Machine Learning and Discrimination

Diana Acosta-Navas
PhD candidate, Harvard Philosophy Department
Adjunct Lecturer in Ethics and Public Policy, Harvard Kennedy School

For Today…

- Discrimination/ wrongful discrimination
- Case Study: *PredPol*
- Disparate treatment vs. Disparate impact
- How predictive policing could wrongfully discriminate
- What contextual considerations are important to determine whether an algorithm wrongfully discriminates?

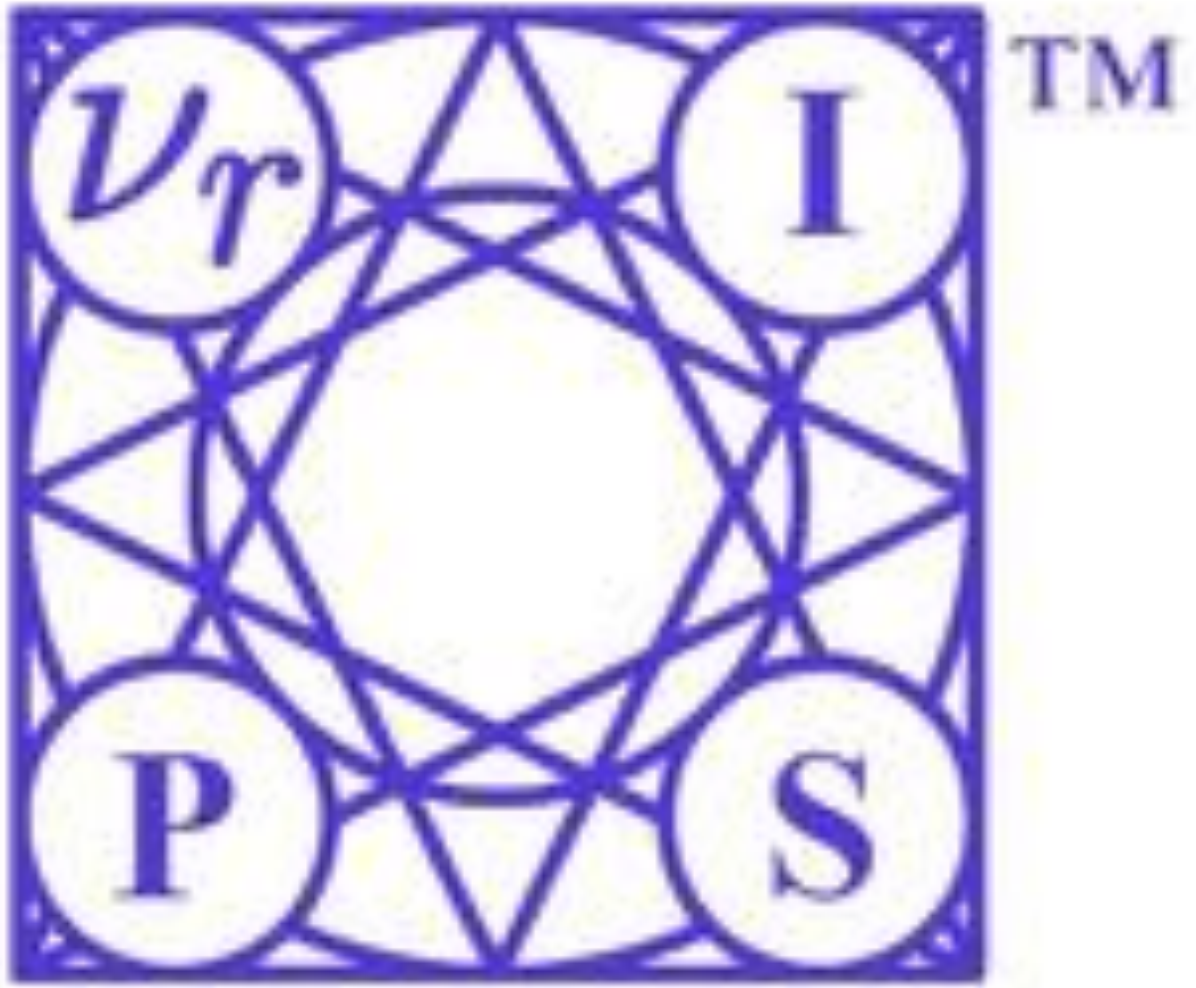
For Today…

- Content warning
- Diversity of perspectives is fundamental
- Respectful and open discussion
- Don’t expect answers, but productive questions
- Polling activities
- Electronic devices

NeurIPS 2020

Broader Impact Statement:

Authors are asked to include a section in their submissions discussing the broader impact of their work, including possible societal consequences —both positive and negative.



Data Mining

Resource allocation

- Benefits
- Opportunities
- Burdens

➢ Data Mining provides a basis on which to allocate these resources

➢ Finding patterns among different people or outcomes to find similarities and differences.

Social Impact

Does an algorithm discriminate?

What do we mean by “discrimination”?

- The action of perceiving, noting, or making a distinction between things.
- Something that enables a distinction to be made; a distinguishing mark, characteristic, or attribute; a difference.
- The power or faculty of observing differences accurately, or of making exact distinctions.
- The treatment of goods, trading partners, etc., on a more or less favourable basis according to circumstances
- Unjust or prejudicial treatment of a person or group, esp. on the grounds of race, gender, sexual orientation, etc.

Yes! That’s its job!

- The action of perceiving, noting, or making a distinction between things.
- Something that enables a distinction to be made; a distinguishing mark, characteristic, or attribute; a difference.
- The power or faculty of observing differences accurately, or of making exact distinctions; discernment.
- The treatment of goods, trading partners, etc., on a more or less favourable basis according to circumstances

Yes! That's its job!

- The action of perceiving, noting, or making a distinction between things.
- Something that enables a distinction to be made; a distinguishing mark, characteristic, or attribute; a difference.
- The power or faculty of observing differences accurately, or of making exact distinctions; discernment.
- The treatment of goods, trading partners, etc., on a more or less favourable basis according to circumstances

Ummm…

Unjust or prejudicial treatment of a person or group, esp. on the grounds of race, gender, sexual orientation, etc.

Wrongful Discrimination

A decision procedure wrongfully discriminates against social group *x* if and only if:

- There is a social group *y* such that the procedure treats the members of *x* less favorably than the members of *y*;
- Part of the explanation for the difference in treatment is their membership in *x* and *y*, respectively; and
- The difference in treatment is not morally justified on independent grounds.

Protected Categories

- Age
- Disability
- National origin
- Race
- Religion
- Sex/Gender
- Sexual Orientation

Case Study: *Predpol*

PredPol

- Location-based prediction of criminality
- Used to deploy law enforcement to the areas where it is most needed.
- PredPol evaluates yesterday's crimes in the context of all crimes occurring over a long time horizon and wide spatial field to calculate accurate probabilities of where and when crime will occur today

"If one can accurately predict where and when crimes will occur, then law enforcement personnel can disrupt those crimes before they happen."

PREDPOL®

PredPol

- Tested against existing practice, PredPol predicts between 1.6-2.5 more crime.
- Increased opportunities to impact crime are therefore of a similar magnitude.

  - Prevention
  - Response

PREDPOL®

PredPol

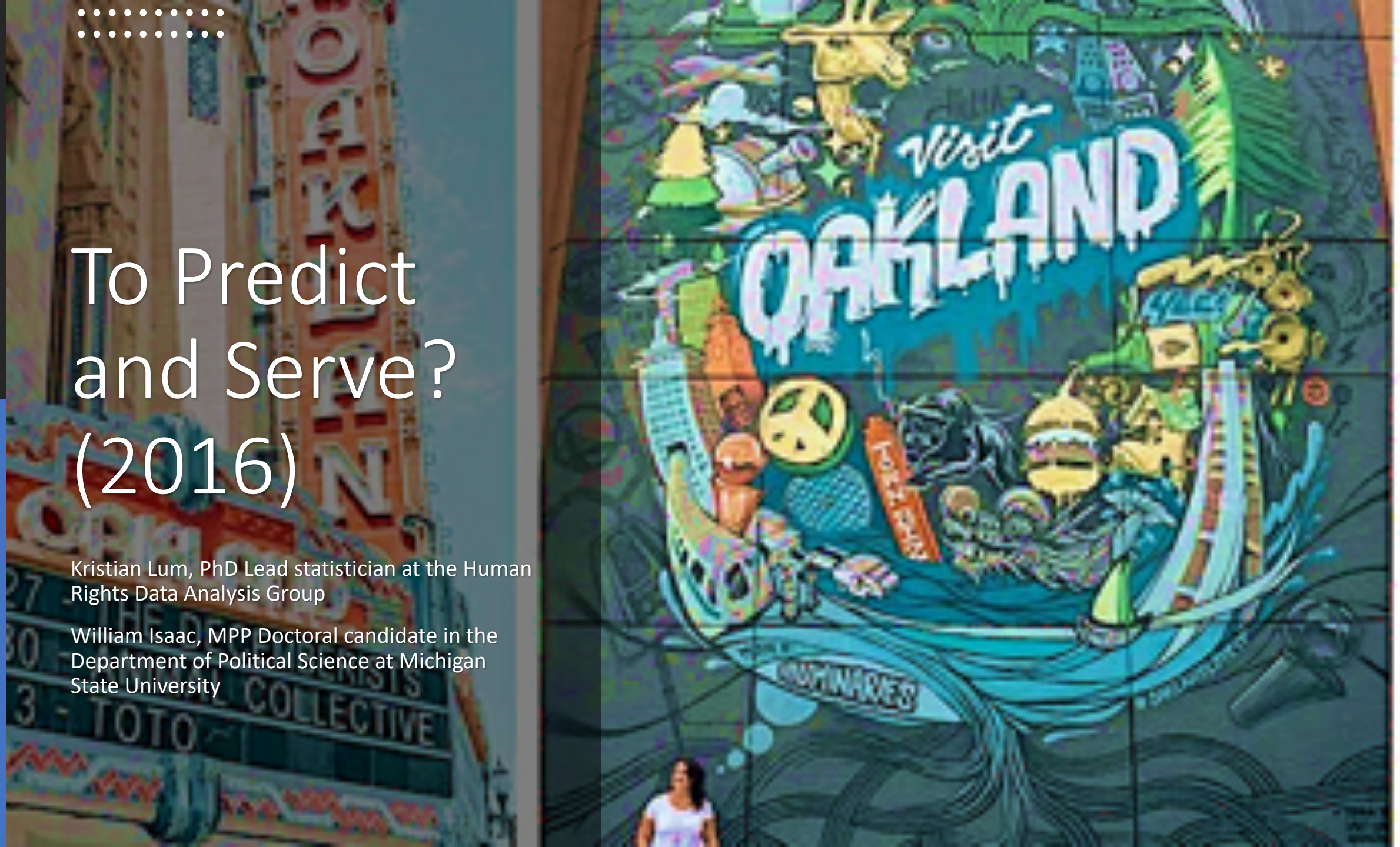
*"PredPol is **not criminal profiling**. It does not use any information about individuals or populations and their characteristics. The patterns inherent in the crimes themselves provide ample information to predict where and when crimes will occur in the future."*

PREDPOL®

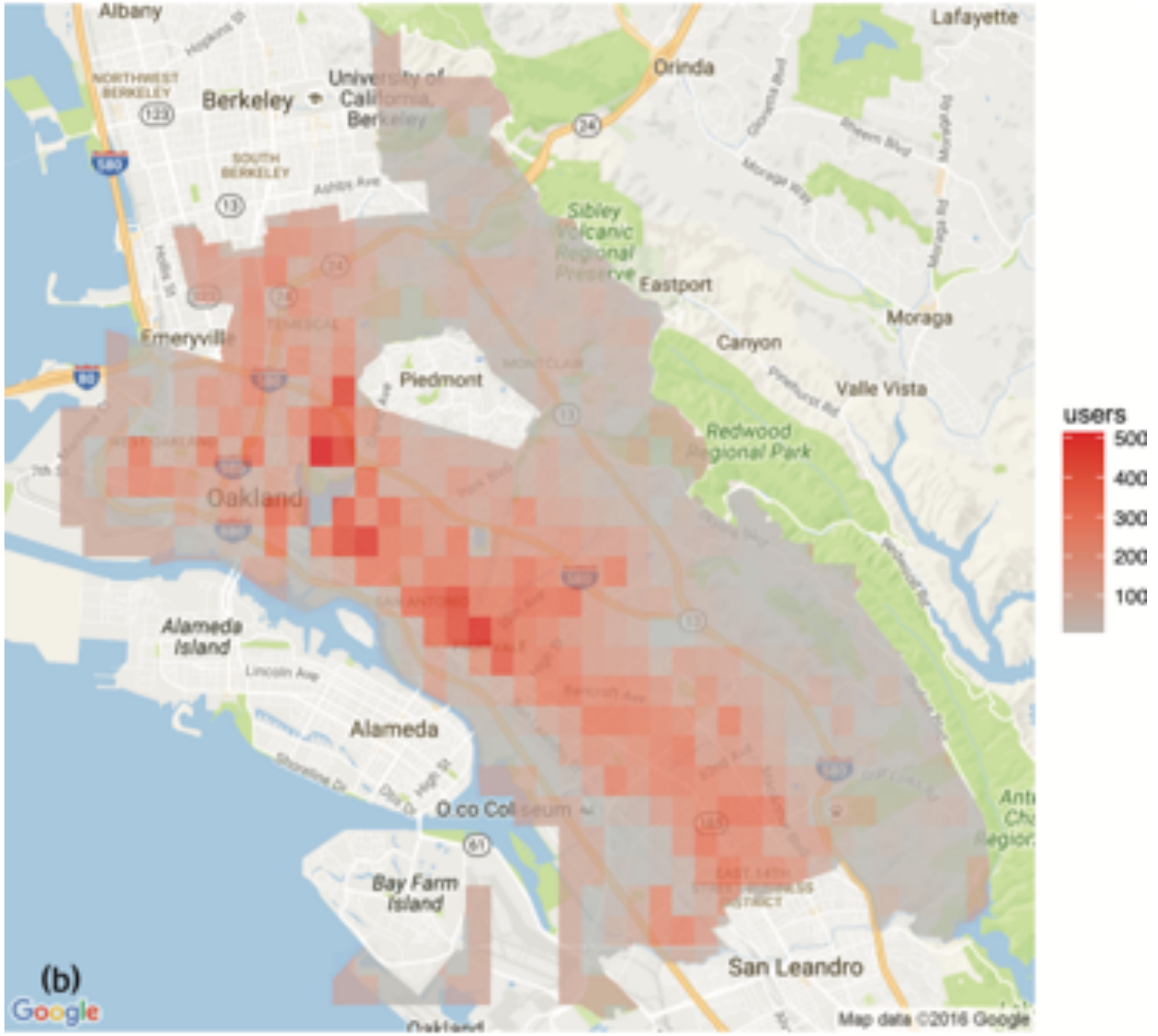
To Predict and Serve? (2016)

Kristian Lum, PhD Lead statistician at the Human Rights Data Analysis Group

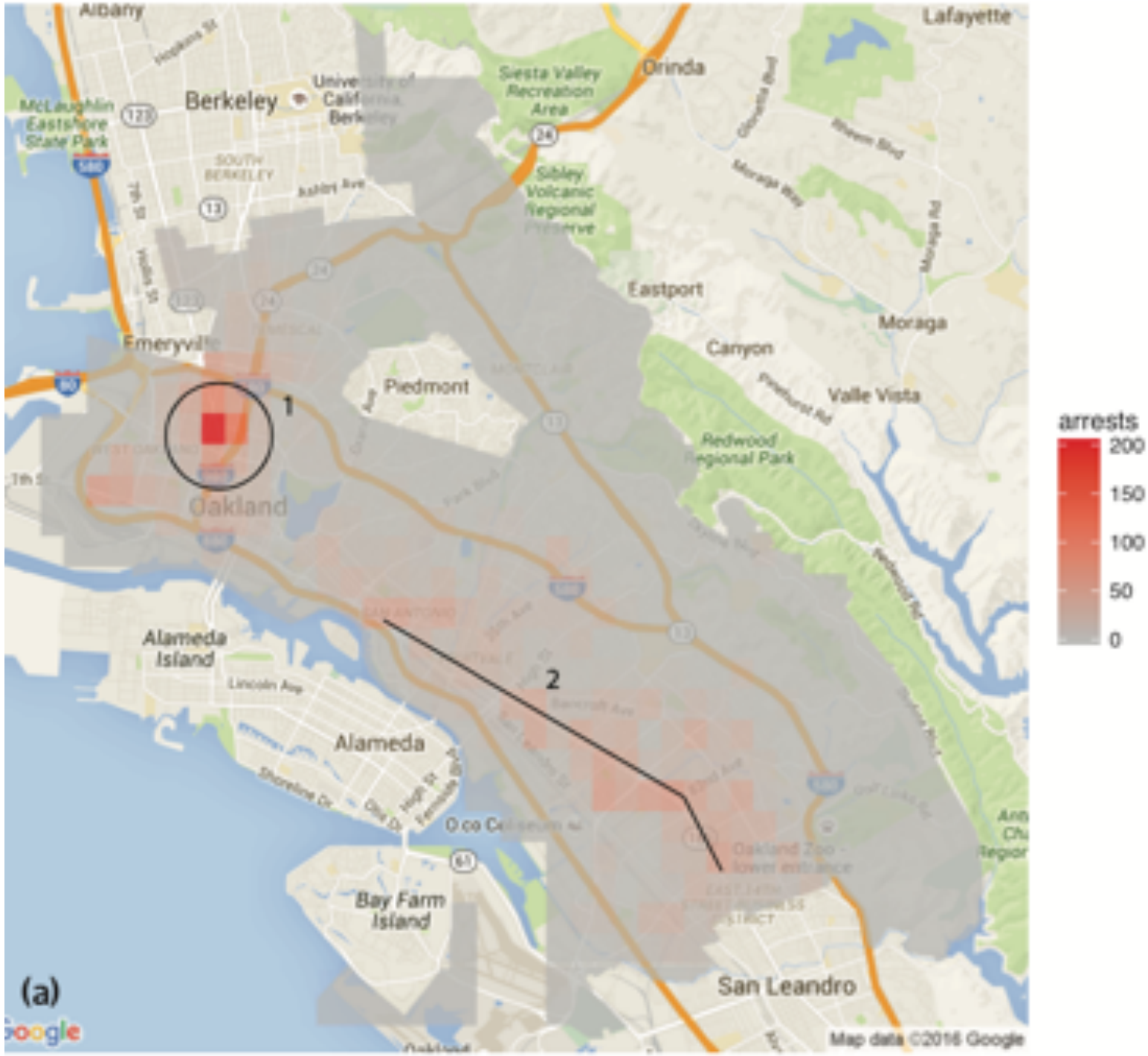
William Isaac, MPP Doctoral candidate in the Department of Political Science at Michigan State University



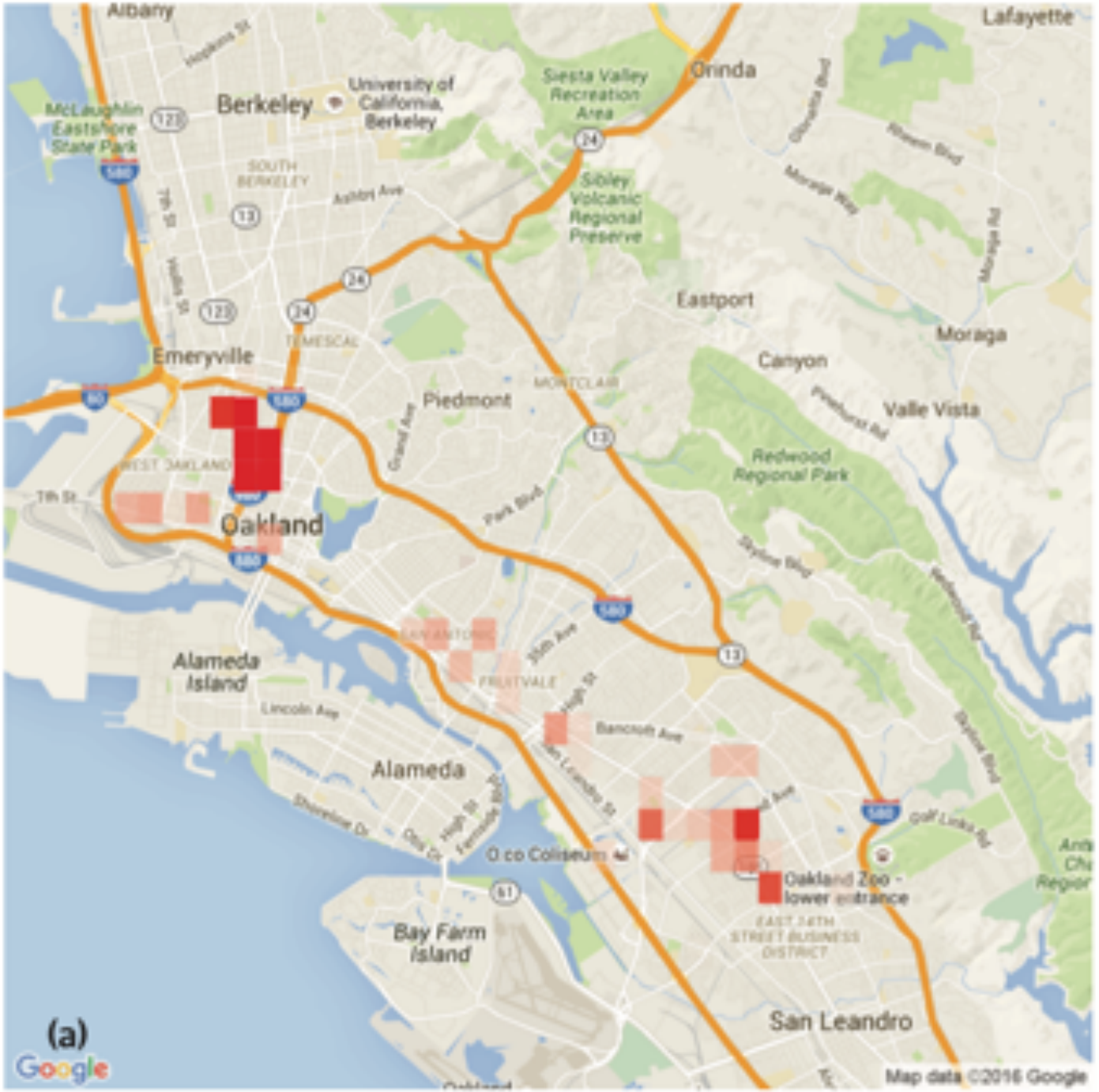
Estimated Drug Users

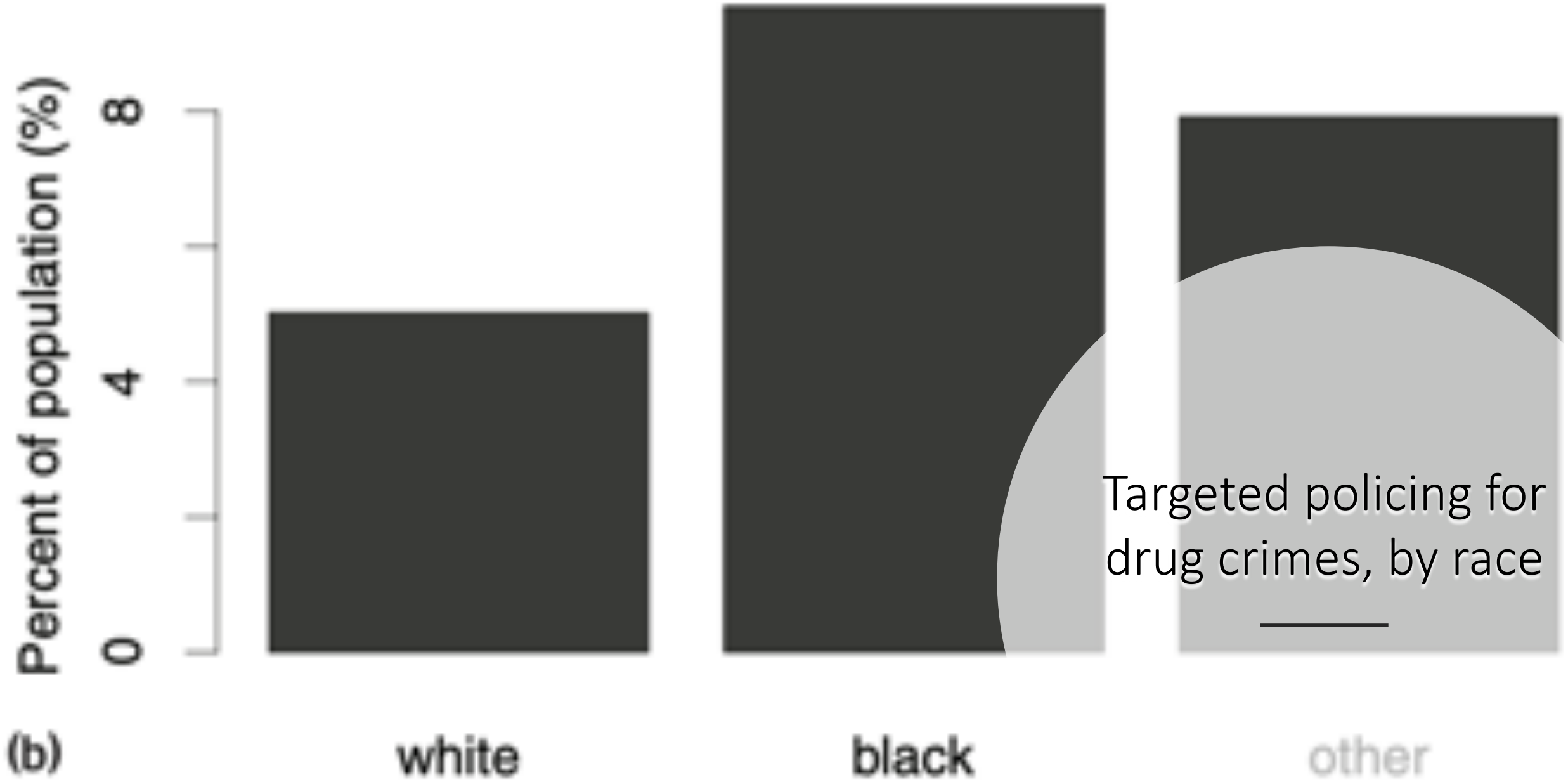


Drug Arrests in 2010

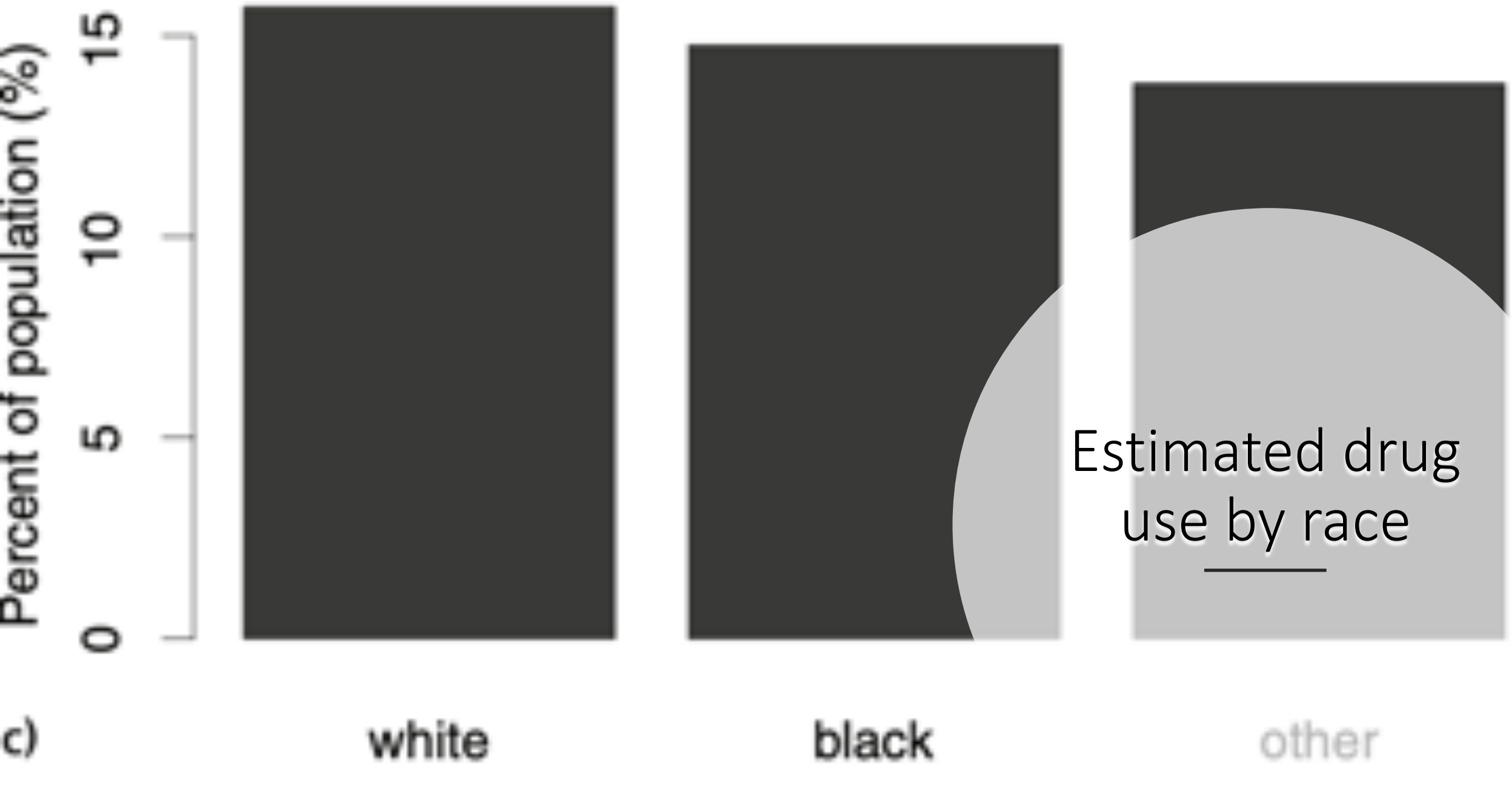


Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland police data

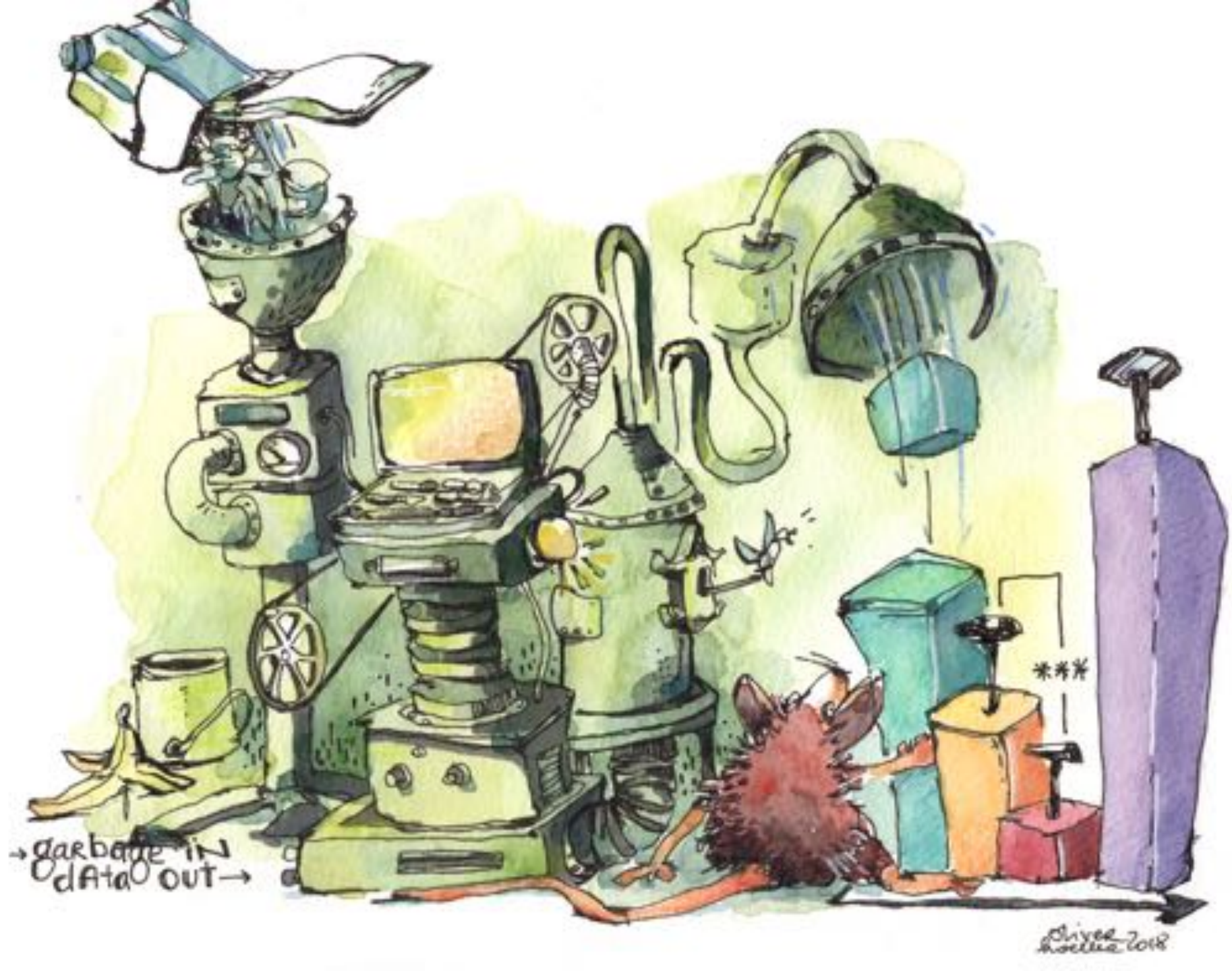




Targeted policing for drug crimes, by race

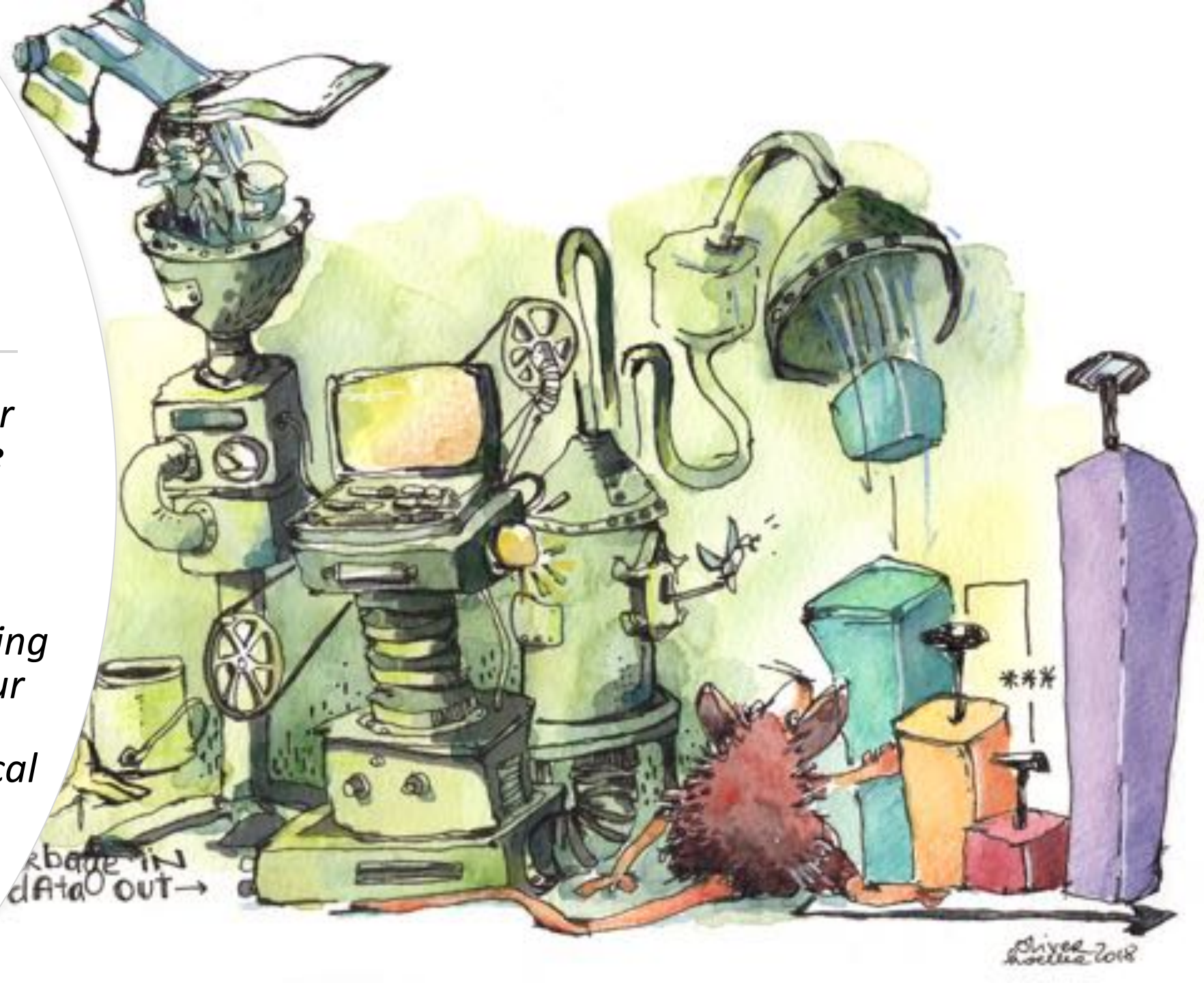

Percent of population (%)
0
5
10
15
white
black
other
Estimated drug use by race

So how does this happen?


→garbage IN
data OUT→
****
Oliver Knolla 2018

“rather than correcting for the apparent biases in the police data, the model reinforces these biases.

The locations that are flagged for targeted policing are those that were, by our estimates, already over-represented in the historical police data”

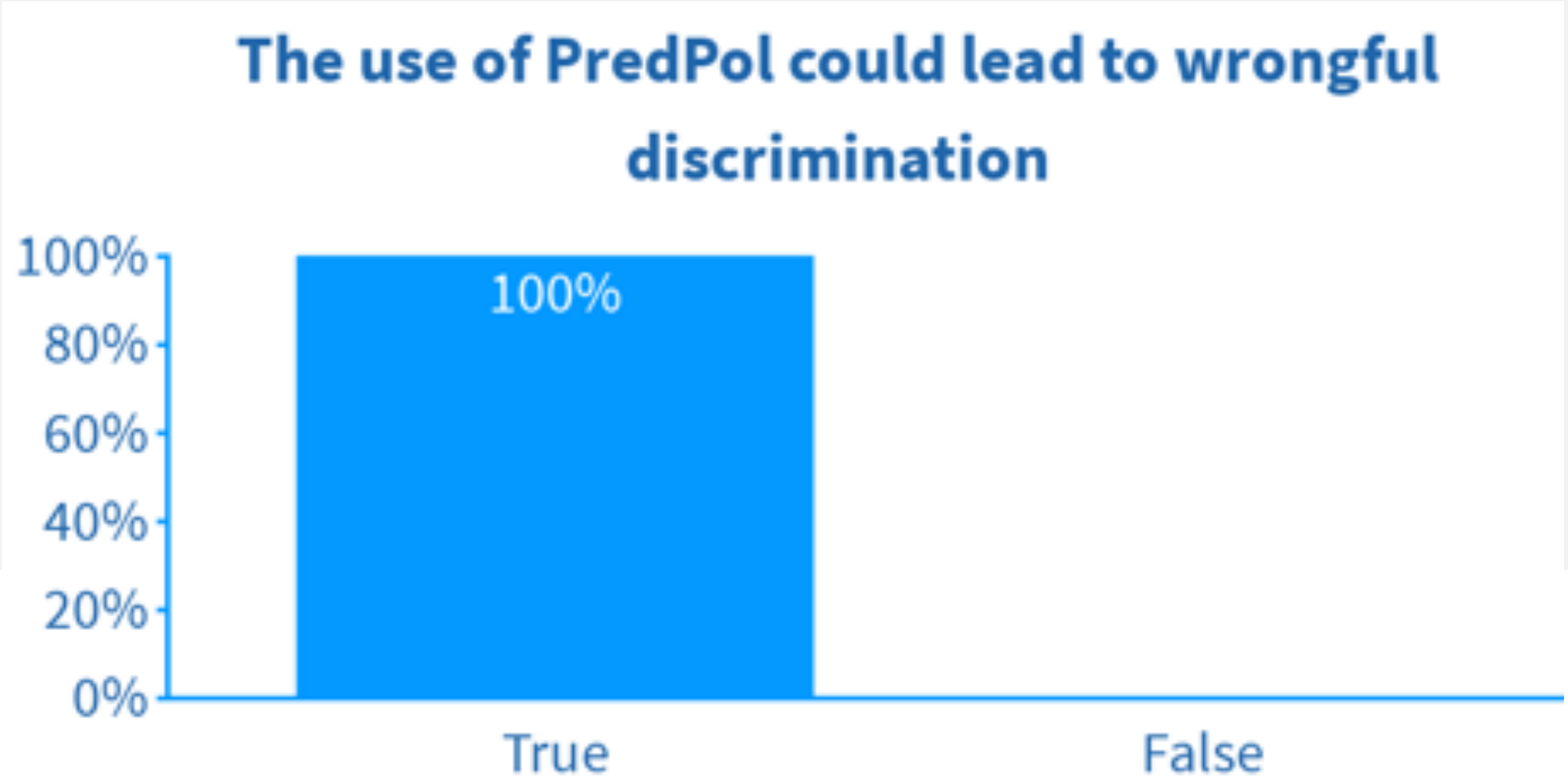


Wrongful Discrimination

A decision procedure wrongfully discriminates against social group *x* if and only if:

- There is a social group *y* such that the procedure treats the members of *x* less favorably than the members of *y*;
- Part of the explanation for the difference in treatment is their membership in *x* and *y*, respectively; and
- The difference in treatment is not morally justified on independent grounds.


Poll

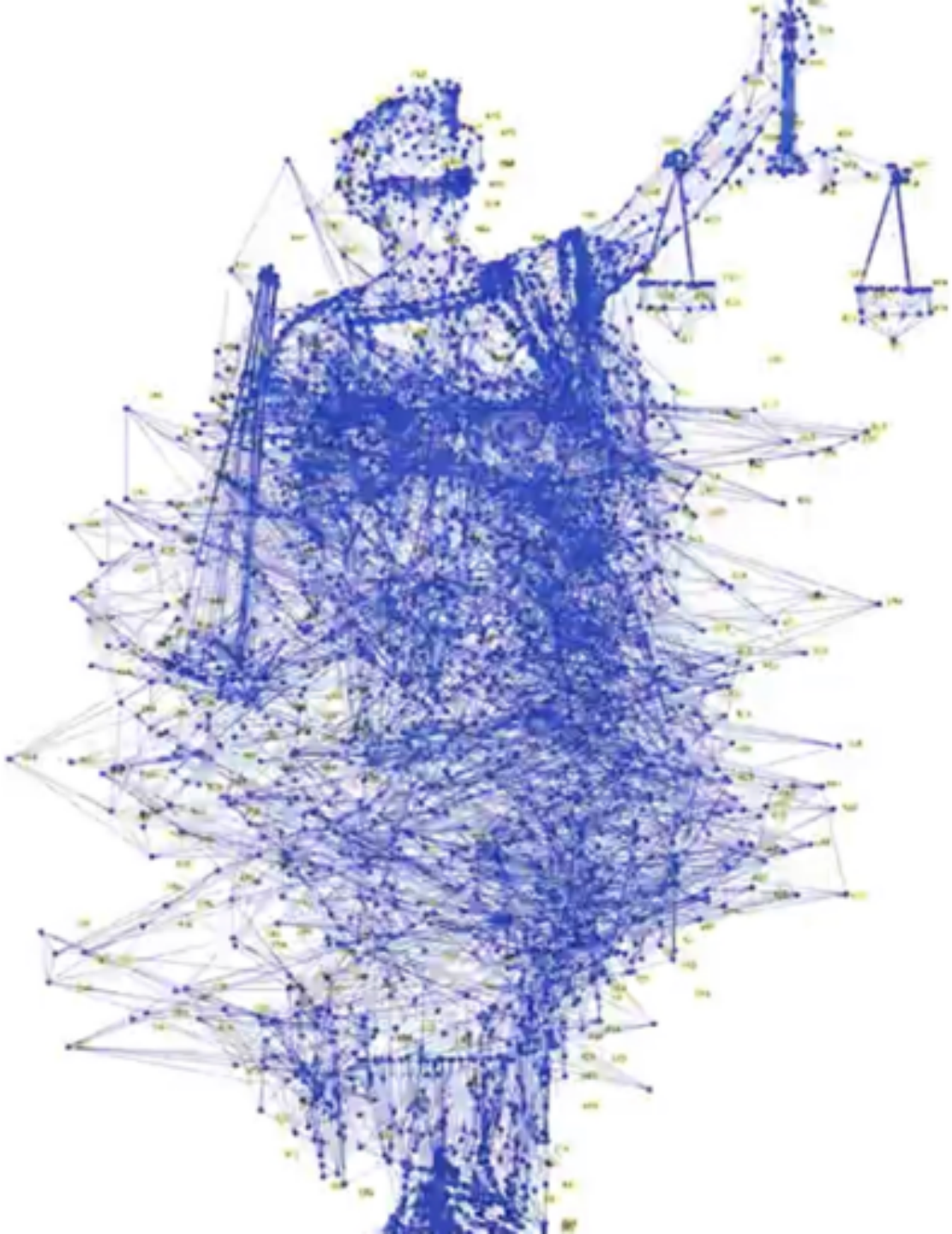

The use of PredPol could lead to wrongful discrimination
100%
80%
60%
40%
20%
0%
100%
True
False


Why?

Discrimination

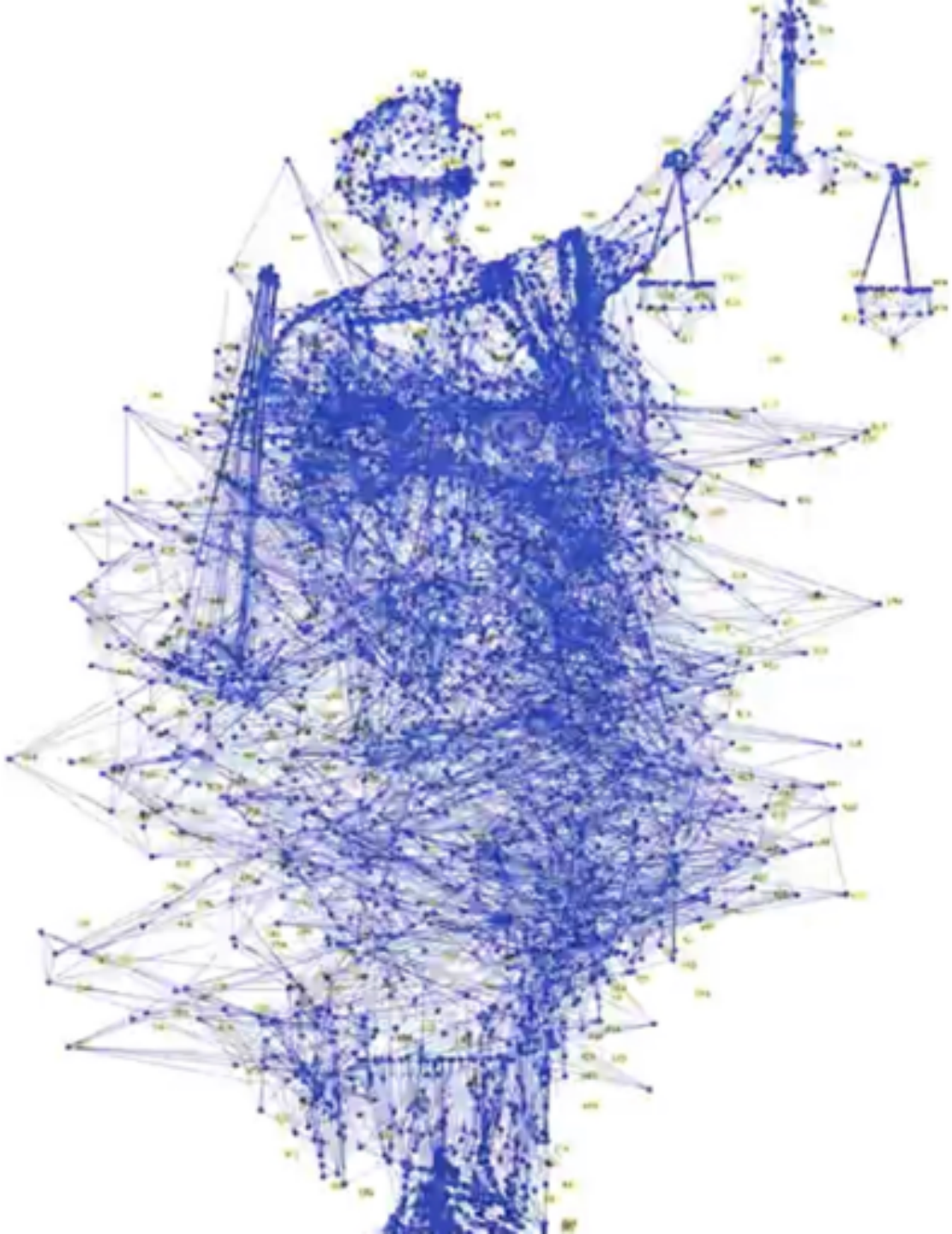
- *Direct discrimination:* discrimination resulting from a negative attitude toward the social group (e.g. animus or indifference)

- *Indirect discrimination:* discrimination that does not result from such an attitude



Disparate treatment

- Individuals are treated differently *because* of their group membership
- *Intent* to discriminate, either by explicitly referring to class membership or not.

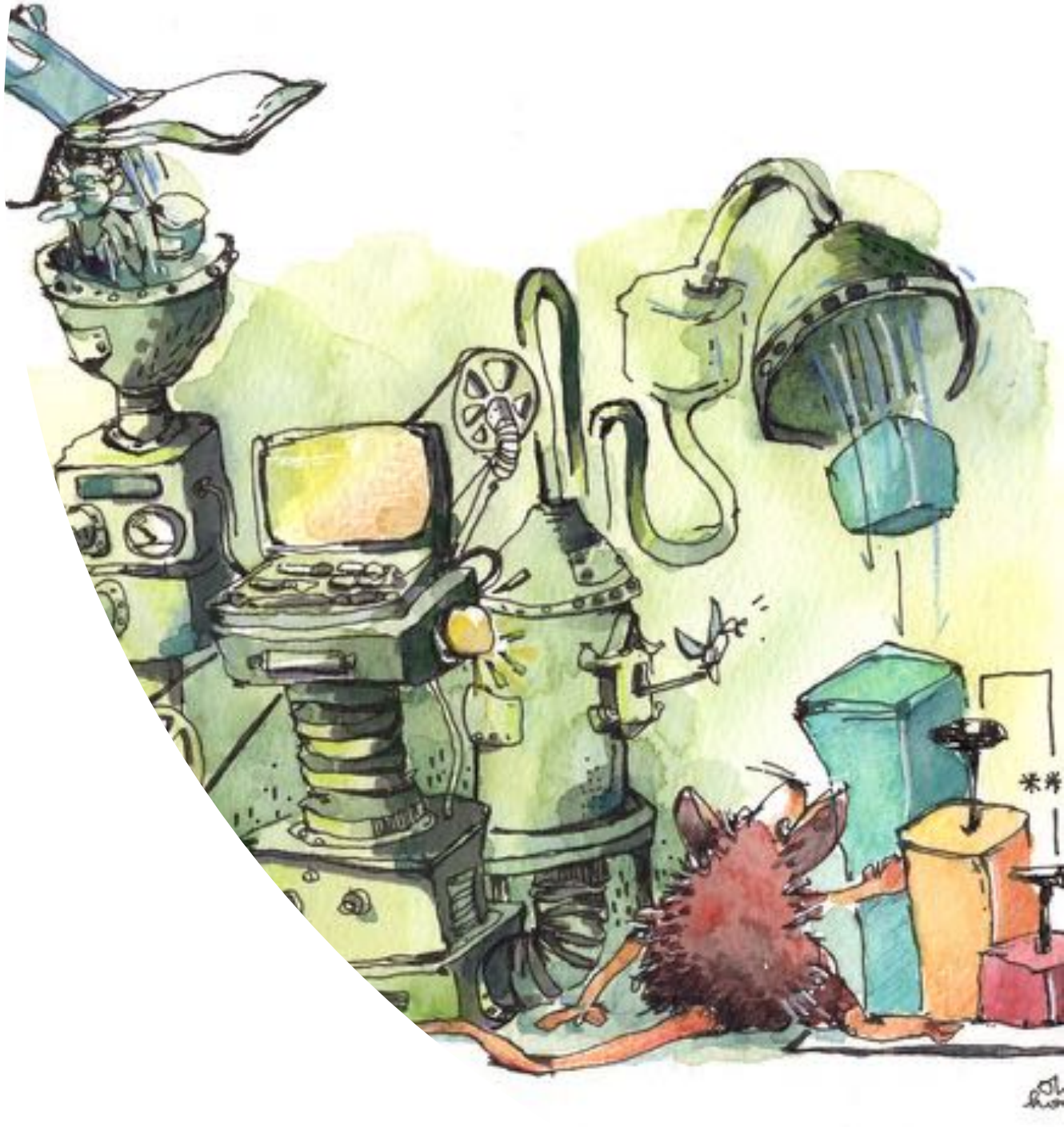


Disparate Impact

- Individuals are treated equally in accordance with a given set of rules and procedures
- Rules and procedures are constructed in a way that favors one group over another
- No intent required

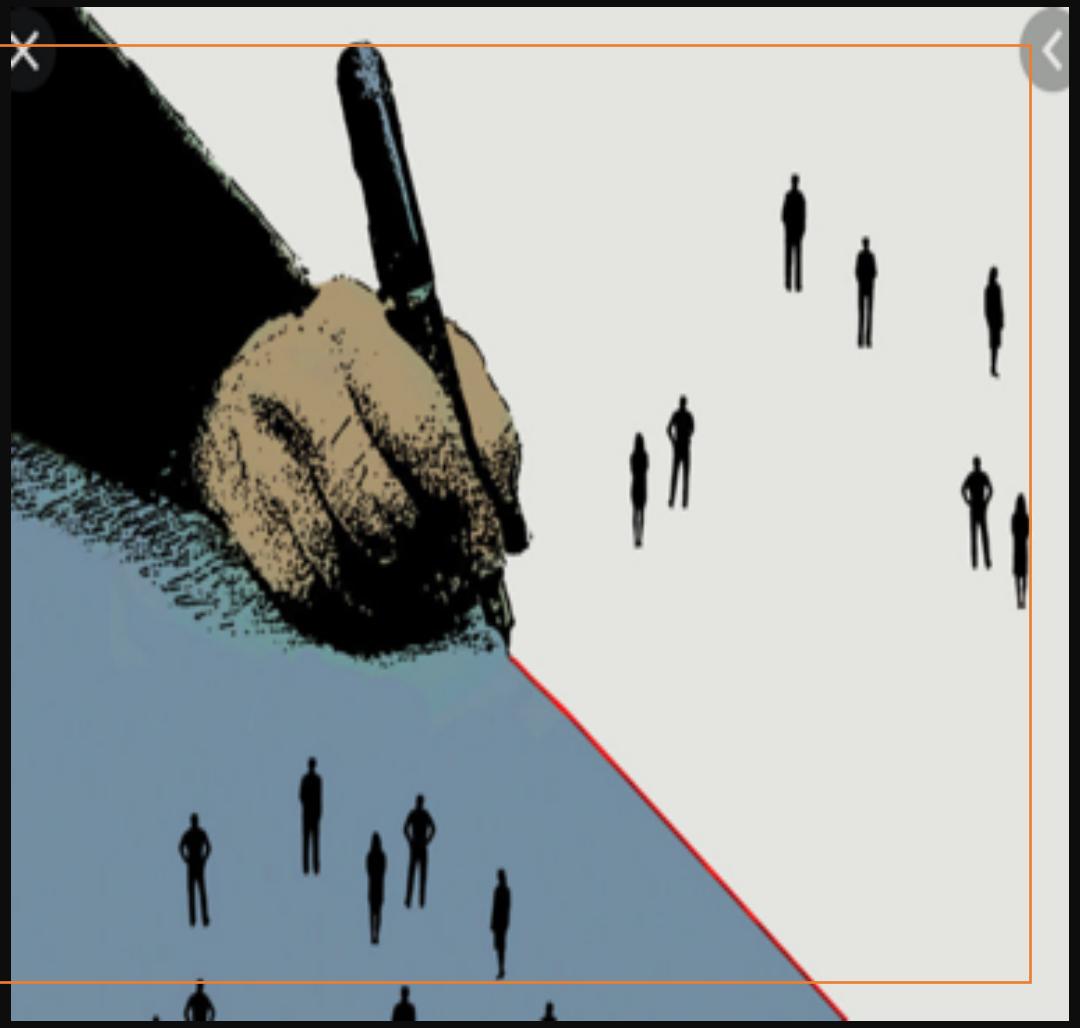
How data can discriminate

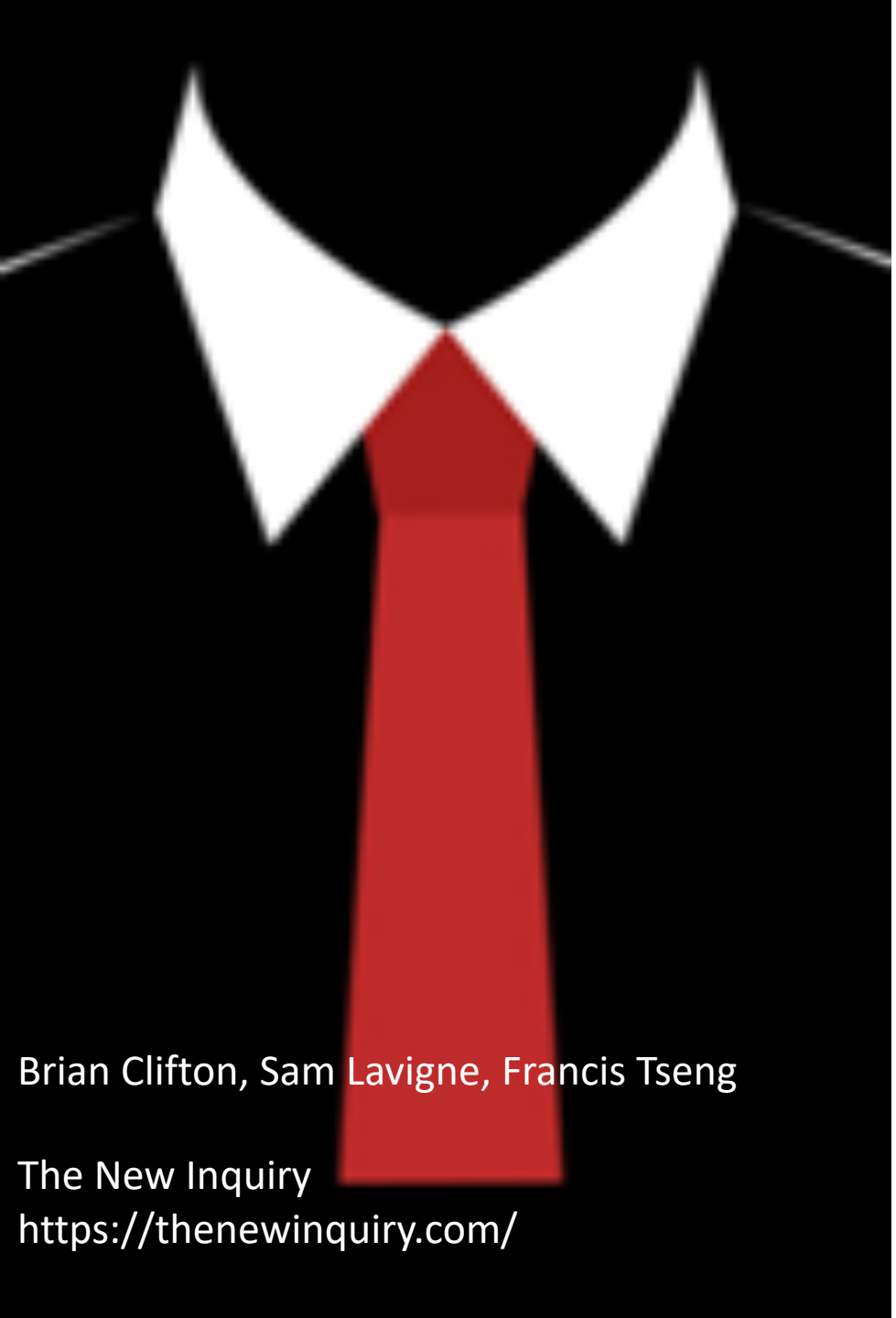
- Defining class labels and target variables
- Training data
  - Treating cases influenced by prejudice as valid examples
  - Drawing inferences from a biased sample
- Proxies: when relevant criteria are (accidentally) used as a proxy for class membership.



Why is Disparate Impact Problematic?

- Epistemic problem
- Feedback loop
- Burden distribution
- Rights may be threatened
- Interaction with other law enforcement practices



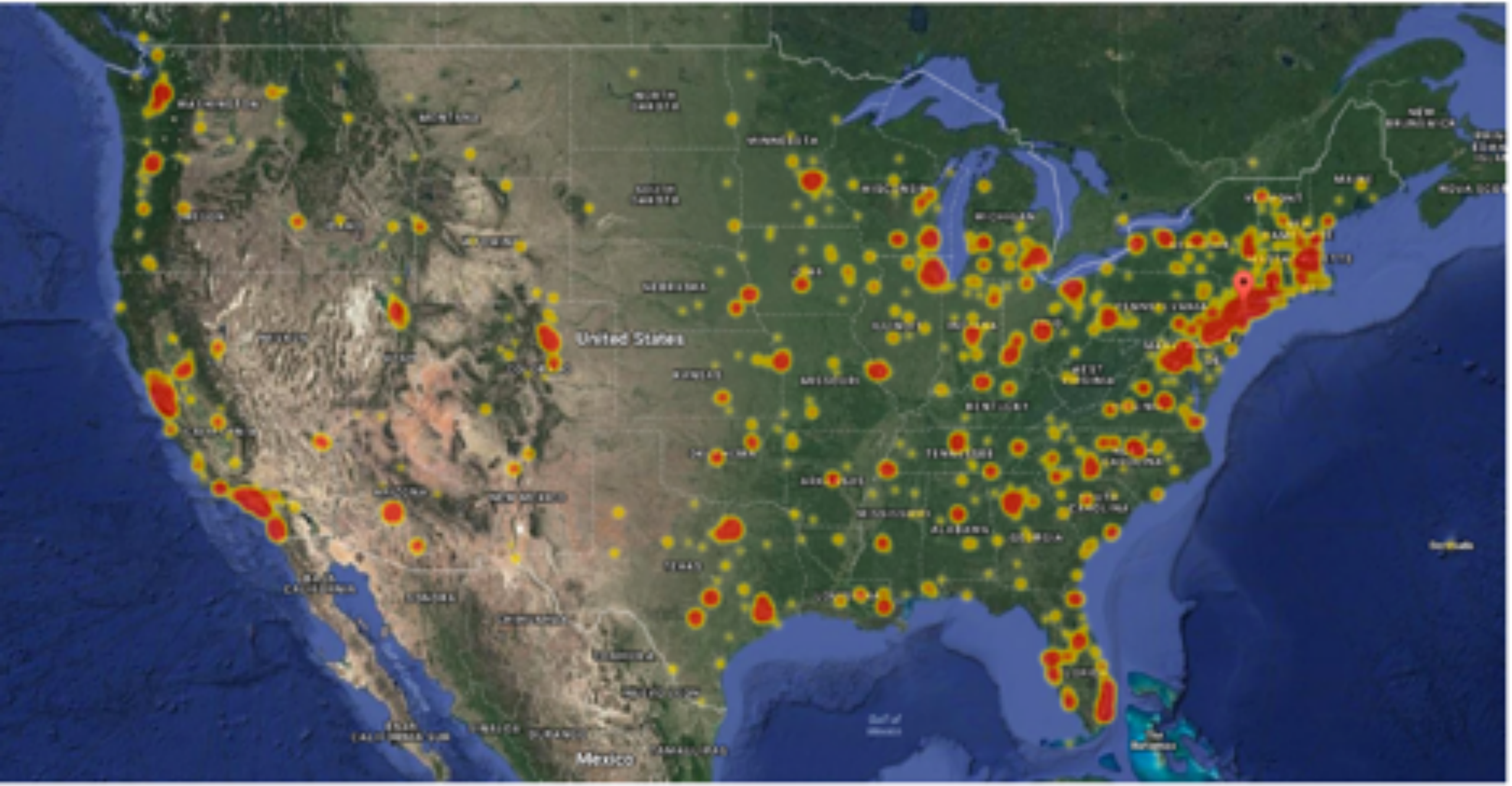


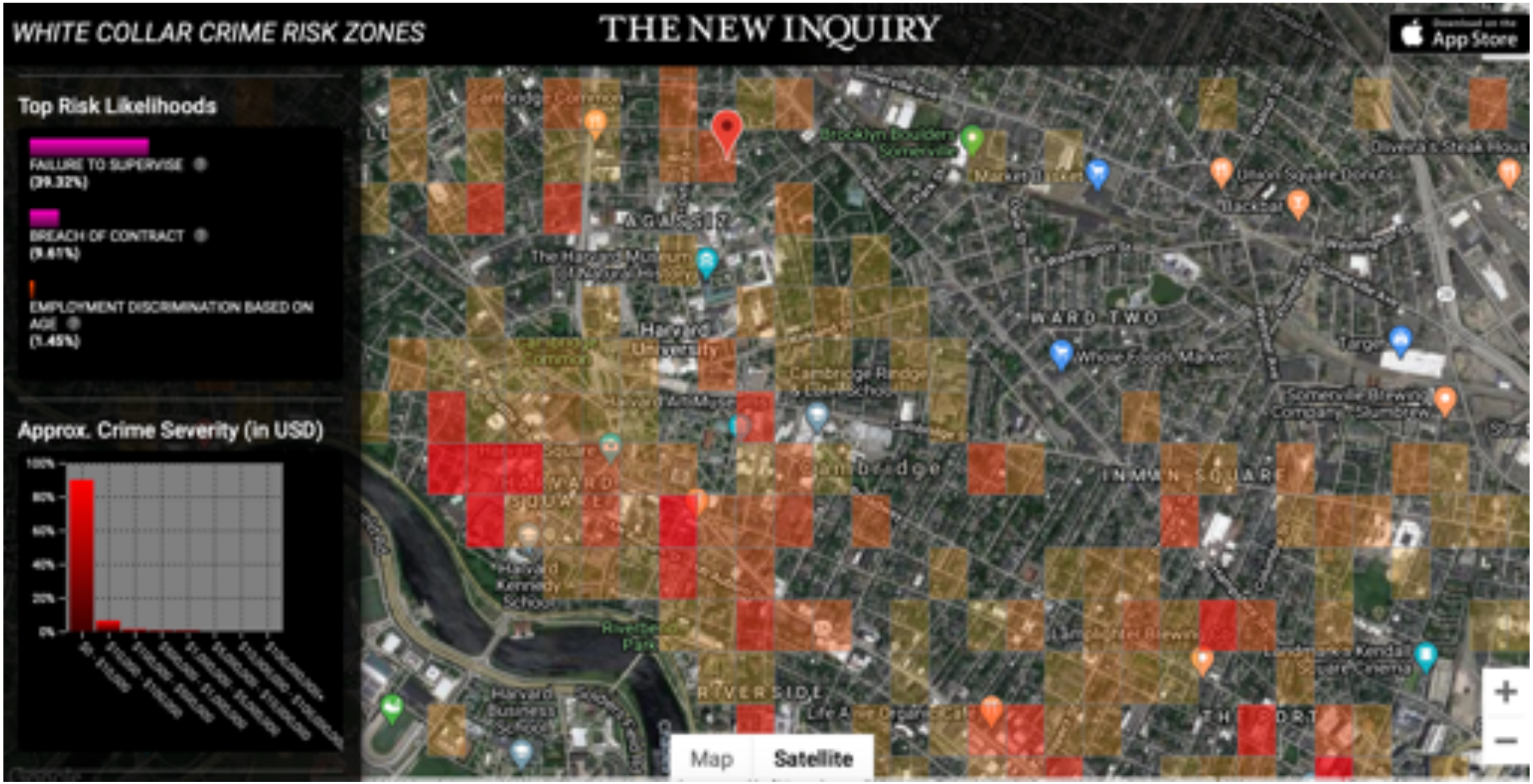
Brian Clifton, Sam Lavigne, Francis Tseng

The New Inquiry
https://thenewinquiry.com/

A Parody: *White Collar Crime Early Warning System*

- Predictive policing algorithm for identifying and assessing the risk of large-scale financial crime at the city block level.
- WCCEWS predicts the likelihood of a white-collar crime occurring within a 76m2 square, which is a 197.37% improvement of precision when compared with other predictive policing algorithms.
- Collected data provided by the Financial Regulatory Authority to compile incidents of financial malfeasance dating back to 1964. On this basis, correlated financial crimes to the location of the perpetrating individual or organization
- Optimized for growth in the policing of high-level financial criminals, with growth measured in higher arrest rates, improved quality of arrests, and higher recovery of funds.


United States
Mexico


WHITE COLLAR CRIME RISK ZONES
THE NEW INQUIRY
App Store
Top Risk Likelihoods
FAILURE TO SUPERVISE
(39.32%)
BREACH OF CONTRACT
(9.61%)
EMPLOYMENT DISCRIMINATION BASED ON AGE
(1.45%)
Approx. Crime Severity (in USD)
100%
80%
60%
40%
20%
0%
Harvard University
Cambridge
WARD TWO
INMAN SQUARE
Harvard Kennedy School
Harvard Business School
RIVERSIDE
Map
Satellite


WHITE COLLAR CRIME RISK ZONES
THE NEW INQUIRY
Top Risk Likelihoods
INACCURATE DATA
(9.87%)
FAILURE TO SUPERVISE
(6.62%)
INCORRECT MARK
(6.58%)
Approx. Crime Severity (in USD)
Cambridge Common
Brooklyn Boulders Somerville
Union Square Donuts
The Harvard Museum of Natural History
Harvard University
WARD TWO
Whole Foods Market
Target
Cambridge Common
Harvard Art Museums
Cambridge Public Library
Somerville Brewing Company (Slumbrew)
Cambridge
INMAN SQUARE
HARVARD SQUARE
Harvard Kennedy School
Riverbend Park
Harvard Business School
RIVERSIDE
Map
Satellite


Poll

The use of WCCEWS could lead to wrongful discrimination

What does WCCEWS reveal about PredPol (and other such systems?)

Some takeaways

- Wrongful discrimination occurs when the members of a group are disadvantaged for being members of that group, absent other moral justifications.
- Disparate impact can constitute wrongful discrimination
  - Epistemic problems
  - Burden distribution
  - Threatening rights
- The problem is how the algorithm operates in a given context
  - Impacted group and how its members are treated
  - How it interacts with other policing practices
  - Whether it can reinforce patterns of discrimination and inequality

Module Evaluation Survey: http://bit.ly/s20cs181

diana_acosta_navas@hks.harvard.edu