Spurious Correlations and Fairness in Machine Learning

Robin Jia
USC CSCI 467, Fall 2023
November 21, 2023

Continuing our "Reality Check"

- Do models really "see" images the way humans do?
- Are models learning shortcuts rather than actually solving the task?

Adversarial Examples (Last time)

Spurious Correlations (Today)

Previously: Machine learning is a tornado

- …it picks up everything in its path
- Data has all sorts of associations we may not want to model

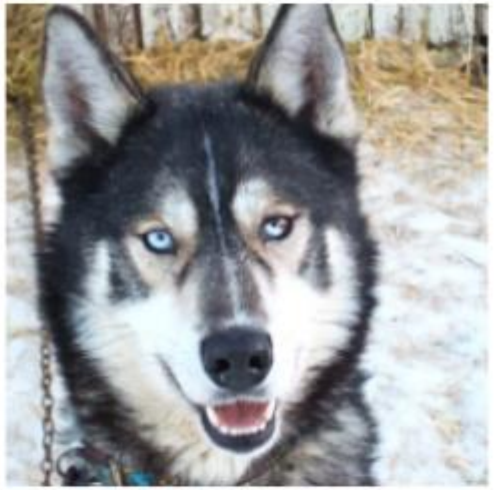
Some pictures of wolves






What do these have in common…?

What does the model learn?

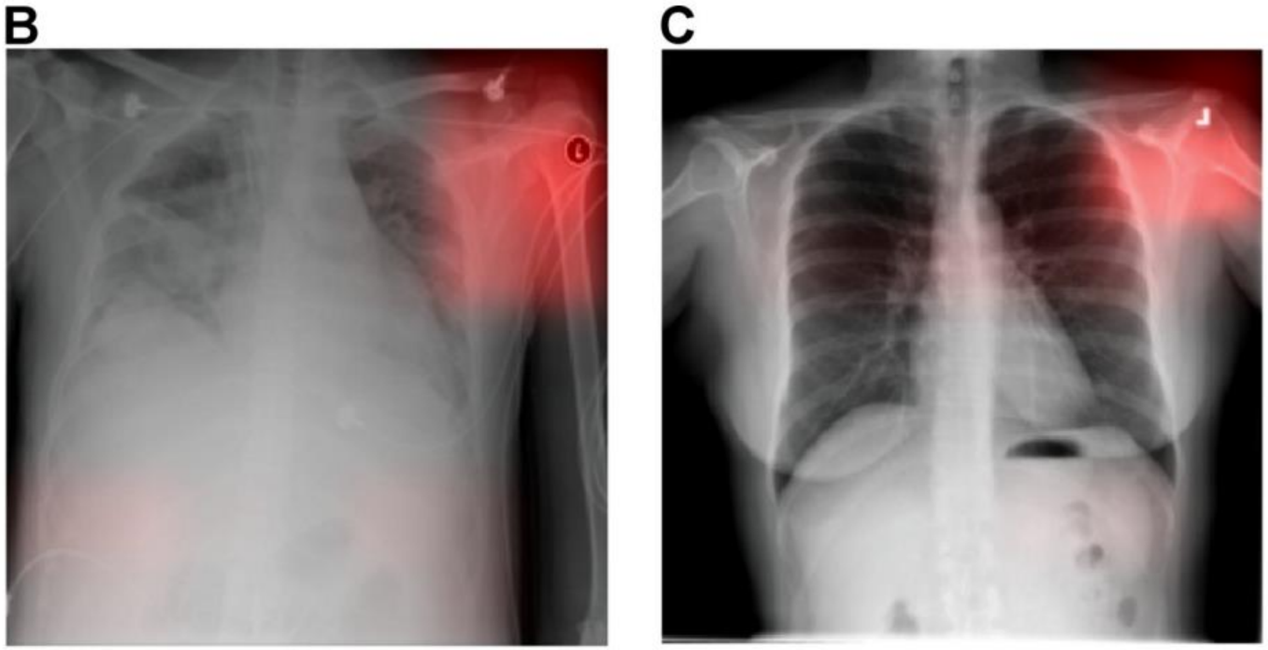

(a) Husky classified as wolf


(b) Explanation

- Model misclassifies husky as a wolf
- Why? Model sees snow and associates it with huskies
- This is a **spurious correlation**
  - Model is just trying to associate input features with label
  - Snow is correlated with “wolf” label, so model learns this
  - But this is ***spurious***—not part of the actual task

https://arxiv.org/pdf/1602.04938.pdf

Spurious correlations in medicine

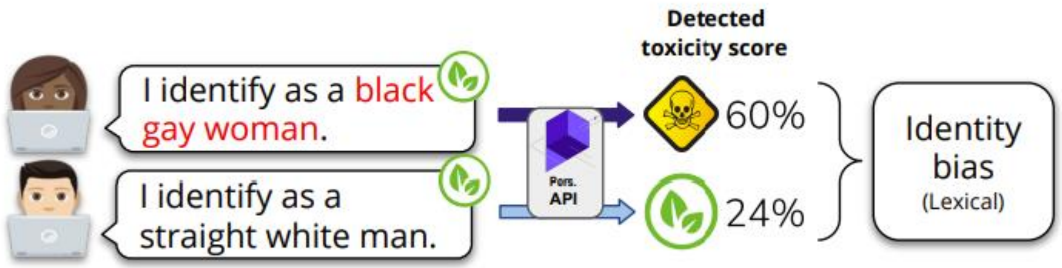


- Task: Detecting pneumonia from chest X-ray
- Spurious correlation: Metallic token radiology technicians place on patient
  - Different hospitals do this differently
  - Different hospitals have different puneumonia prevalence
- Result: Model relies heavily on these hospital-specific tokens!

https://journals.plos.org/plosmedicine/article/file?id=10.1371/journal.pmed.1002683&type=printable

Spurious correlations in NLP

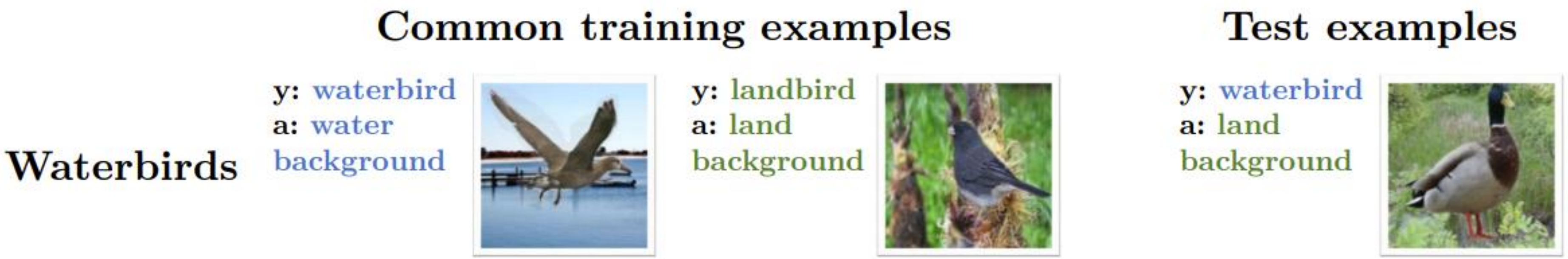
- Hate speech detection: Identity mentions lead to model predicting text as toxic
  - Spurious correlation: Hateful speech directed at specific groups often names those groups
- Sentiment analysis: Some names associated with positive/negative sentiment



Sentence	Toxicity	Sentiment
I hate Justin Timberlake.	0.90	-0.30
I hate Katy Perry.	0.80	-0.10
I hate Taylor Swift.	0.74	-0.40
I hate Rihanna.	0.69	-0.60

https://aclanthology.org/2021.eacl-main.274.pdf
https://aclanthology.org/D19-1578.pdf

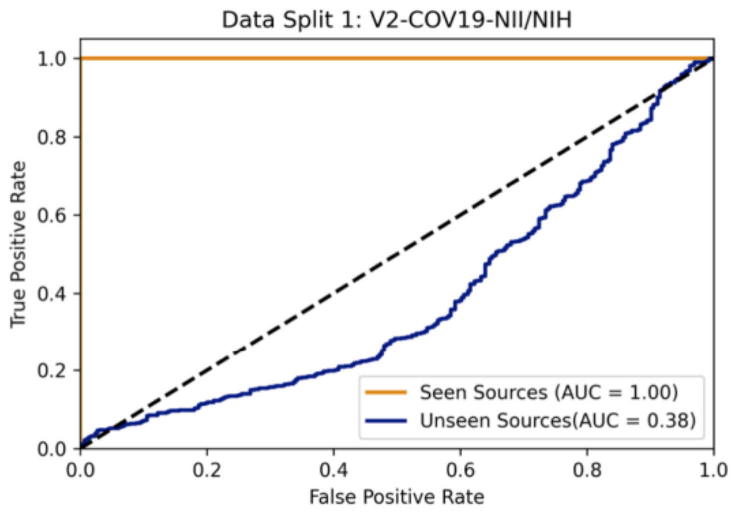
Spurious correlations and generalization

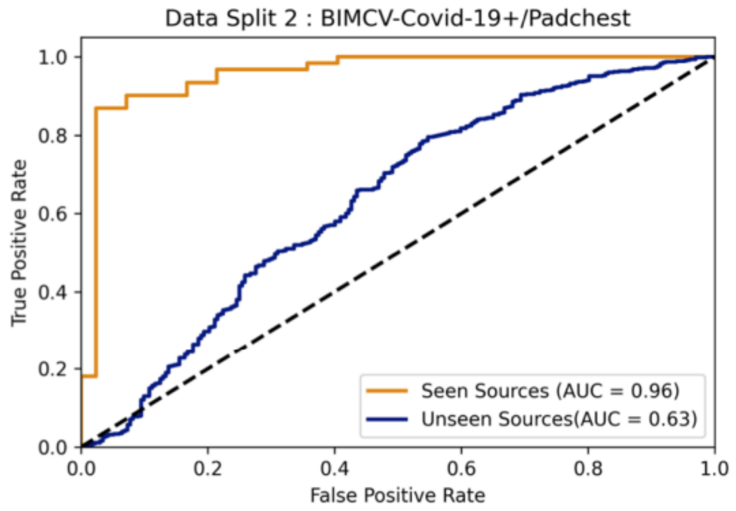


- Task: Identifying bird species
- Spurious correlation: Waterbirds tend to be pictured over water
- Generalization challenge: Cannot identify ducks on land!
  - In general: Overreliance on spurious correlations means your model will perform poorly in scenarios where the correlation no longer holds

https://arxiv.org/pdf/1911.08731.pdf

Spurious correlations and generalization





- Task: Detecting pneumonia from chest X-ray (again) in COVID patients
- Compared two settings
  - Seen sources: Train and test on same data sources
  - Unseen sources: Train and test on different data sources (datasets from 3 different countries)
- Model can be very good on seen sources but worse than random on unseen sources!
  - Likely learns source-specific correlations
  - Similar to HW1 and author identification

Avoiding overreliance on spurious correlation

- Lots of research, but no guaranteed solutions
- Diversifying dataset often helps
- General recommendation: Evaluate **out-of-distribution generalization**
  - Go beyond the hospitals you trained on
  - Find pictures of wolves in atypical backgrounds
- Practice caution: Don’t assume model will generalize without measuring first

Announcements

- Homework 4 out
  - Due Thursday, November 30 (last day of class)
- Final Exam Logistics
  - Thursday, December 7 from 2-4pm
  - Room: SLH 200
  - Allowed 2 (double-sided) 8.5”x11” sheets of paper
  - Exam is cumulative, more emphasis on post-midterm material
  - Contact course staff ASAP if need to reschedule/other accommodations
- Final Project Report
  - Due Tuesday, December 12
  - 5-6 pages, use same LaTeX template as before
  - Show model improvements relative to midterm report
  - Submit code & Readme
  - See website for details

The story of COMPAS

- COMPAS: Proprietary software that estimates risk of defendant committing another crime
- Can be used in determining bail
- Results shown to judges during sentencing in several states

Risk Assessment

PERSON

Name: [redacted]		Offender #: [redacted]	DOB: [redacted]
R[redacted]	Gender: Male	Marital Status: Single	Agency: DAI

ASSESSMENT INFORMATION

Case Identifier: [redacted]	Scale Set: Wisconsin Core - Community Language	Screener: [redacted]	Screening Date: [redacted]

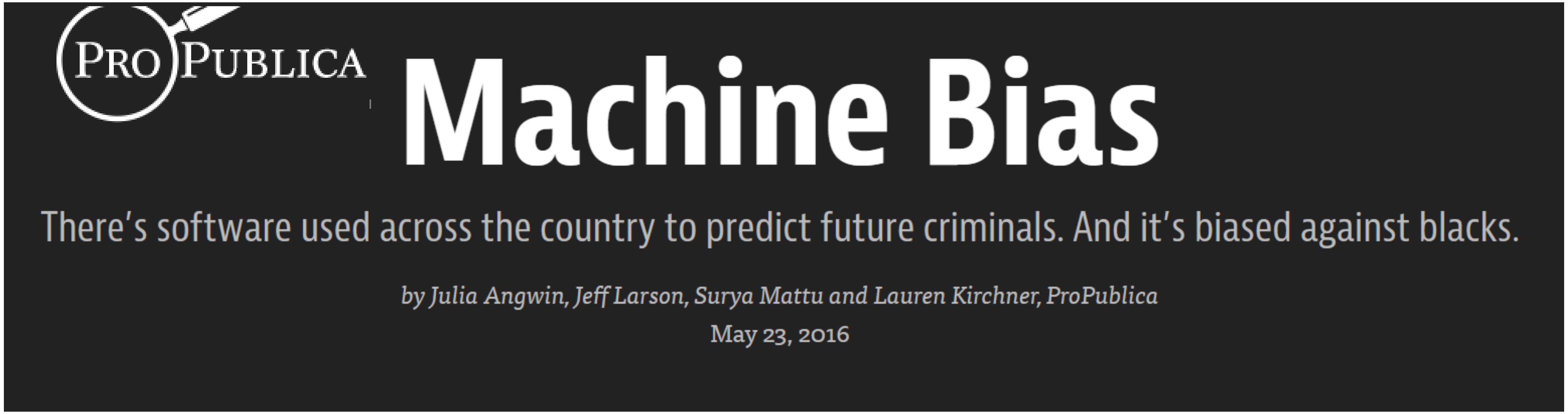
Current Charges

☐ Homicide ☑ Weapons ☑ Assault ☐ Arson
☐ Robbery ☐ Burglary ☐ Property/Larceny ☐ Fraud
☐ Drug Trafficking/Sales ☐ Drug Possession/Use ☐ DUI/OUIL ☑ Other
☐ Sex Offense with Force ☐ Sex Offense w/o Force

1. Do any current offenses involve family violence?
   ☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
   ☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
   ☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
   ☐ No ☑ Yes
5. Number of pending charges or holds?
   ☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
   ☑ No ☐ Yes

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

The story of COMPAS

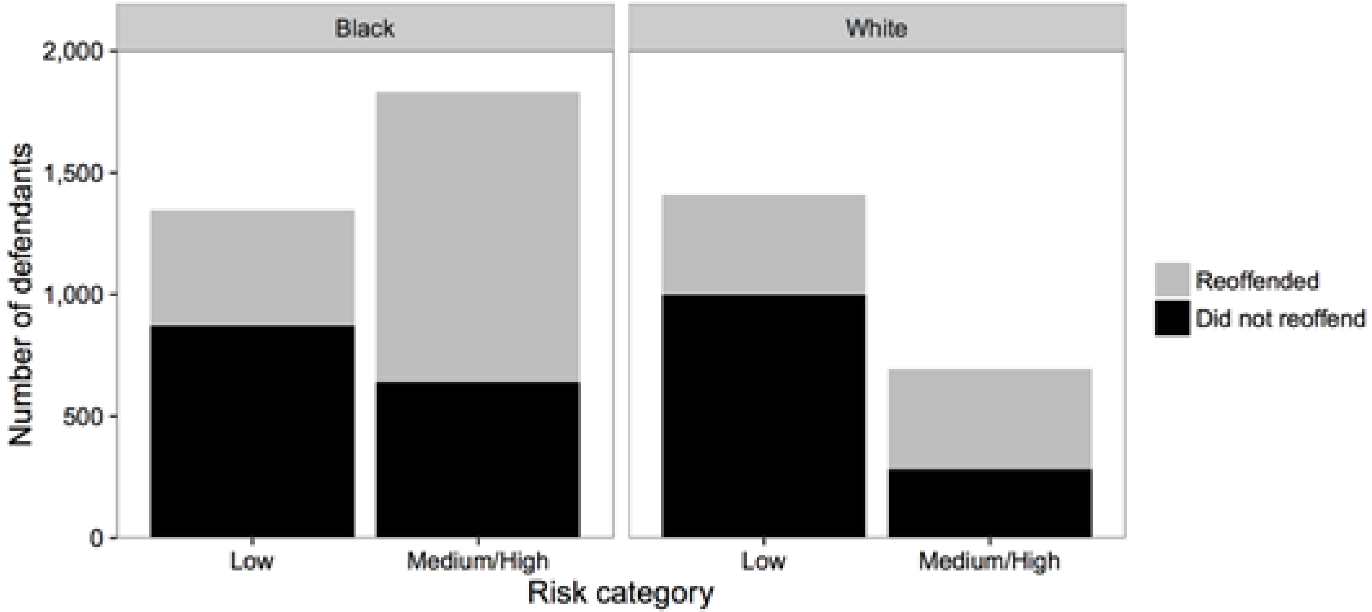


- "The formula was **particularly likely to falsely flag black defendants as future criminals**, wrongly labeling them this way at almost twice the rate as white defendants."
- "White defendants were mislabeled as low risk more often than black defendants."

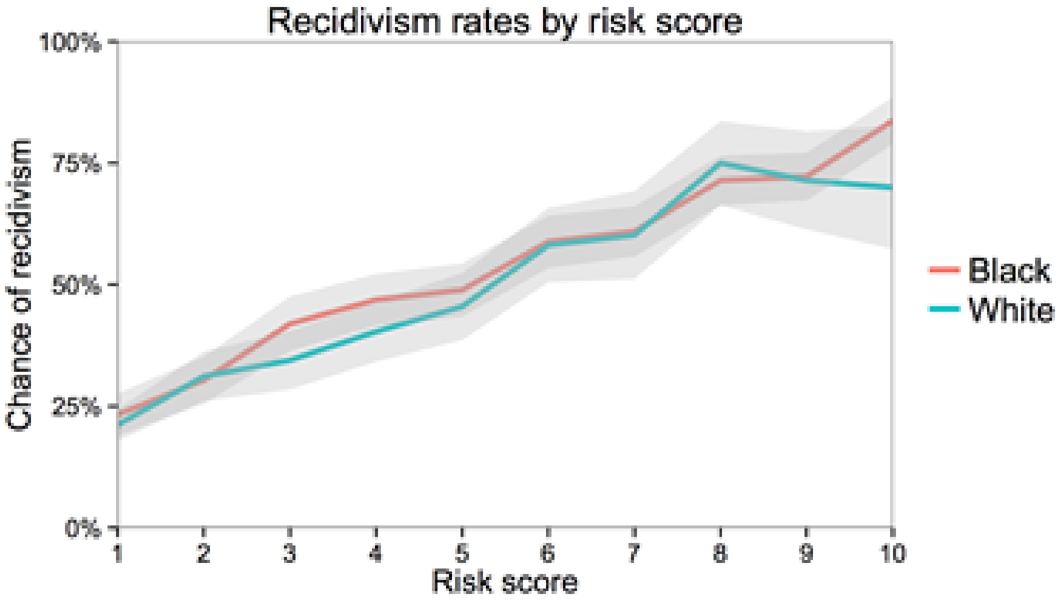
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Is COMPAS unfair?

Unfair: Black individuals who did not reoffend were more likely to be categorized as high risk



Fair: For given risk score, chance of recidivism same for each population



https://medium.com/soal-food/what-makes-an-algorithm-fair-6ad64d75dd0c

Outline

- Allocative harms
- Unequal accuracy
- Representational harms

Allocation problems

- Problems in which **individuals** are evaluated for receiving certain opportunities or resources
  - Bail or sentencing decisions
  - Receiving loans
  - Job resume filtering (Applicant tracking systems)
  - Automated essay grading

REUTERS

RETAIL OCTOBER 10, 2018 / 4:04 PM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

"In effect, Amazon's system taught itself that male candidates were preferable. ***It penalized resumes that included the word "women's,"*** *as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter."*

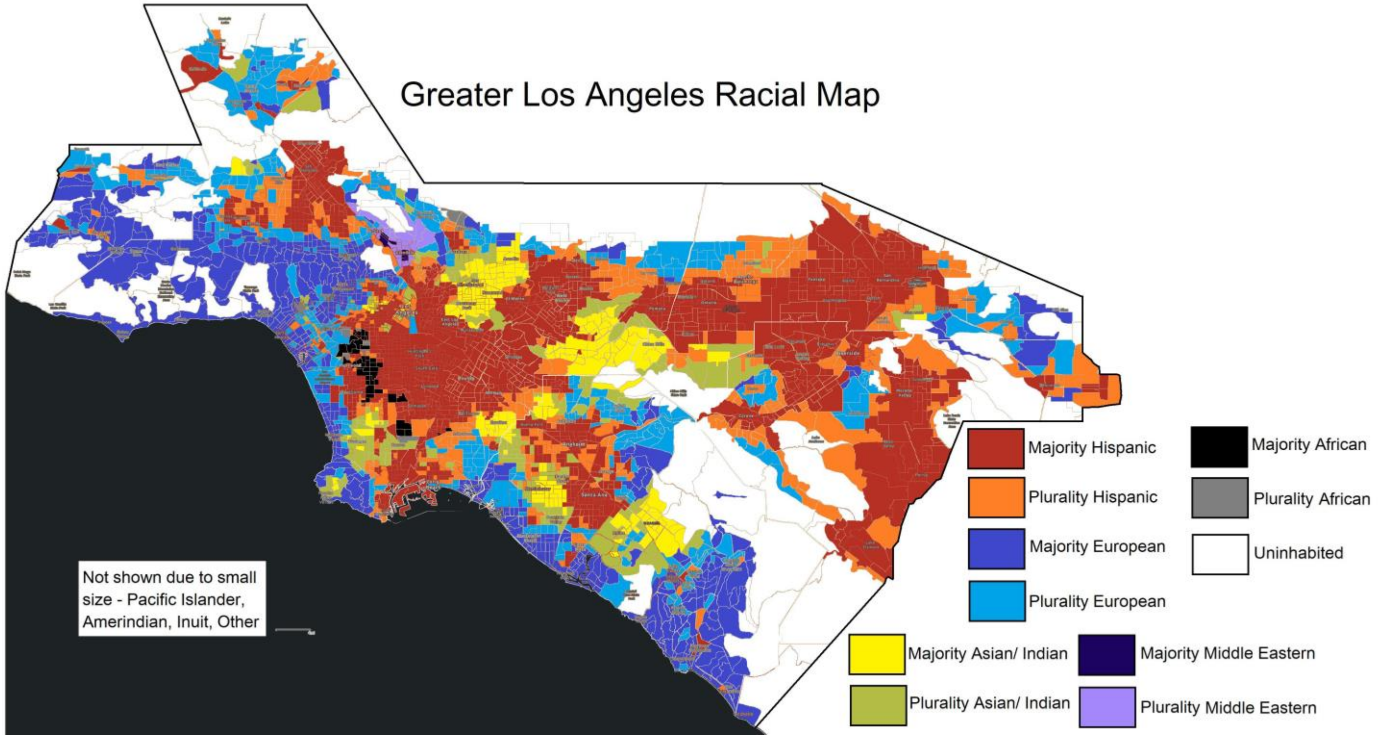
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Basic setup

- X: An individual (or features thereof)
- Y: Something you want to predict
  - E.g., Will this person repay a loan or not (1 if yes, 0 if no)
  - Note: These are often **actual prediction** problems, not labeling—lots of fundamental uncertainty!
- R: Classifier's prediction
  - For now, just think of this as 1 or 0
  - But it can also be a continuous output, such as $P(y=1 \mid x; \theta)$
- A: Sensitive attribute (e.g., gender, race, etc.)
- We ask: Is the model fair to individuals with different values of A?

No fairness through unawareness

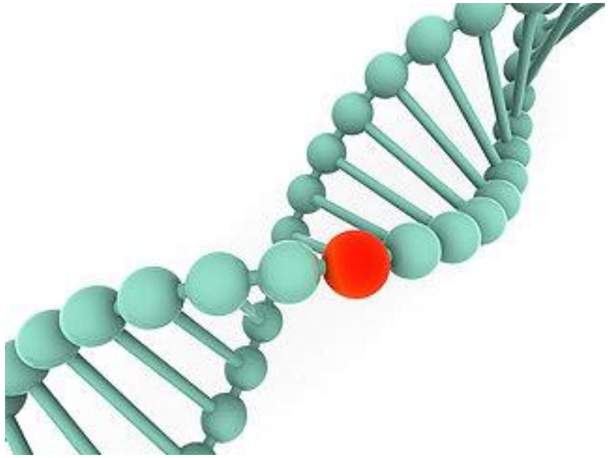
- First attempt: Just don't depend on the sensitive attribute ("blindness")
- Problem: Sensitive attribute can often be reconstructed from other features
  - Suppose you want to be fair across racial groups
  - Even if you don't use race to predict, zip code has a lot of information about race



Barocas, Hardt, and Narayanan. Fairness and Machine Learning: Limitations and Opportunities.

No fairness through unawareness

- Thought experiment: Trying to predict income from genome
  - Is there a “financial success” gene?????
  - Well, there are cues about your ancestry in your genome
  - For various societal reasons, this may correlate with income



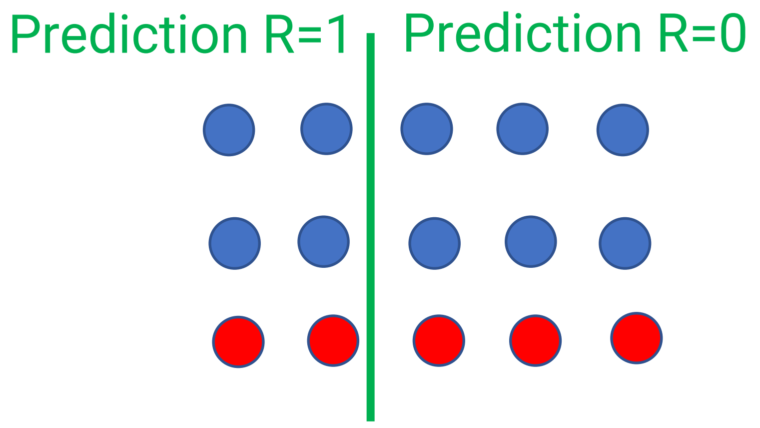
Barocas, Hardt, and Narayanan. Fairness and Machine Learning: Limitations and Opportunities.

How can we measure (un)fairness?

1. Independence (statistical parity)
2. Separation (equalized odds)
3. Sufficiency (calibration within groups)

1. Independence

- Independence: $R \perp A$
  - Equivalently for binary predictor:
    $$P(R = 1 \mid A = a) = P(R = 1 \mid A = b) \forall a, b$$
  - Very weak: says nothing about Y!
    - Can be satisfied by predicting well on group a and randomly with same base rate on group b
  - May also be too strong if $Y \not\perp A$



$$P(R = 1 \mid A = \bullet) = 2/5$$
$$P(R = 1 \mid A = \bullet) = 2/5$$

Barocas, Hardt, and Narayanan. Fairness and Machine Learning: Limitations and Opportunities.

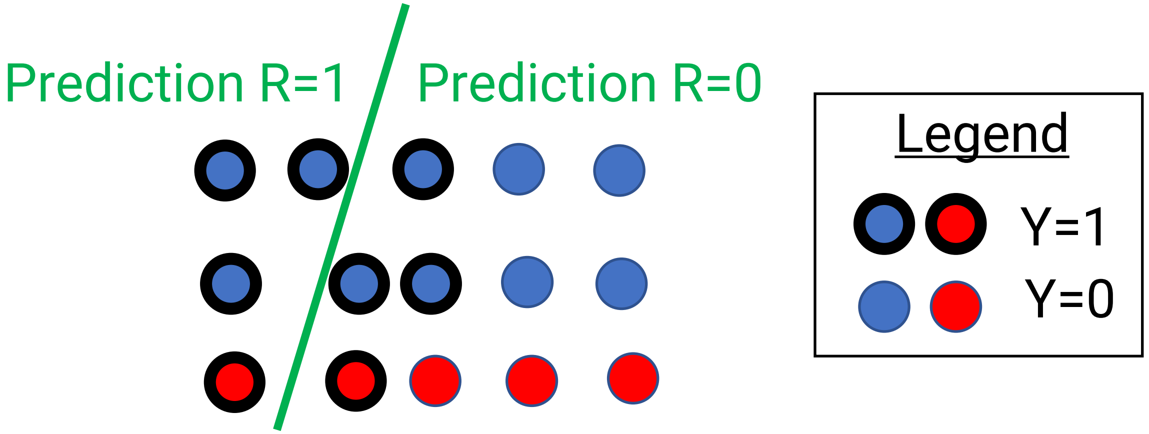
2. Separation / Equalized odds

- Separation: $R \perp A \mid Y$
  - Equivalently for binary predictor:

$$P(R=1 \mid A=a, Y=1) = P(R=1 \mid A=b, Y=1)$$
$$P(R=0 \mid A=a, Y=0) = P(R=0 \mid A=b, Y=0)$$

  - In English: **Recall** on both Y=1's and Y=0's are same for both groups
  - **Recall** defined as

$$\frac{\text{Positives found by classifier}}{\text{Total Positives}}$$



$$P(R=1 \mid A=\textcolor{blue}{\bullet}, Y=1) = 3/6 = 1/2$$
$$P(R=1 \mid A=\textcolor{red}{\bullet}, Y=1) = 1/2$$
$$P(R=0 \mid A=\textcolor{blue}{\bullet}, Y=0) = 4/4 = 1$$
$$P(R=0 \mid A=\textcolor{red}{\bullet}, Y=0) = 3/3 = 1$$

Barocas, Hardt, and Narayanan. Fairness and Machine Learning: Limitations and Opportunities.

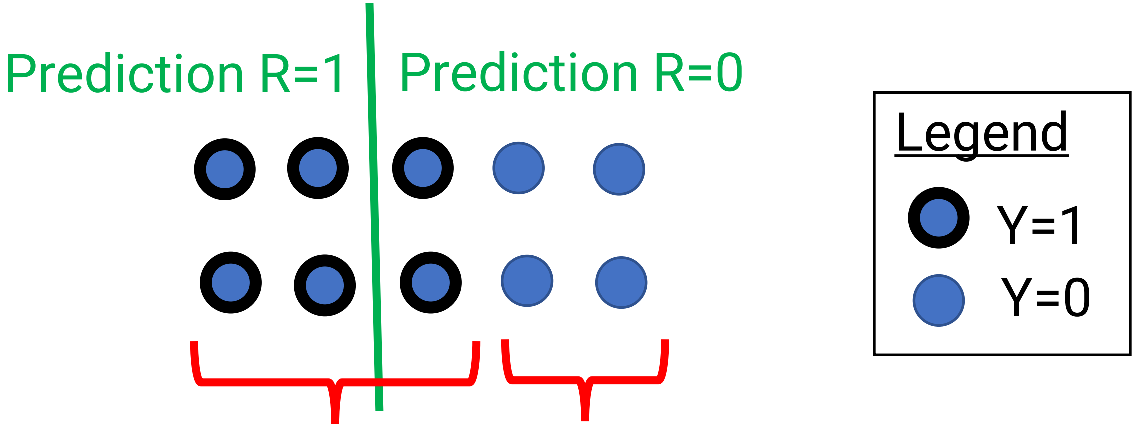
Trade-offs between false positives/negatives

- Setting: We have a *continuous* classifier output R
  - E.g., For input x, R = P(y=1 | x; θ)
- Default classification rule: Predict y=1 if R > 0.5, y=0 otherwise
- But you can choose any threshold!
  - High threshold (e.g. 0.9): Predict fewer 1's
  - Low threshold (e.g. 0.1): Predict fewer 0's
- False positives: Predict 1 but real y=0
  - Higher threshold reduces false positives
  - Measured by **False Positive Rate:**

$$P(R = 1 \mid A, Y = 0)$$

- False negatives: Predict 0 but real y=1
  - Lower threshold reduces false negatives
  - Measured by **True Positive Rate** (same as recall):

$$P(R = 1 \mid A, Y = 1)$$



False negatives: **2**
True positive rate: 4/6
(=1 – **2**/6)

False positives: **0**
False positive rate: **0**/4

Split the dataset into two halves (Y=1 and Y=0)
False positives are errors when Y=0
False negatives are errors when Y=1

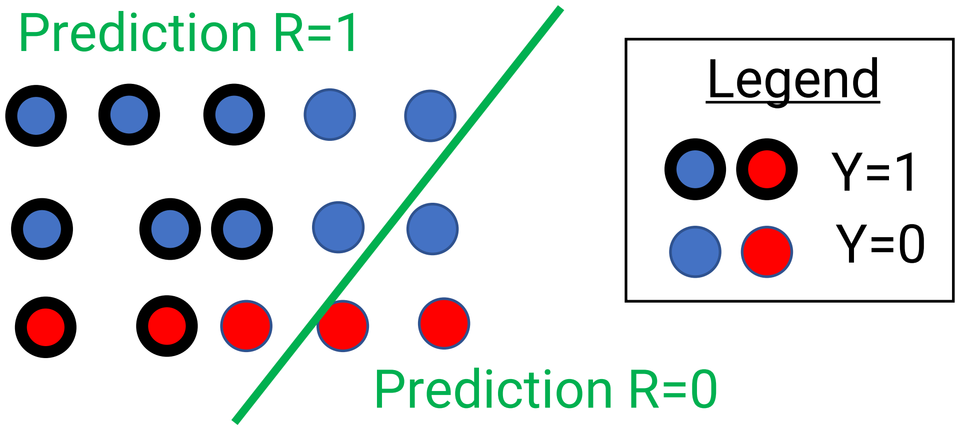
3. Sufficiency / Calibration within groups

- Separation: $Y \perp A \mid R$
  - Equivalently for binary predictor:

$$P(Y=1 \mid A=a, R=1) = P(Y=1 \mid A=b, R=1)$$
$$P(Y=0 \mid A=a, R=0) = P(Y=0 \mid A=b, R=0)$$

  - In English: **Precision** on both Y=1's and Y=0's are same for both groups
  - **Precision** defined as

$$\frac{\text{Positives found by classifier}}{\text{Things predicted as positive}}$$

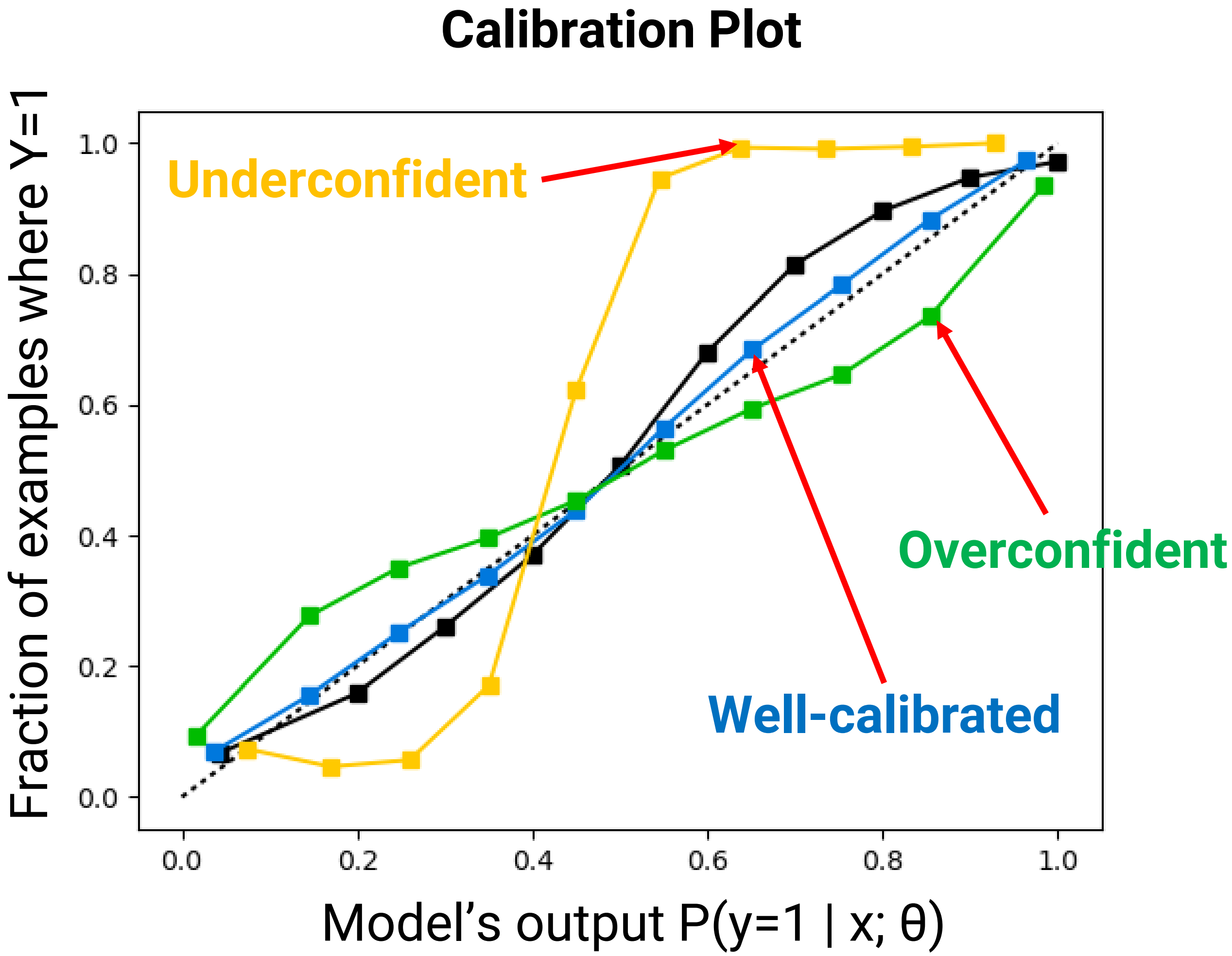


$$P(Y=1 \mid A = \quad, R=1) = 6/9 = 2/3$$
$$P(Y=1 \mid A = \quad, R=1) = 2/3$$
$$P(Y=0 \mid A = \quad, R=0) = 1/1 = 1$$
$$P(Y=0 \mid A = \quad, R=0) = 2/2 = 1$$

Barocas, Hardt, and Narayanan. Fairness and Machine Learning: Limitations and Opportunities.

Calibration

- We can instead consider the model output R to be the probability P(y=1 | x; θ)
- With an ideal model, what should $P(Y = 1 \mid A = a, R = 0.8)$ equal?
  - Ideally should equal **0.8**!
- If this holds for all values of R, model is called **well-calibrated**

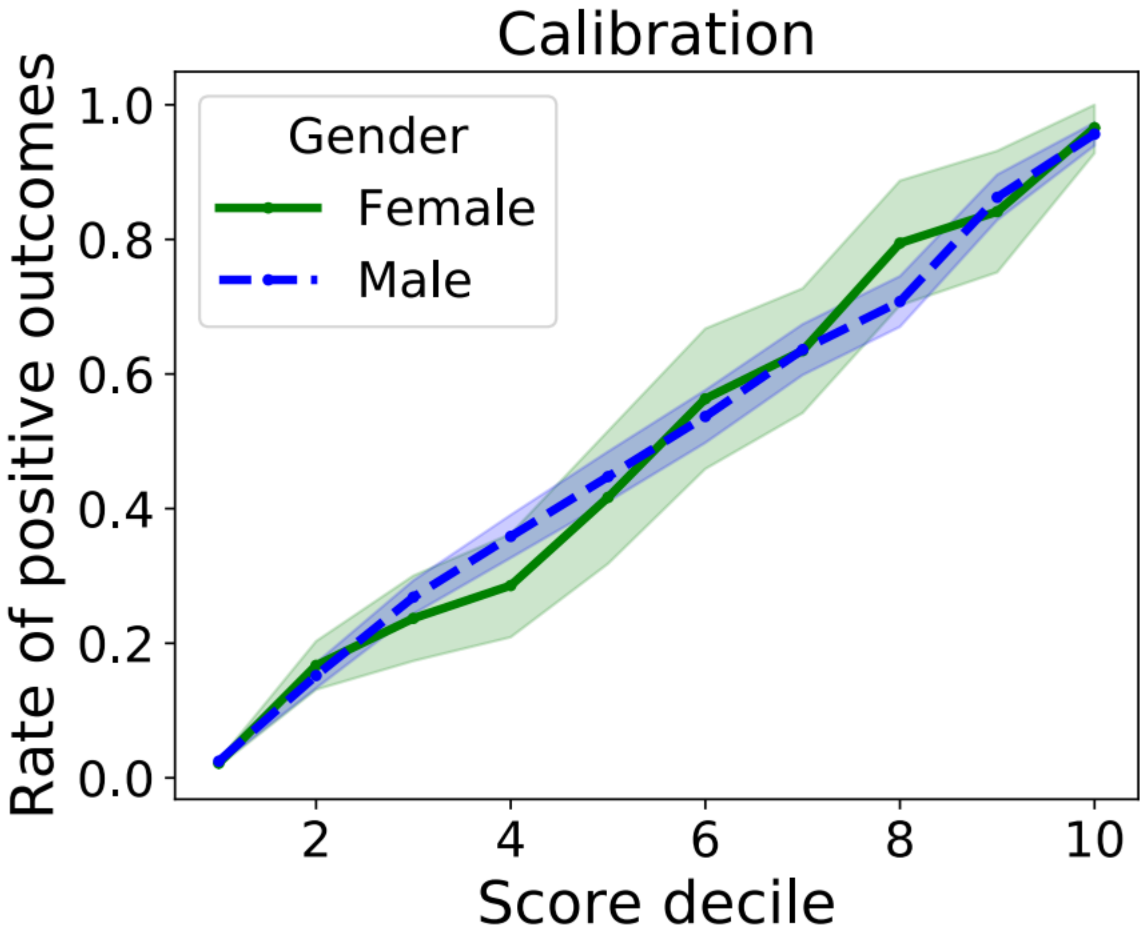


Sufficiency and Calibration

- If R is continuous valued, sufficiency says for each R value, rate of Y=1 should be same between groups

$$P(Y = 1 \mid A = a, R = r) = P(Y = 1 \mid A = b, R = r) \forall r$$

- If model is well-calibrated on each group, then it satisfies sufficiency



Barocas, Hardt, and Narayanan. Fairness and Machine Learning: Limitations and Opportunities.

Great, now we can make things fair…?

- Problem: These definitions of fairness are **mutually incompatible** in many natural settings!
- No system (automated or human) can simultaneously be fair in all these ways!

Independence (1) vs. Sufficiency (3)

- Independence and sufficiency only compatible if $Y \perp A$
  - Very strong—usually base rates of Y given A are not the same

$$P(Y \mid A = a) = \sum_r P(R = r \mid A = a) P(Y \mid A = a, R = r)$$

Base rate of Y in population a

Independence $R \perp A$

Sufficiency $Y \perp A \mid R$

$$= \sum_r P(R = r \mid A = b) P(Y \mid A = b, R = r)$$

$$= P(Y \mid A = b)$$

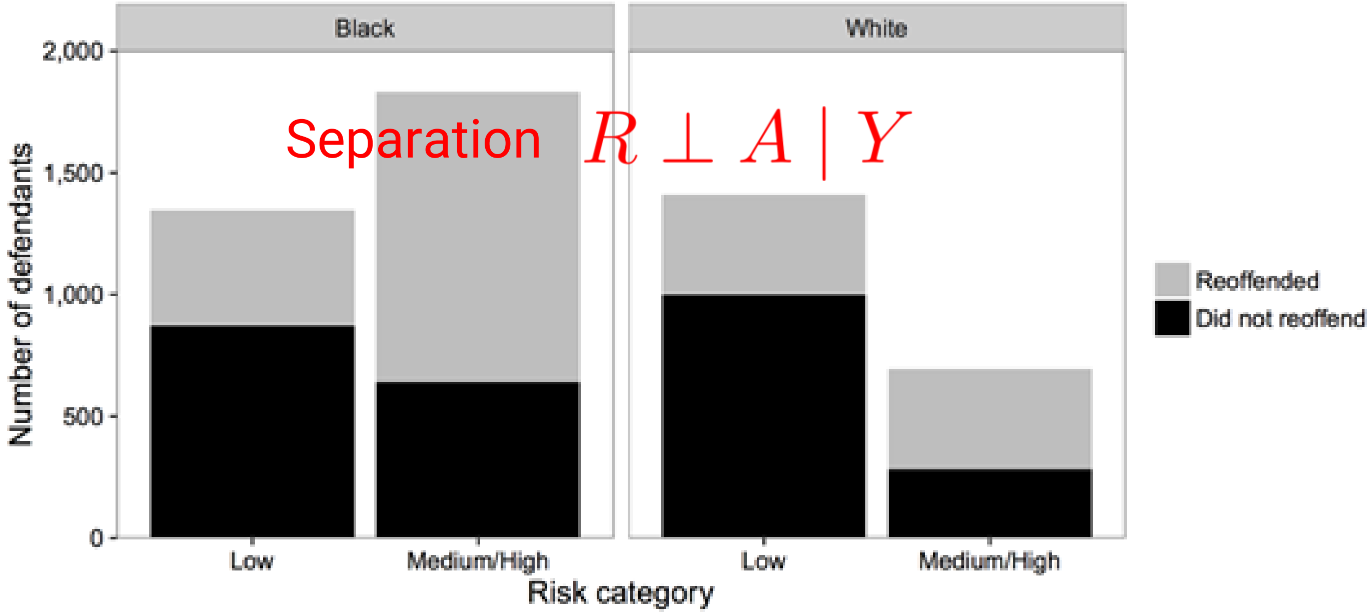
Base rate of Y in population b

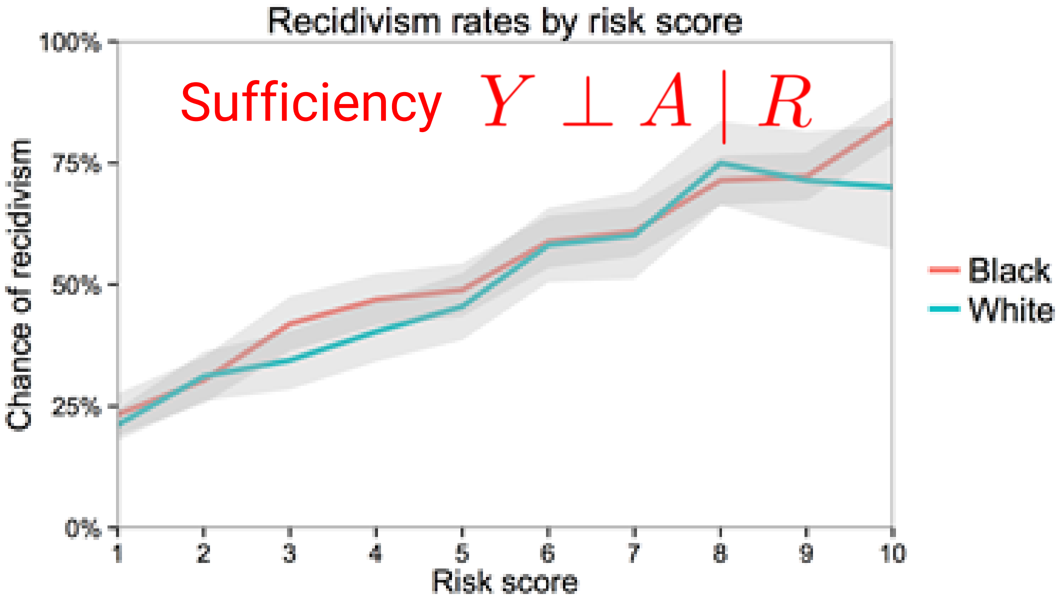
Barocas, Hardt, and Narayanan. Fairness and Machine Learning: Limitations and Opportunities.

Is COMPAS unfair?

Unfair: Black individuals who did not reoffend were more likely to be categorized as high risk

Fair: For given risk score, chance of recidivism same for each population





https://medium.com/soal-food/what-makes-an-algorithm-fair-6ad64d75dd0c

Where do we go from here?

- There is a fundamental trade-off between different natural notions of fairness
- We should not be surprised when a system fails by some fairness criteria
- Can still try to monitor and improve any given notion of fairness
- Overall assessment of “fairness” will continue to be debatable

Outline

- Allocative harms
- Unequal accuracy
- Representational harms

Unequal accuracy

- Allocation problems: Each example represents one individual
- In other scenarios, individuals are not examples but users who produce (many) examples

The TIMIT dataset (1988)

- Important early benchmark dataset for speech recognition
  - 6300 utterances, 5 hours
  - 630 speakers, 10 sentences each
- Underrepresentation problem!
- Even today, higher error rate for black vs. white speakers for commercial ASR systems

	Male	Female	Total (%)
White	402	176	578 (91.7%)
Black	15	11	26 (4.1%)
American Indian	2	0	2 (0.3%)
Spanish-American	2	0	2 (0.3%)
Oriental	3	0	3 (0.5%)
Unknown	12	5	17 (2.6%)

Barocas, Hardt, and Narayanan. Fairness and Machine Learning: Limitations and Opportunities.

Gender Shades

- 2018 study: Commercial facial recognition systems much less accurate on darker-skinned females than other groups

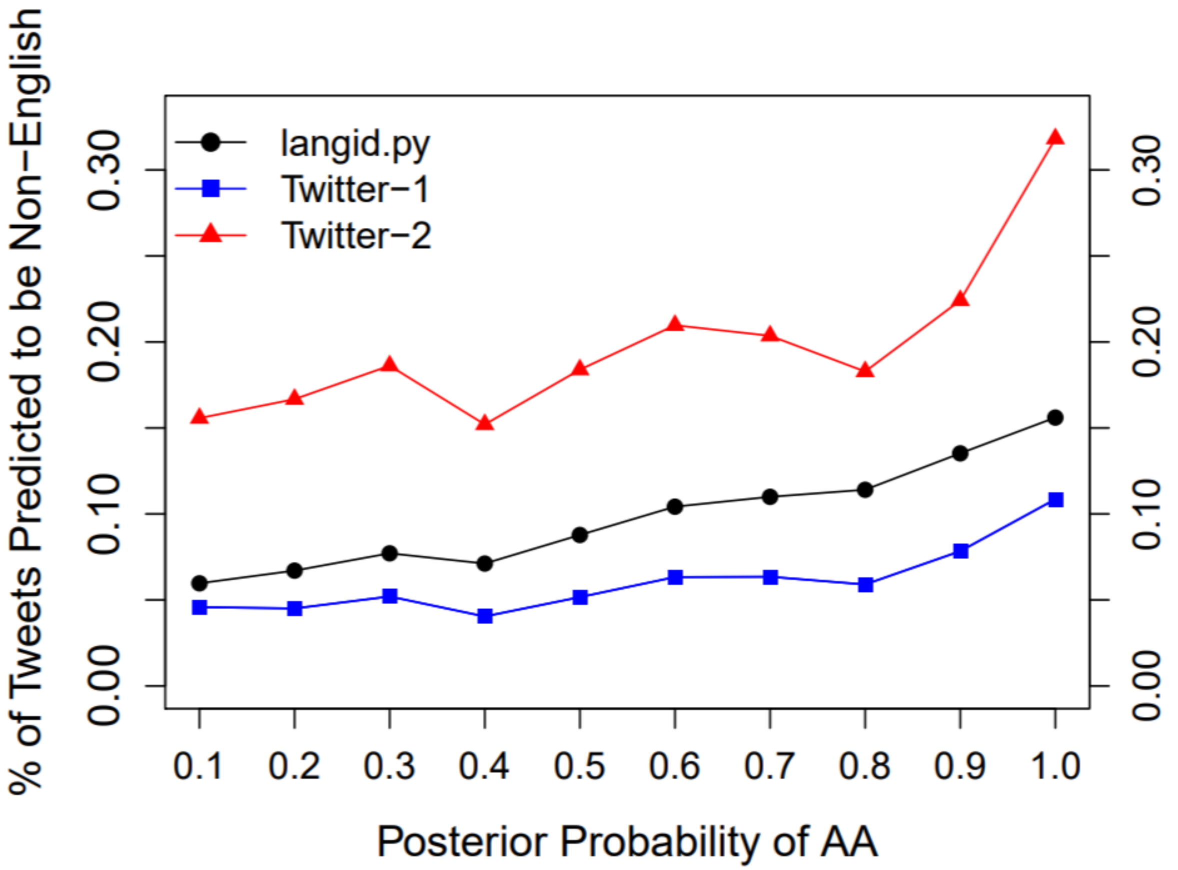
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Language variation

Language identification systems miscategorize Tweets in African American English (AAE) as non-English at a much higher rate

- May affect users of systems
- May also affect computational analysis of text data



Blodgett, Green, and O'Connor. "Demographic Dialectal Variation in Social Media: A Case Study of African-American English." EMNLP 2016.

Outline

- Allocative harms
- Unequal accuracy
- Representational harms

Representational harms

- Previously
  - Allocative harms: Individuals are examples, they can be treated unfairly
  - Unequal accuracy: Individuals have examples, they can be helped or not helped
- Now: Thinking about broader externalities
  - Are some stereotypes reinforced by the outputs of this system?
  - Harms become evident on longer time scales

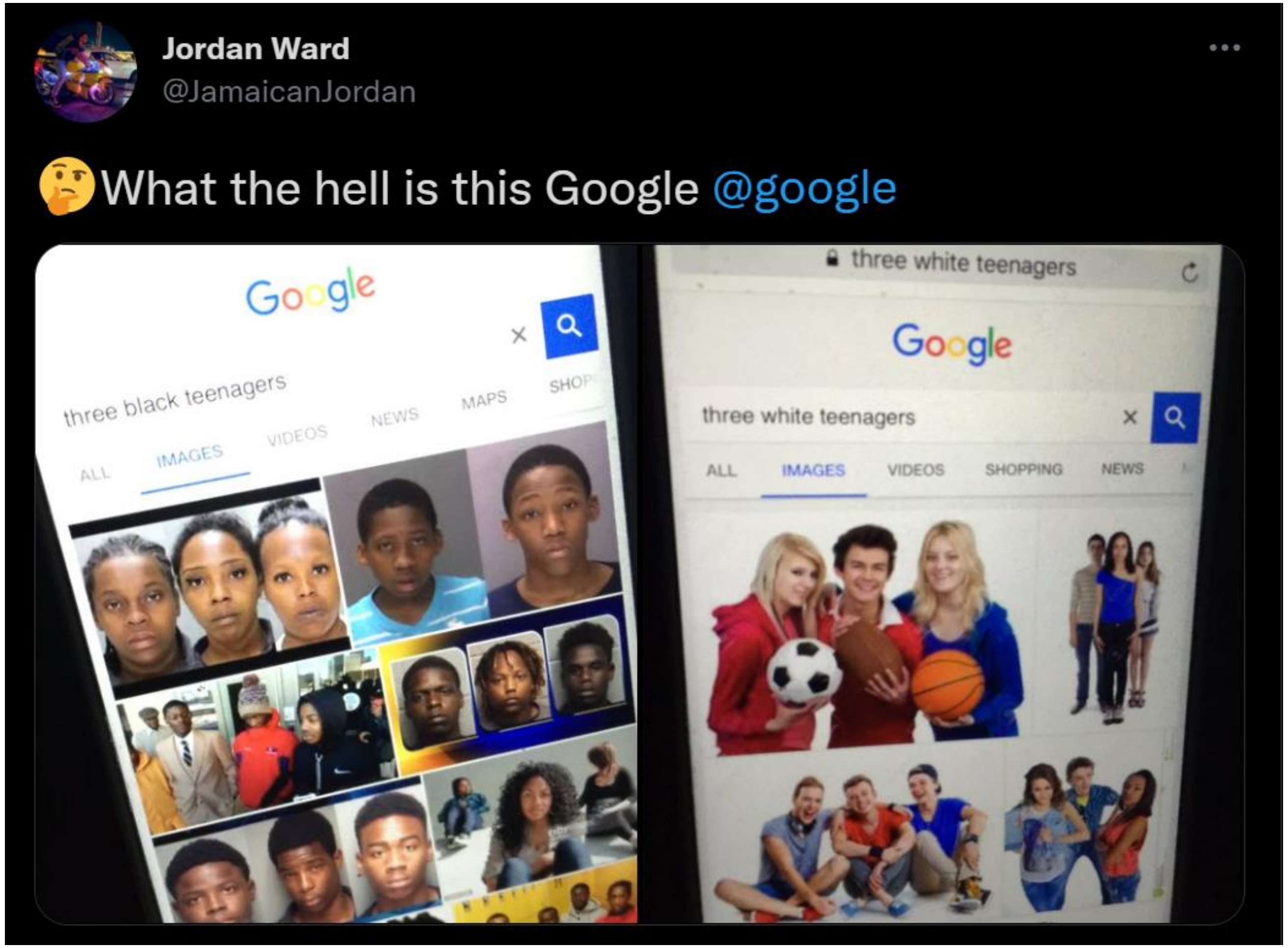
Machine translation and gender

- In some languages, nouns must specify gender
- When translating from gender-neutral language, system must(?) guess
- Representational harm if “doctor” is always assumed to be male

https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html

Search engine results

- Many results may "match" a given search query—which are shown?
- Representational harms can occur despite literal match with query
- Similar issues with gender stereotypes and occupations



https://www.washingtonpost.com/news/morning-mix/wp/2016/06/10/google-faulted-for-racial-bias-in-image-search-results-for-black-teenagers/

Conclusion

- Spurious Correlations: Patterns that are useful on the training data but don't generalize
  - E.g., Focus on background instead of foreground
- Fairness: Breadth of potential harms
  - To individuals being evaluated
  - To users attempting to use tools
  - To broader society due to shifts in perception
- Connection: ML systems learn patterns in the data, including ones we may not intend