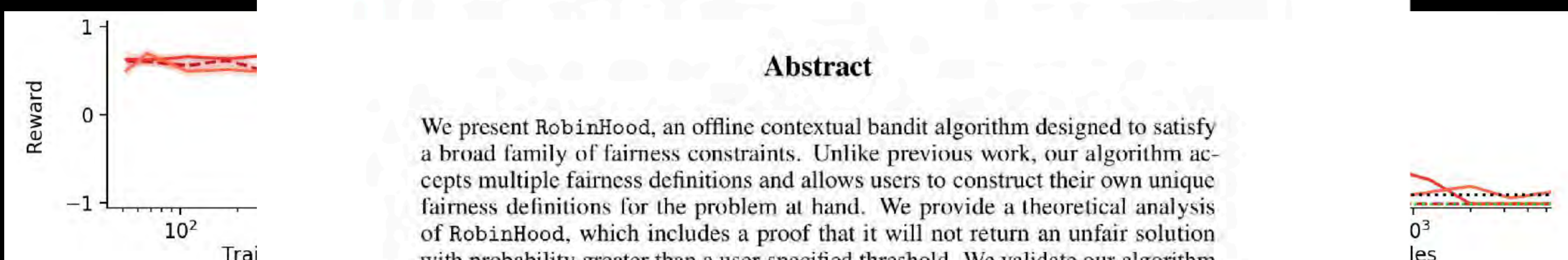
# Fairness: Delayed Impact

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

# RobinHood: Fair Contextual Bandits

Instead ... idivism,
these r...
to mini...




## Offline Contextual Bandits with High Probability Fairness Guarantees


Blossom Metevier[UM] Stephen Giguere[UM] Sarah Brockman[UM] Ari Kobren[UM]
Yuriy Brun[UM] Emma Brunskill[S] Philip S. Thomas[UM]
[UM]College of Information and Computer Sciences, University of Massachusetts Amherst
[S]Computer Science Department, Stanford University



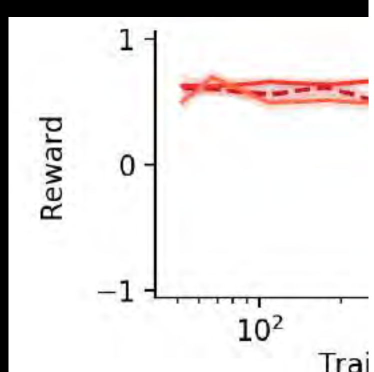
### Abstract

We present RobinHood, an offline contextual bandit algorithm designed to satisfy a broad family of fairness constraints. Unlike previous work, our algorithm accepts multiple fairness definitions and allows users to construct their own unique fairness definitions for the problem at hand. We provide a theoretical analysis of RobinHood, which includes a proof that it will not return an unfair solution with probability greater than a user-specified threshold. We validate our algorithm on three applications: a tutoring system in which we conduct a user study and consider multiple unique fairness definitions; a loan approval setting (using the Statlog German credit data set) in which well-known fairness definitions are applied; and criminal recidivism (using data released by ProPublica). In each setting, our algorithm is able to produce fair policies that achieve performance competitive with other offline and online contextual bandit algorithms.


Metevier, Giguere, Brockman, Kobren, Brun, Brunskill, Thomas. Offline Contextual Bandits with High Probability Fairness Guarantees. NeurIPS 2019.

# RobinHood:
# Fair Contextual Bandits

Insteac idivism,
these r
to mini




## Offline Contextual Bandits with High Probability Fairness Guarantees

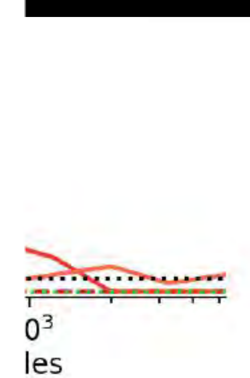

Blossom Metevier[UM] Stephen Giguere[UM] Sarah Brockman[UM] Ari Kobren[UM]
Yuriy Brun[UM] Emma Brunskill[S] Philip S. Thomas[UM]

[UM]College of Information and Computer Sciences
University of Massachusetts Amherst

[S]Computer Science Department
Stanford University


### Abstract

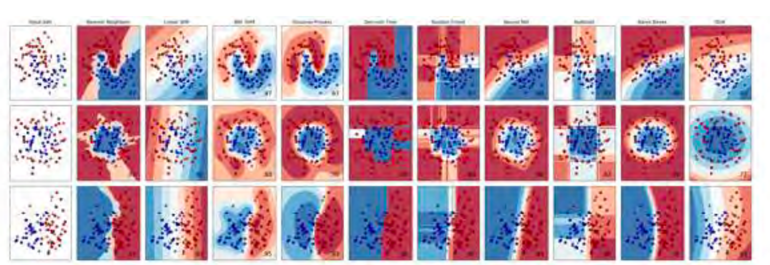

We present RobinHood, an offline contextual bandit algorithm designed to satisfy a broad family of fairness constraints. Unlike previous work, our algorithm accepts multiple fairness definitions and allows users to construct their own unique fairness definitions for the problem at hand. We provide a theoretical analysis of RobinHood, which includes a proof that it will not return an unfair solution with probability greater than a user-specified threshold. We validate our algorithm on three applications: a tutoring system in which we conduct a user study and consider multiple unique fairness definitions; a loan approval setting (using the Statlog German credit data set) in which well-known fairness definitions are applied; and criminal recidivism (using data released by ProPublica). In each setting, our algorithm is able to produce fair policies that achieve performance competitive with other offline and online contextual bandit algorithms.



Metevier, Giguere, Brockman, Kobren, Brun, Brunskill, Thomas. Offline Contextual Bandits with High Probability Fairness Guarantees. NeurIPS 2019.

# Empowering Data Scientists



## scikit-learn

*Machine Learning in Python*

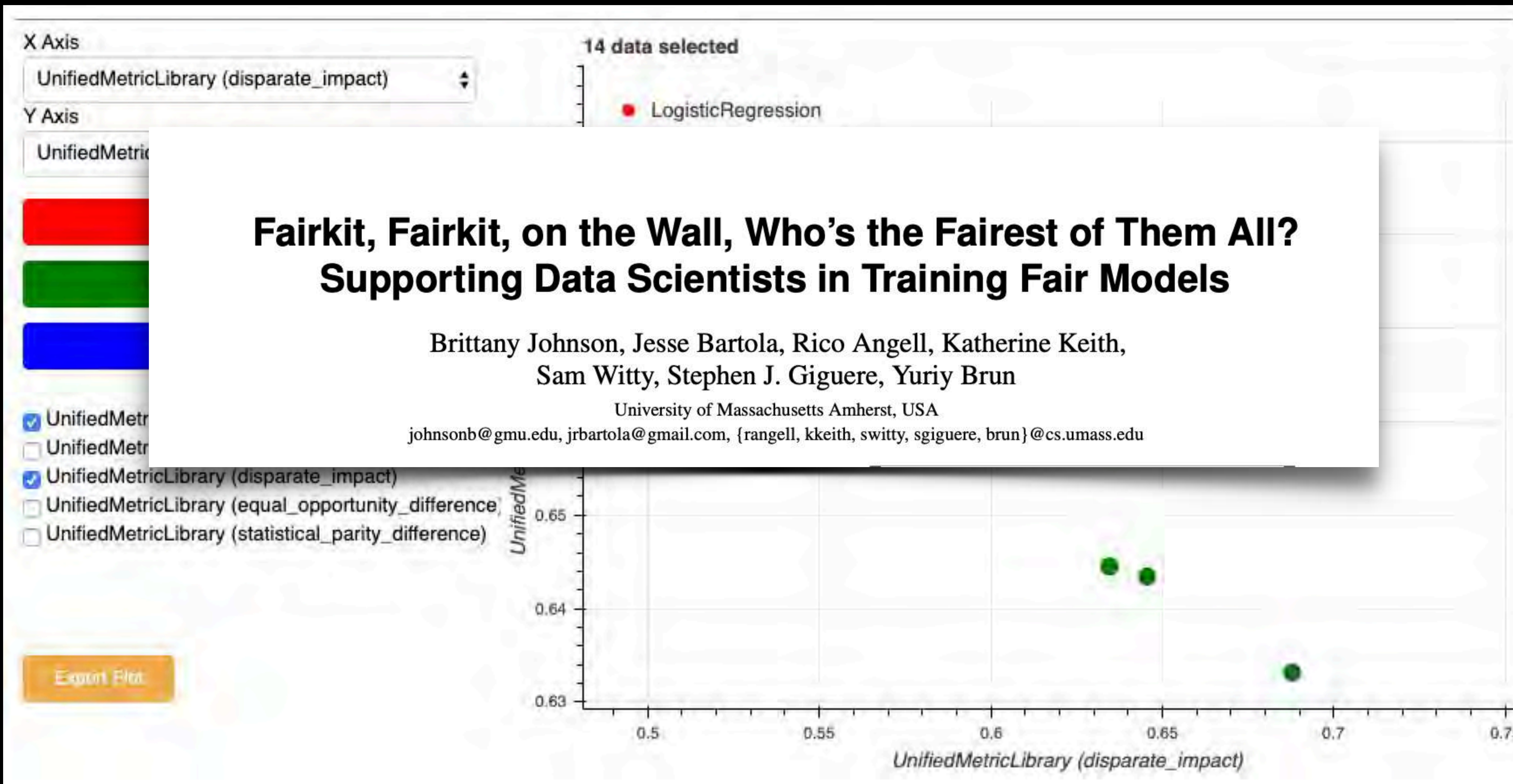
- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

https://scikit-learn.org

IBM’s AI Fairness 360 adds fairness metrics, fairness-aware algorithms, datasets

http://aif360.mybluemix.net/

# fairkit-learn




**Fairkit, Fairkit, on the Wall, Who's the Fairest of Them All? Supporting Data Scientists in Training Fair Models**

Brittany Johnson, Jesse Bartola, Rico Angell, Katherine Keith,
Sam Witty, Stephen J. Giguere, Yuriy Brun
University of Massachusetts Amherst, USA
johnsonb@gmu.edu, jrbartola@gmail.com, {rangell, kkeith, switty, sgiguere, brun}@cs.umass.edu

# What we need:

Help stakeholders understand what needs to be enforced.

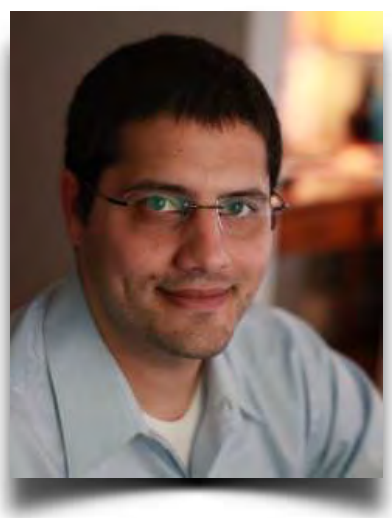
Provide components that enforce these properties themselves.

Help validate systems adhere to the to-be-enforced properties.

Help visualize behavior to reason about the properties.

# Research Landscape

safe machine learning



Philip Thomas

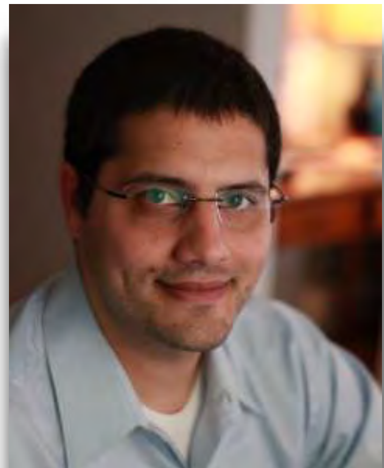
Emma Brunskill

Michael Kearns

and many many more

# Research Landscape

## safe machine learning


Philip Thomas


Emma Brunskill


Michael Kearns

and many many more

## data-based fairness


Alexandra Meliou

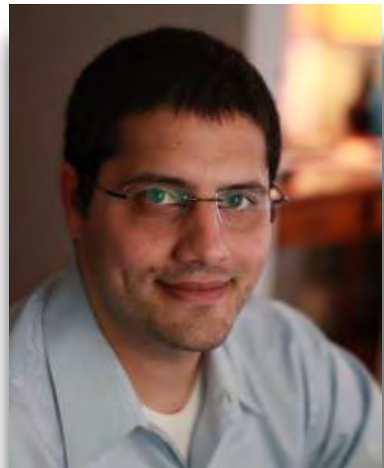

Dan Suciu


Bill Howe


Julia Stoyanovich

# Research Landscape

## safe machine learning



Philip Thomas

Emma Brunskill

Michael Kearns

and many many more

## data-based fairness

Alexandra Meliou

Dan Suciu

Bill Howe

Julia Stoyanovich

## software engineering

Brittany Johnson

Jin Song Dong

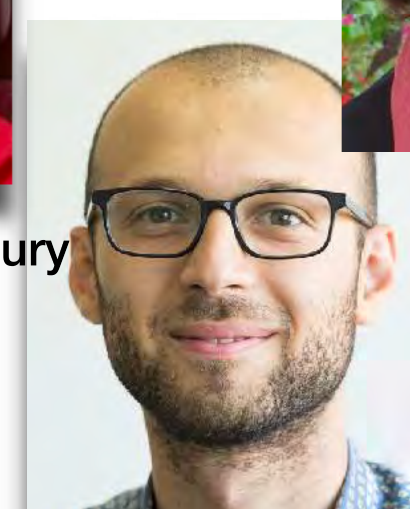
Abhik Roychoudhury

Julia Rubin

Jon Whittle

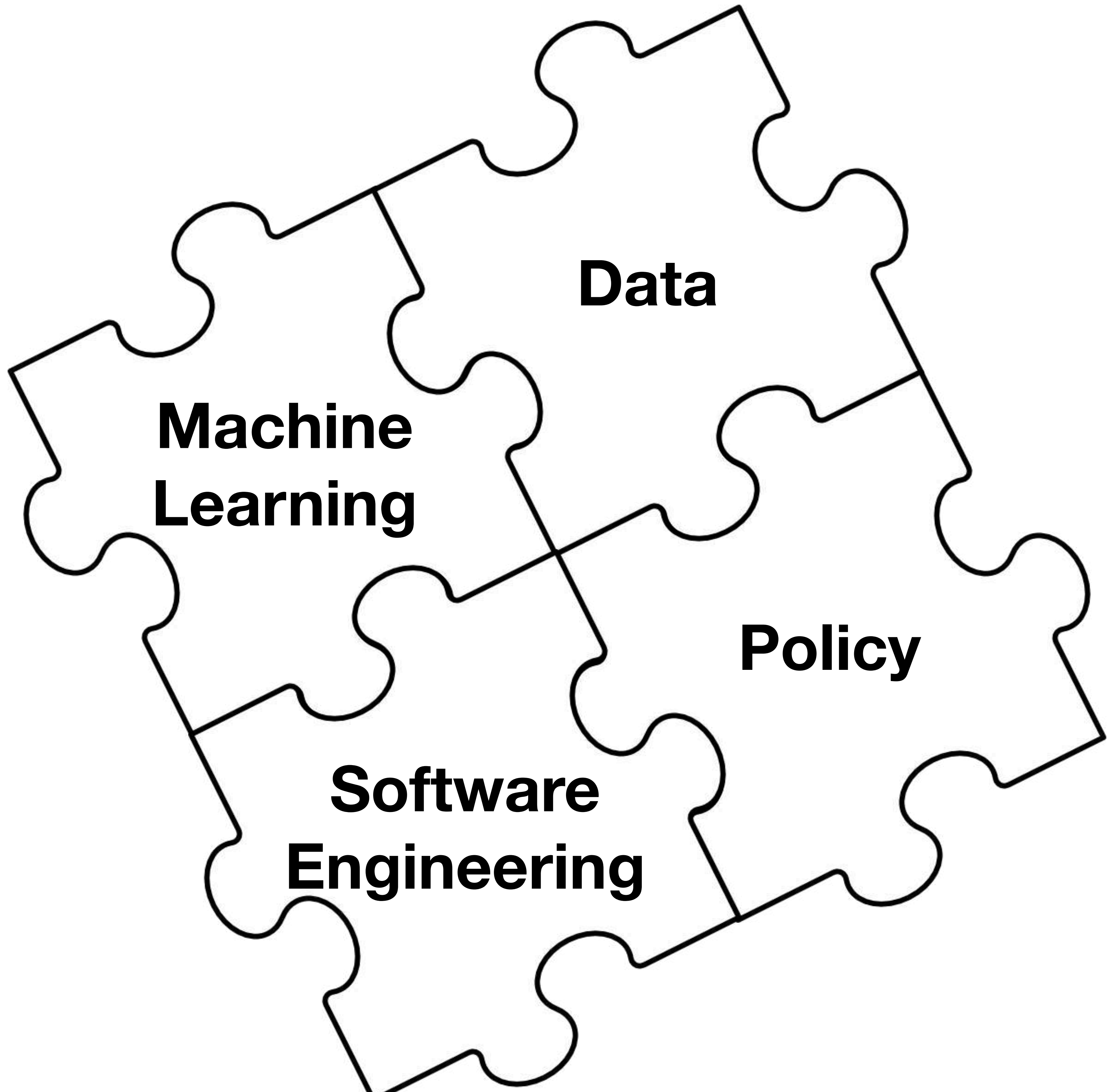
Sudipta Chattopadhyay

Tim Menzies

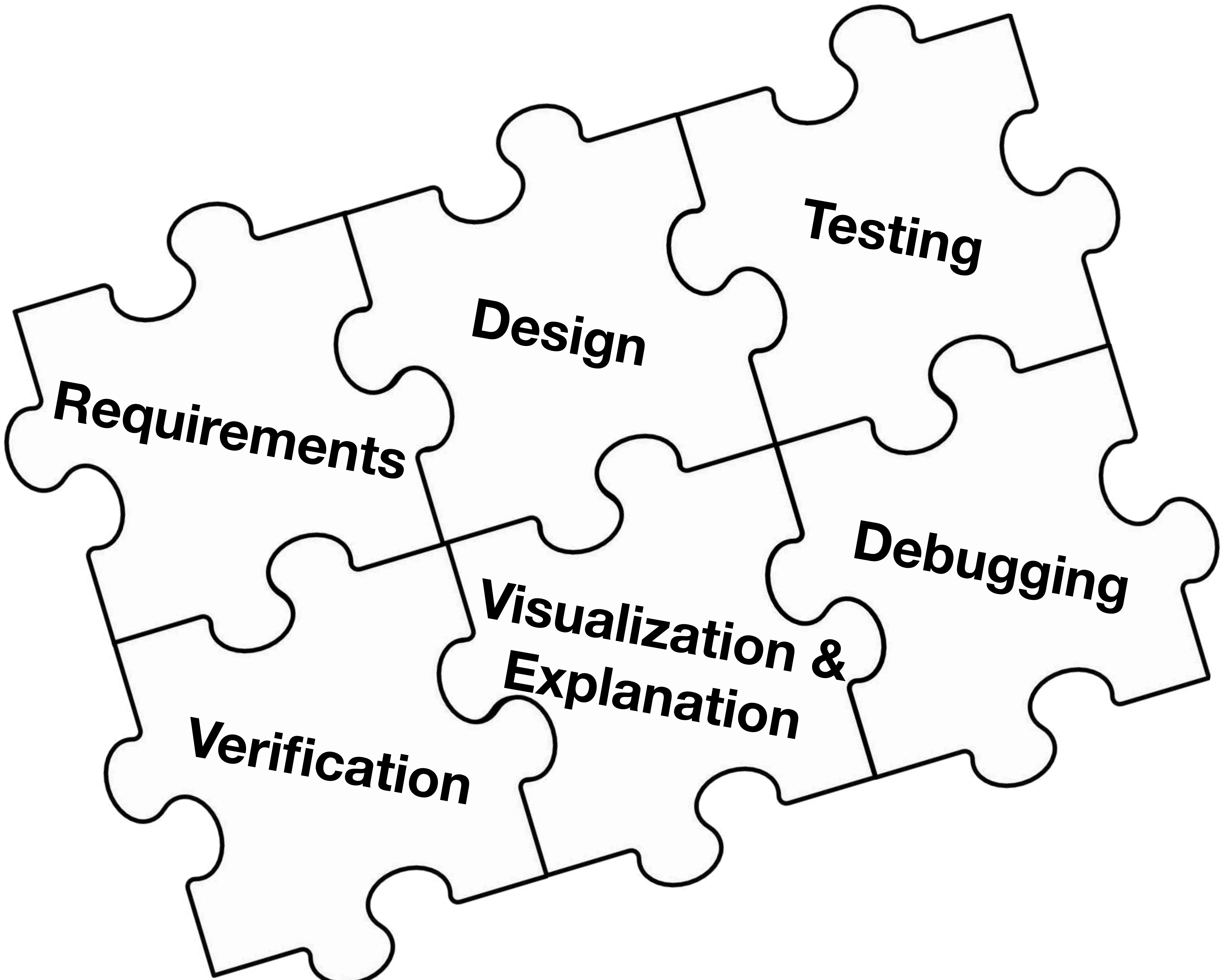


Aws Albarghouthi

# Research Landscape



# Research Landscape (SE)



# Contributions

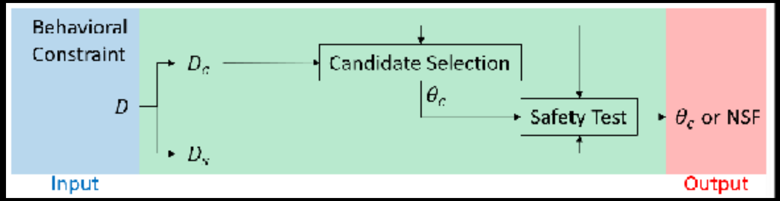
http://fairness.cs.umass.edu

 → LOAN → **causal bias measure** ←  ← LOAN

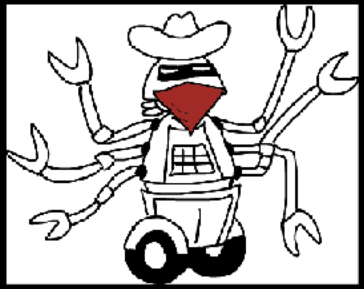
complex inputs

many causes of software bias

tools for data scientists




provable guarantees



delayed impact

Rico Angell

Brittany Johnson

Stephen Giguere

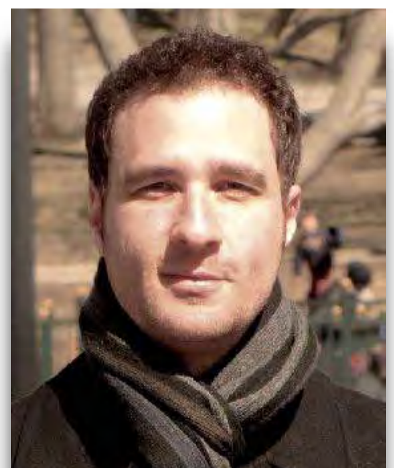
Sarah Brockman

Blossom Metevier

Sainyam Galhotra

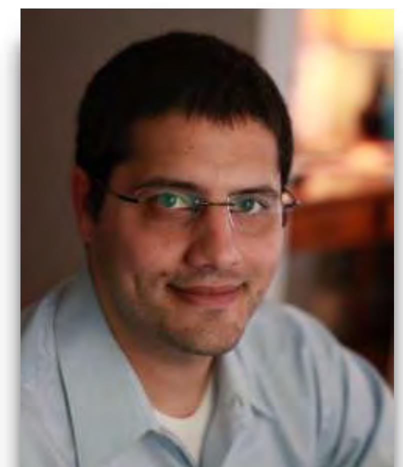
Alexandra Meliou

Andy Barto



Bruno Castro da Silva

Emma Brunskill



Philip Thomas

Yuriy Brun

# http://fairness.cs.umass.edu

https://tinyurl.com/FairnessPaper

http://doi.org/10.1126/science.aag3311

NSF Google ORACLE UMass Amherst

*"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."*

The Networking and Information Technology Research and Development (NITRD) Program

**Mailing Address:** NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

**Physical Address:** 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov