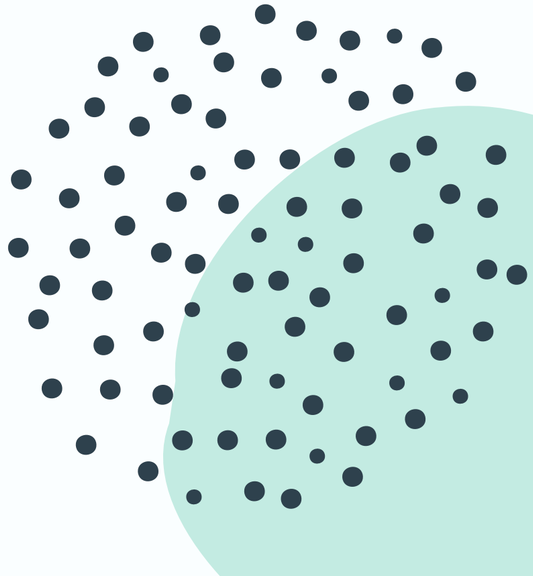
TESTING DISCRIMINATION

Responsible Machine Learning
13/04/2021

Anna Boros, Krzysztof Osesik



Agenda

PRESENTATION STRUCTURE:

1. INTRO: Questions about fairness in the context of ML and decision making systems
2. Testing discrimination in human-decision making systems
3. Testing discrimination in algorithmic systems
4. OUTRO: Discussion

Intro

Presentation based on

Fairness and machine learning

Limitations and Opportunities

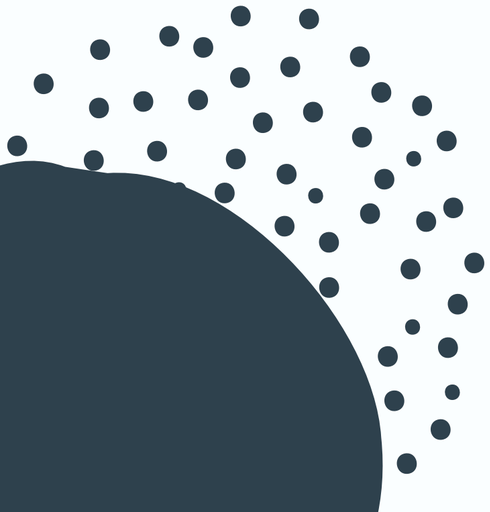
by

Solon Barocas, Moritz Hardt, Arvind Narayanan

Chapter 5 - "Testing discrimination in practice"

available at https://fairmlbook.org/

Testing discrimination in human-decision making systems



Testing discrimination

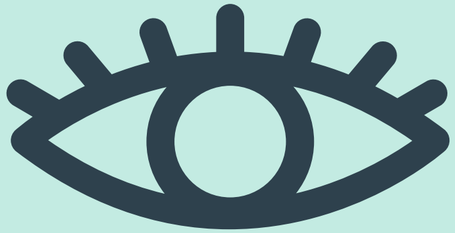
1. Conducting audit studies
2. Testing the impact of blinding
3. Revealing extraneous factors in decisions
4. Conducting purely observational tests
5. Outcome-based testing
6. Overcoming 'selective labels' issue

THE AUDIT STUDY

Field experiment

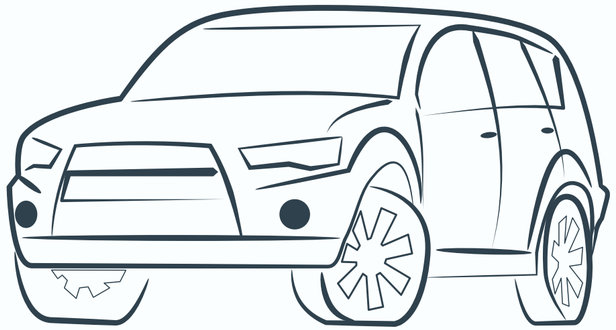
as opposed to the lab simulation it actually happens on a "living" decision-making system, ideally allows for establishing the treatment effect and to control/manipulate variables

ATTEMPT TO TEST BLINDNESS FOR SENSITIVE ATTRIBUTES EXAMPLE: RACE



experiment with 150 dealerships

38 testers visiting in pairs (difference in terms of race/gender), usually within few days of each other



minimizing differences

- 28-32 years old,
- had 3-4 years of postsecondary education,
- “were subjectively chosen to have average attractiveness”
- every aspect of their verbal or nonverbal behavior was governed by a script
- memorize responses to a long list of questions they were likely to encounter
- wore similar ‘yuppie’ sportswear and drove to the dealership in similar rented cars

Findings

- Bargaining people would get final offers $ 1 100 higher if they were black (on average)
- Results hold valid when taking into account initial offers, percantage mark-up, bargaining strategy, noise related to being in certain pair

- **What does it mean to be the same in terms of everything but race?**
- **Is accent always a component of racial identity?**
- **Is controlling for many attributes (minimizing differences) the only valuable approach?**

Complexity of defining and testing blindness: the CV case

'By creating pairs of resumes that were identical except for the name (it was suppoused to signal race), they found that white names were 50% more likely to result in a callback than black names.'

- **Preference for more common names?**
- **Inference of higher socioeconomic status of candidates with certain names?**
- **Can results be applied to the real-world issues?**

BLINDING

'natural experiment'

being able to observe difference between the cases where some sensitive characteristic is known or not, e.g. difference-in-differences approach - happens by either observing the existing difference in decision-making procedures or establishing such procedural difference

USING BLINDING EXAMPLE:
ARTICLES SUBMISSIONS

double-blinded vs single-blinded reviews

- some journals assing reviewer at random but also conceal identity of researcher
- what is the rate of accepted female-authored manuscripts?

- **Why change to double-blinded reviews happened in journals? New editors? New policies encouraging female authors?**
- **Spill-over effects?**
- **Serial correlation of outcomes?**

- **No ontological troubles.**

EXTRANEOUS FACTORS IN DECISIONS

arbitrariness

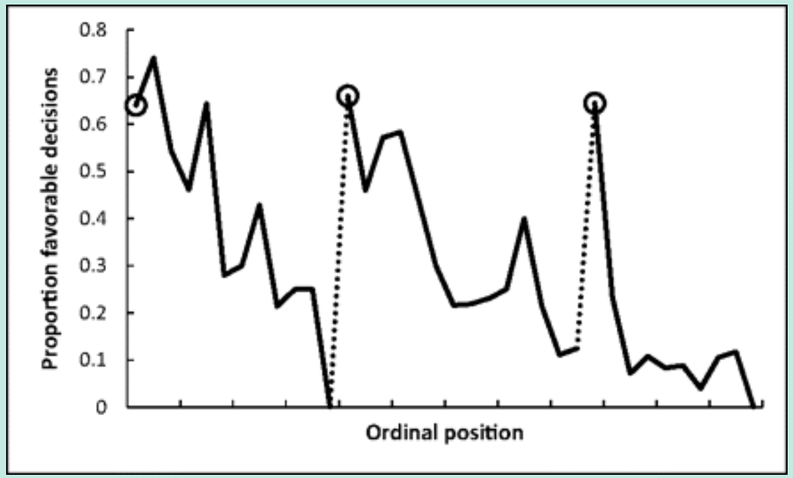
decisions an be influenced by factors that are unrelated to the issue at hand and characteristics of people whom these decisions concern/affect

EXTRANEOUS FACTORS
EXAMPLE: JUVENILE COURT

judging after the lost game

- in Louisiana juvenile courts the week following a loss suffered by the Louisiana State University football team, judges imposed 7% longer sentences on average
- The effect was driven entirely by judges who got their undergraduate degrees at LSU

- **'hungry judges'?:**

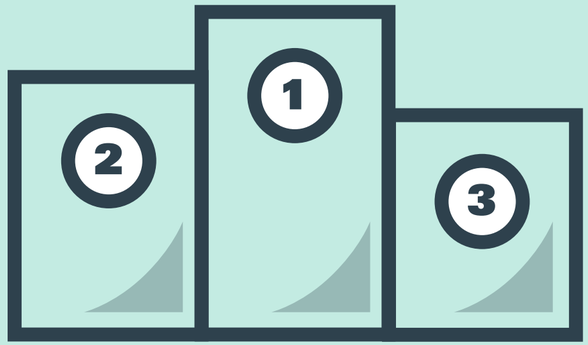


- **prisoners without an attorney are presented last within each session, and tend to prevail at a much lower rate !**

IMPACT OF DECISIONS AND INTERVENTIONS

decisions' impact on outcomes

it can be easily inferred from a bank example - setting higher interest rate for people who are more likely to default makes the default even more likely (self-confirmation of criterions)



REGRESSION DISCONTINUITY

scholarships' impact

- looking for differences in earnings between students close to the cut-off point who were/ were not awarded a scholarship

- **Is GPA the only criterion when awarding scholarships?**
- **Can students take tests multiple times? Is it costly?**
- **What are the reasons for students resigninations from scholarships? Maybe better College could be attended?**
- **Is there a lot of paperwork?**

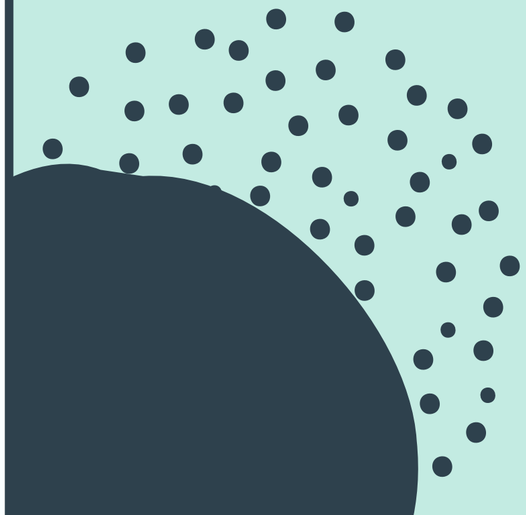
PURELY OBSERVATIONAL TESTS

regression analysis

it is checked whether attributes other than the sensitive one can collectively explain the decision - if not, one can say that decision-maker used the sensitive attribute

however...

It must be recognized that with sufficiently large dataset, the sensitive attribute could be reconstructed using other attributes.



- **Controlling for potentially causal variables (conditional parity)**
- **Normative considerations**

OUTCOME-BASED TESTS

sufficiency, predictive parity

sufficiency: the true outcome is independent of the sensitive attribute given the prediction

predictive parity **(outcome test)**: rate of positive (negative) outcomes for each group (defined by sensitive attribute) is the same

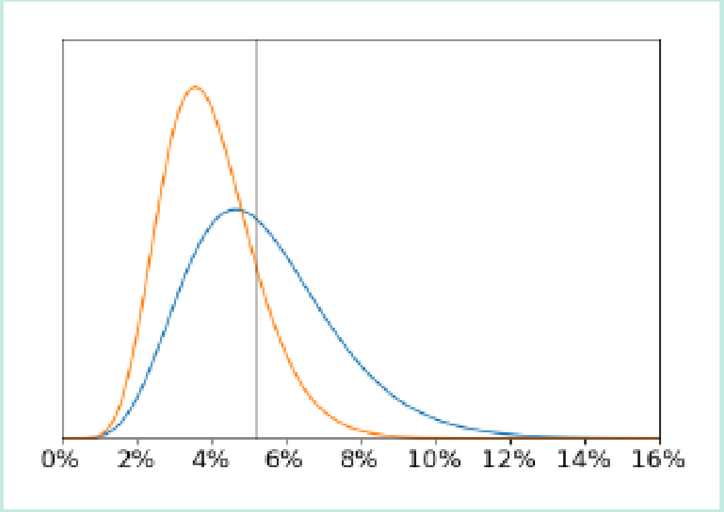
Can we say that predictive parity implies sufficiency?

OUTCOME-BASED TESTS

infra-marginality

the measurement is aggregated over samples that are far from the decision threshold

resolution: narrow attention to samples close to the threshold with a set of characteristics X' that need not coincide with X observed by decision-maker



threshold test

ERROR RATE PARITY

Separation and selective labels

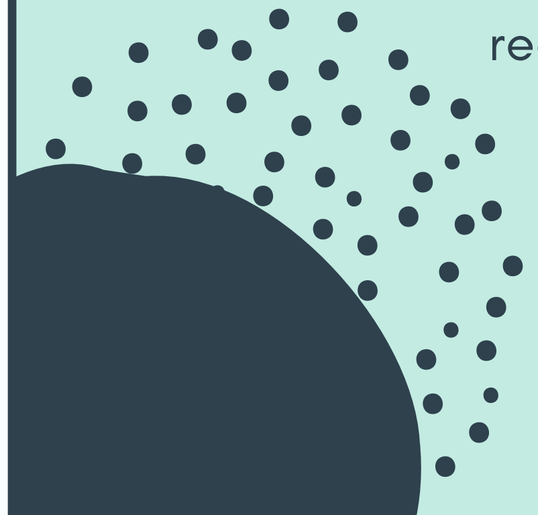
***should be considered especially with human judgement, not predicting future as such**

separation: the prediction and sensitive attribute are independent given the true outcome

...but 'selective labels'

labels can only be observed for the 'positive' decisions (applicants who received a loan -we will not know if others would default)

SOLTUION EXAMPLE: POSITIVE LABELS TO SOME NEGATIVE PREDICTIONS



taste-based

irrational prejudice for a group, suboptimal decisions

statistical

distribution of target variable varies by group, optimal decisions

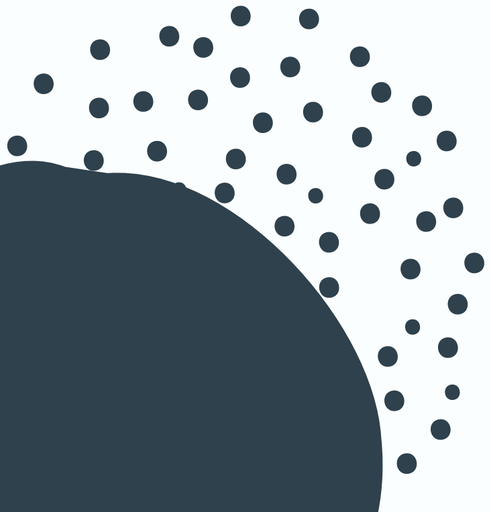
Distinction is not so important as it may be hard. Understanding the cause of discrimination and its source (e.g. in an institution) in a domain-specific and mechanism-oriented way is vital.

Undearstanding discrimination

Qualitative methods:

Ethnographic studies (*Pedigree:* leissure activities, *Inside Graduate Admissions*: cool hobby)

Testing discrimination in algorithmic systems



Background

- First cases of discrimination on a massive scale were identified already in 1950s in the United States.
- A matching algorithm took preferences of applicants for medical residency programs as input and produces an assignment of applicants to hospitals that optimizes mutual desirability.
- The algorithm discriminated against couples wished to stay together (they could not state their preferred hospital in absolute terms)
- **Solution** : "leading member" (potentially discriminative ?)

Discrimination in algoritms - cases

- Natural language processing
- Computer vision systems
- Recommendation systems
- Ad targeting
- Online marketplaces

Natural language processing

- Natural language processing algoritms are prone for discrimination
- 2016 study found that they had substantially more false negatives for tweets written in African-American English (AAE)
- 13.2 % of AAE tweets were classified as non-English compared to 7.6 % of “white-aligned” English tweets.
- Source of discrimination: **underrepresentation** in the training data.

Bias in NPL systems

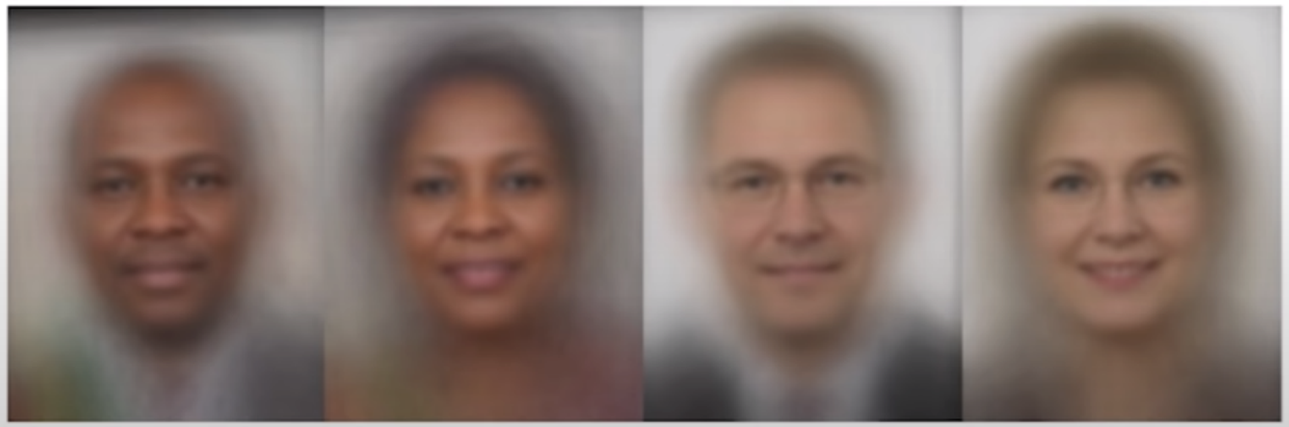
- nurse - "her patient"
- doctor - "his patient"

- Source: MIT 6.S191: AI Bias and Fairness, https://www.youtube.com/watch?v=wmyVODy_WD8

Computer vision

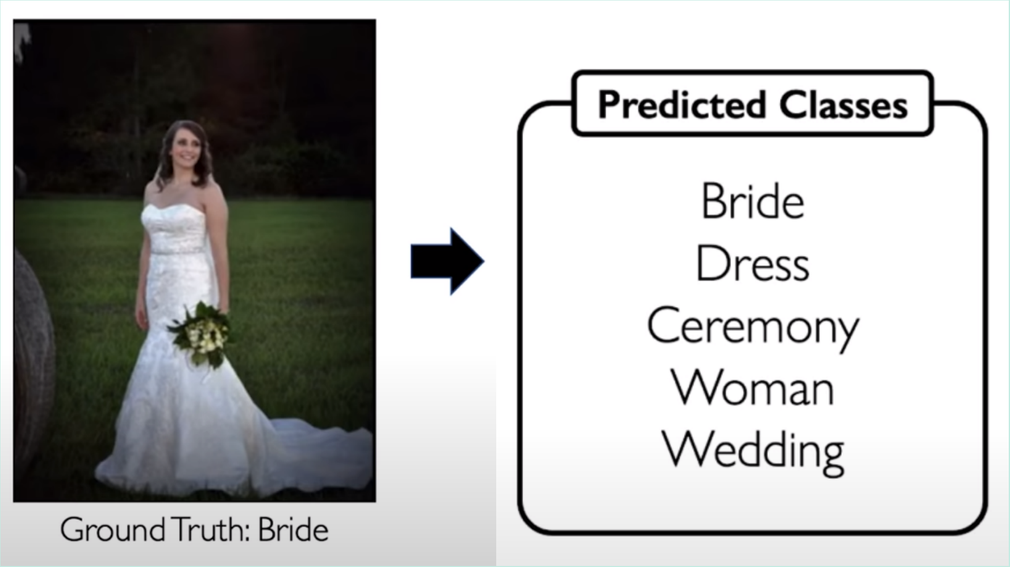
- Computer vision systems often exhibit disparities in performance based on gender, race, skin tone, and other attributes
- First, they perform better for men than for women
- Secondly, they perform better for lighter faces than for darker faces

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%



- Source: MIT 6.S191: AI Bias and Fairness, https://www.youtube.com/watch?v=wmyVODy_WD8

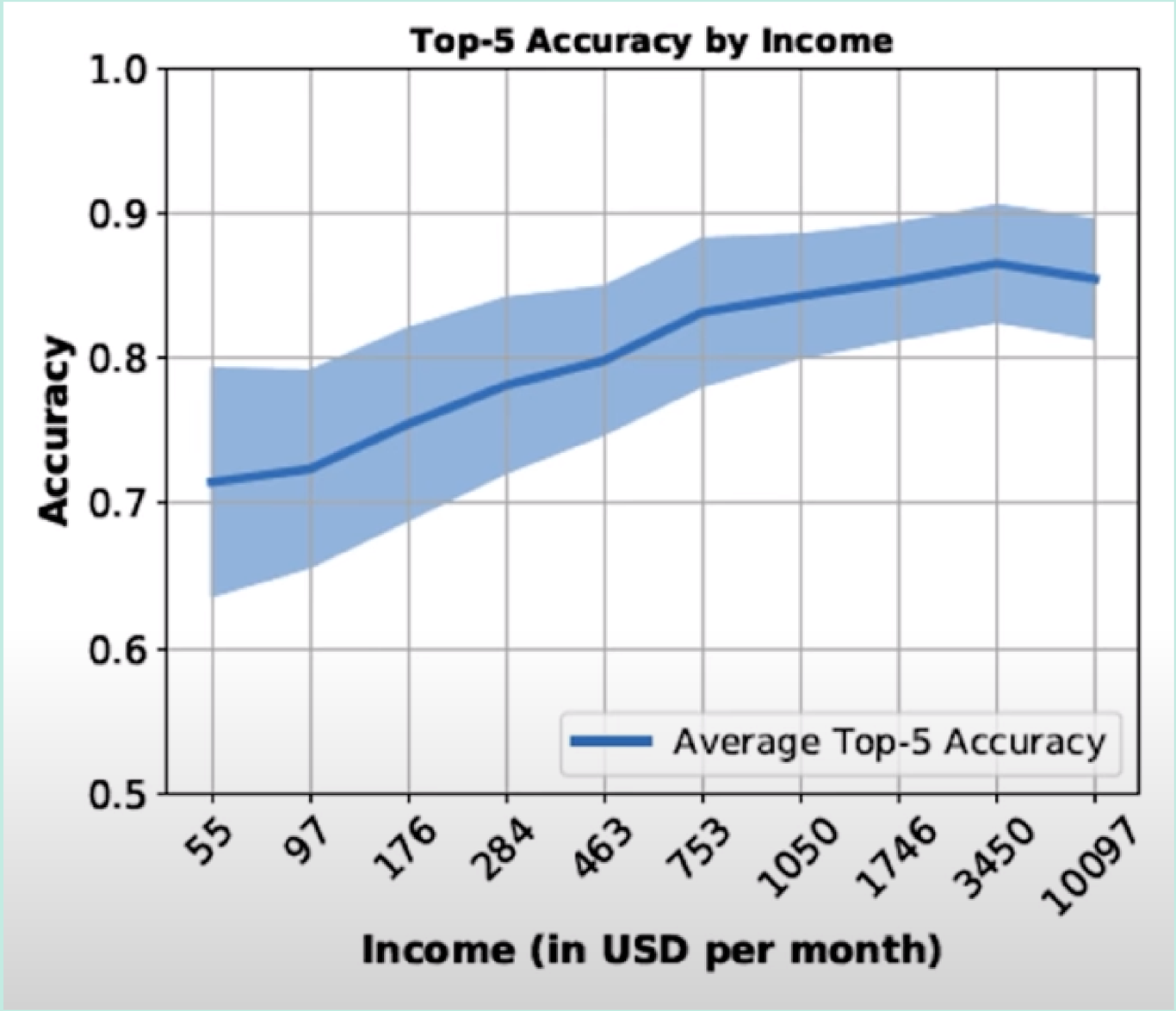
Bias in image classification



- Source: MIT 6.S191: AI Bias and Fairness, https://www.youtube.com/watch?v=wmyVODy_WD8

UNDERREPRESENTATION PROBLEM

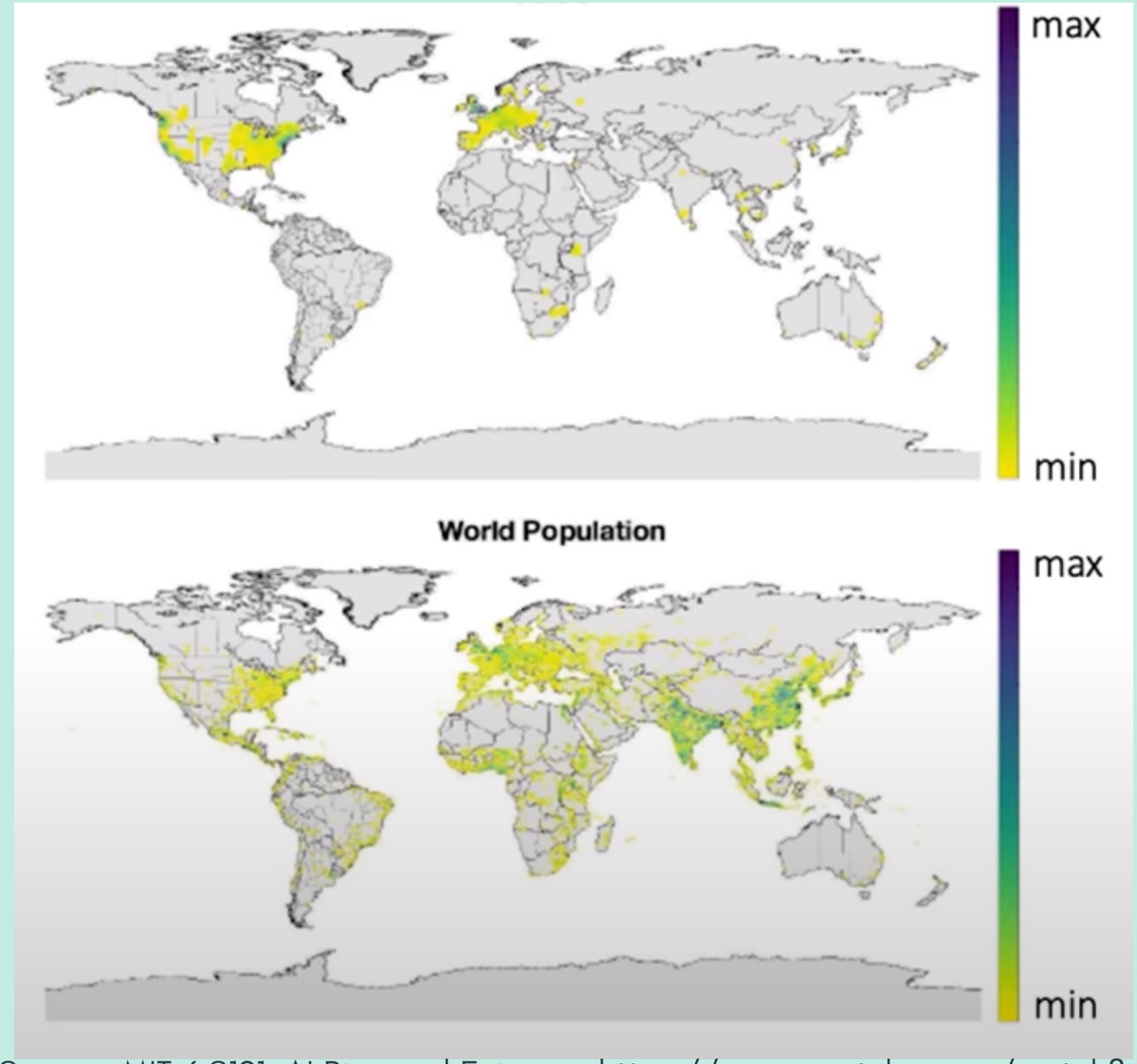
Bias correlation with Income



- Source: MIT 6.S191: AI Bias and Fairness, https://www.youtube.com/watch?v=wmyVODy_WD8

UNDERREPRESENTATION PROBLEM

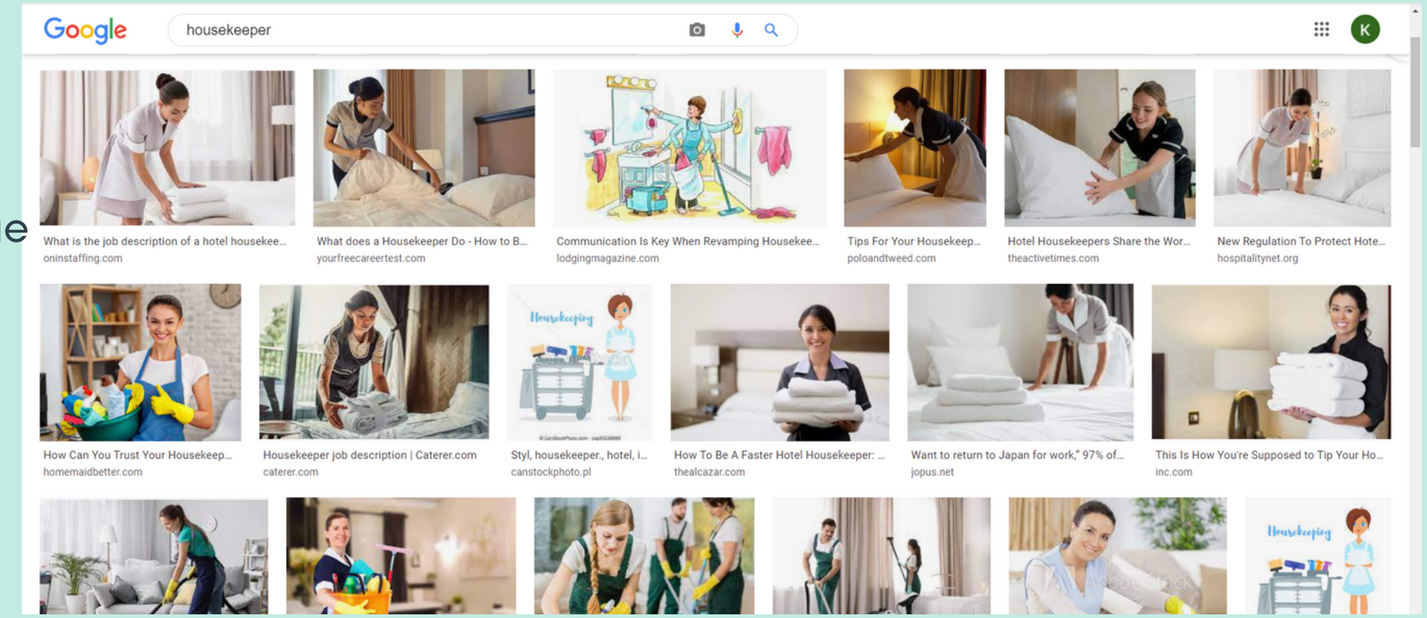
Bias correlation with Geography



- Source: MIT 6.S191: AI Bias and Fairness, https://www.youtube.com/watch?v=wmyVODy_WD8

Recommendation systems

- "users like me who liked this item also liked....." (discrimination towards minority groups - underrepresentation)
- Amplification of cultural stereotypes
  - e.g. search for "housekeeper" in google
    - do the search results present the true ratio of occupations by gender ?
    - even, if they do, does not it amplify cultural stereotypes ?
  - reinforcment loop



Ad targeting

- Might be discriminative, becasue :
  - firms target specified groups of people based on their ZIP codes
  - firms select potential clinets based on their past behavior (click rates, visited pages, etc.) and other characetristics (e.g. age, gender)
- In a study, Google showed different employment related ads to men and women
- In addition, social media (e.g. Facebook) allow advertisers to learn **more information** it has inferred or purchased about a user than it will allow the user himself to access (e.g. Cambridge Analytica)

Online marketplaces

- Algorithms are not key in online marketplaces, product characteristics are more important
- However, buy and sell decisions in many instances are finally made by a human and they might be

  prone to discrimination practices (knowingly or unknowingly)
- Algorithms, systems and social media might facilitate and amplify discriminative actions by people
  - "Would you buy a car from somebody whose Facebooks posts you do not like ?"
  - "Would you rent a flat to somebody from a another country / ethnic minority ?"
- **Potential solution**: withhold sensitive information on counterparties (e.g. Airbnb)

Research methods

- Real and fake users / accounts
  - real users offer more variability and reliability, but are less controled
  - fake users are easier to implmenet, but sometimes are blocked by algorithms
- Lab studies of programming codes
  - algorithms are reproduced in a lab, but researchers usually do not have full dataset

Thank you
&
Discussion