10-301/601: Introduction to Machine Learning Lecture 16 – Societal Impacts of ML

Henry Chai & Matt Gormley & Hoda Heidari

03/18/24

ML in Societal Applications

8 WAYS MACHINE LEARNING WILL IMPROVE EDUCATION

BY MATTHEW LYNCH / JUNE 12, 2018 / 5

≡ techworld FROM IDG Features Technology Innovation Partner Zone the techies

Home › Features › Emerging tech & innovation Features

Researcher explains how algorithms can create a fairer legal system

Deep learning is being used to predict critical COVID-19 cases

Artificial Intelligence and Accessibility: Examples of a Technology that Serves People with Disabilities

The New Y

TheUpshot

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

By Claire Cain Miller

SCI-FI VISIONS

Your Future Doctor May Not be Human. This Is the Rise of AI in Medicine.

From mental health apps to robot surgeons, artificial intelligence is already changing the practice of medicine.

20 JAN 2017 | Insight

Kevin Petrasic | Benjamin Saul

Algorithms and bias: What lenders need to know

The algorithms that power fintech may discriminat[e …] can be difficult to anticipate—and financial instituti[ons …] accountable even when alleged discrimination is c[…] unintentional.

HOME > STRATEGY

Artificial intelligence is slated to disrupt 4.5 million jobs for African Americans, who have a 10% greater likelihood of automation-based job loss than other workers

Allana Akhtar Oct 7, 2019, 12:57 PM

Misinformation on coronavirus is proving highly contagious

By DAVID KLEPPER July 29, 2020

ACLU

Email address | ZIP code

BECOME A MEMBER / RENEW / TAKE ACT

ISSUES KNOW YOUR RIGHTS DEFENDING OUR RIGHTS BLOGS ABOUT

SPEAK FREELY

How Facebook Is Giving Sex Discrimination in Employment Ads a New Life

By Galen Sherwin, ACLU Women's Rights Project

SEPTEMBER 18, 2018 | 10:00 AM

The Switch

Wanted: The 'perfect babysitter.' Must pass AI scan for respect and attitude.

The New York Times

I.R.S. Changes Audit Practice That Discriminated Against Black Taxpayers

The agency will overhaul how it scrutinizes returns that claim the earned-income tax credit, which is aimed at alleviating poverty.

MEDICAL MALAISE

If you're not a white male, artificial intelligence's use in healthcare could be dangerous

By Robert David Hart • July 10, 2017

ProPublica

Donate

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Societal Goals

Foster:

- Productivity and efficiency gains
- Innovation and economic growth
- Due process
  - Consistency
  - Traceability
  - Making choices & biases evident
- ...

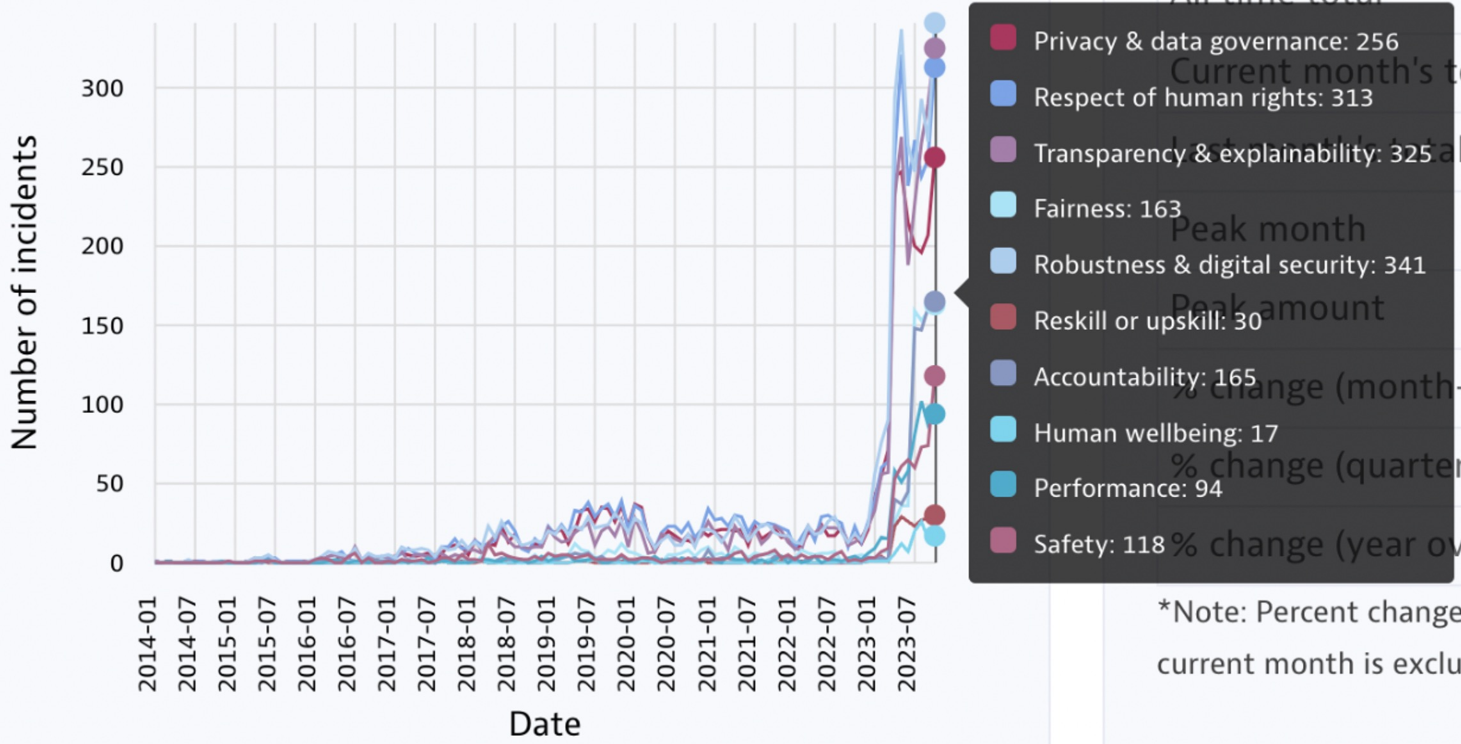
Mitigate:

- Violations of human rights
  - Justice, equity, and non-discrimination
  - Privacy and non-surveillance
  - Freedom of communication and expression
  - Economic freedom
- Negative impact on human flourishing and wellbeing
  - Loss of human sovereignty and control
  - Human cognitive abilities
  - ...

AI Incidents on the Rise

Summary visualisations

Evolution of incidents by AI principle



Summary statistics

	Incidents	Articles
All time total	6264	36345
Current month's total	317	1768
Last month's total	616	3227
Peak month	2023-10	2023-10
Peak amount	616	3227
% change (month-over-month)	23.2	51.22
% change (quarter-over-quarter)	13.01	13.87
% change (year over year)	961.58	690.9

*Note: Percent change is calculated based on preceding full months (i.e. the current month is excluded).

Principles

- Fairness
- Accountability
- Transparency
- Safety and reliability
- Privacy
- ...

Safe and Effective Systems

Algorithmic Discrimination Protections

Data Privacy

Notice and Explanation

Human Alternatives, Consideration, and Fallback

Beyond Principles

Concerns around **impact**:

- Economic (IP, Antitrust, labor market effects)
- Sustainability and environmental
- Eroding democratic values
  - misinformation and disinformation

Concerns around the **process**:

- Human sovereignty, autonomy, agency, self-determination
  - Participation
  - Recourse / appeal
  - Mental health
- ...

Unfairness and Discrimination

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

(Outcome) Unfairness

Formal Principle of Distributive Justice:

"Equals should be treated equally, and unequals unequally, in proportion to relevant similarities and differences." [Aristotle, …, Feinberg'1973]

↓

Working Definition of Outcome Unfairness:

Disparate or unequal allocation of harm/benefit across socially salient, but morally irrelevant groups of people.

Mathematical Notions of Fairness

- **Group** notions
  - Statistical parity
  - Equality of accuracy
  - Equality of false positive/false negative rates
  - Equality of positive/negative predictive value
- **Individual** notions
  - Treat similar individuals similarly.
- **Counterfactual** notions

Statistical/Demographic Parity

- Equal **selection rate** across different groups:

$$P[\hat{Y} = 1 \mid S = s_1] = P[\hat{Y} = 1 \mid S = s_2]$$

- Equal Employment Opportunity Commission:

> *"A selection rate for any race, sex, or ethnic group which is less than four-fifths (or 80%) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of [discrimination]."*

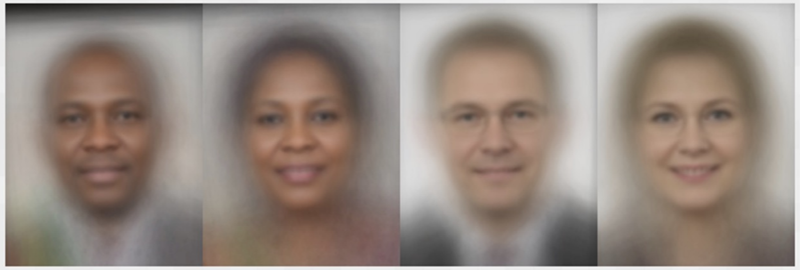
Equality of Accuracy

- Equality of the prediction accuracy (L) across groups:

$$E[L(\hat{y}, y) \mid S = s_1] = E[L(\hat{y}, y) \mid S = s_2]$$

- **Example:** Gender shades (Buolamwini et al.'18)

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%



Equality of FPR/FNR

- Equality of the **False Positive Rate (FPR)** across groups:

$$P[\hat{Y}=1 \mid Y=0, S=s_1]=P[\hat{Y}=1 \mid Y=0, S=s_2]$$

- Equality of the **False Negative Rate (FNR)** across groups:

$$P[\hat{Y}=0 \mid Y=1, S=s_1]=P[\hat{Y}=0 \mid Y=1, S=s_2]$$

- Equality of **Odds**: equal FNR and FPR simultaneously

Equality of PPV/NPV

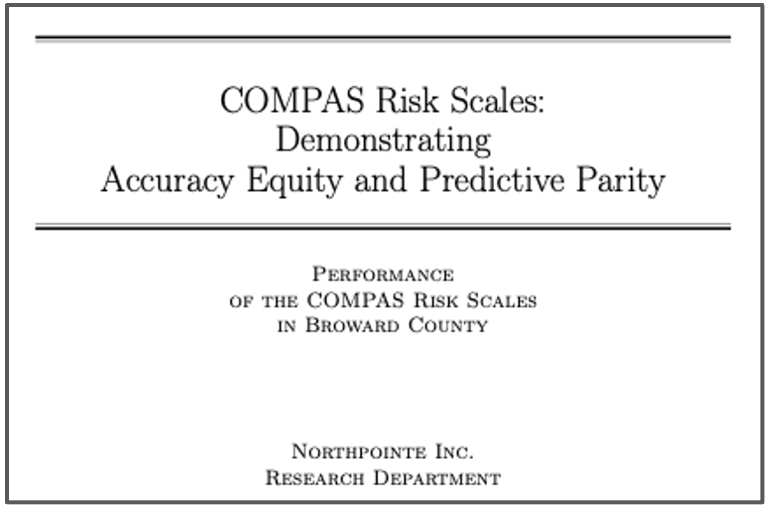
- Equality of the **Positive Predictive Value (PPV)**

$$P[Y=1 \mid \hat{Y}=1, S=s_1] = P[Y=1 \mid \hat{Y}=1, S=s_2]$$

- Equality of the **Negative Predictive Value (NPV)**

$$P[Y=0 \mid \hat{Y}=0, S=s_1] = P[Y=0 \mid \hat{Y}=0, S=s_2]$$

- **Predictive Value Parity (PVP):** equal PPV and NPV simultaneously



COMPAS Risk Scales:
Demonstrating
Accuracy Equity and Predictive Parity

PERFORMANCE
OF THE COMPAS RISK SCALES
IN BROWARD COUNTY

NORTHPOINTE INC.
RESEARCH DEPARTMENT

Common Pros and Cons

- Ignoring possible correlation between Y and S.
- Allowing for trading off different types of error.
- Not considering practical considerations.
    - e.g., High accuracy difficult to attain for small groups
- ...

Summary of Fairness Notions w. Confusion Matrix

For each group s, form:

	$\hat{Y}=0$	$\hat{Y}=1$
Y=0	a (true negative)	b (false positive)
Y=1	c (false negative)	d (true positive)

- Statistical parity = Equality of $\frac{b+d}{a+b+c+d}$
- Equality of accuracy = Equality of $\frac{a+d}{a+b+c+d}$
- Equality of FPR/FNR = Equality of $\frac{b}{a+b} / \frac{c}{c+d}$
- Equality of PPV/NPV = Equality of $\frac{d}{d+b} / \frac{a}{a+c}$

across all s.

Individual vs. Group Fairness

- Treating people as individuals, regardless of their group membership.
- Disparate Treatment:

  "Similarly situated individuals must be treated similarly."

- Similarity must be defined *with respect to the task at hand.*

  Example: movie casting vs. employment decisions in tech sector

Formalizing Individual Fairness

(Dwork et al. 2012):

- $d(\mathbf{x}_i,\mathbf{x}_j)$: a metric defining distance between two individuals
- D: a measure of distance between distributions
- A randomized classifier h mapping $\mathbf{x}$ to $\Delta_h(\mathbf{x})$ satisfies the (D, d)-Lipschitz property if $\forall \mathbf{x}_i$, $\mathbf{x}_j$,

$$D(\Delta_h(\mathbf{x}_i), \Delta_h(\mathbf{x}_j)) \le d(\mathbf{x}_i , \mathbf{x}_j).$$

Several problems with the Formulation

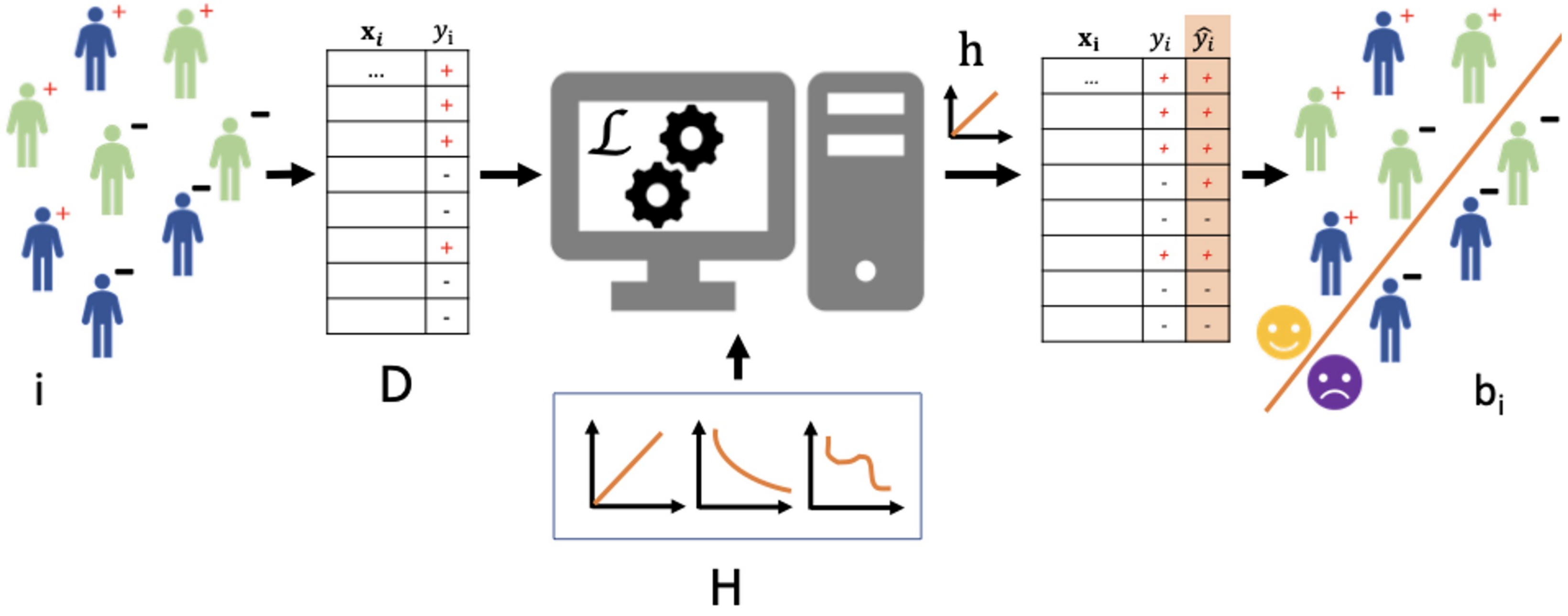
- Does not treat **dis**similar individuals **differently**.
- How should we pick d and D?
- Applicable to probabilistic models, only.
- Computationally expensive ($O(n^2)$ pairwise constraints)
- …

Myth: Data and ML Tools Are Neutral!

- Translating high-level goals into data is not neutral.
- Data at best reflect the current state of the world.
- Learning algorithms pick up the patterns in data.
- Predictive models make errors.
- Deployment in real-world may have unforeseen consequences.

Simplified ML Pipeline

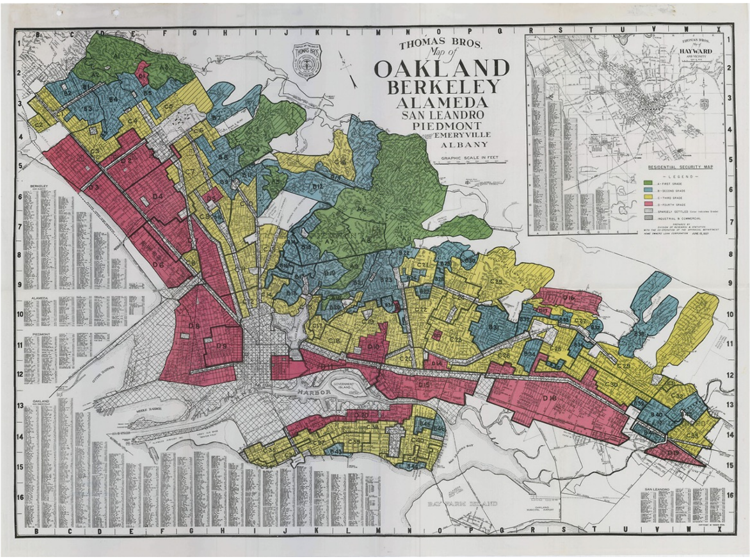
1. Task definition → Choosing (**x**,y)
2. Data collection → Collecting D
3. Model specification → choosing H
4. Model fitting/training → choosing and optimizing for L
5. Deployment in real-world → translating y^ into decisions leading to b_i : D'



Task Definition

Feature selection (**x**)

- Different statistical properties (e.g., SAT score)
- Omitted variable bias (e.g., SAT prep courses)
- Proxies (e.g., redlining)



Task Definition

Choice of the target variable (y)

- Ambiguous target (e.g., “good employee” vs. “positive annual evaluations”);
- Proxy target (e.g.,“commit a crimes” vs. “is rearrested”)
- Discretization (e.g., binary gender classification)

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc’s (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Data Collection

Sample selection bias (D)

- Under/over-representation (e.g., street bumps app)
- Less data from the minority (e.g., accents in speech recognition)
- Outdated instances (e.g., hiring decisions for IT positions)

Boston releases Street Bump app that automatically detects potholes while driving

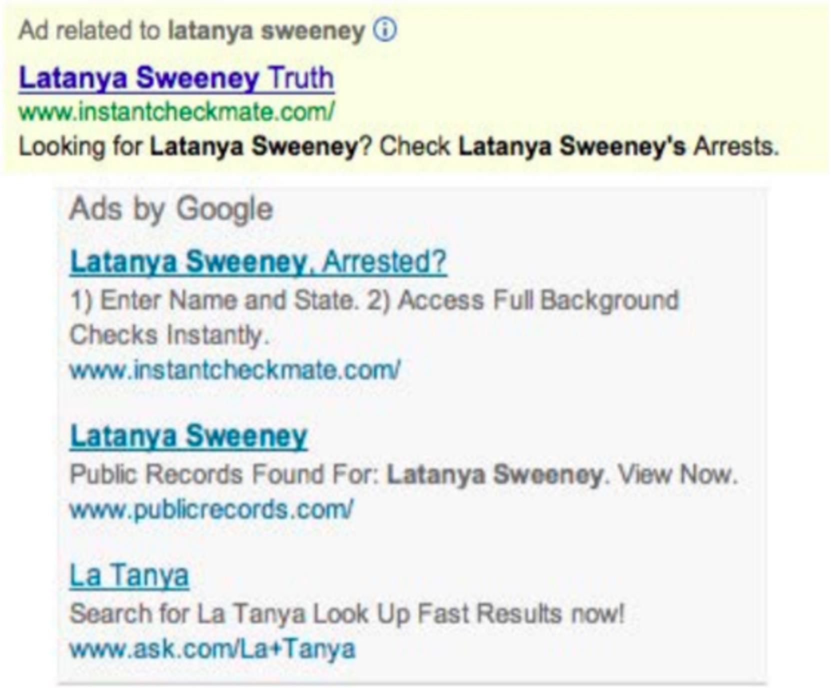
By DAILY MAIL REPORTER

PUBLISHED: 00:37 GMT, 21 July 2012 | **UPDATED:** 01:01 GMT, 21 July 2012

Data Collection

Data encoding past or existing injustices and prejudices

- Google queries for black-sounding names



Data Collection

Measurement bias (x)

- e.g., assessing levels of pain

INSIGHTS | DIVERSITY AND INCLUSION | HEALTH CARE | MEDICAL EDUCATION

How we fail black patients in pain

Model Specification

Simplified setting:

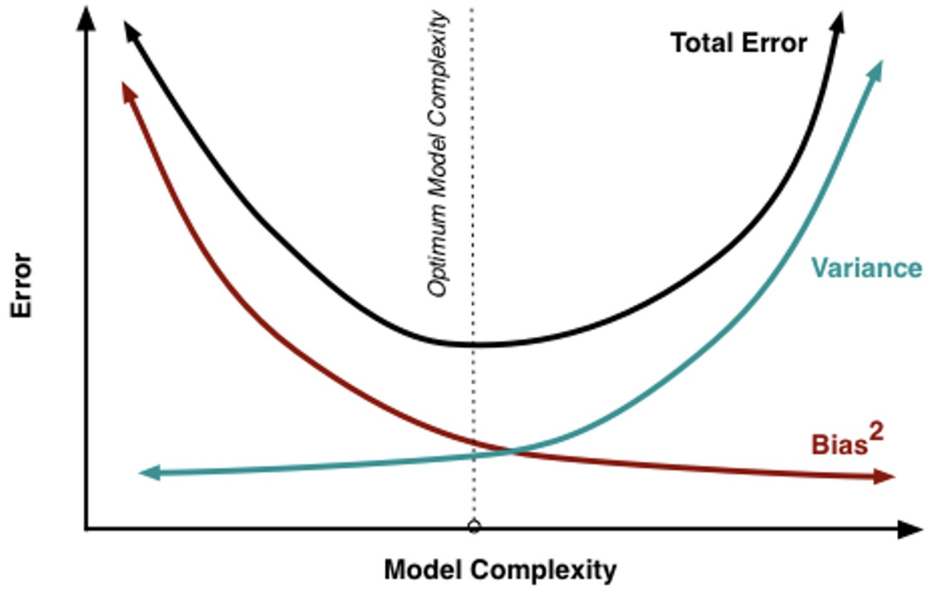
- f^*, the underlying model ($y_i = f^*(x_i) + \varepsilon_i$).
- $h^* \in H$, the best available hypothesis.
- $h = \arg\min_{h' \in H} L(D, h')$, the best model on finite sample
- For the sake of concreteness, let's for now assume $s \in \{A, D\}$,

$$\text{Unfairness} = E[(h(x) - y)^2 \mid s = D] - E[(h(x) - y)^2 \mid s = A]$$

Model Specification

$E[(h(x)-y)^2 \mid s] = E[\,(h(x)-h^*(x)+h^*(x)-f^*(x)+f^*(x)-y)^2 \mid s]$

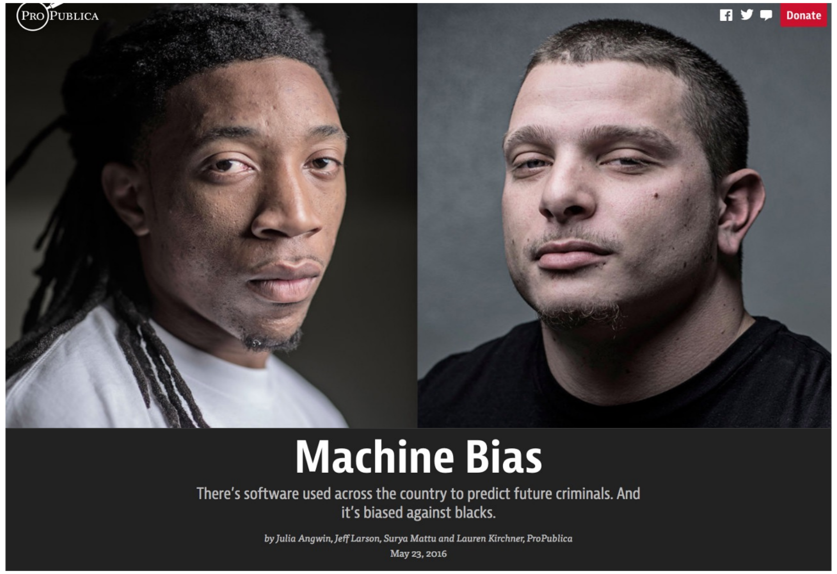
- Inherent uncertainty: $E[\,(f^*(x)-y)^2 \mid s\,] = Var[\varepsilon \mid s]$.
- Approximation error (choice of H): $E[\,(f^*(x) - h^*(x))^2 \mid s]$.
- Estimation error: $E[\,(h^*(x) - h(x))^2 \mid s]$



Model Training

Choice of objective function (L)

- Defining the cost or utility to be optimized
- Choice of the regularizer
- Optimization



Deployment Consequences

Feedback loops, e.g.,

- Observe if “crime rate is high” only if there is enough policing.
- Observe if “paid back the loan” only if loan granted.
- Observe if “committed a crime” only if released on bail.

Biased policing is made worse by errors in pre-crime algorithms

4 October 2017 , updated 27 April 2018

By Matt Reynolds

Deployment Consequences

Mismatch between training and deployment populations

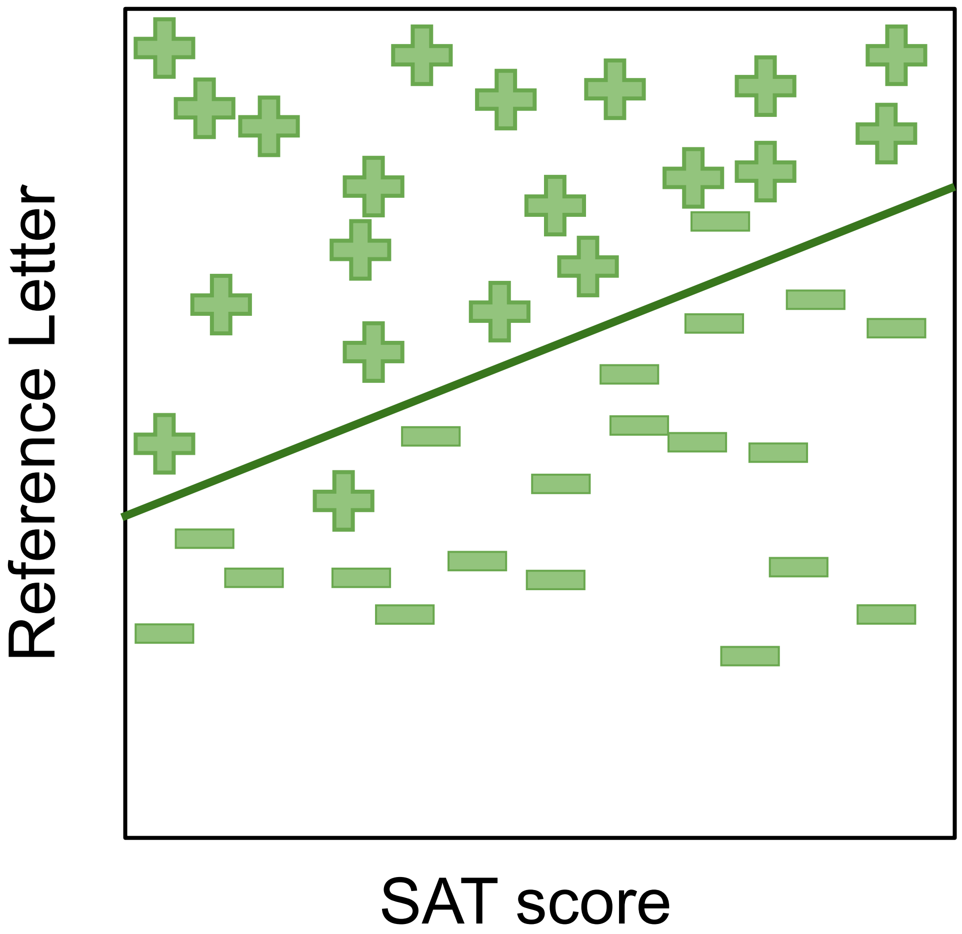
- Different population (e.g., facial recognition)
- Drifting populations (e.g., predictive policing)

Deployment Consequences

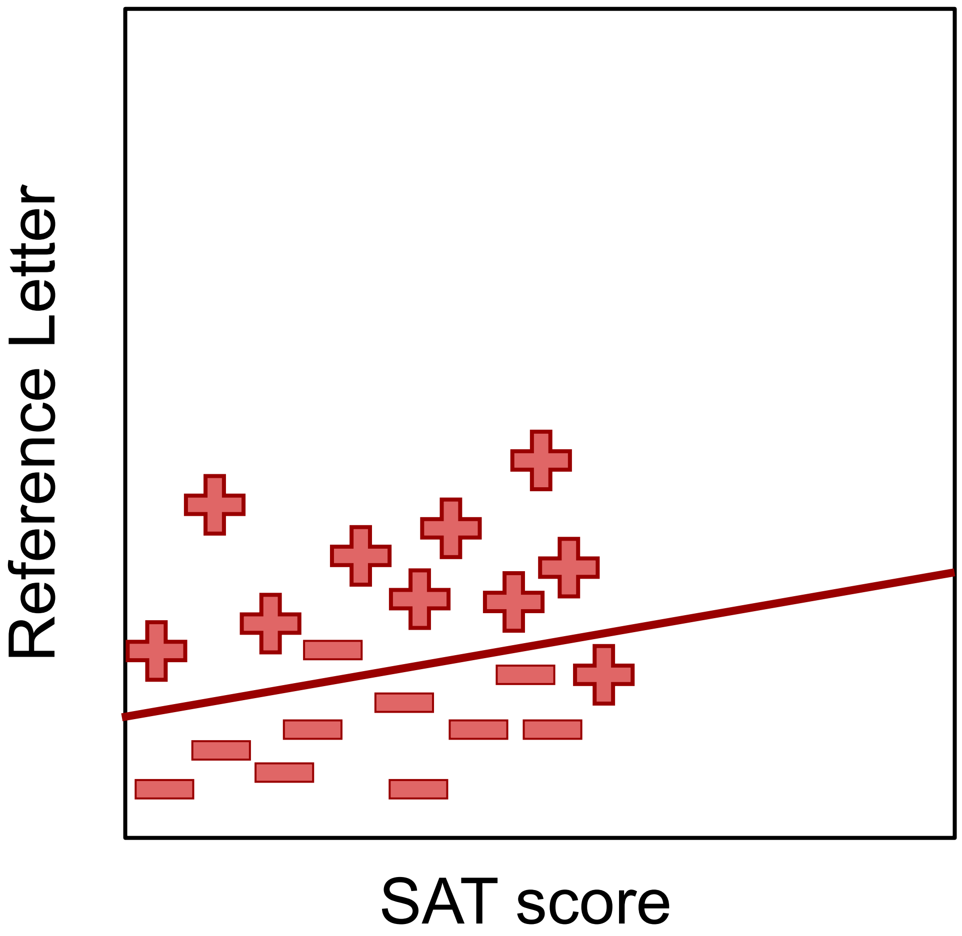
Adverse strategic response

- Gaming the system
- Unintended use or adversarial attacks (e.g., Tay.ai)

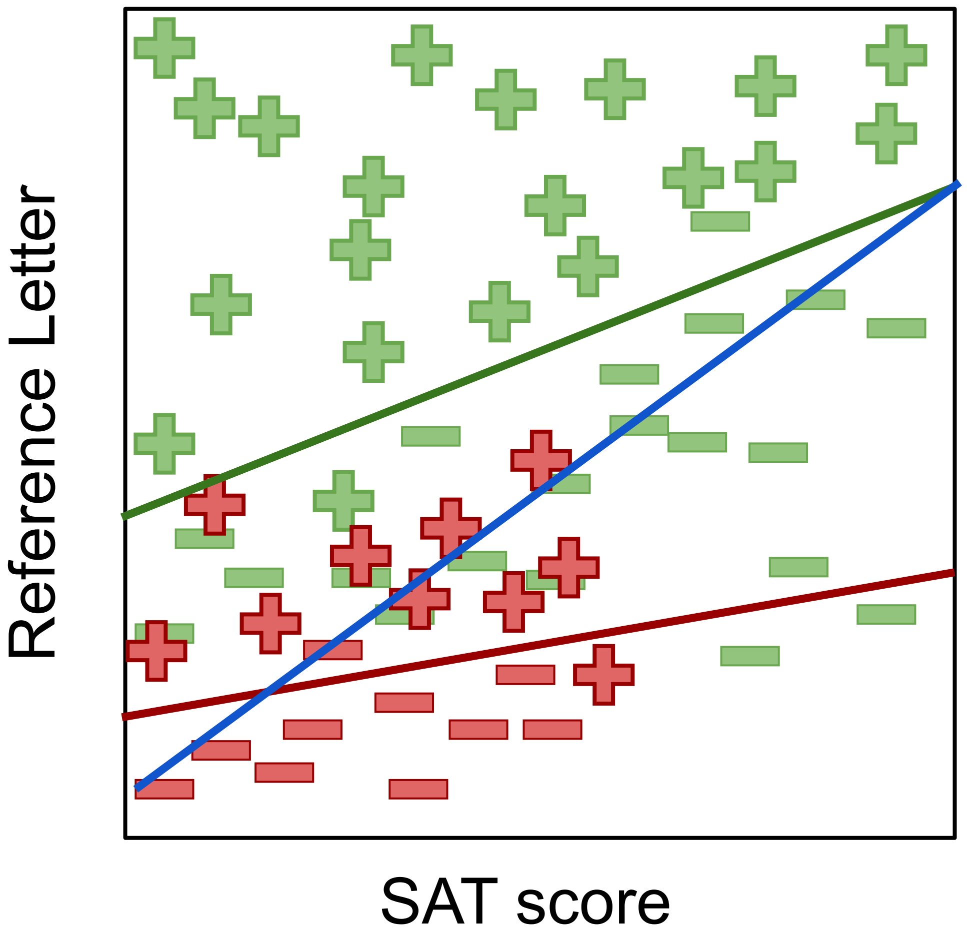
Example: College Admissions



Example: College Admissions



Example: College Admissions



Example: College Admissions

Evident biases:

- Less data from the minority (i.e., red)
- Different statistical correlation (i.e., SAT score with success)
- Disparate error distribution
- Omitted variable bias (i.e., group membership)

Potential biases:

- Labels in the dataset may be biased against reds.
- Measurement bias (i.e., strength of letter)
- Discouraging red students

…

Objectives

- Awareness of the common societal/ethical concerns surrounding the use of AI in society
- Familiarity with existing notions of fairness and their limitations
    - Mathematical definitions
    - How to compute them using the confusion matrix
- Ability to hypothesize causes of unfairness in a given application