Algorithmic Bias and Discrimination

CS16: Introduction to Data Structures & Algorithms
Spring 2020

Outline

- Potential and limitations of machine learning
- Sources of unfairness
- Bias in computer vision
- ML in criminal risk assessment
- Fairness Criteria
- Feedback Loops

Forbes

Billionaires Innovation

The New York Times

ADVERTISEMENT

The New York Times

Account

TECHNOLOGY

The New York Times

PLAY THE

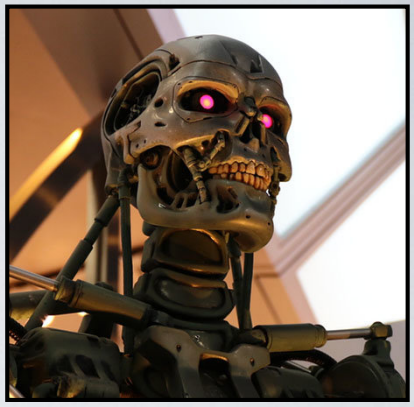
ADVERTISEMENT

Turing Award Won by 3 Pioneers in Artificial Intelligence

Forget go

3

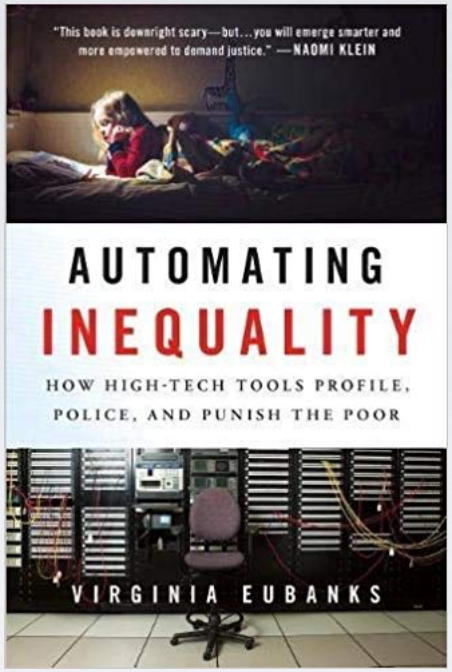
Machine Learning

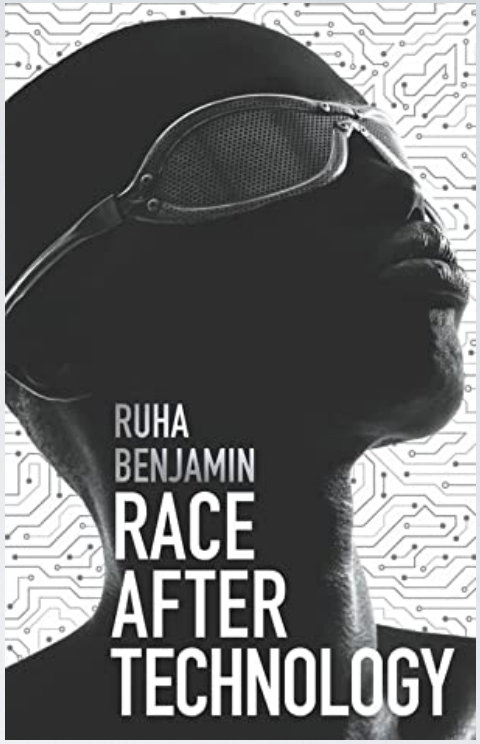


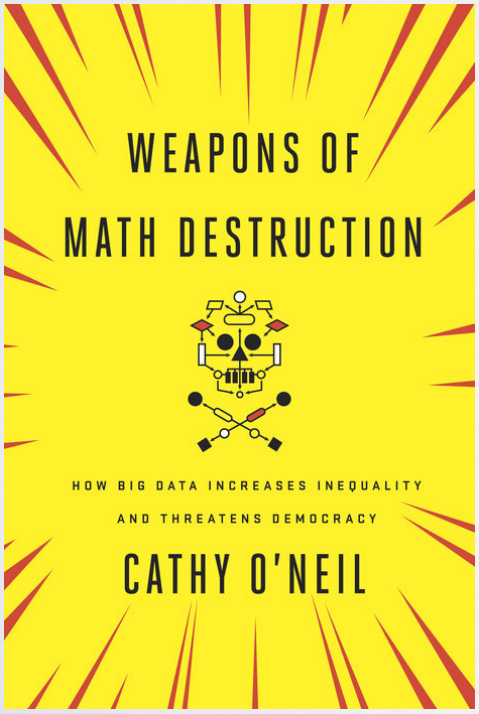
- ML introduces new ethical concerns
  - Loss of jobs (increasing number of jobs can be automated)
  - Automated warfare (ML models decide who lives and who dies)
    - flash wars: what if autonomous weapons decide to engage in warfare and it escalates so fast we can't stop it
  - Accountability (who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

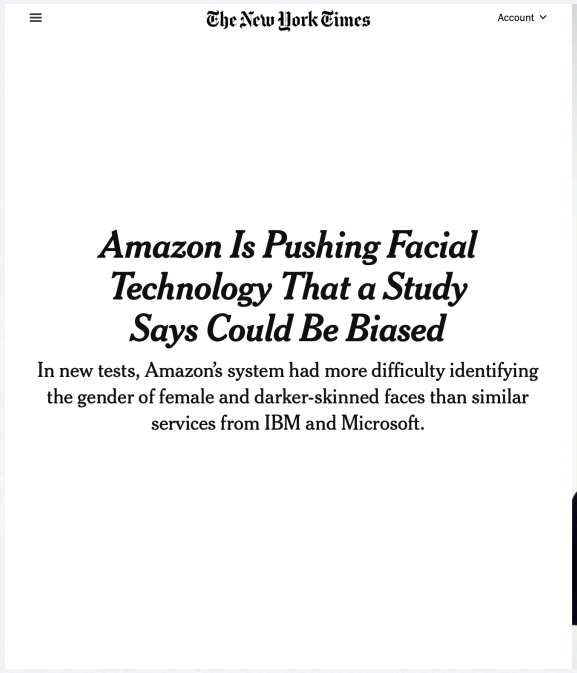
ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)

A computer science conference with a cross-disciplinary focus that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.









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?

- 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)
- Griggs v. Duke Power Co.
  - Case involving Willie Griggs, man who filed a class-action lawsuit on behalf of himself and other black employees
  - Company required that all employees who wanted to work transfer to higher positions achieve a minimum score on aptitude tests and have high school diploma
  - Duke's policy did, in fact, discriminate a protected class of people, even if unintentional

Discrimination

- **Domain** specific
- not **arbitrary** - historically and systematically, we have used these determiners as bases for unjustified adverse treatment in the past

Sources of unfairness

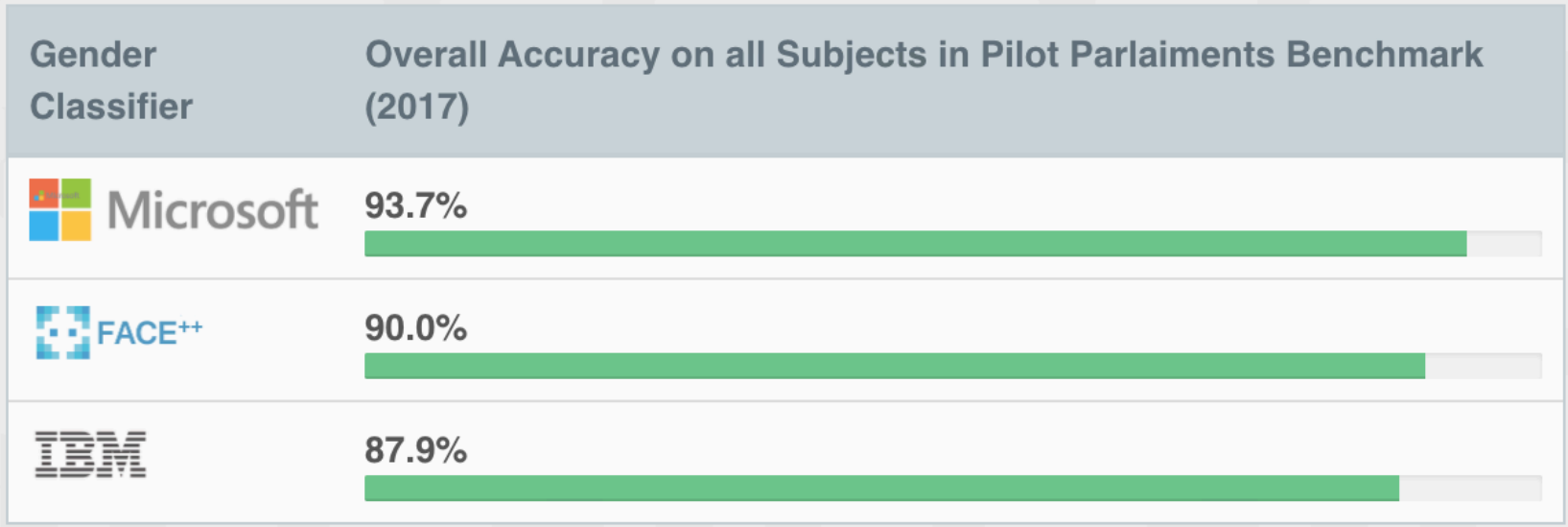
"How can machine learning wind up being unfair without any explicit wrongdoing?"

Gender Shades

- Evaluates the accuracy of AI powered gender and racial classification products
- Led by a team of researchers at MIT Media Lab
- Bias is defined as "having practical differences in gender classification error rates between groups"

Gender Shades

- Used Pilot Parliaments Benchmark
  - 1270 images consisting of subjects selected from 3 African countries and 3 European countries, grouped by gender, skin type, and intersection of both
- How did IBM, Microsoft, and Face++ AI and computer vision products for classifying gender do across the board?



Pretty well! Right?

Accuracy Rates Across Groups

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Oh… what's wrong here?

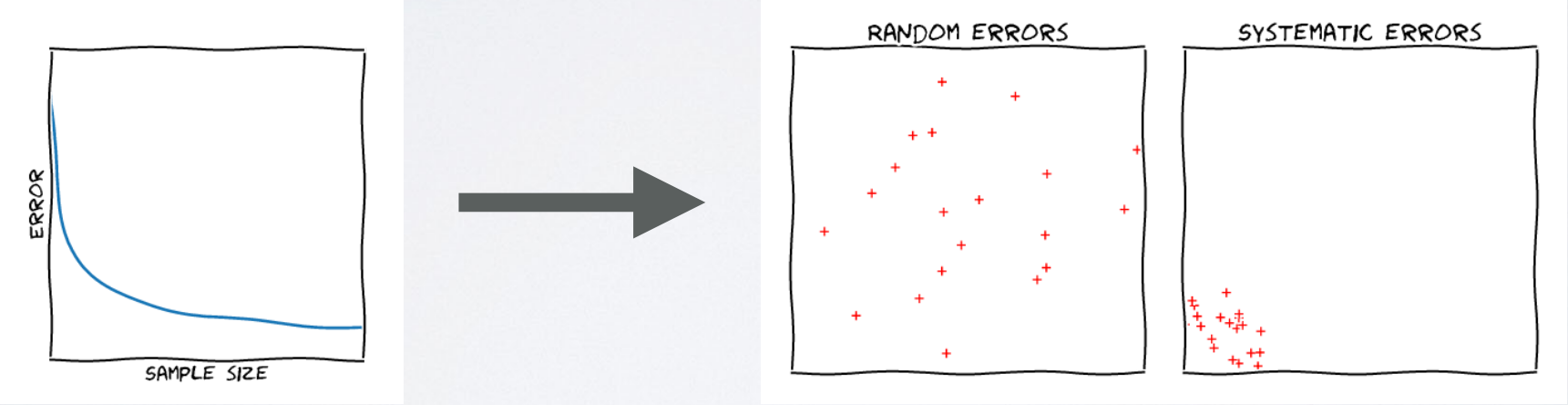

93.6%
of error
Microsoft
ERROR
34.4%
DIFFERENCE
IBM

"A company might justify the market readiness of a classifier by presenting performance results in aggregate. Yet a gender and phenotypic breakdown of the results shows that performance differs substantially for distinct subgroups. Classification is 8% - 20% worse on female than male subjects and 11% -19% worse on darker than lighter subjects."

– Buolamwini et al., MIT Media Lab

Sample size disparity

- Generally, the more data the better!
- But what if we have less data for minority groups?
  - general tendency for automated decisions to favor dominant group



- Overall low error =/= equal distribution of error rates

"The lesson is that statistical patterns that apply to the majority might be invalid within a minority group."

– Moritz Hardt, UC Berkeley

Biases in data

- Collection
  - Demographic, geographic, behavioral, temporal
- Pre-existing biases
  - Gender roles in text and images, racial stereotypes

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 classification	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 #1

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?

Depends on your definition of fairness…

Fairness Criteria

- Turns out we can define ≈ 20 different notions of algorithmic fairness
  - *error balance rate*, *predictive parity*, *calibration*, *statistical parity*, …,
  - 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***

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 male resumes)
- Why do you think this happened?

Calibration and Error Balance Rate

- ***Calibration*** - requires that outcomes are independent of protected attribute
  - ex: for any prediction, the probability of the applicant being qualified should be the same for women and men
- ***error balance rate*** - requires that both groups have the same probability of being classified as a false positive or false negative
  - ex: the probability of hiring an unqualified applicant should be the same for both women and men

Setup

- ***X*** features of an individual
- ***A*** sensitive attribute (gender, class, sexual orientation)
- ***C = C(X, A)*** classifier mapping X and A to some prediction
- ***Y*** actual outcome
- Using these factors, how can we come up with a definition of what's fair?

Demographic Parity

- Say company **Faceuberzon** uses a hiring classifier
- Suppose C and A are binary variables where:
  - if **C=1**, Faceuberzon hires, and if **C=0**, "Thank you for applying."
  - if **A** (say, age)**=1**, elderly, and if **A = 0**, young
- Classifier C satisfies *demographic parity* if:

$$\mathbf{P(C=1 \mid A=1) = P(C=1 \mid A=0)}$$

*Conditional probability notation: P(M=m|N=n) is the probability M=m given that N=n

- In other words, the probability Faceuberzon's classifier will hire the elderly should be the same as the probability Faceuberzon will hire the young

Equality of Opportunity

- Equalizes the *true positive rates* between the two groups, requiring the Faceuberzon to hire from both groups at an equal rate among applicants who are qualified
- Classifier C satisfies *equality of opportunity* (also known as true positive parity) if:

$$\mathbf{P(C=1 \mid Y=1, A=1) = P(C=1 \mid Y=1, A=0)}$$

*Y is the actual outcome

- In other words, the probability that Faceuberzon's classifier will hire someone, given that they are elderly and qualified, is the same as the probability Faceuberzon will hire someone that is young and qualified

What if we could prevent the algorithm from looking at protected attributes such as race, color, religion, gender, etc.?

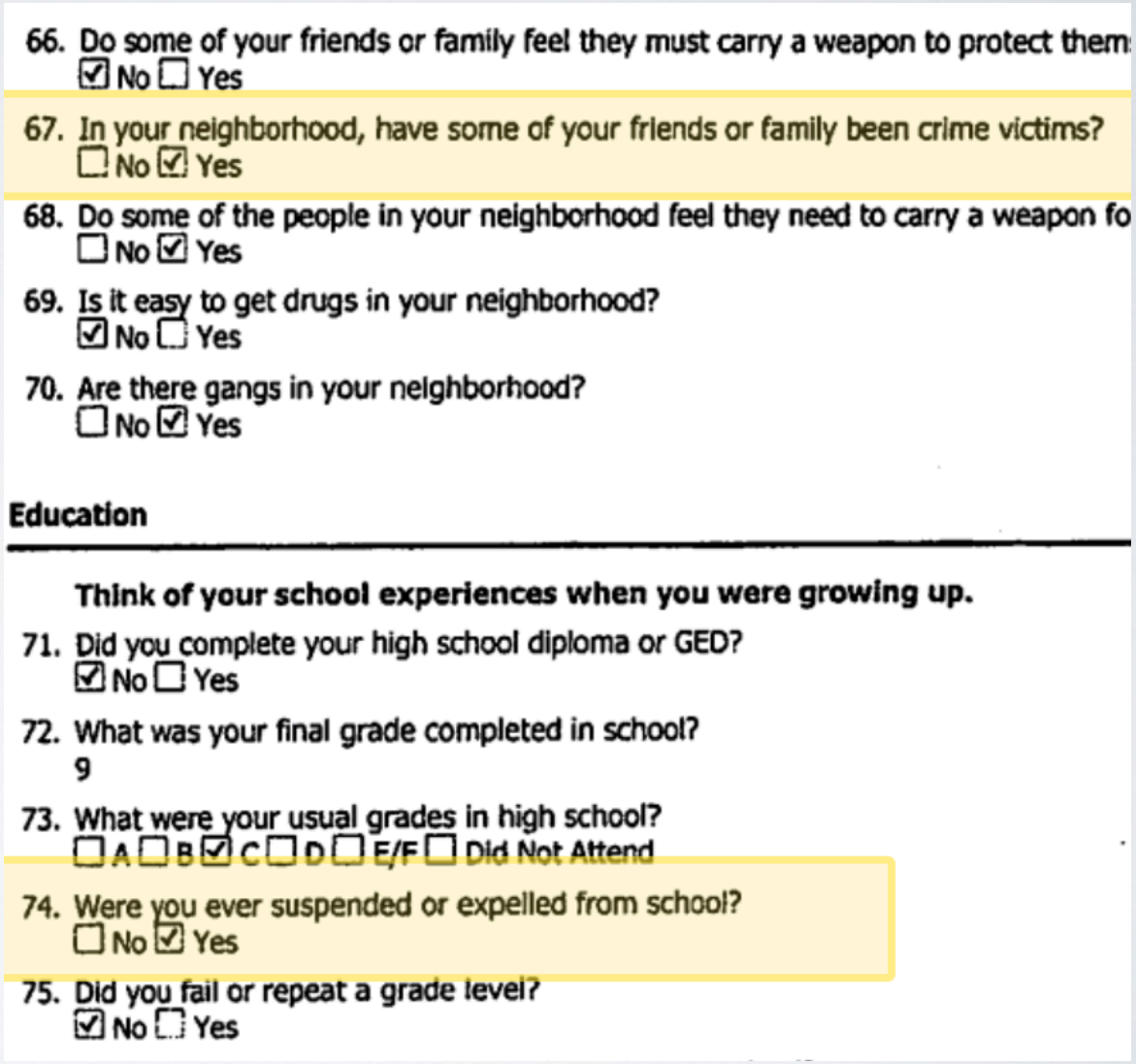
… a solution?

Fairness through blindness

- Problem: there are always ways of predicting hidden protected attributes through other features in the dataset
- ***Redundant encodings***

Survey used for COMPAS

- Did not explicitly ask about things like race or location… but



66. Do some of your friends or family feel they must carry a weapon to protect them
☑ No ☐ Yes

67. In your neighborhood, have some of your friends or family been crime victims?
☐ No ☑ Yes

68. Do some of the people in your neighborhood feel they need to carry a weapon fo
☐ No ☑ Yes

69. Is it easy to get drugs in your neighborhood?
☑ No ☐ Yes

70. Are there gangs in your neighborhood?
☐ No ☑ Yes

Education

Think of your school experiences when you were growing up.

71. Did you complete your high school diploma or GED?
☑ No ☐ Yes

72. What was your final grade completed in school?
9

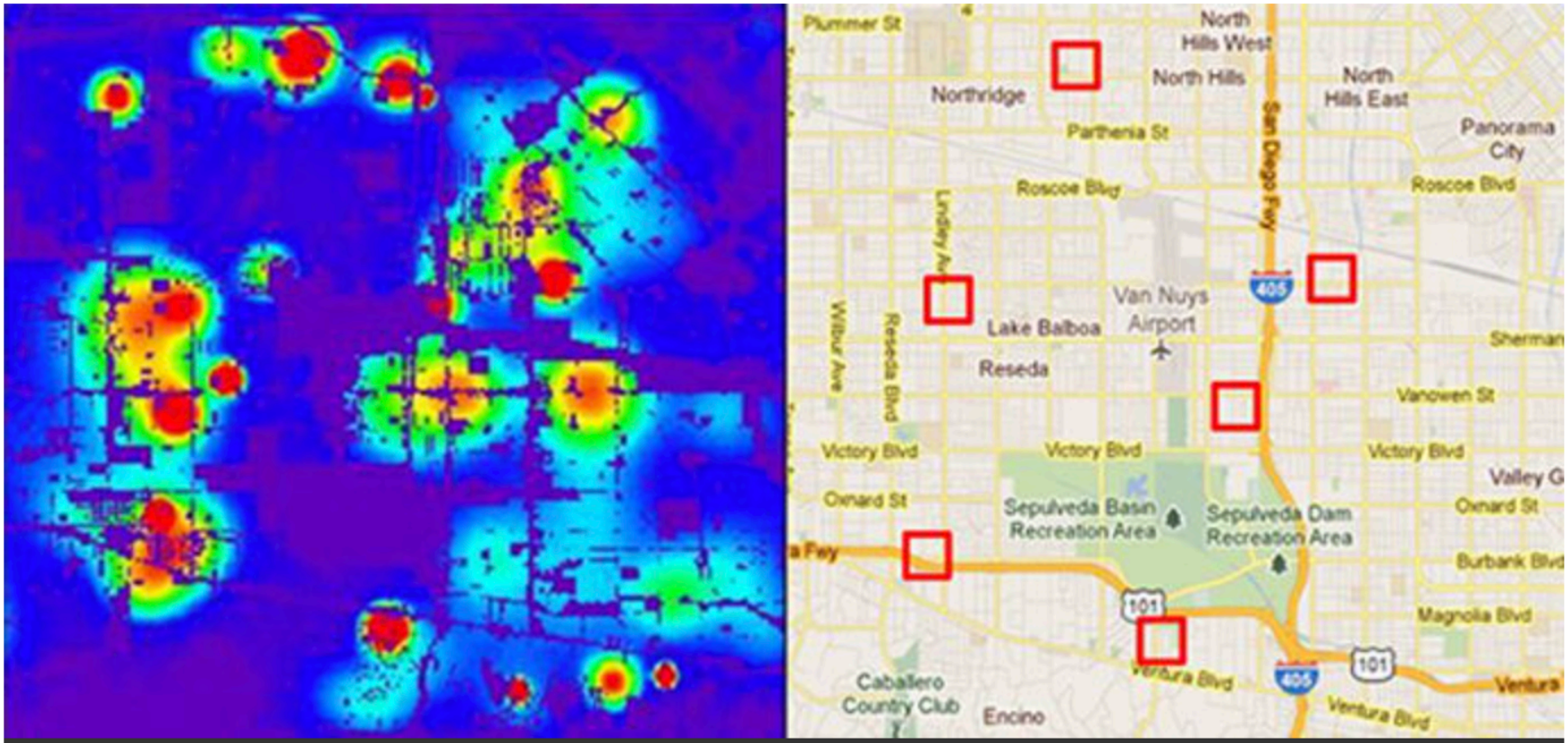
73. What were your usual grades in high school?
☐ A ☐ B ☑ C ☐ D ☐ E/F ☐ Did Not Attend

74. Were you ever suspended or expelled from school?
☐ No ☑ Yes

75. Did you fail or repeat a grade level?
☑ No ☐ Yes

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

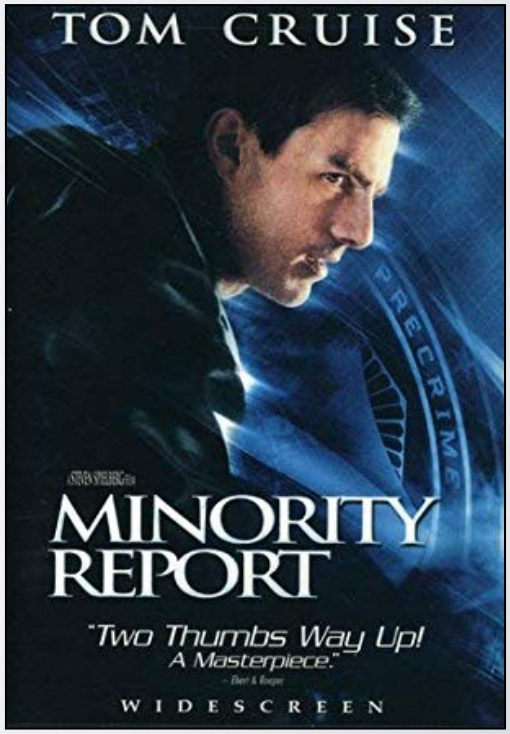
The software is supposed to make policing more fair and accountable. But critics say it still has a way to go.

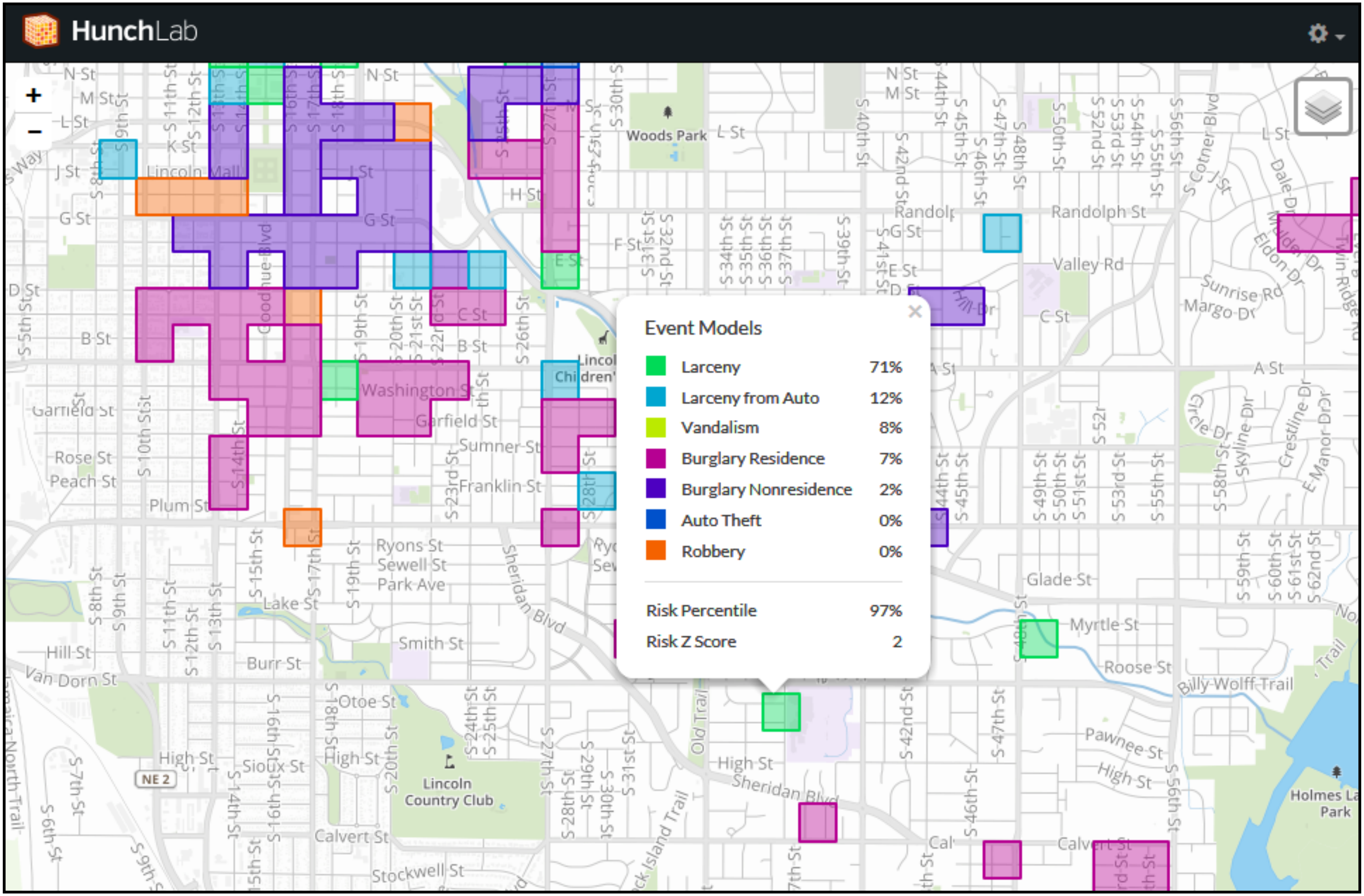


Predictive policing is built around algorithms that identify potential crime hotspots.. (PredPol)

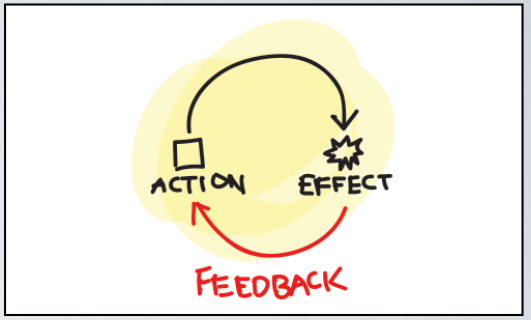
Predictive Policing

- 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



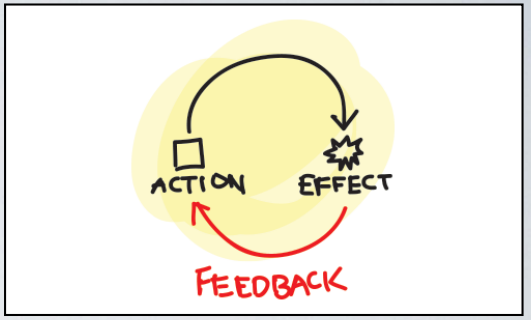

HunchLab
Event Models
Larceny 71%
Larceny from Auto 12%
Vandalism 8%
Burglary Residence 7%
Burglary Nonresidence 2%
Auto Theft 0%
Robbery 0%
Risk Percentile 97%
Risk Z Score 2
Woods Park
Lincoln Mall
Lincoln Country Club
Washington St
Sheridan Blvd
Billy Wolff Trail

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



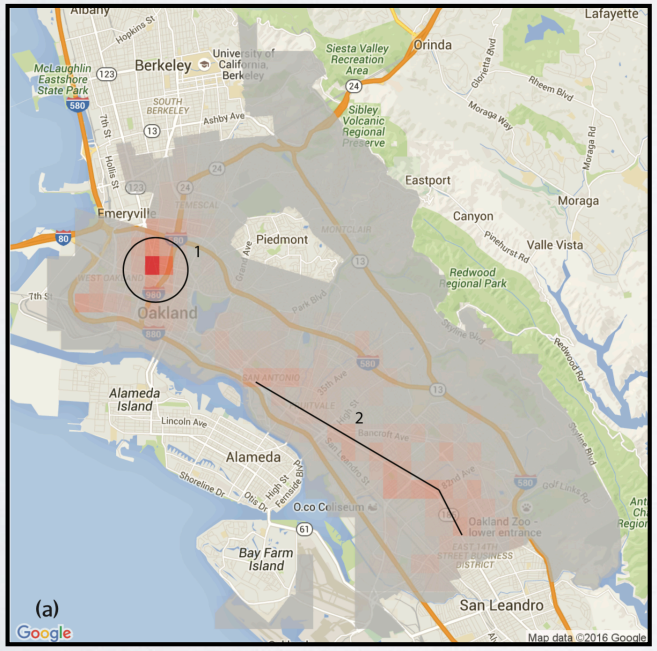
- Predictive policing
  - split historical crime data into training and test sets
  - use training data to train model & use test data to test model
  - use model to predict crime at some location
  - if crime did not occur, model was wrong and update it accordingly
  - if crime did occur, model was 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

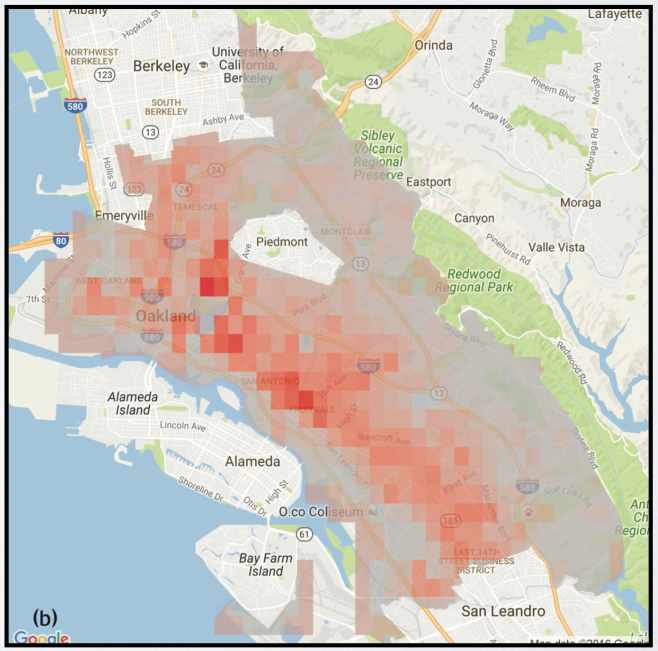
- If training data is biased…
  - …then model will make biased decisions…
  - …which are used to create new training data…
  - …and model will make more biased decisions…
- Do predictive policing systems suffer from feedback loops?

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



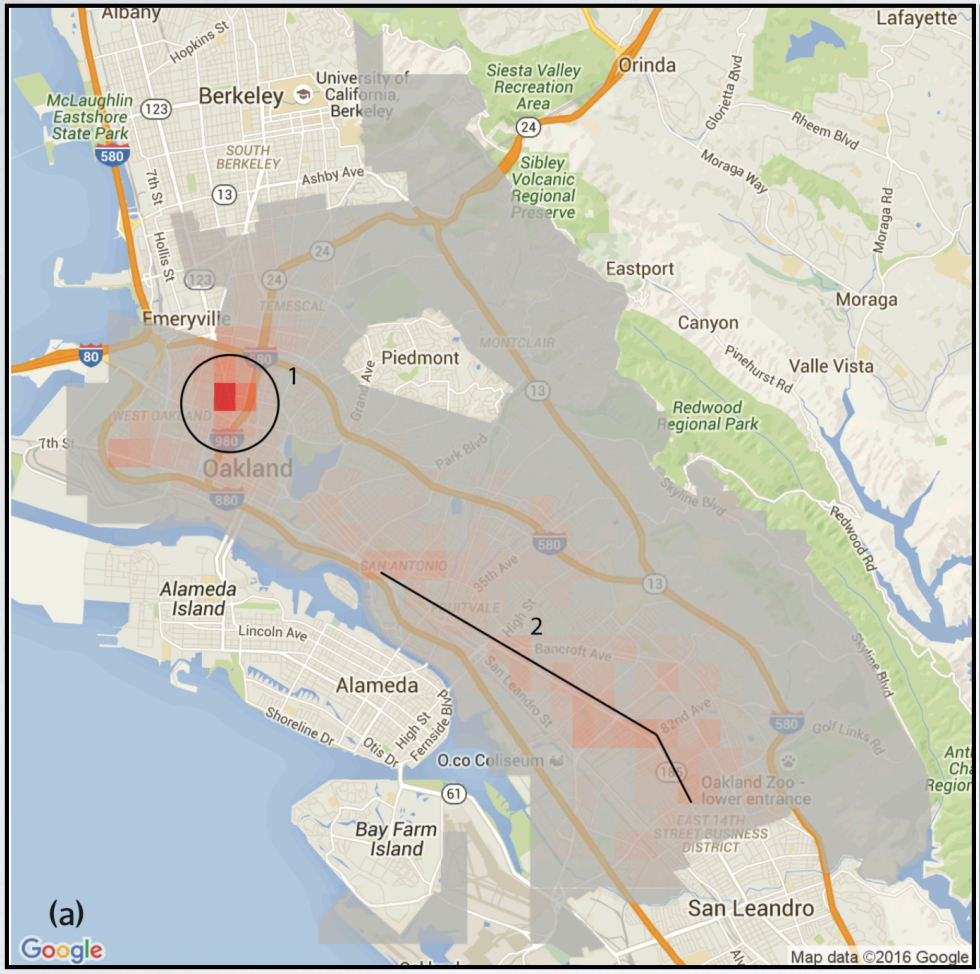
Police records of drug arrests



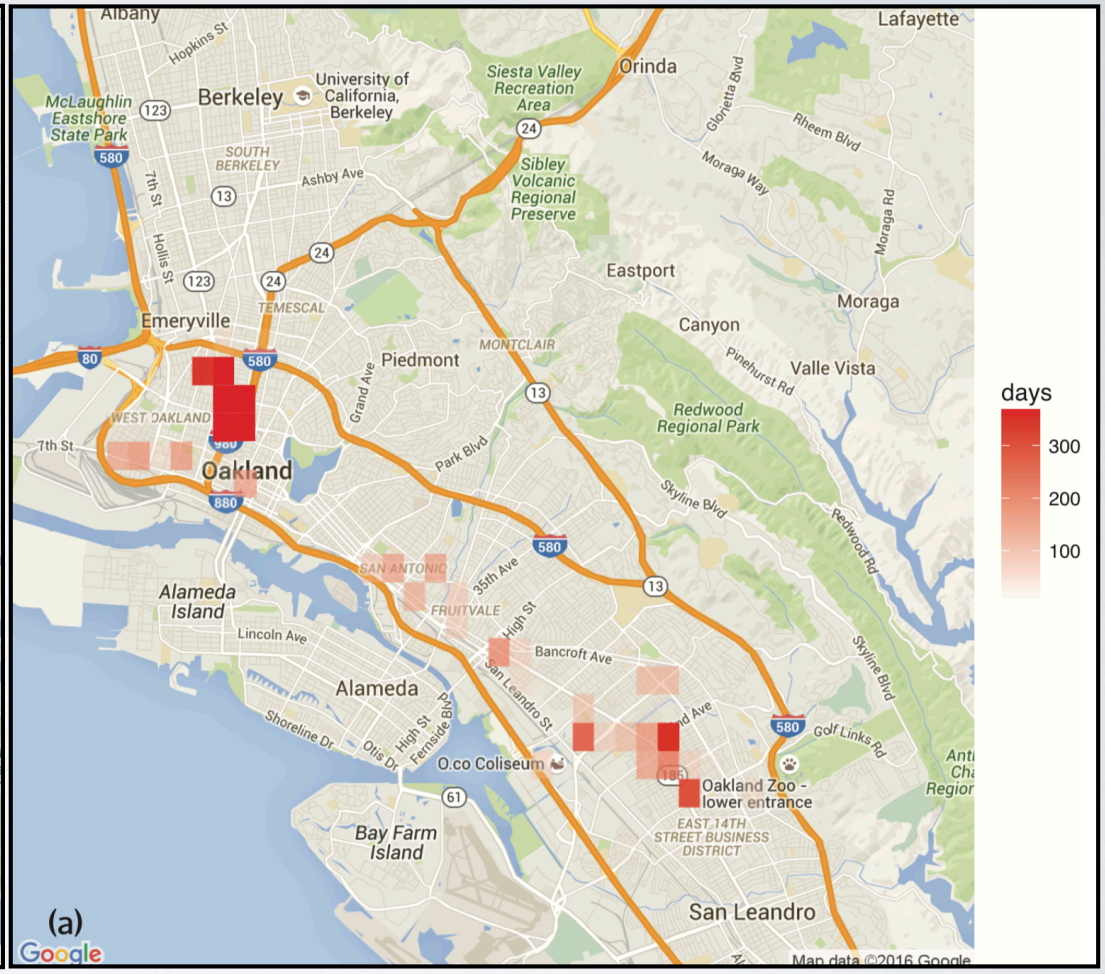
Estimated drug use from public health data

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
  - (only define others 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

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

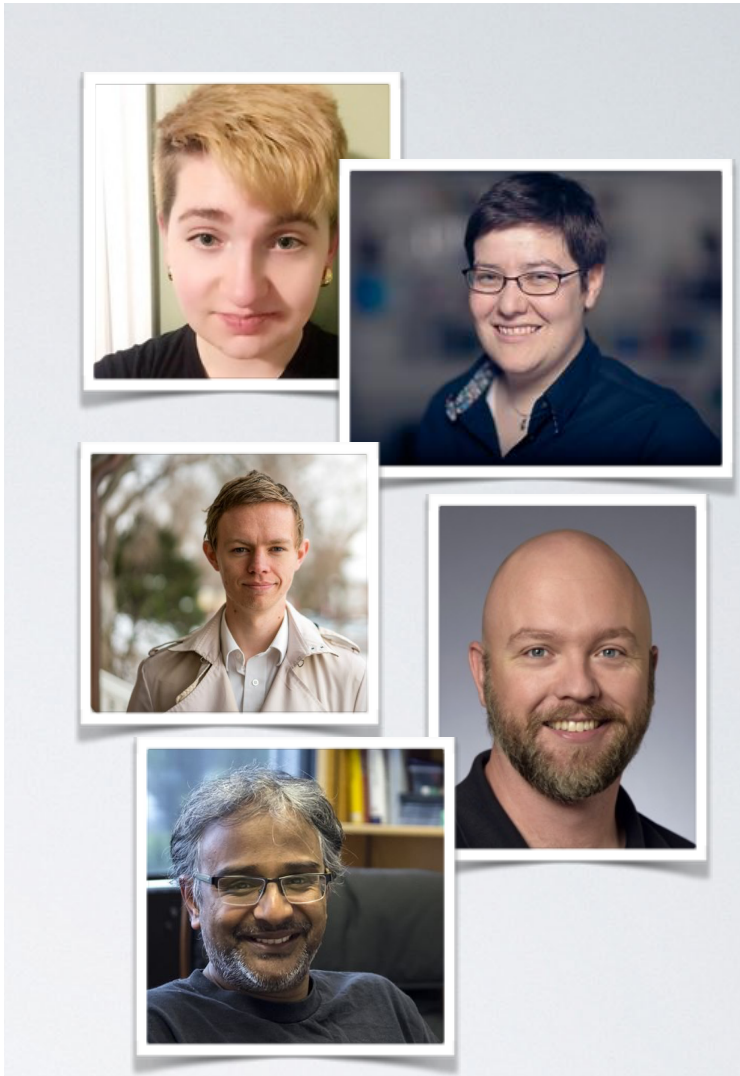
Machine learning models are increasingly being used to make real-world decisions, such as who to hire, who should receive a loan, where to send police, and who should receive parole. These deployed models mostly use traditional batch-mode machine learning, where decisions are made and observed results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

Questions?

References

- Buolamwini, J. & Gebru, T.. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR 81:77-91
- Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 4356–4364
- Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In Proceedings of the International Workshop on Software Fairness (FairWare '18). Association for Computing Machinery, New York, NY, USA, 1–7. DOI:https://doi.org/10.1145/3194770.3194776
- https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- https://fairmlclass.github.io/
- http://blog.mrtz.org/2016/09/06/approaching-fairness.html
- https://bair.berkeley.edu/blog/2018/05/17/delayed-impact/
- https://mrtz.org/nips17/
- https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de#.llzo69u3p
- https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb
- http://www.bu.edu/articles/2018/algorithmic-fairness/
- https://bair.berkeley.edu/blog/2018/05/17/delayed-impact/