Berkeley
UNIVERSITY OF CALIFORNIA

DAYLIGHT
SECURITY
RESEARCH LAB

ML Fairness Bootcamp

Day 1: Investigating bias

Nick Merrill

Credits

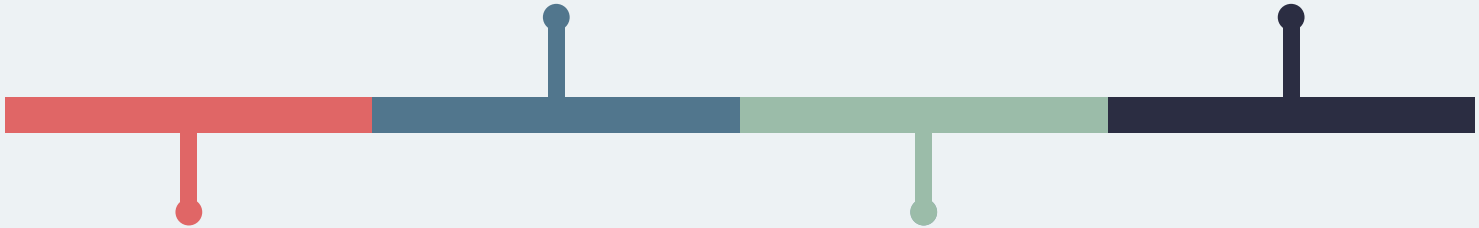
Daylight Security Research Lab, UC Berkeley

The barrier to entry for creating unfair systems is much lower than barrier to identifying bias in systems. The Daylight Lab has created interactive labs to help students identify (and ameliorate) bias in ML models. This mini-lecture series draws from their research, with credit to the following:

- Nick Merrill – director of the Daylight Research Lab, researcher at the Center for Long Term Cybersecurity
- Research Assistants Inderpal Kaur, Samuel Greenberg, Jasmine Zhang

Learn more at **https://daylight.berkeley.edu/mlfailures**.

Bootcamp Timeline



Lab 1

Bias in Healthcare

Lab 2

Gender Bias in Hiring

Lecture 1

Introducing Bias

Lecture 2

Mitigating Bias

Table of contents

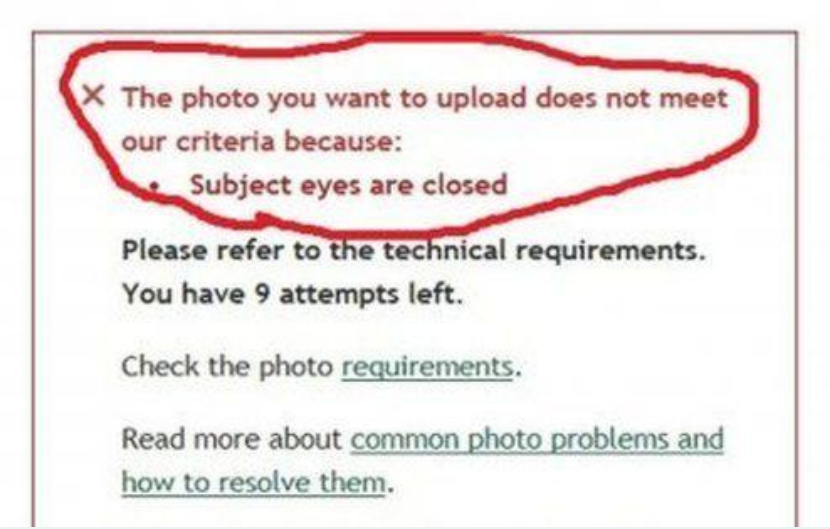
01

What's the problem?

A **ML Failure** occurs when a ML system does something unanticipated and/or undesirable.

... in Facial Recognition,

an ML Failure... makes it hard for someone of East Asian descent to get their passport.

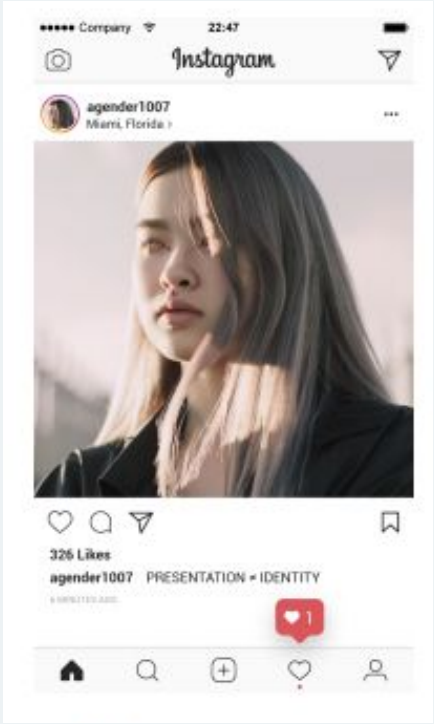


... in Gender Classification,

an ML Failure leads to... someone being misgendered.

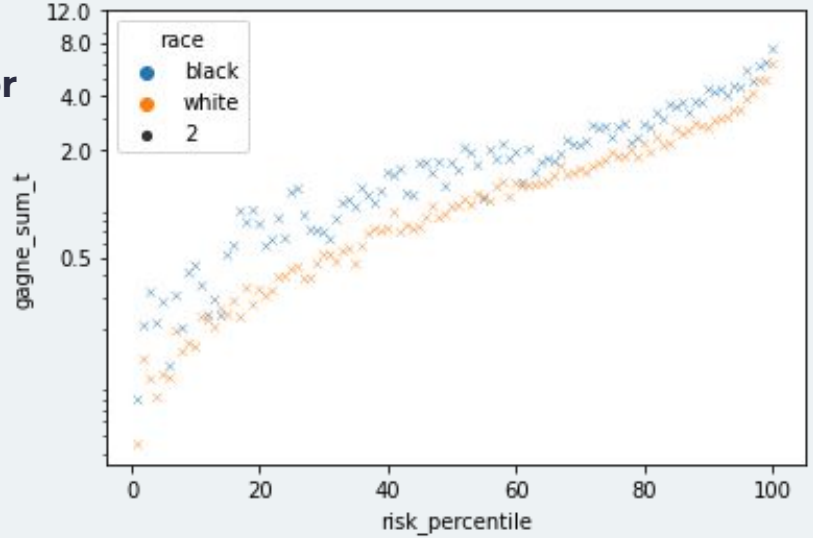
```
...
"age": {
    "min": 20,
    "max": 23,
    "score": 0.923144
},
"face_location": {
    "height": 494,
    "width": 428,
    "left": 327,
    "top": 212
},
"gender": {
    "gender": "FEMALE",
    "gender_label": "female",
    "score": 0.9998667
}
```

```
{
  "class": "woman",
  "score": 0.813,
  "type_hierarchy": "/person
  /female/woman"
},
{
  "class": "person",
  "score": 0.806
},
{
  "class": "young lady (heroine)",
  "score": 0.504,
  "type_hierarchy": "/person/female
  /woman/young lady (heroine)"
}
...
```



... in Healthcare,

an ML Failure leads to... worse medical care for Black patients than for white patients.



Going Back to Our Definition ...

"A ML Failure occurs when a ML system does something **unanticipated** and/or **undesirable**."

Undesirable to whom?

- Designers?
- Users?
- Shareholders?

Unanticipated by whom?

- Designers?
- Users?
- Shareholders?

ML bias is a sociotechnical problem.

Technical approaches are **remediations** – not solutions.

We are learning how to **understand** bias, and putting some **tools in our toolbox** to make bias less harmful.

02

Why is it a big deal?

There are **protected attributes** that algorithms impede upon.

Isn’t the point of ML to discriminate?

Yes, to a point. In ML bias, we are interested in *preventing* discrimination based on “socially salient qualities that have **served as the basis for unjustified and systematically adverse treatment** in the past.”
(Hardt, 2017)

Publicly Regulated Domains

There's legal precedent in the US that protects people from discrimination in various activities.

- Credit (Equal Credit Opportunity Act)
- Education (Civil Rights Act of 1964; Education Amendments of 1972)
- Employment (Civil Rights Act of 1964)
- Housing (Fair Housing Act)
- 'Public Accommodation' (Civil Rights Act of 1964)
- Race; Color; Religion; National origin (Civil Rights Act of 1964)
- Citizenship; Age; Sex; Familial status; Pregnancy; Disability status; Veteran status; Genetic Information

Example: Canada's AI Policies

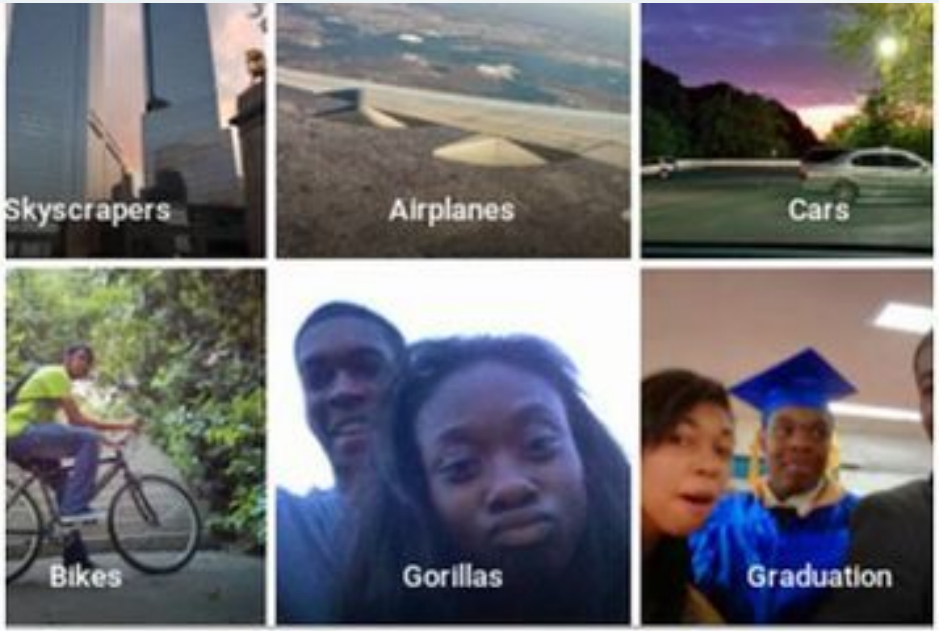
Different countries look at discrimination differently. Canada treats AI differently!

Deep Dive into the Civil Rights Act

There's legal precedent for evaluating whether treatment is biased in US labor law. We'll see later how these ideas apply in machine learning context!

- Disparate treatment
  - Formal: Explicitly considering class membership, even if it's relevant
  - Informal: Purposefully attempting to discriminate without direct reference to class membership
- Disparate impact
  - Four-fifths rule: Selection rate for a certain group is less than 80% of rate for group with highest selection rate, there is adverse impact.
  - Alternative practice: Could the same goal be achieved using a different procedure that would result in a smaller disparity?
    - Alternative practice

Beyond the law: bias in images



Beyond the law: bias in images

Fairness is a contested concept

We don’t just combat ML failures with the law. People are combating bias through many different means.

- Worker organizing
  - Google, Alphabet Workers Union
- Design & business decisions
  - Office of Responsible AI @ Microsoft
- Moral imperatives
- Equity, inclusion, belonging
- How can we do better?

03

How does it happen?

Bias can occur at **all points in the cycle** by various different stakeholders.

Types of Errors

Imbalanced Data Sets

→ Amazon's hiring model
→ Google's CV model

Removing Sensitive Features

→ If a credit score model 'removes' race, the model will learn racial discrimination

Lurking Bias in Features

→ Facial recognition based on white faces
→ Credit score & race

Poorly Framed Problems

→ Predicting if someone is a criminal based on their face

Errors Lead to …

… Wrongful Convictions

Break out groups

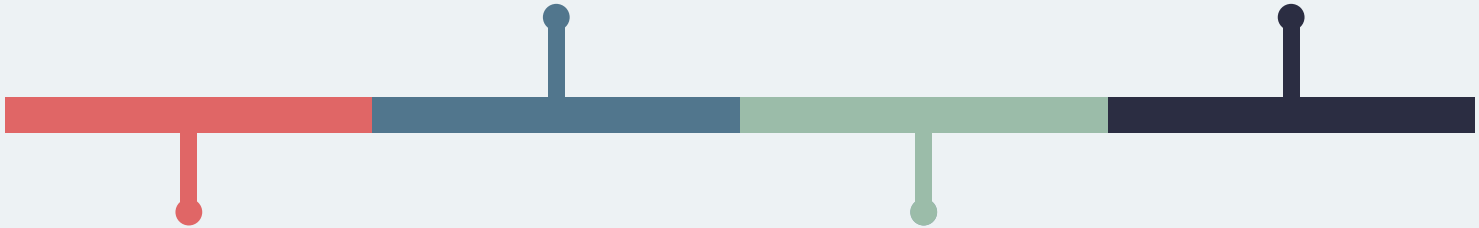
Designers aren’t trained to look.

Users are too trusting of systems.

Watchdogs and protection agencies are overloaded.

Fairness is itself a contested concept.

Bootcamp Timeline



Lab 1

Bias in Healthcare

Lab 2

Gender Bias in Hiring

Lecture 1

Introducing Bias

Lecture 2

Mitigating Bias

RESEARCH ARTICLE

ECONOMICS

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]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

There is growing concern that algorithms may reproduce racial and gender disparities via the people building them or through the data used to train them (*1*–*3*). Empirical work is increasingly lending support to these concerns. For example, job search ads for highly paid positions are less likely to be presented to women (*4*), searches for distinctively Black-sounding names are more likely to trigger ads for arrest records (*5*), and image searches for professions such as CEO produce fewer images of women (*6*). Facial recognition systems increasingly used in law enforcement perform worse on recognizing faces of women and Black individuals (*7*, *8*), and natural language processing algorithms encode language in gendered ways (*9*).

Empirical investigations of algorithmic bias, though, have been hindered by a key constraint: Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them. Instead, researchers must work "from the outside," often with great ingenuity, and resort to clever workarounds such as audit studies. Such efforts can document disparities, but understanding how and why they arise—much less figuring out what to do about them—is difficult without researcher-created algorithms (*10*–*13*). Without an algorithm's training data, objective function, and prediction methodology, we can only guess as to the actual mechanisms for the important algorithmic disparities that arise.

In this study, we exploit a rich dataset that provides insight into a live, scaled algorithm deployed nationwide today. It is one of the largest and most typical examples of a class of commercial risk-prediction tools that, by industry estimates, are applied to roughly 200 million people in the United States each year. Large health systems and payers rely on this algorithm to target patients for "high-risk care management" programs. These programs seek to improve the care of patients with complex health needs by providing additional resources, including greater attention from trained providers, to help ensure that care is well coordinated. Most health systems use these programs as the cornerstone of population health management efforts, and they are widely considered effective at improving outcomes and satisfaction while reducing costs (*14*–*17*). Because the programs are themselves expensive—with costs going toward teams of dedicated nurses, extra primary care appointment slots, and other scarce resources—health that rely on past data to build a predictor of future health care needs.

Our dataset describes one such typical algorithm. It contains both the algorithm's predictions as well as the data needed to understand its inner workings: that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise. It should be emphasized that this algorithm is not unique. Rather, it is emblematic of a generalized approach to risk prediction in the health sector, widely adopted by a range of for- and non-profit medical centers and governmental agencies (*21*).

Our analysis has implications beyond what we learn about this particular algorithm. First, the specific problem solved by this algorithm has analogies in many other sectors: The predicted risk of some future outcome (in our case, health care needs) is widely used to target policy interventions under the assumption that the treatment effect is monotonic in that risk, and the methods used to build the algorithm are standard. Mechanisms of bias uncovered in this study likely operate elsewhere. Second, even beyond our particular finding, we hope that this exercise illustrates the importance, and the large opportunity, of studying algorithmic bias in health care, not just as a model system but also in its own right. By any standard—e.g., number of lives affected, life-and-death consequences of the decision—health is one of the most important and widespread social sectors in which algorithms are already used at scale today, unbeknownst to many.

Data and analytic strategy

Working with a large academic hospital, we identified all primary care patients enrolled in risk-based contracts from 2013 to 2015. Our primary interest was in studying differences between White and Black patients. We formed race categories by using hospital records, which are based on patient self-reporting. Any patient who identified as Black was considered to be Black for the purpose of this analysis. Of the

Background

Healthcare providers use medical **"risk scores"** to prioritize those who need care urgently.

Risk scoring in hospitals:

- Algorithm in this lab is pervasive
  - Applied to ~200MM people in the US every year
- Produced by private companies
  - Aims to lower cost, improve patient outcomes

What features did the algorithm use?

- Demographics (age, sex)
- Insurance type
- Diagnosis & procedure codes
- Medication
- **Medical costs**

Algorithm specifically excludes race.
(RED FLAG!)

Medical cost acts as a proxy of need of but also access to care.

Thus, it has become a **proxy of systemic anti-Black racism** in healthcare, since race is excluded.

Initial Discussion Questions

Who designed the risk scoring algorithm?	• Private company • Evaluated by hospital managers
Who uses the risk scoring algorithm?	• Nurses, medical workers on the ground.
What is cost of errors across stakeholders?	• Spending too much on a healthy patient? • Spending too little on a sick patient? • Err on side of spending too little or too much? (Doctors? Patients? Insurers?)

Let’s get started!

https://colab.research.google.com/drive/1yYHoLqbM5in4T801mQ083XGpRqHwtFhm?usp=sharing

Run in browser: File > Save a copy in Drive

Run as Jupyter notebook: File > Download > Save as .ipynb (Requires wget, python3 and pip)