Rediet Abebe, Cornell University

Computational Interventions to Improve Access to Opportunity

Time: ***Monday April 29th, 12:00pm***
Location: ***CIT 368***

Algorithmic Fairness

CS16: Introduction to Data Structures & Algorithms
Spring 2019

Outline

- Potential and limitations of machine learning
- ML in hiring decisions
- ML in criminal risk assessment
- ML in predictive policing

Forbes Billionaires Innovation Le

10,352 views | Feb 25, 2019, 08:21am

The New York Times

ADVERTISEMENT

The New York Times

Account

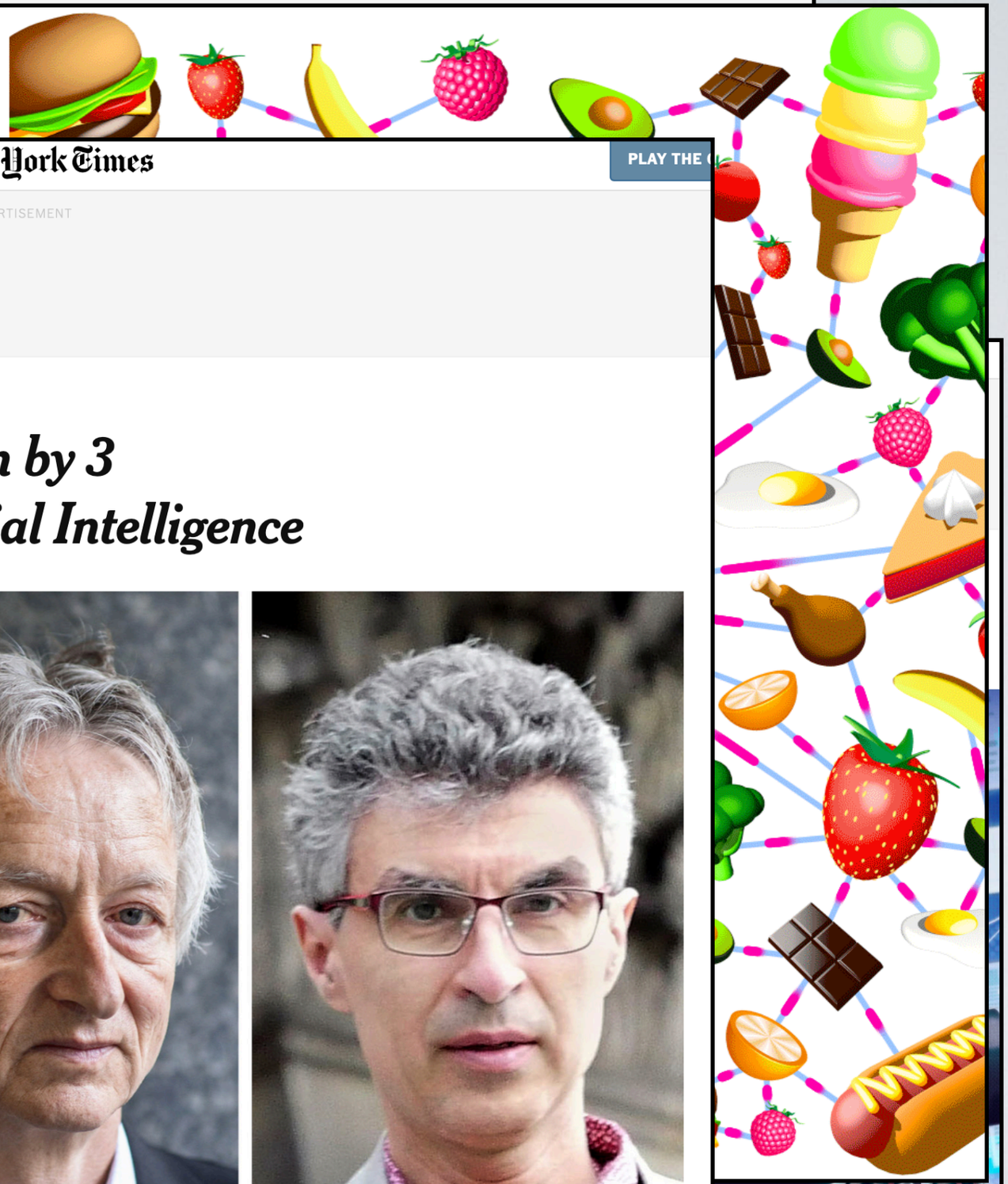
TECHNOLOGY

The New York Times

PLAY THE

ADVERTISEMENT

Turing Award Won by 3 Pioneers in Artificial Intelligence

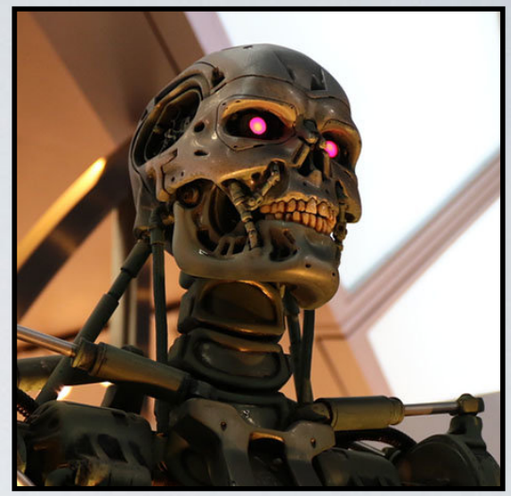


Forget go

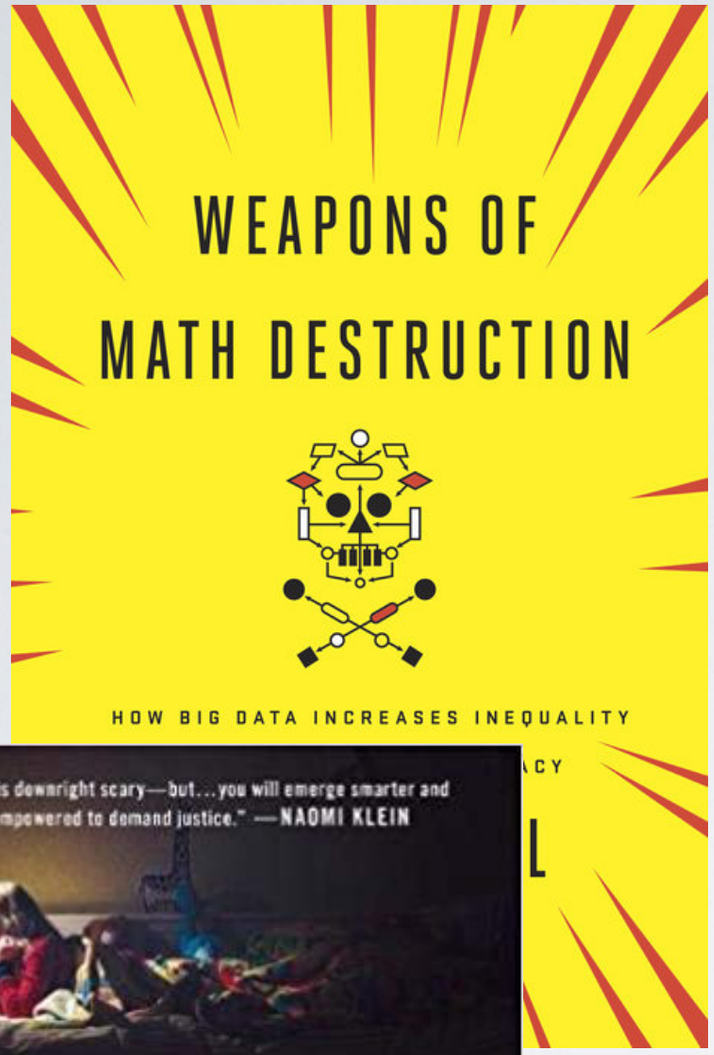
Machine Learning

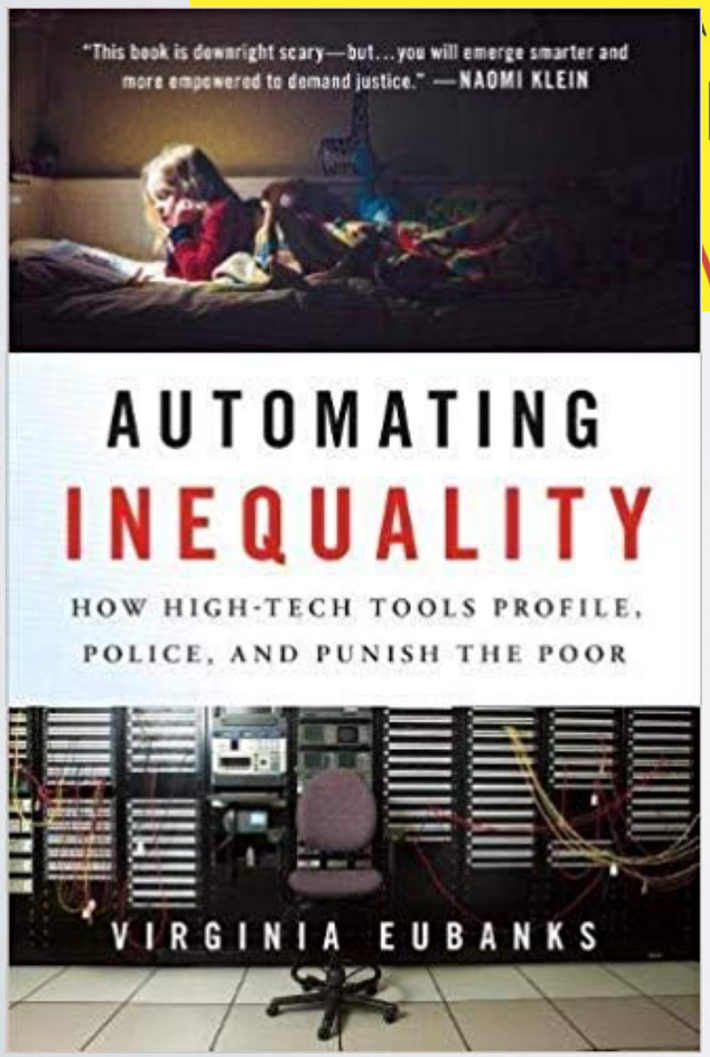
- Algorithms and machine learning can do amazing things
  - computer vision, robots, self-driving cars
  - detect diseases
  - generate Chopin-like music and create cool art
  - …
- Also, lots of jobs in ML and data science
  - according to LinkedIn these are top **2** fastest growing jobs

Machine Learning



- ML introduces new ethical concerns
  - Loss of jobs (increasing number of jobs can be automated)
  - Automated warfare (ML models decide who lives and who dies)
    - flash wars: what if autonomous weapons decide to engage in warfare and it escalates so fast we can't stop it
  - Accountability (who is responsible if a self-driving car crashes)
  - Fairness (are ML models fair? or do they discriminate?)





Amazon Is Pushing Facial Technology That a Study Says Could Be Biased

In new tests, Amazon's system had more difficulty identifying the gender of female and darker-skinned faces than similar services from IBM and Microsoft.

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

What's Going On?

- Machine learning models are being deployed everywhere
  - self-driving cars, insurance, criminal justice system, policing, education, healthcare, …
- Being used in the real-world to make decisions that affect people's lives
  - these decisions don't always seem "fair"
- This leads to several important questions
  - what do we mean by bias and fairness?
  - why are these models biased? where does the bias come from?
  - can we design ML models that are fair?

What is Discrimination?

- How would you define discrimination?
- Wikipedia says
  - "Discrimination consists of treatment of an individual or group, based on their actual or perceived membership in a certain group or social category, in a way that is worse than the way people are usually treated"
  - But this kind of fuzzy
- Let's look to the law for a more precise definition

What is Discrimination?

- The Civil Rights Act of 1964
  - Title VII (Equal Employment Opportunity) "prohibits discrimination by covered employers on the basis of race, color, religion, sex or national origin."

- Title VII can be violated in 2 ways
  - disparate treatment: employer's actions were motivated by discriminatory intent
  - disparate impact: employer's actions were discriminatory in its effect; *even if there was no discriminatory intent*

Examples of Disparate Impact

- A zoning ordinance that limits the type of residence could disproportionally impact people with disabilities
- Condo rules that ban signs and other materials in hallways would disproportionally impact observant Jews (who could not post a mezuzah)
- In 2006 (post Katrina) the Parish of St. Bernard in New Orleans passed a law prohibiting residents to rent their homes to anyone except blood relatives
  - since **96%** of residents were White this disproportionally affected everyone else

Machine Learning for Hiring

- In 2014, Amazon created team to explore automated hiring
  - "*an engine where I'm going to give you* **`100`** *resumes, it will spit out the top five, and we'll hire those*" — Reuters
  - recognizing the top resumes done using a ML model
- Team created **`500`** models for various kinds of jobs
- Models were trained on **`10`** years of resumes submitted to Amazon
- What do you think happened and why?

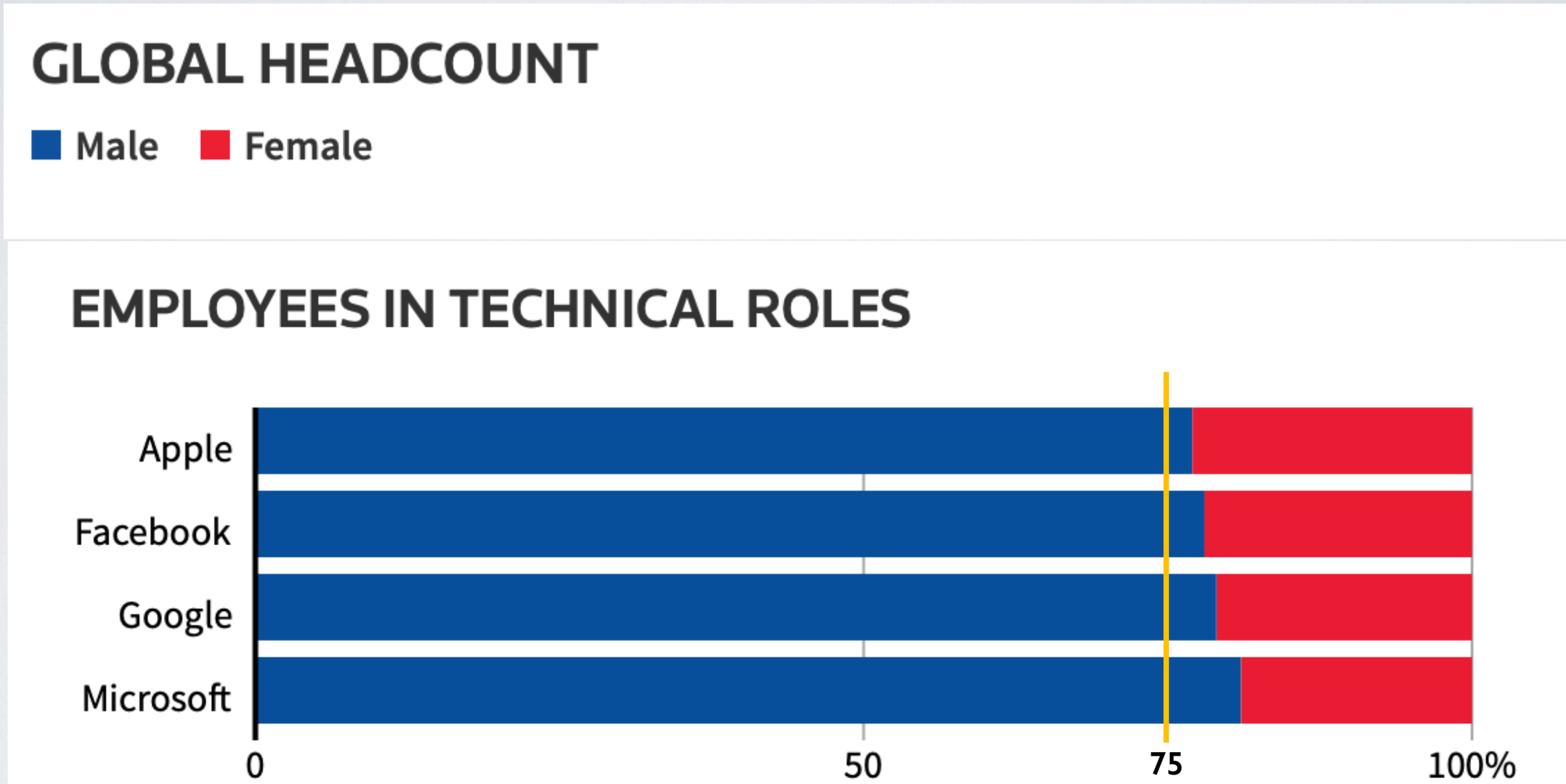
Machine Learning for Hiring

- How did the models do?
  - downgraded resumes that included "women's"
  - downgraded graduates from two all-women universities
  - upgraded resumes that included "executed" and "captured" (more commonly found in men's resumes)
- Why do you think this happened?

Machine Learning for Hiring

- Amazon's training data most likely included examples composed of
  - a resume and a yes/no hiring decision as the classification
- What are some problems with this training data?
- Gender imbalance
  - tech industry is highly gender imbalanced
  - training data likely had very few resumes of women…
  - …and a lot of resumes for men
  - not enough women examples for model to make good decisions

Gender Imbalance in Tech Industry



Note: Amazon does not disclose the gender breakdown of its technical workforce.

Source: Latest data available from the companies, since 2017.

Machine Learning for Hiring

- What other problems could have occurred?
- What if Amazon's hiring practices are gender-biased?
  - then the classifications in the training data would be biased
  - during training, the model would learn those biases as well
- In 2015, Amazon noticed these problems and shut the project down
- But many other companies are using ML for hiring
  - Goldman Sachs uses tool to match candidates to division that "fits best"
  - LinkedIn can rank candidates based on fit for a job advertisement

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

Criminal Risk Assessment Tools

- COMPAS by Northpointe predicts
  - Risk of new violent crime
  - Risk of general recidivism
  - Pretrial risk (failure to appear)

From Lecture 15

ProPublica Study

- In **2016** ProPublica conducted a study of COMPAS
  - **7000** arrests in Broward County, FL
  - between **2013** and **2014**
- OK predictions for *all* crimes (misdemeanors included)
  - **61%** of people labeled high risk committed new crimes
- But unreliable for *violent* crimes
  - **20%** of people labeled high risk committed new *violent* crimes

From Lecture 15

ProPublica Study

- Found significant *racial* disparities
- Out of people labeled high risk but didn't re-offend
  - `44.9%` were African American
  - `23.5%` were White
- Out of people labeled low risk but did re-offend
  - `28%` were African American
  - `47.7%` were White
- Study accounted for
  - Criminal history, age and gender

From Lecture 15

False Positives & False Negatives

- A model that classifies inputs into two classes can fail in two ways
- False positives
  - the model claims input is positive when input is negative
  - ex: person labeled as high-risk but person does not re-offend
- False negatives
  - the model claims input is negative when input is positive
  - ex: person labeled as low-risk but person does re-offend

False Positive Rate & False Negative Rate

Input	true clasification	model classification
x_1	F	F
x_2	T	F
x_3	T	T
x_4	F	T
x_5	T	F
x_6	F	T

Activity #2

1 min

ProPublica Study

- ProPublica showed that COMPAS *fails* differently for different groups
  - false positive rate of African Americans is **44.9%** and of whites is **23.5%**
  - false negative rate of African Americans is **28%** and of whites is **47.7%**
- Study shows that COMPAS violates the *error rate balance* property
  - error rate balance: for groups g_1, g_2
    - $FPR(g_1)=FPR(g_2)$ and $FNR(g_1)=FNR(g_2)$
- Northpointe countered that COMPAS satisfies the *predictive parity* property
  - predictive parity ≈ when considering people labeled high risk, the probability they re-offended is the same no matter which group they belonged to
- So is COMPAS biased or not?

Algorithmic Fairness

- Turns out we can define ≈ 20 different notions of algorithmic fairness
  - *error balance rate*, *predictive parity*, *calibration*, *statistical parity*, …, and *disparate impact*
- Why so many?
  - they *seem* to capture different intuitions we have about fairness
  - some are more appropriate to certain situations than others
  - some are related and some are even contradictory!
    - Kleinberg, Mullainathan & Raghavan, and Chouldechova proved that…
    - …no model can satisfy both *calibration* and *error rate balance*

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

Alexandra Chouldechova *

Last revised: February 2017

Abstract

Recidivism prediction instruments (RPI's) provide decision makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time. While such instruments are gaining increasing popularity across the country, their use is attracting tremendous controversy. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. This paper discusses several fairness criteria that have recently been applied to assess the fairness of recidivism prediction instruments. We demonstrate that the criteria cannot all be simultaneously satisfied when recidivism prevalence differs across groups. We then show how disparate impact can arise when a recidivism prediction instrument fails to satisfy the criterion of error rate balance.

Keywords: disparate impact; bias; recidivism prediction; risk assessment; fair machine learning

1 Introduction

Risk assessment instruments are gaining increasing popularity within the criminal justice system, with versions of such instruments being used or considered for use in pre-trial decision-making, parole decisions, and in some states even sentencing [1, 2, 3]. In each of these cases, a high-risk classification—particularly a high-risk misclassification—may have a direct adverse impact on a criminal defendant's outcome. If the use of RPI's is to become commonplace, it is especially important to ensure that the instruments are free from discriminatory biases that could result in unethical practices and inequitable outcomes for different groups.

In a recent widely popularized investigation conducted by a team at ProPublica, Angwin et al. [4] studied an RPI called COMPAS[a], concluding that it is biased against black defendants. The

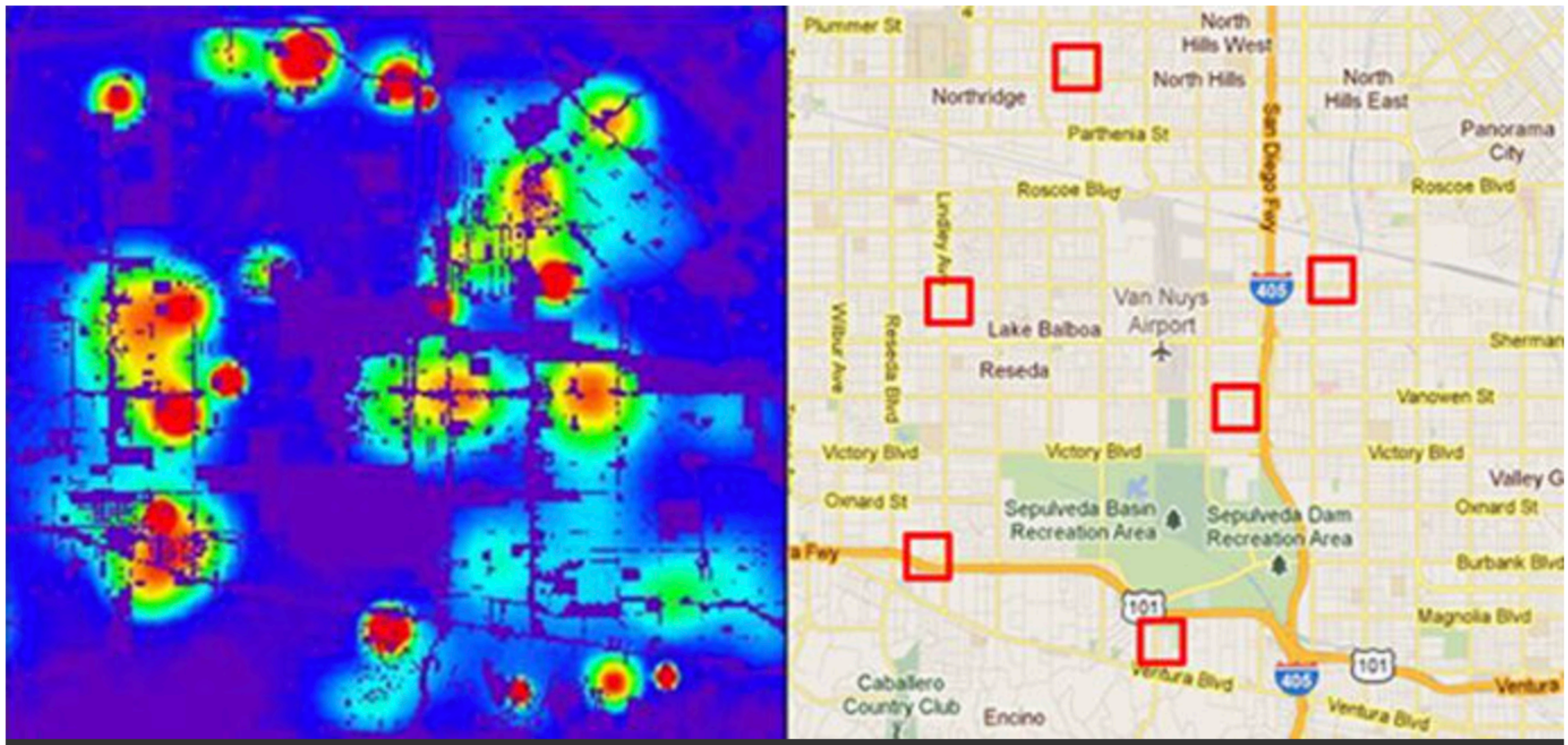
*Heinz College, Carnegie Mellon University

[a]COMPAS [5] is a risk assessment instrument developed by Northpointe Inc.. Of the 22 scales that COMPAS provides, the Recidivism risk and Violent Recidivism risk scales are the most widely used. The empirical results in this paper are based on decile scores coming from the COMPAS Recidivism risk scale.

- Chouldechova *also* proved that
- if a model violates *error rate balance*…
- …then it must violate *disparate impact*

Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?

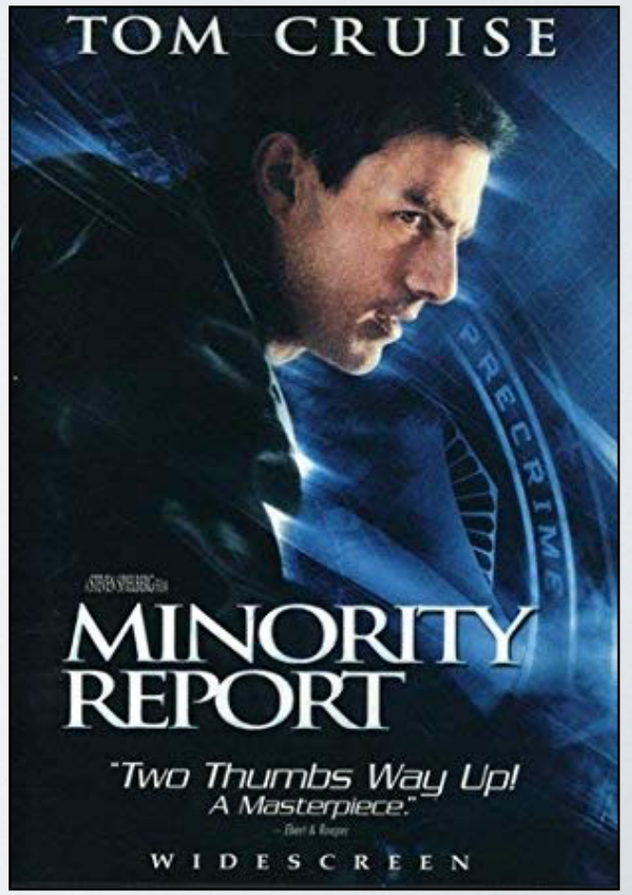
The software is supposed to make policing more fair and accountable. But critics say it still has a way to go.



Predictive policing is built around algorithms that identify potential crime hotspots.. (PredPol)

Predictive Policing

- Models that use historical crime data to predict crimes at various locations
- Used by police departments in
  - California, Washington, South Carolina, Arizona, Tennessee, Illinois
- Multiple systems available
  - PredPol (used by LAPD)
  - HunchLab
  - IBM

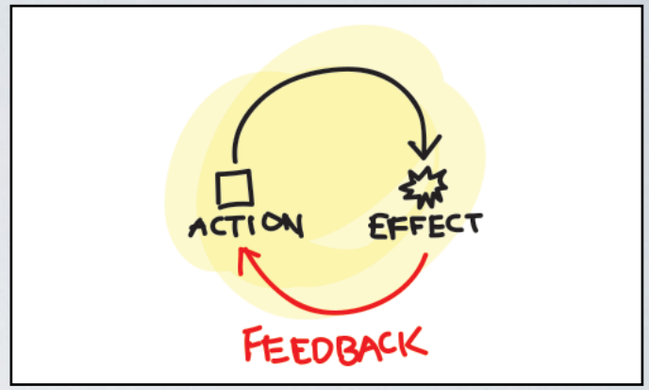


HunchLab

Event Models

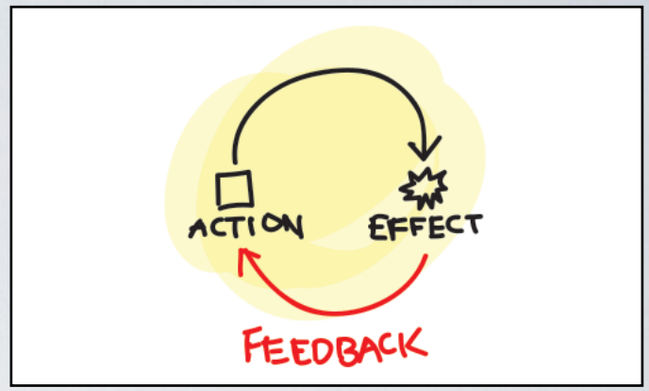
Larceny	71%
Larceny from Auto	12%
Vandalism	8%
Burglary Residence	7%
Burglary Nonresidence	2%
Auto Theft	0%
Robbery	0%
Risk Percentile	97%
Risk Z Score	2

Training Machine Learning Models



- "Traditional" model training
  - split examples into training data & test data
  - use training data to train model & test data to test model
  - use model on new input to generate a prediction
- In practice, models are sometimes used with *feedback*
  - use training data to train model & test data to test model
  - use model on new input to generate a prediction
  - after prediction fails/succeeds, use that knowledge to update model

Training Machine Learning Models



- Predictive policing
  - split historical crime data into training and test sets
  - use training data to train model & use test data to test model
  - use model to predict crime at some location
    - if crime did not occur, model was wrong and update it accordingly
    - if crime did occur, model was right and update it accordingly
- The model's decisions are affecting its own training
  - if model indicates high probability of crime at some location…
  - …and officer is sent there then we are more likely to see crime at location
  - the feedback used to update the model is influenced by its decisions

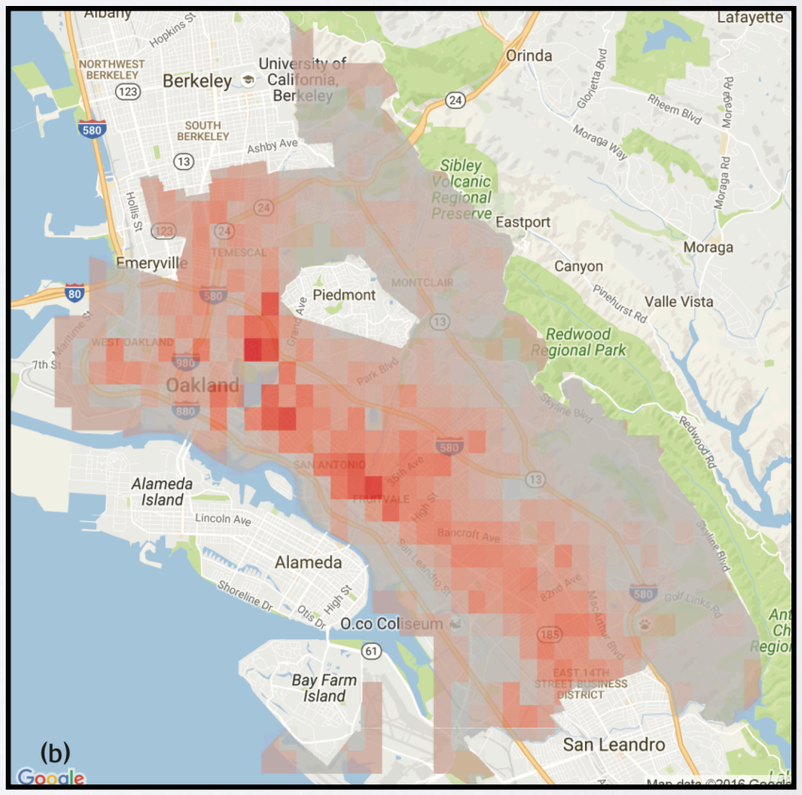
Feedback Loops

- If training data is biased…
  - …then model will make biased decisions…
  - …which are used to create new training data…
  - …and model will make more biased decisions…
- Do predictive policing systems suffer from feedback loops?

Feedback Loops in PredPol

- Kristian Lum and William Isaac decided to study this question
- First, they argued that police crime data is biased by comparing
  - Oakland police department records of drug arrests in 2010
  - to estimates of drug use from public-health data

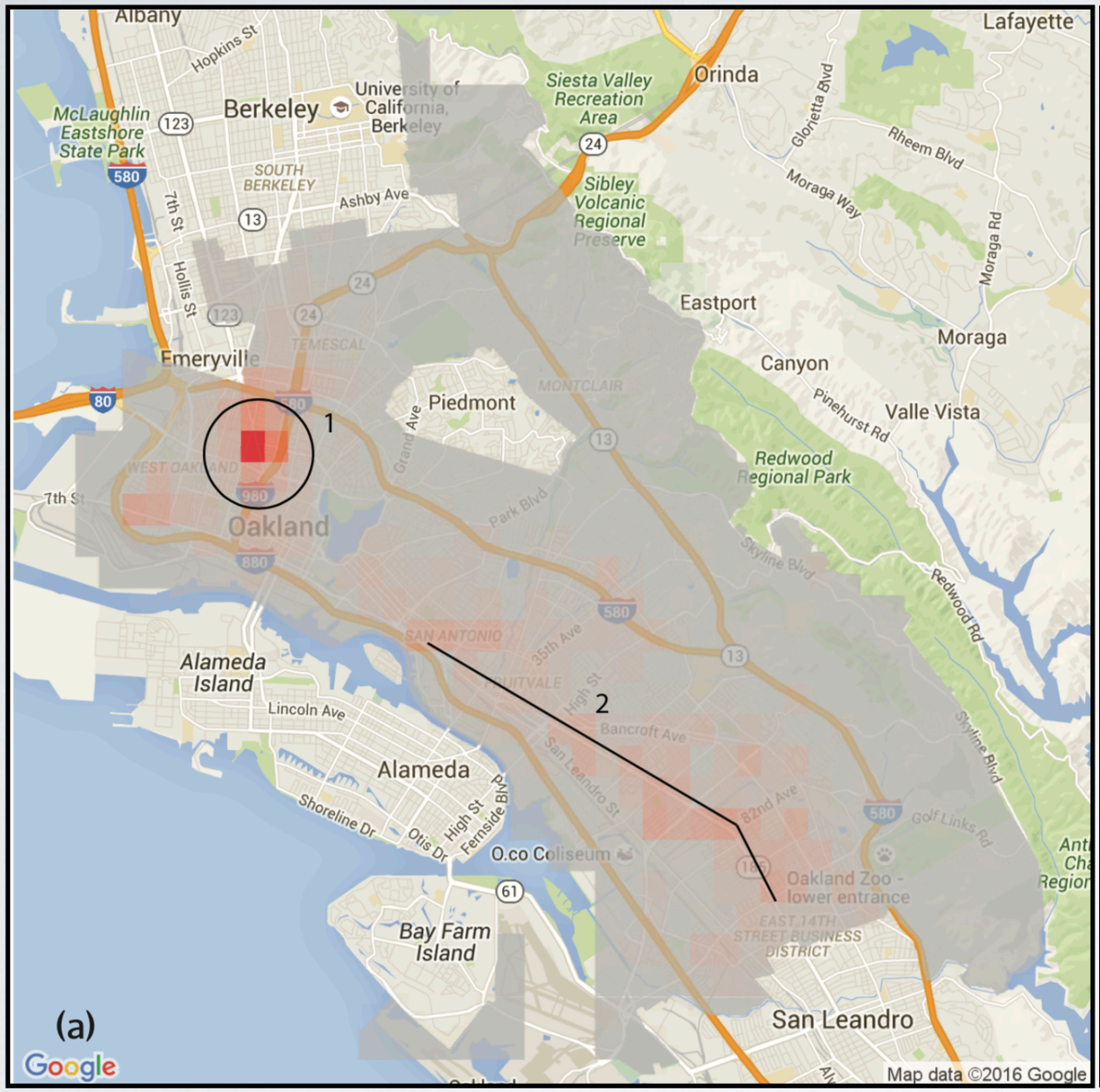
Police records of drug arrests



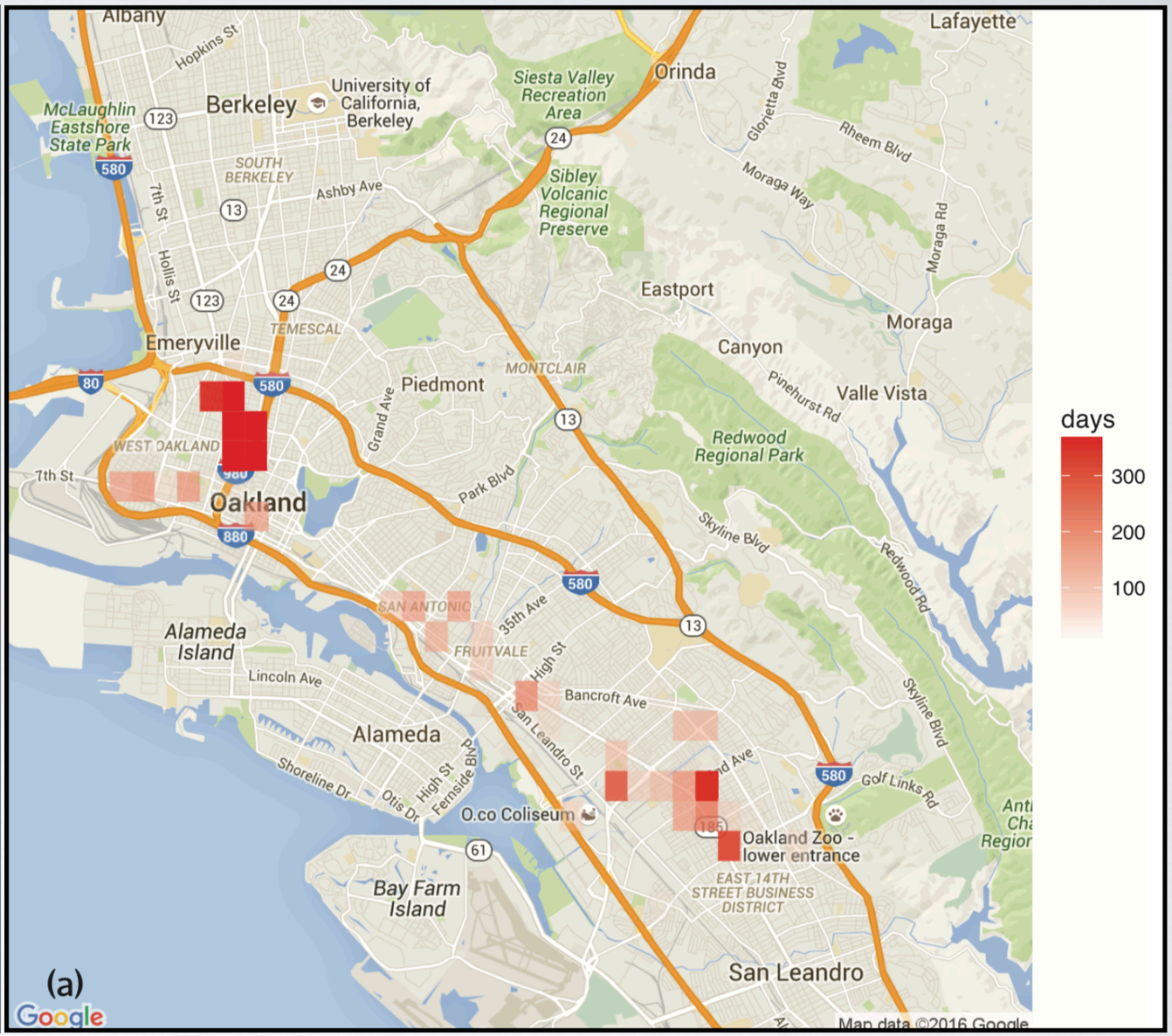
Estimated drug use from public health data

Feedback Loops in PredPol

- Then simulated PredPol on Oakland PD data
  - PredPol predicts crime rates across city for the next day
  - Areas with highest rates are flagged & receive more officers next day
  - Ran predictions for every day of 2011
    - For each location, counted how many days it would be flagged
- Findings
  - PredPol *increased* the bias already found in Oakland PD data
  - Most flagged areas were the ones already over-represented in data



Police records of drug arrests



Number of days flagged by PredPol

Feedback Loops in PredPol

- Also found that PredPol affected different groups differently
  - Drug use is roughly equal among races but
  - simulations showed that PredPol would cause Black people to be targeted by police at **`2x`** the rate as White people and others at **`1.5x`** times the rate
- (study defined others only as non-white and non-black)

How to Handle Feedback Loops?

- Lum and Isaac's work was eye opening!
- Motivates the question
  - "can we do anything about these feedback loops"
- Last year, professors and undergrads (!) from
  - U. of Utah, Harverford and U. of Arizona
  - studied and answered this question

Proceedings of Machine Learning Research 81:1–12, 2018 Conference on Fairness, Accountability, and Transparency

Runaway Feedback Loops in Predictive Policing*

Danielle Ensign DANIPHYE@GMAIL.COM
University of Utah

Sorelle A. Friedler SORELLE@CS.HAVERFORD.EDU
Haverford College

Scott Neville DROP.SCOTT.N@GMAIL.COM
University of Utah

Carlos Scheidegger CSCHEID@CSCHEID.NET
University of Arizona

Suresh Venkatasubramanian† SURESH@CS.UTAH.EDU
University of Utah

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

Predictive policing systems are increasingly used to determine how to allocate police across a city in order to best prevent crime. Discovered crime data (e.g., arrest counts) are used to help update the model, and the process is repeated. Such systems have been empirically shown to be susceptible to runaway feedback loops, where police are repeatedly sent back to the same neighborhoods regardless of the true crime rate.

In response, we develop a mathematical model of predictive policing that proves why this feedback loop occurs, show empirically that this model exhibits such problems, and demonstrate how to change the inputs to a predictive policing system (in a black-box manner) so the runaway feedback loop does not occur, allowing the true crime rate to be learned. Our results are quantitative: we can establish a link (in our model) between the degree to which runaway feedback causes problems and the disparity in crime rates between areas. Moreover, we can also demonstrate the way in which *reported* incidents of crime (those reported by residents) and *discovered* incidents of crime (i.e. those directly observed by police officers dispatched as a result of the predictive policing algorithm) interact: in brief, while reported incidents can attenuate the degree of runaway feedback, they cannot entirely remove it without the interventions we suggest.

Keywords: Feedback loops, predictive policing, online learning.

1. Introduction

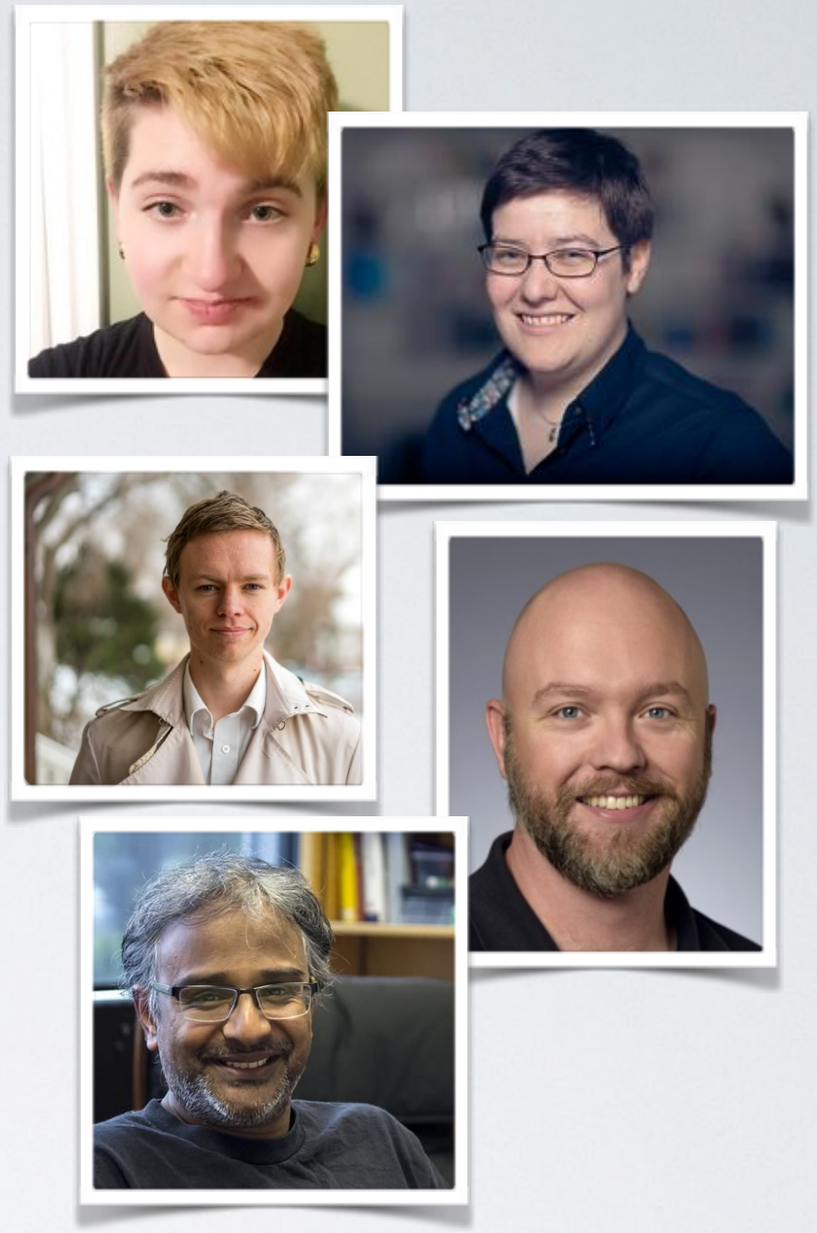
Machine learning models are increasingly being used to make real-world decisions, such as who to hire, who should receive a loan, where to send police, and who should receive parole. These deployed models mostly use traditional batch-mode machine learning, where decisions are made and observed results supplement the training data for the next batch.

However, the problem of *feedback* makes traditional batch learning frameworks both inappropriate and (as we shall see) incorrect. Hiring algorithms only receive feedback on people who were hired, predictive policing algorithms only observe crime in neighborhoods they patrol, and so on. Decisions made by the system influence the data that is fed to it in the future. For example, once a decision has been made to patrol a certain neighborhood, crime discovered in *that* neighborhood will be fed into the training apparatus for the next round of decision-making.

In this paper, we focus on predictive policing – an important exemplar problem demonstrating

* This research was funded in part by the NSF under grants IIS-1633387, IIS-1513651, and IIS-1633724. Code for our urn simulations can be found at https://github.com/algofairness/runaway-feedback-loops-src.

† Corresponding author.



How to Handle Feedback Loops?

- Using advanced techniques from statistics
  - they showed mathematically that PredPol is vulnerable to feedback loops
  - found a strategy that provably fixes PredPol's biases

- Suppose model sends police to
  - location A `90%` of the time
  - location B `10%` of the time
- Update training data as follows
  - if crime occurs in location A, ignore this example with prob `.9`
  - if crime occurs in location B, ignore this example with prob `.1`

Algorithms

- In 16 you've learned how to *design*, *analyze* and *implement* algorithms
- You learned the algorithmic foundations of most of CS
  - big-O, worst-case analysis, amortized analysis and expected analysis
  - recursion, dynamic programming, hash tables, binary trees, priority queues, sorting algorithms, shortest paths, minimum spanning trees, decision trees and neural networks
- You now know how to design fast algorithms
  - this is a valuable skill that you will use throughout your career

Algorithms

- You've seen examples of powerful algorithms
  - Seamcarve, PageRank, ID3, Multi-Layer Perceptrons
- And you've seen examples of harmful algorithms
  - COMPAS, PredPol
- Don't forget that ultimately your algorithms impact people
  - sometimes in direct ways and sometimes in indirect ways
- Always be mindful of that and think about
  - the positive impact of your work
  - but also the (potentially) negative impact of your work