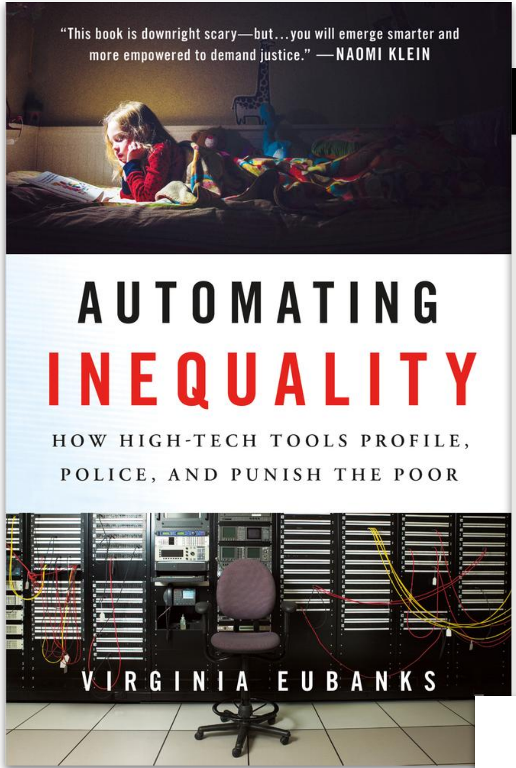
The University of Texas at Austin
McCombs School of Business

Sources and Consequences of Algorithmic Bias

Maria De-Arteaga

Assistant Professor
Information, Risk and Operations Management Department
University of Texas at Austin

The University of Texas at Austin
McCombs School of Business



Health

Racial bias in a medical algorithm favors white patients over sicker black patients

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / A YEAR AGO

REUTERS

Amazon scraps secret AI recruiting tool that showed bias against women

Outline

1. Taxonomy of sources of bias.

2. Bias in victim crime reporting and its effect on predictive policing (*FAccT'21*).

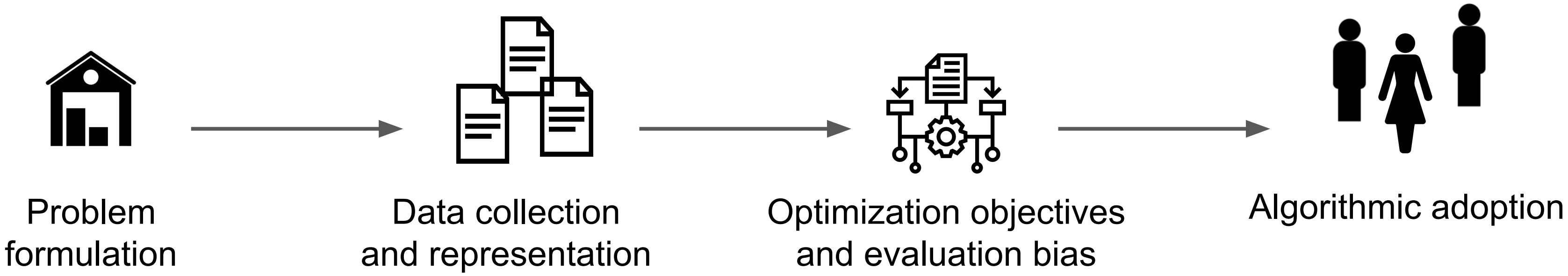
The effect of differential victim crime reporting on predictive policing systems

Nil-Jana Akpinar
nakpinar@stat.cmu.edu
Department of Statistics and Data Science & Machine Learning Department
Carnegie Mellon University

Maria De-Arteaga
Information, Risk, and Operations Management Department
McCombs School of Business
University of Texas at Austin

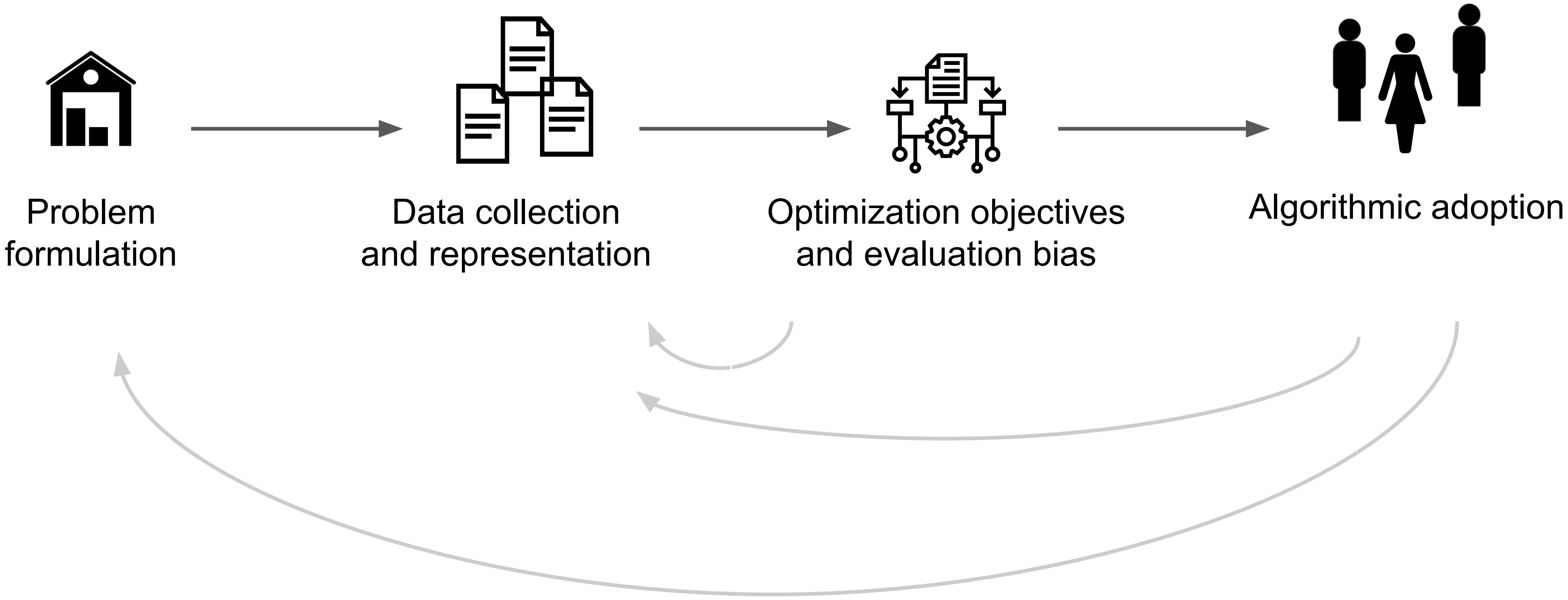
Alexandra Chouldechova
Heinz College & Department of Statistics and Data Science
Carnegie Mellon University

Sources of bias

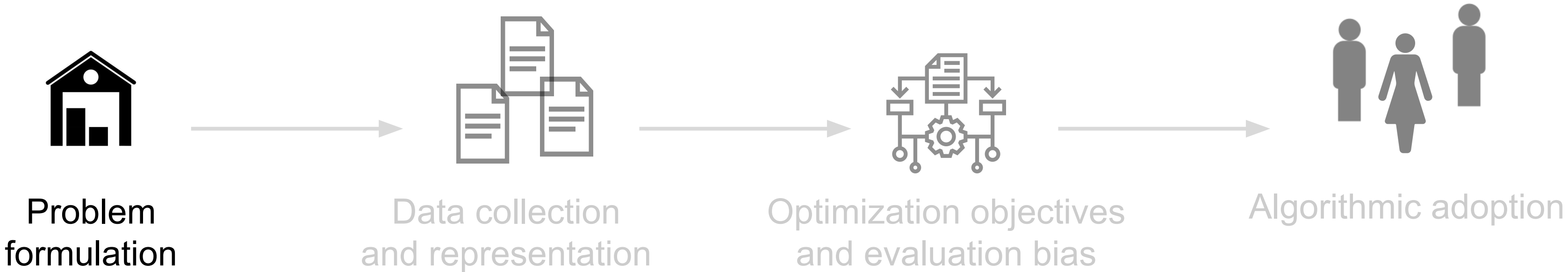


Working paper, joint with: Stefan Feuerriegel & Maytal Saar-Tsechansky. *Algorithmic Fairness in Business Analytics: Directions for Research and Practice*.

Sources of bias



Sources of bias



Bias ingrained in underlying assumptions

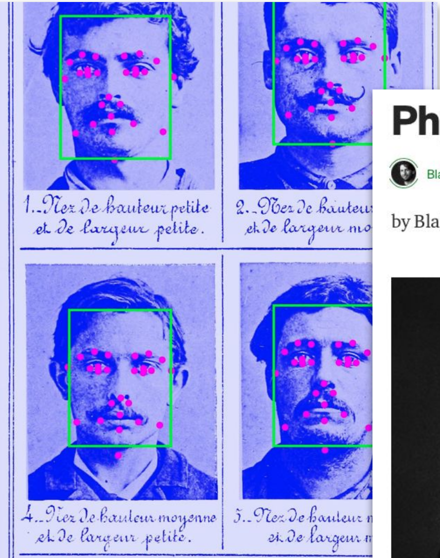
The New York Times

Account

Opinion

The Racist History Behind Facial Recognition

When will we finally learn we cannot predict people's character from their appearance?



Physiognomy's New Clothes

Blaise Aguera y Arcas May 6, 2017 · 38 min read

by Blaise Agüera y Arcas, Margaret Mitchell and Alexander Todorov

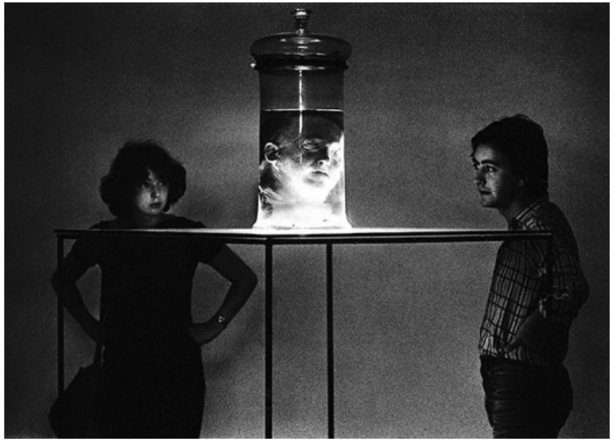
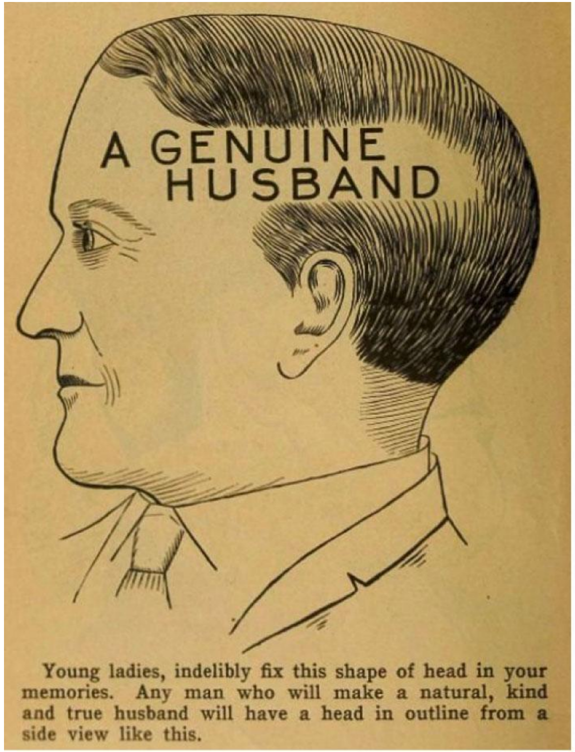
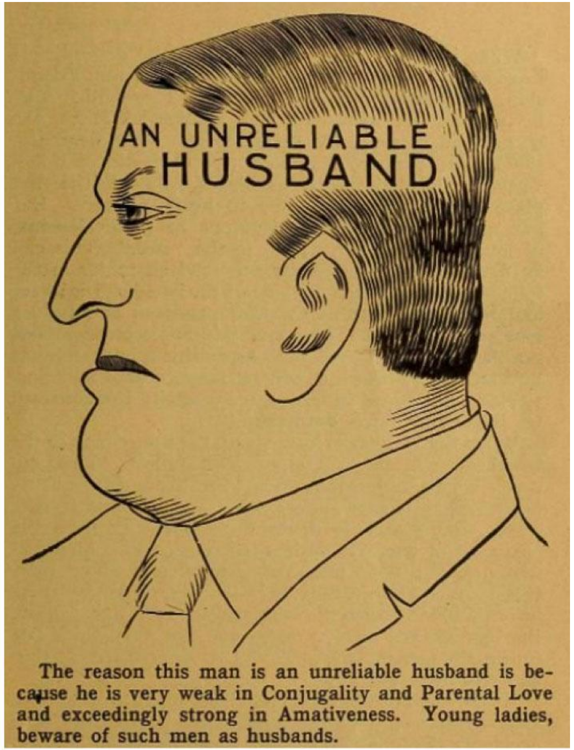


Figure 1. A couple viewing the head of Italian criminologist Cesare Lombroso preserved in a jar of formalin at an exhibition in Bologna, 1978. (Photo by Romano Cagnoni/Hulton Archive/Getty Images)





Pages from "Vaught's Practical Character Reader," a phrenology book published in 1902.Source: Archive.org

FACEPTION
Facial Personality Analytics

Our Technology | Verticals | About us | News & Events | Blog | Contact us

FACEPTION IS A FACIAL PERSONALITY ANALYTICS TECHNOLOGY COMPANY

We reveal personality from facial images at scale to revolutionize how companies, organizations and even robots understand people and dramatically improve public safety, communications, decision-making, and experiences.

FACEPTION
Facial Personality Analytics

Our Technology | Verticals | About us | News & Events | Blog | Contact us

OUR CLASSIFIERS

Utilizing advanced machine learning techniques we developed and continue to evolve an array of classifiers. These classifiers represent a certain persona, with a unique personality type, a collection of personality traits or behaviors. Our algorithms can score an individual according to their fit to these classifiers.

HIGH IQ

Endowed with a reasoning skills, like logic, spatial skills. Self-made people, free-thinkers and entrepreneurs. Exceptionally gifted, tend to be less socially oriented, value truth, facts and logic more than emotional relations. Creative and independent minded, with exceptional concentration abilities, a high intellect and mental capacity

Academic Researcher

Endowed with sequential thinking, high analytical abilities, a multiplicity of ideas, deep thoughts and seriousness. Creative, with a high concentration ability, high mental capacity, and interest in data and information.

Professional Poker Player

Endowed with a high concentration ability, perseverance and patience. Goal-oriented, analytical, with a dry sense of humor. Silent, devoid of emotion and emotional expression, strict and sharp minded, with high critical perception.

Bingo Player

Endowed with a high mental ceiling, high concentration, adventurousness, and strong analytical abilities. Tends to be creative, with a high originality and imagination, high conservation and sharp senses.

Brand Promoter

Endowed with a high self-confidence, authoritative, charismatic and magnetic personality, with high intellect and high verbal ability. Tends to be kind, sociable and direct, and very practical.

White-Collar Offender

Tends to have a low self-esteem, a high IQ and charisma. Anxious, tensed and frustrated, competitive, ambitious and dominant. Usually loves to take risks and have a dry sense of humor.

Terrorist

Suffers from a high level of anxiety and depression. Introverted, lacks emotion, calculated, tends to pessimism, with low self-esteem, low self image and mood swings.

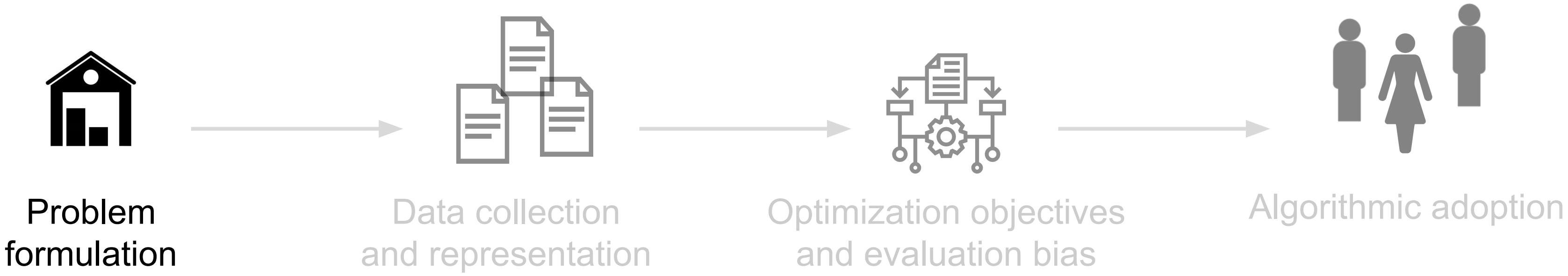
Pedophile

Suffers from a high level of anxiety and depression. Introverted, lacks emotion, calculated, tends to pessimism, with low self-esteem, low self image and mood swings.

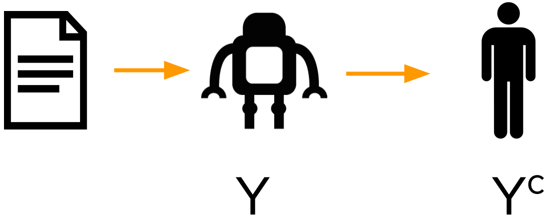
Bias ingrained in underlying assumptions

- AI ethics is broader than AI fairness.
- Some tasks are *inherently* biased, grounded on discriminatory assumptions.
- Misleading to talk about “fairness” in these context.
- Accuracy and good AI performance can also be harmful!

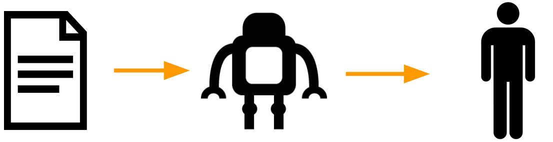
Sources of bias



Proxies and mismeasurement errors



Proxies and mismeasurement errors



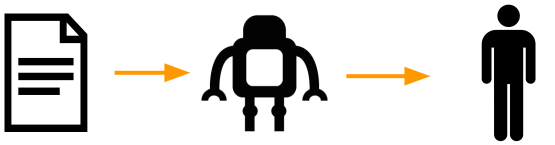
Y^c

Complex, multi-faceted,
(and sometimes contested) outcome

Examples:

- Patients’ healthcare needs
- Prospective employees potential

Proxies and mismeasurement errors



Y $\qquad\qquad\qquad$ Y^c

Quantifiable, imperfect proxy

Examples:

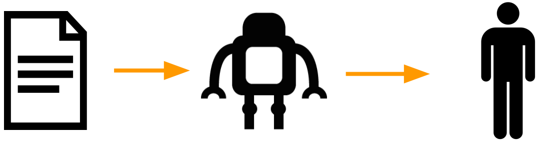
- Healthcare spending
- Job promotions

Complex, multi-faceted, (and sometimes contested) outcome

Examples:

- Patients' healthcare needs
- Prospective employees potential

Proxies and mismeasurement errors



Y Y^c

Quantifiable, imperfect proxy

Examples:

- Healthcare spending
- Job promotions

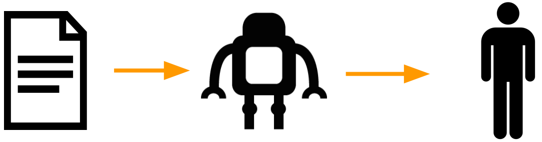
Complex, multi-faceted, (and sometimes contested) outcome

Examples:

- Patients' healthcare needs
- Prospective employees potential

Mind the gap!

Proxies and mismeasurement errors



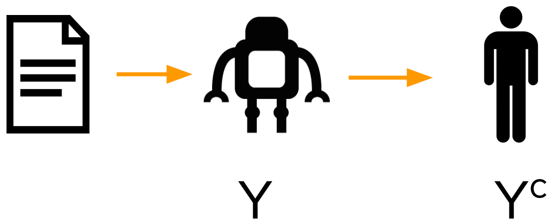
Y Y^c

Mind the gap!

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

*"Bias occurs because the algorithm **uses health costs as a proxy for health needs. Less money is spent on Black patients** who have the same level of need, and the algorithm thus f**alsely concludes that Black patients are healthier** than equally sick White patients."*



Mind the gap!

Misleading comparisons

Erroneously assuming the human is engaged in the same predictive task

ML "better than" the human expert with respect to artificial task

Self-fulfilling prophecies

Evaluating with respect to proxy

Incorrect predictions erroneously assessed as correct when assuming proxy is "ground truth"

Problem formulation

- Decisions involved:
  - What is the overarching goal of the system?
  - What is the mechanism of entry into the population subjected to the algorithm?
  - What is the space of possible decisions?

Prediction-Based Decisions and Fairness:
A Catalogue of Choices, Assumptions, and Definitions

Shira Mitchell
Civis Analytics
sam942@mail.harvard.edu

Eric Potash
University of Chicago
epotash@uchicago.edu

Solon Barocas
Microsoft Research and Cornell University
sbarocas@cornell.edu

Alexander D'Amour
Google Research
alexdamour@google.com

Kristian Lum
University of Pennsylvania
kl1@seas.upenn.edu

Problem formulation: how does it matter?

Algorithms that solve seemingly the same task can be embedded within entirely different problem formulations, which directly impacts fairness considerations: what burdens and benefits are allocated?

Problem formulation: how does it matter?

Algorithms that solve seemingly the same task can be embedded within entirely different problem formulations, which directly impacts fairness considerations: what burdens and benefits are allocated?

Predicting Student Loan Default for the University of Texas at Austin

Elizabeth Herr

Larry Burt

Predicting and Deterring Default with Social Media Information in Peer-to-Peer Lending

Ruyi Ge, Juan Feng, Bin Gu & Pengzhu Zhang

Problem formulation: how does it matter?

Algorithms that solve seemingly the same task can be embedded within entirely different problem formulations, which directly impacts fairness considerations: what burdens and benefits are allocated?

Predicting Student Loan Default for the University of Texas at Austin

Elizabeth Herr

Larry Burt

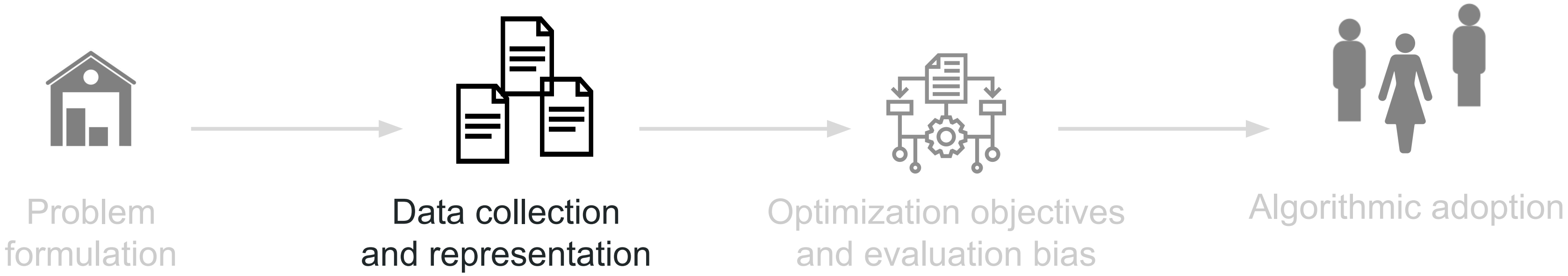
→ *"Interventions that focus on student persistence and academic success were seen as the primary actions needed to help prevent student loan default".*

Predicting and Deterring Default with Social Media Information in Peer-to-Peer Lending

Ruyi Ge, Juan Feng, Bin Gu & Pengzhu Zhang

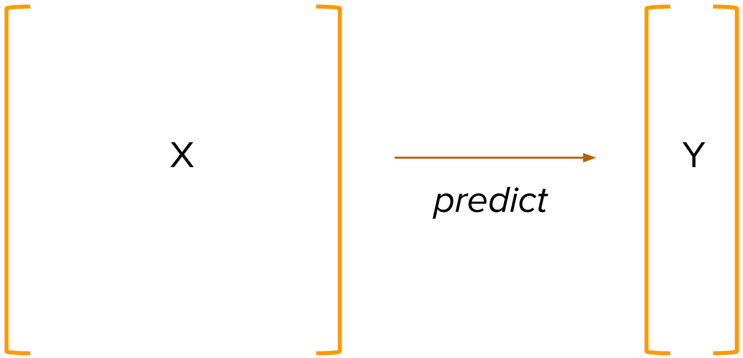
← *"deterring borrowers with social stigma and shaming on online social media could be a low-cost enforcement option for Chinese P2P lending platforms".*

Sources of bias



Sampling bias

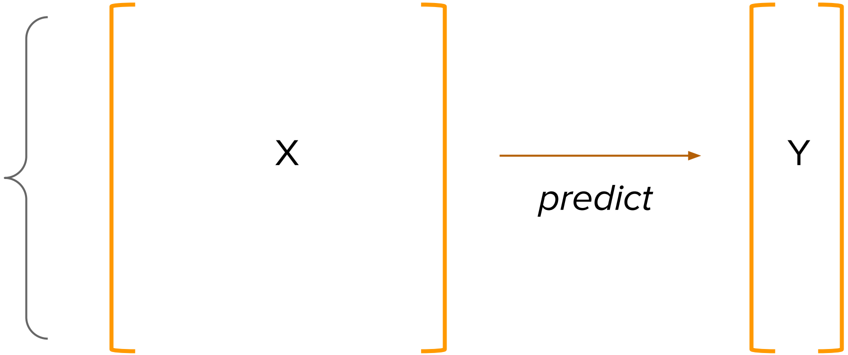
- ML algorithms often trained with convenient, inexpensive data.
- Big data ≠ good data.



Sampling bias

- ML algorithms often trained with convenient, inexpensive data.
- Big data ≠ good data.

Sampling bias:
Who is represented?

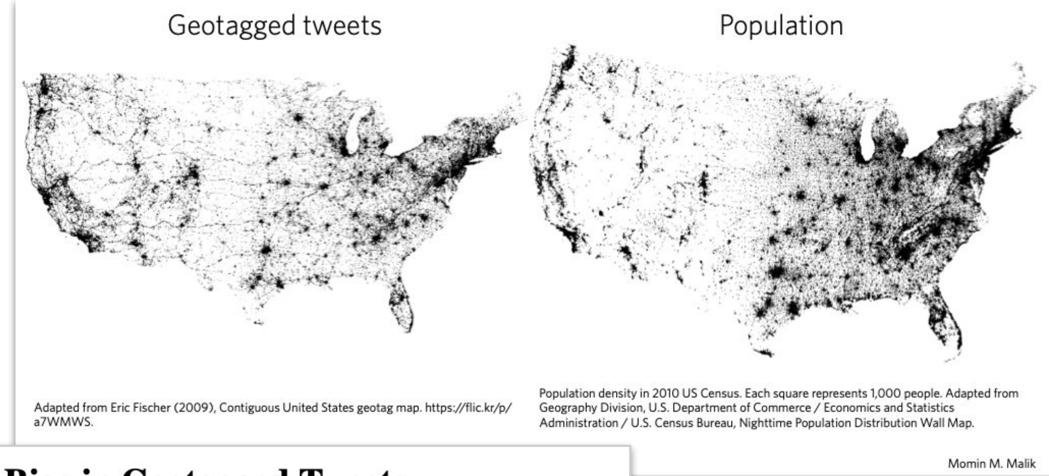


Sampling bias

- ML algorithms often trained with convenient, inexpensive data.
- Big data ≠ good data.

Why it matters?

May result in bias in disaster response (among other tasks).



Population Bias in Geotagged Tweets

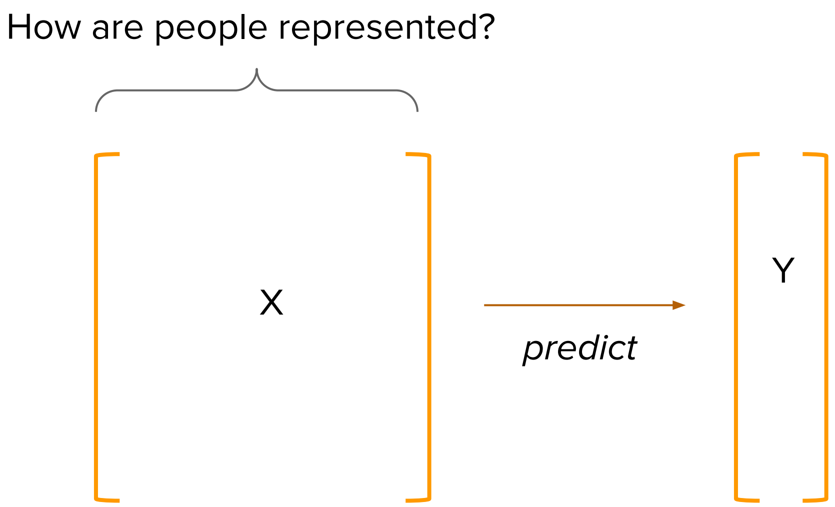
Momin M. Malik, Hemank Lamba, Constantine Nakos, and **Jürgen Pfeffer**
Institute for Software Research
School of Computer Science
Carnegie Mellon University

Sampling bias

- ML algorithms often trained with convenient, inexpensive data.
- Big data ≠ good data.
- Sources of sampling bias:
    - Access to technology and resources
    - Previously served/underserved communities
    - Trust in authorities (reporting)
- Invisibilization in sampling bias often *compounds* previous injustices.

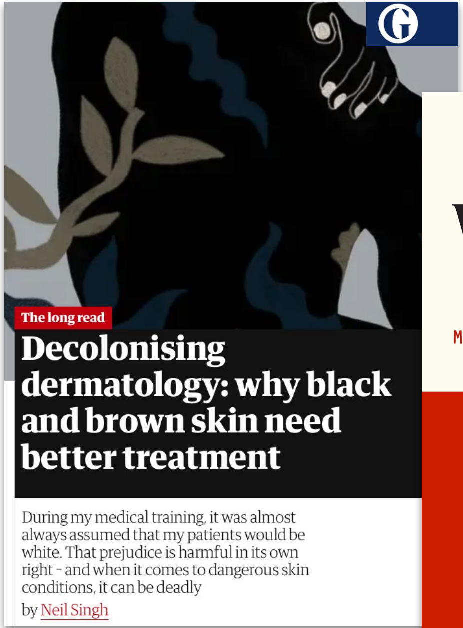
Differential subgroup validity

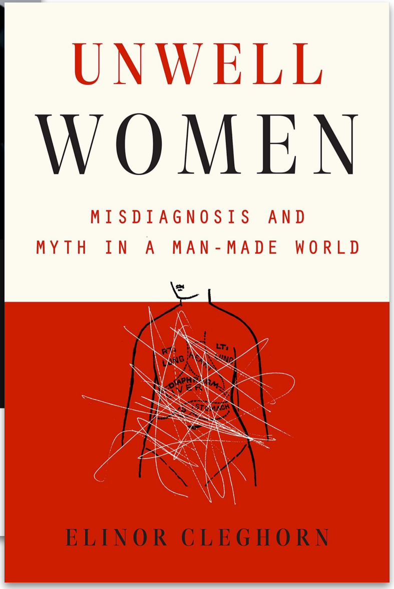
- Predictive power of features collected may differ across subpopulations.



Differential subgroup validity

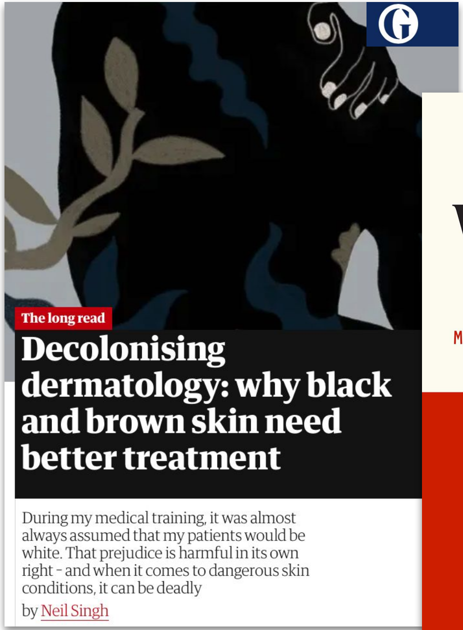
- Predictive power of features collected may differ across subpopulations.
- In healthcare: symptoms that are studied, taught, and recorded may only hold diagnostic power for some.





Differential subgroup validity

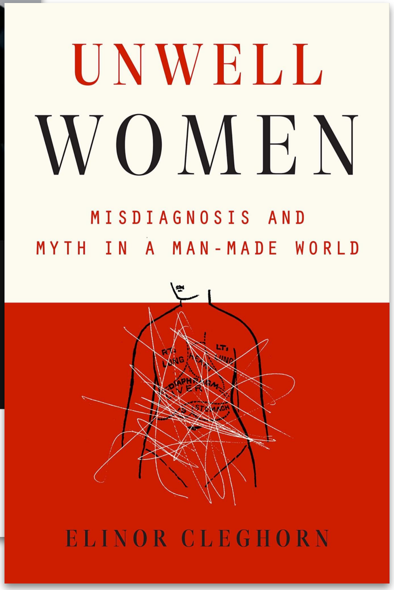
- Predictive power of features collected may differ across subpopulations.
- In healthcare: symptoms that are studied, taught, and recorded may only hold diagnostic power for some.
- Choice may be informed by, and only hold predictive power, in some cultural contexts.
    - Example: number of credit cards as a positive signal for “creditworthiness”.


The long read

Decolonising dermatology: why black and brown skin need better treatment

During my medical training, it was almost always assumed that my patients would be white. That prejudice is harmful in its own right – and when it comes to dangerous skin conditions, it can be deadly

by Neil Singh

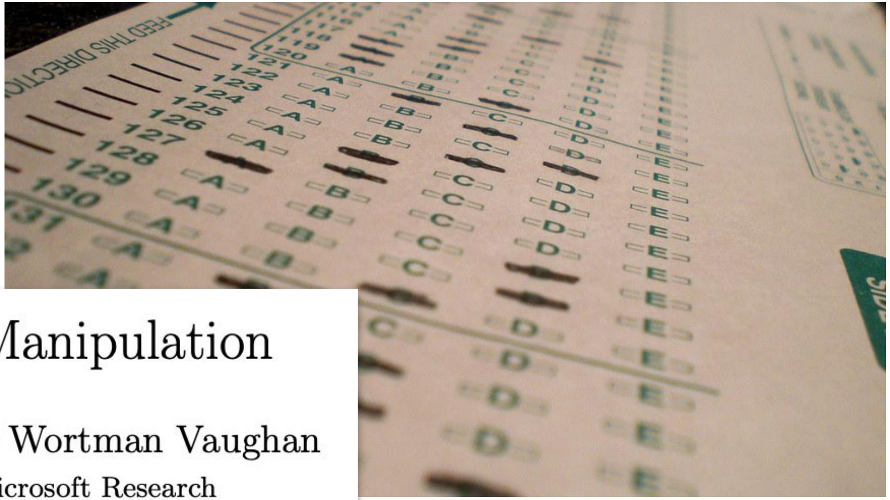


Differential subgroup validity

- Ability to adapt to a certain choice of features may also differ across groups.

Differential subgroup validity

- Ability to adapt to a certain choice of features may also differ across groups.
- Strategic adaptation to incentives is not possible for everyone.
- Standardized tests: incentivize students to invest in tutoring and retake tests, but not everyone can do this.



The Disparate Effects of Strategic Manipulation

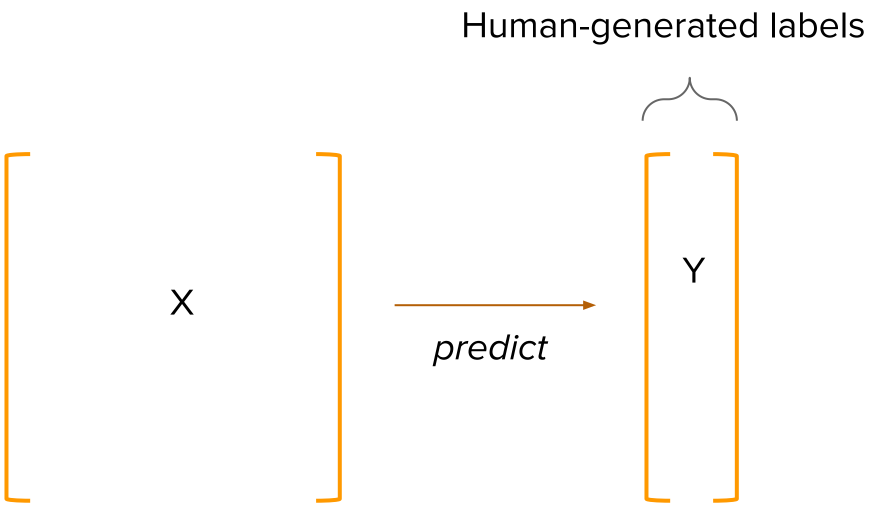
Lily Hu
Harvard University

Nicole Immorlica
Microsoft Research

Jennifer Wortman Vaughan
Microsoft Research

Human assessments encoded as labels

- Target labels are often termed “ground truth”, but they may also encode biases.
- Whose views are encoded and valued?



Human assessments encoded as labels

- Target labels are often termed “ground truth”, but they may also encode biases.
- Whose views are encoded and valued?

Implicitly or explicitly predicting HR and managerial reviews perpetuates past biases

REUTERS

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Human assessments encoded as labels

- Target labels are often termed “ground truth”, but they may also encode biases.
- Whose views are encoded and valued?

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / A YEAR AGO

REUTERS

Amazon scraps secret AI recruiting tool that showed bias against women

The NEW ENGLAND JOURNAL of MEDICINE

Perspective

Machine Learning and the Cancer-Diagnosis Problem — No Gold Standard

Adewole S. Adamson, M.D., M.P.P., and H. Gilbert Welch, M.D., M.P.H.

Lack of ground truth pervasive across tasks.

Sources of bias

Problem formulation

Data collection and representation

Optimization objectives and evaluation bias

Algorithmic adoption

Optimization objectives and evaluation bias

- What is the optimization objective? How do we evaluate performance?
- Even if data is representative of the target population, a choice of performance metric centered on “overall performance” is centering majority populations.

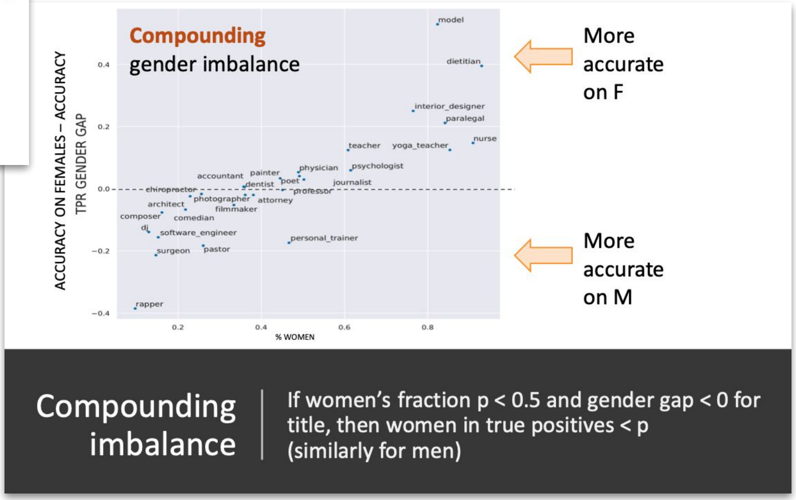
Optimization objectives and evaluation bias

- What is the optimization objective? How do we evaluate performance?
- Even if data is representative of the target population, a choice of performance metric centered on "overall performance" is centering majority populations.

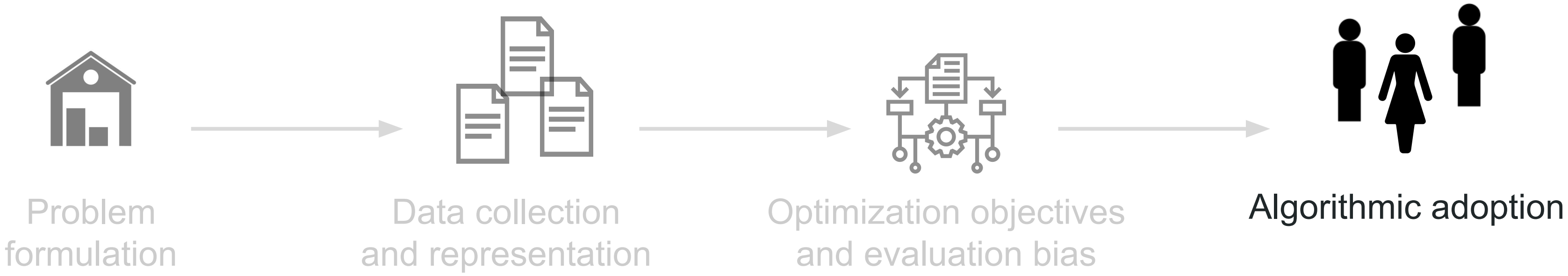
Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting

Maria De-Arteaga[1], Alexey Romanov[2], Hanna Wallach[3], Jennifer Chayes[3], Christian Borgs[3], Alexandra Chouldechova[1], Sahin Geyik[4], Krishnaram Kenthapadi[4], Adam Tauman Kalai[3]

[1]*Carnegie Mellon University,* [2]*University of Massachusetts Lowell,* [3]*Microsoft Research,* [4]*LinkedIn*



Sources of bias



Algorithmic adoption

- How are algorithmic recommendations integrated into decisions?
- Regardless of “fairness properties” of the predictions, non-uniform adherence to recommendations may lead to unfair decisions.

Algorithmic adoption

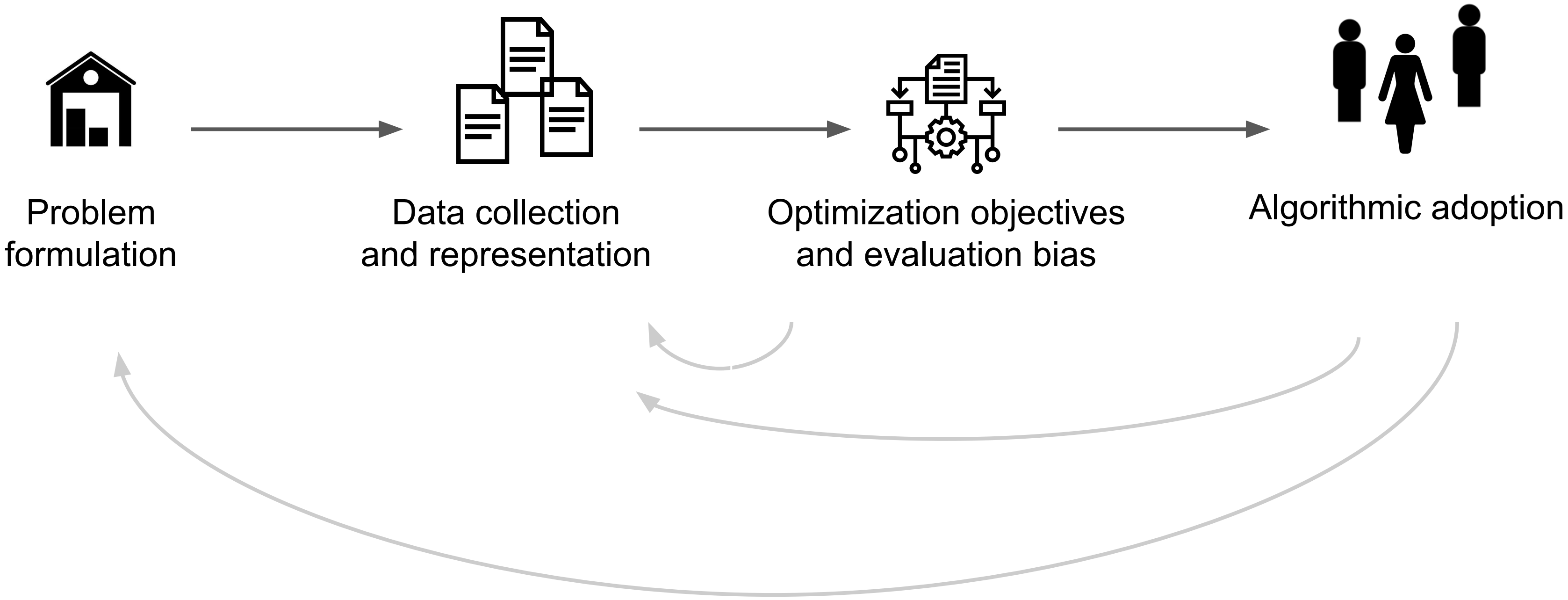
- How are algorithmic recommendations integrated into decisions?
- Regardless of “fairness properties” of the predictions, non-uniform adherence to recommendations may lead to unfair decisions.

If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions*

Alex Albright†

*“I show that this [raw racial disparities in initial bond] increase was **not simply a consequence of different risk scores** by race. Rather, the **recommended default was also more likely to be overridden** (in favor of harsher bond conditions) for black defendants than similar white defendants”.*

Sources of bias: Multiple may be present and interact with each other



Rethinking the bias problem

- General framing: The bias is in the data.

Rethinking the bias problem

- General framing: ~~The bias is in the data~~.
- Alternative framing: The bias is in the gap between the question that motivates us and the question that we are answering.

Rethinking the bias problem

- General framing: ~~The bias is in the data~~.
- Alternative framing: The bias is in the gap between the question that motivates us and the question that we are answering.
- Centering decisions, rather than predictions, allows us to:
  - Reimagine the questions that we are answering.
  - Rethink how we integrate algorithms into decision-making pipelines.
  - Scrutinize proxy objectives.
  - Evaluate ML-assisted decisions, not ML predictions.
  - Anticipate long-term, contextual risks.

Outline

1. Taxonomy of sources of bias.

2. **Bias in victim crime reporting and its effect on predictive policing (*FAccT'21*).**

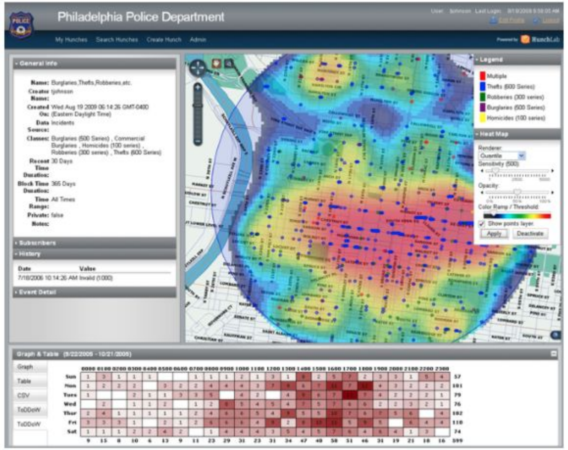
The effect of differential victim crime reporting on predictive policing systems

Nil-Jana Akpinar
nakpinar@stat.cmu.edu
Department of Statistics and Data Science & Machine Learning Department
Carnegie Mellon University

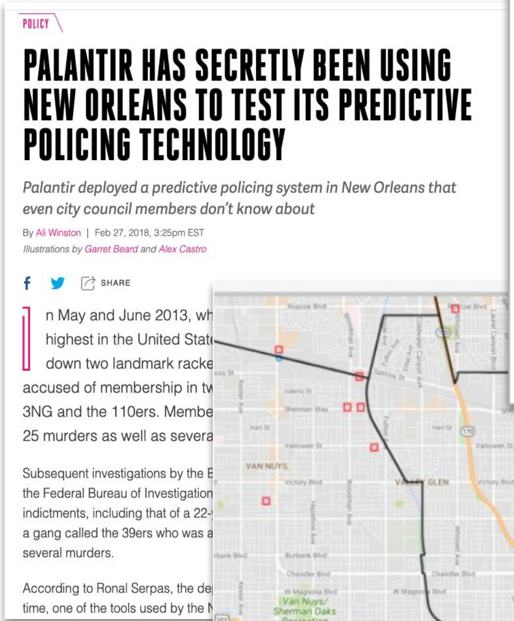
Maria De-Arteaga
Information, Risk, and Operations Management Department
McCombs School of Business
University of Texas at Austin

Alexandra Chouldechova
Heinz College & Department of Statistics and Data Science
Carnegie Mellon University

Predictive policing increasingly deployed across cities and countries



Azavea blog: Why we sold HunchLab (2019)



POLICY

PALANTIR HAS SECRETLY BEEN USING NEW ORLEANS TO TEST ITS PREDICTIVE POLICING TECHNOLOGY

Palantir deployed a predictive policing system in New Orleans that even city council members don't know about

By Ali Winston | Feb 27, 2018, 3:25pm EST

Illustrations by Garret Beard and Alex Castro

SHARE

In May and June 2013, wh
highest in the United State
down two landmark racke
accused of membership in tw
3NG and the 110ers. Membe
25 murders as well as severa

Subsequent investigations by the E
the Federal Bureau of Investigation
indictments, including that of a 22-
a gang called the 39ers who was a
several murders.

According to Ronal Serpas, the de
time, one of the tools used by the N

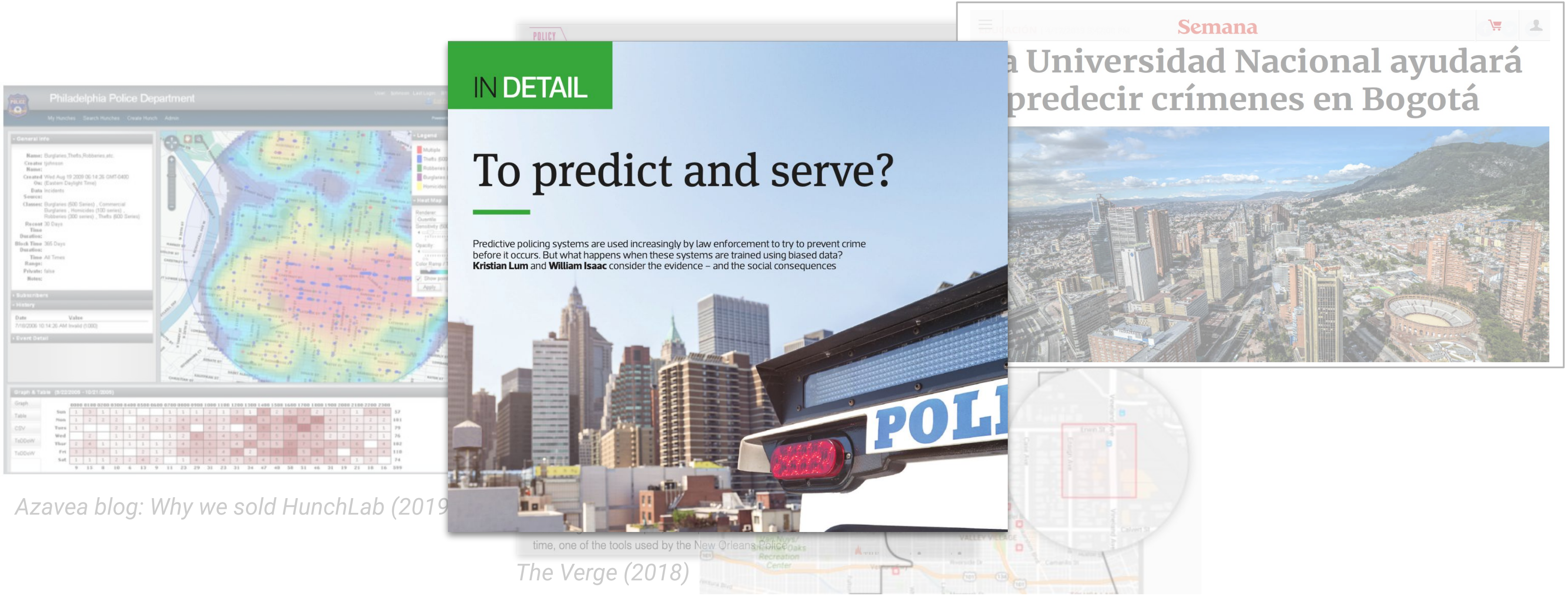
The Verge (2018)

PredPol website (2021)

Semana

La Universidad Nacional ayudará a predecir crímenes en Bogotá

Modern predictive policing systems have come under scrutiny due to a lack of transparency and concerns about biased outcomes.



Azavea blog: Why we sold HunchLab (2019

The Verge (2018)

PredPol website (2021)

Critics have demonstrated the potential for dangerous feedback loops when using arrest data.

- **Lum and Isaac [1]** demonstrate how using data on drug arrests in Oakland, CA as inputs to the PredPol predictive policing algorithm would result in high concentrations of policing in racial and ethnic minority neighborhoods.

- **Ensign et al. [2]** use a generalized Pólya urn model to theoretically analyze how feedback loops in arrest-based predictive policing systems arise.

[1] *Kristian Lum and William Isaac. 2016. To predict and serve? Significance (2016)*
[2] *Danielle Ensign et al. 2018. Runaway Feedback Loops in Predictive Policing. Conference on Fairness, Accountability, and Transparency (FAT* 2018)*

Critics have demonstrated the potential for dangerous feedback loops when using arrest data. Proponents argue other types of data are used.

- **Lum and Isaac [1]** demonstrate how using data on drug arrests in Oakland, CA as inputs to the PredPol predictive policing algorithm would result in high concentrations of policing in racial and ethnic minority neighborhoods.

- **Ensign et al. [2]** use a generalized Pólya urn model to theoretically analyze how feedback loops in arrest-based predictive policing systems arise.

PredPol [3]

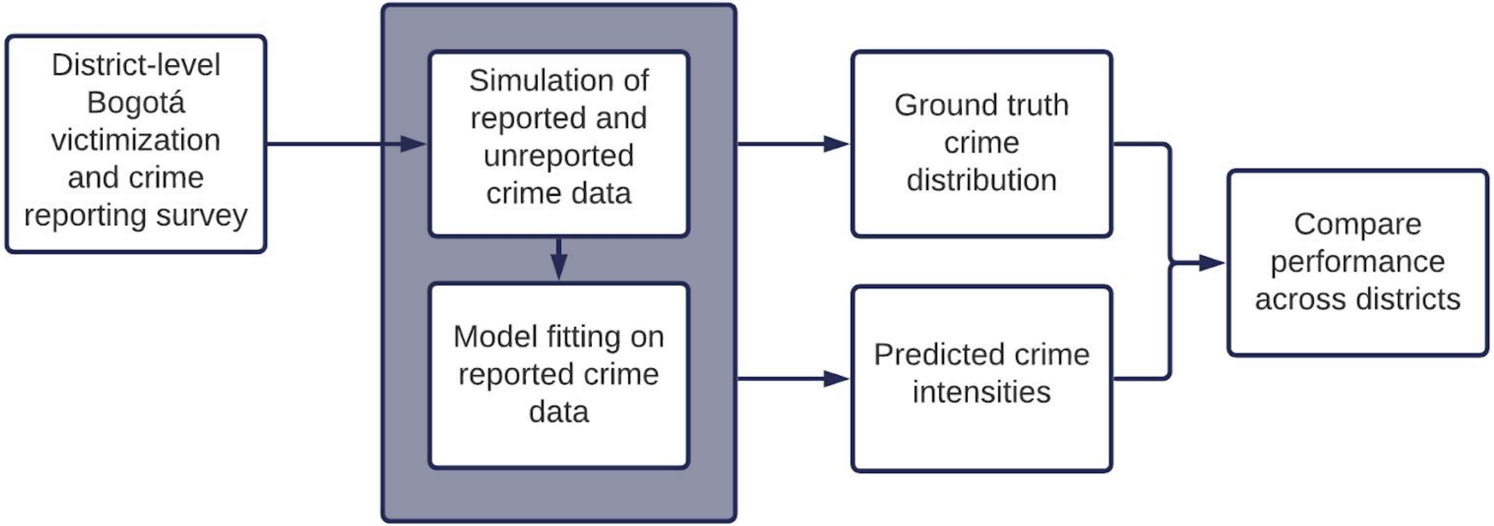
"unbiased nature of [...] algorithm"

"data collected and analyzed is primarily victim data"

"excludes drug related offenses and traffic citation data from its predictions to remove officer bias"

[1] *Kristian Lum and William Isaac. 2016. To predict and serve? Significance (2016)*
[2] *Danielle Ensign et al. 2018. Runaway Feedback Loops in Predictive Policing. Conference on Fairness, Accountability, and Transparency (FAT* 2018)*
[3] *PredPol 2017. Machine Learning and Policing.* https://blog.predpol.com/machine- learning-and-policing. [Online; accessed 1/20/21]

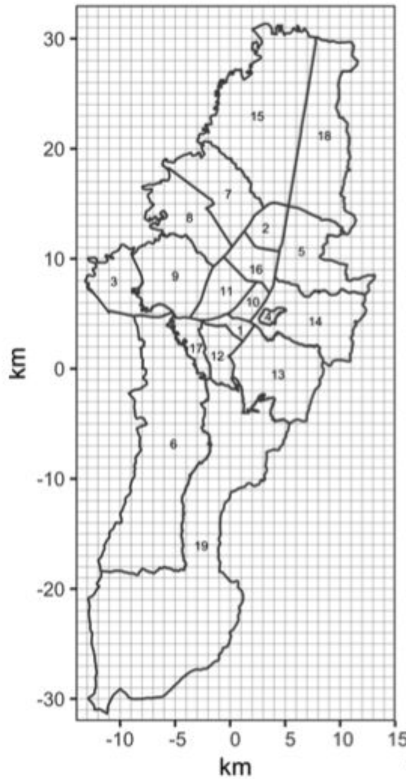
We demonstrate how differential victim crime reporting can lead to geographical outcome disparities in hot spot prediction with no arrest data used.



Self-exciting spatio-temporal point-process models (used by PredPol)

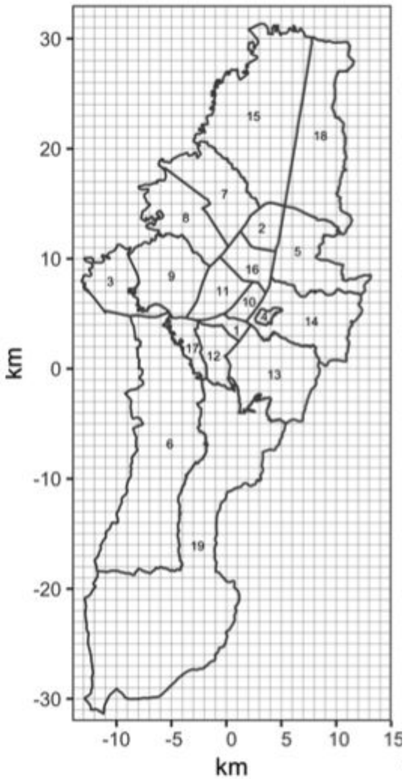
Our analysis is based on a crime simulation patterned after district-level crime statistics for Bogotá, Colombia.

- District-level victimization and victim crime reporting rates collected by Bogotá's chamber of commerce, Cámara de Comercio de Bogotá (CCB) in 2014.
- ~10,000 participants from all socio-economic statuses and all 19 urban districts.



Our analysis is based on a crime simulation patterned after district-level crime statistics for Bogotá, Colombia.

- District-level victimization and victim crime reporting rates collected by Bogotá's chamber of commerce, Cámara de Comercio de Bogotá (CCB) in 2014.
- ~10,000 participants from all socio-economic statuses and all 19 urban districts.



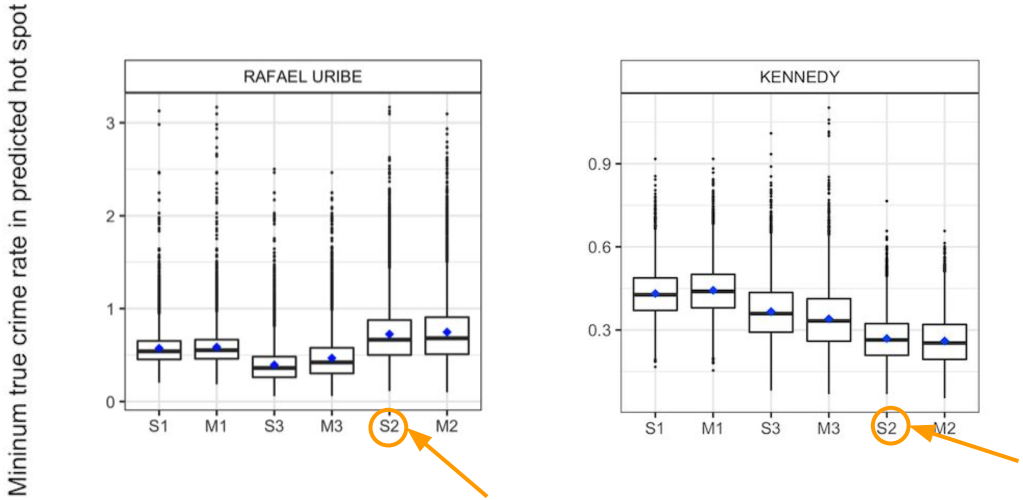
Victimization rate

Reporting rate

Id	District	Pop.	Vict. rate	Rep. rate
1	Antonio Nariño	109,176	15 %	33 %
2	Barrios Unidos	243,465	12 %	22 %
3	Bosa	673,077	13 %	26 %
4	Candelaria	24,088	12 %	22 %
5	Chapinero	139,701	9 %	28 %
6	Ciudad Bolívar	707,569	8 %	17 %
7	Engativá	887,080	11 %	20 %
8	Fontibón	394,648	10 %	19 %
9	Kennedy	1,088,443	13 %	28 %
10	Los Mártires	99,119	17 %	25 %
11	Puente Aranda	258,287	14 %	32 %
12	Rafael Uribe Uribe	374,246	12 %	15 %
13	San Cristóbal	404,697	13 %	21 %
14	Santa Fe	110,048	17 %	17 %
15	Suba	1,218,513	5 %	19 %
16	Teusaquillo	153,025	14 %	19 %
17	Tunjuelito	199,430	17 %	23 %
18	Usaquén	501,999	18 %	13 %
19	Usme	457,302	9 %	33 %

When trained on only reported crime data, some districts require more than double the crime rate of other districts to have their cells selected as hot spots.

When trained on only reported crime data, some districts require more than double the crime rate of other districts to have their cells selected as hot spots.



S1 = all crime SEPP

S2 = reported crime SEPP

S3 = rescaled reported crime SEPP

M1 = all crime MAVG

M2 = reported crime MAVG

M3 = rescaled reported crime MAVG

Minimum expected crimes required to predict hot spot on average: 0.27
victim crime reporting rate 28%

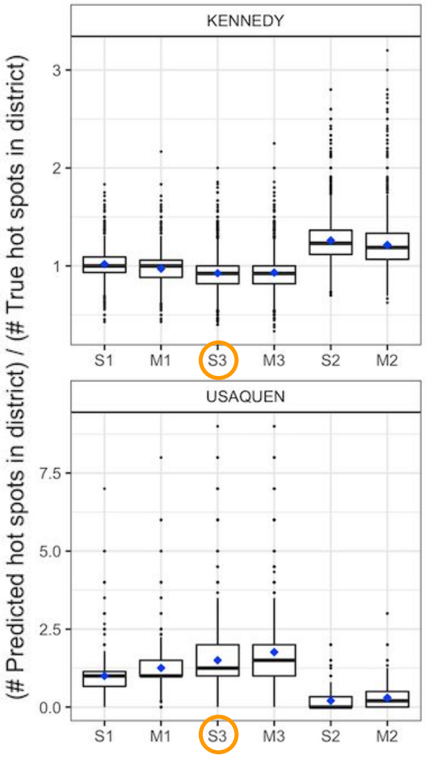
Minimum expected crimes required to predict hot spot on average: 0.73
victim crime reporting rate 13%

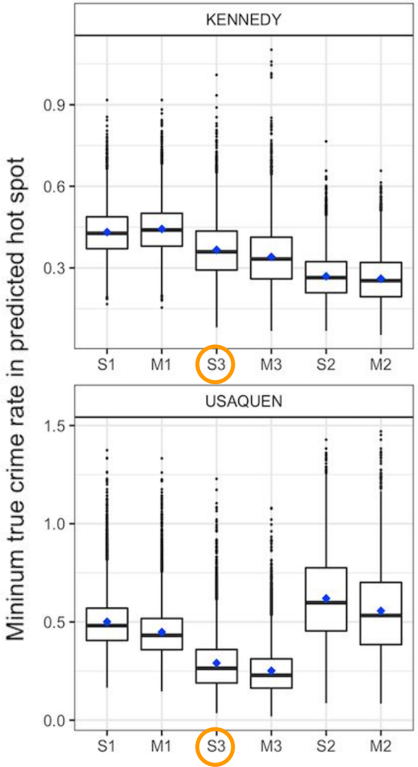
On average the minimum true crime rate that leads to a predicted hot spot in Rafael Uribe Uribe is 2.7 times the minimum crime rate required in Kennedy

Rescaling according to victim reporting rates as solution?

Rescaling according to victim reporting rates as solution? Only alleviates the problem partly.

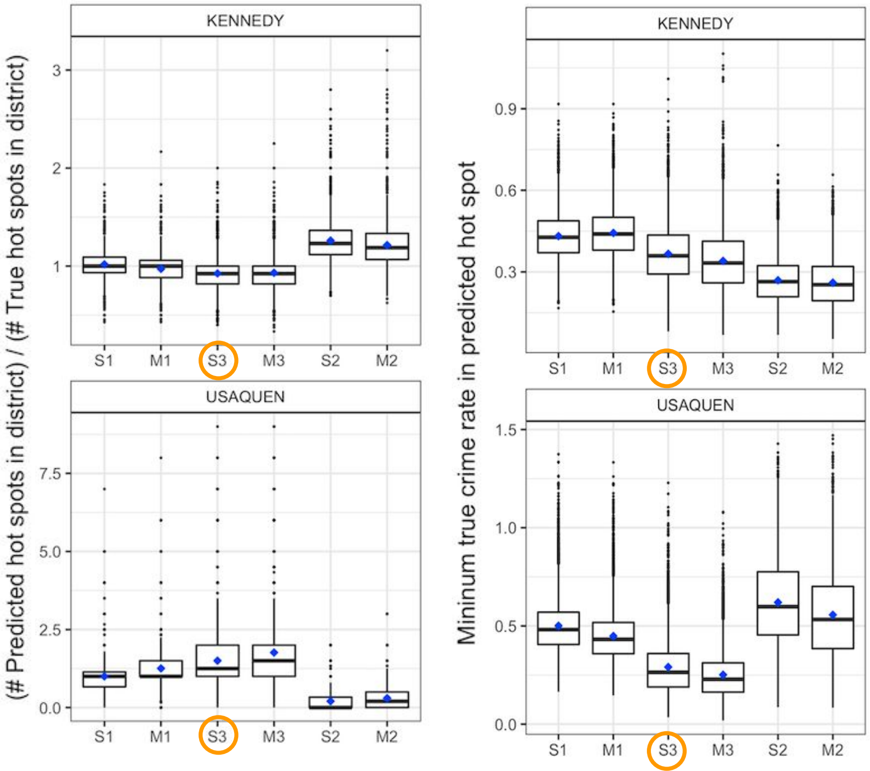
Rescaling according to victim reporting rates as solution? Only alleviates the problem partly.





- Rescaled model (S3) often closer to the full data model (S1) than crime-report model (S2).
- **Problem:** Rescaling increases predictions in district by **fixed factor irrespective of cell-specific crime**. The wrong cells can be selected in consequence.
  - ➢ Misallocation within district.

Rescaling according to victim reporting rates as solution? Only alleviates the problem partly.



- Rescaled model (S3) often closer to the full data model (S1) than crime-report model (S2).

- **Problem:** Rescaling increases predictions in district by **fixed factor irrespective of cell-specific crime**. The wrong cells can be selected in consequence.
  - ➢ Misallocation within district.

- In order to recover the cell-wise true crime distribution, a cell-by-cell rate of victim crime reporting would be required which is unattainable in practice.

Not specific to SEPP: A within-cell exponentially weighted moving average model of crime counts leads to very similar results.

Figure. Equity measures for hot spot selection in Bogotá districts. Each data point represents a distinct evaluation day (189 days) in a given simulation run (50 runs).

S1 = all crime SEPP

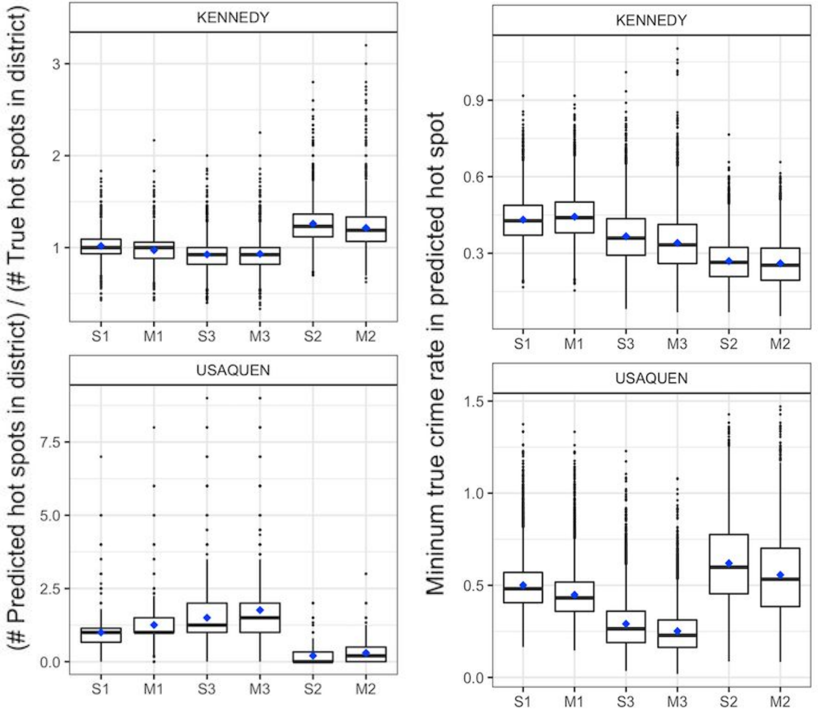
S2 = reported crime SEPP

S3 = rescaled reported crime SEPP

M1 = all crime MAVG

M2 = reported crime MAVG

M3 = rescaled reported crime MAVG



Findings

Differences in victim crime reporting rates can lead to geographical bias in common hot spot prediction algorithms even when no data from arrests or police initiated contact is used.

Findings

Differences in victim crime reporting rates can lead to geographical bias in common hot spot prediction algorithms even when no data from arrests or police initiated contact is used.

Implications

This can lead to misallocation of police patrols in the form of over-policing of some neighborhoods while areas with effectively higher crime rates are under-policed.

Findings

Differences in victim crime reporting rates can lead to geographical bias in common hot spot prediction algorithms even when no data from arrests or police initiated contact is used.

Implications

This can lead to misallocation of police patrols in the form of over-policing of some neighborhoods while areas with effectively higher crime rates are under-policed.

Non-solution

It is unclear how this problem could be mitigated.
Rescaling predicted crime rates by surveyed victim crime reporting rates also increases noise and can can make singling out specific cells within a district hard.

Findings

Differences in victim crime reporting rates can lead to geographical bias in common hot spot prediction algorithms even when no data from arrests or police initiated contact is used.

Implications

This can lead to misallocation of police patrols in the form of over-policing of some neighborhoods while areas with effectively higher crime rates are under-policed.

Non-solution

It is unclear how this problem could be mitigated.
Rescaling predicted crime rates by surveyed victim crime reporting rates also increases noise and can can make singling out specific cells within a district hard.

Socio-technical context

Victim crime reporting rates are known to be driven by socio-economic factors, types of crime and other demographics. More work is needed for an in-depth discussion of the interplay between predictive disparities and these factors in the Bogotá context.

Thanks!

@mariadearteaga