Fairness in Machine Learning, Differential Privacy

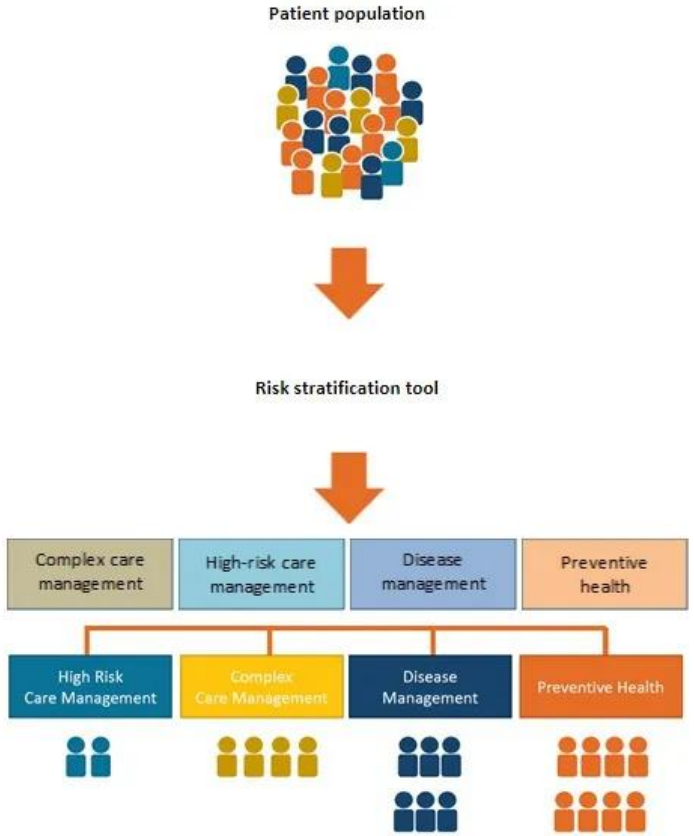
Robin Jia

USC CSCI 467, Fall 2026

November 23, 2026

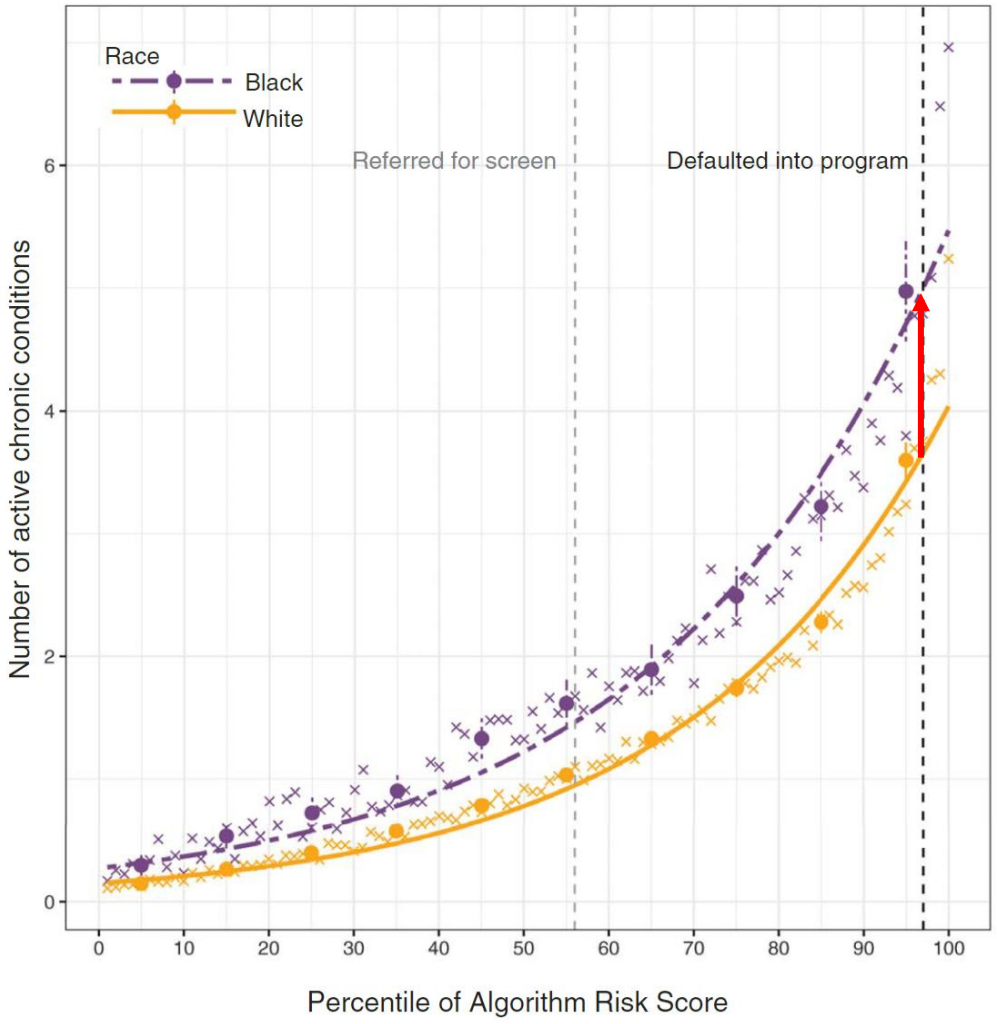
Insurance Risk Models

- Insurance companies must decide which patients are eligible for expensive high-risk care management programs
- Priority given to patients with greatest future care needs
- Thus: Insurance companies use **algorithms designed to predict future care needs**
- ML problem: Given information about patient right now, predict how much medical care they will need in the future



Obermeyer et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 2019.

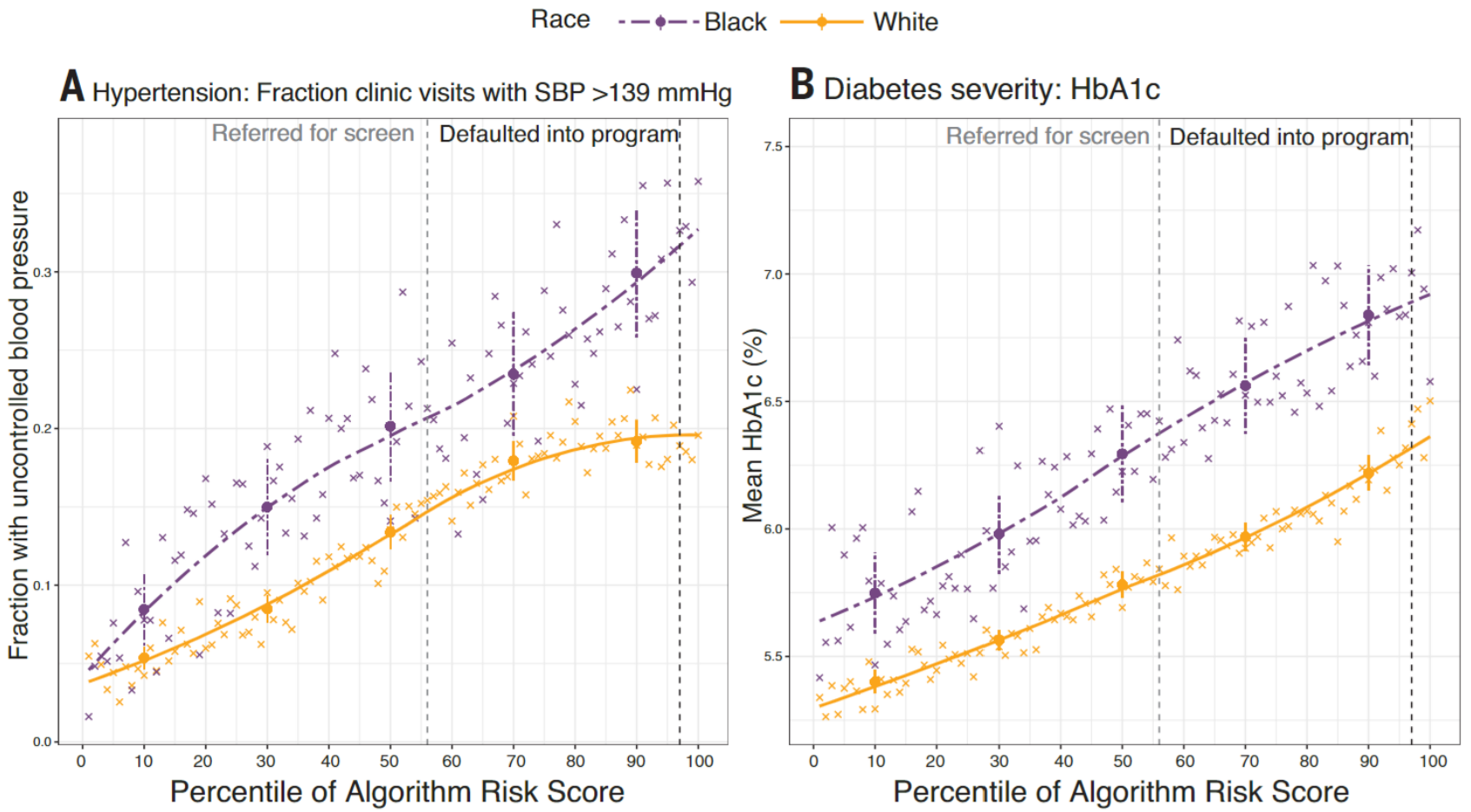
Risk Models are Biased



- Study done on deployed risk prediction tool used to assess 200 million people each year
- At the same score, black patients have more chronic conditions than white patients
- Black patients have to be much sicker to get defaulted into the care management program
  - Threshold = 97th percentile risk score

Obermeyer et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 2019.

Risk Models are Biased

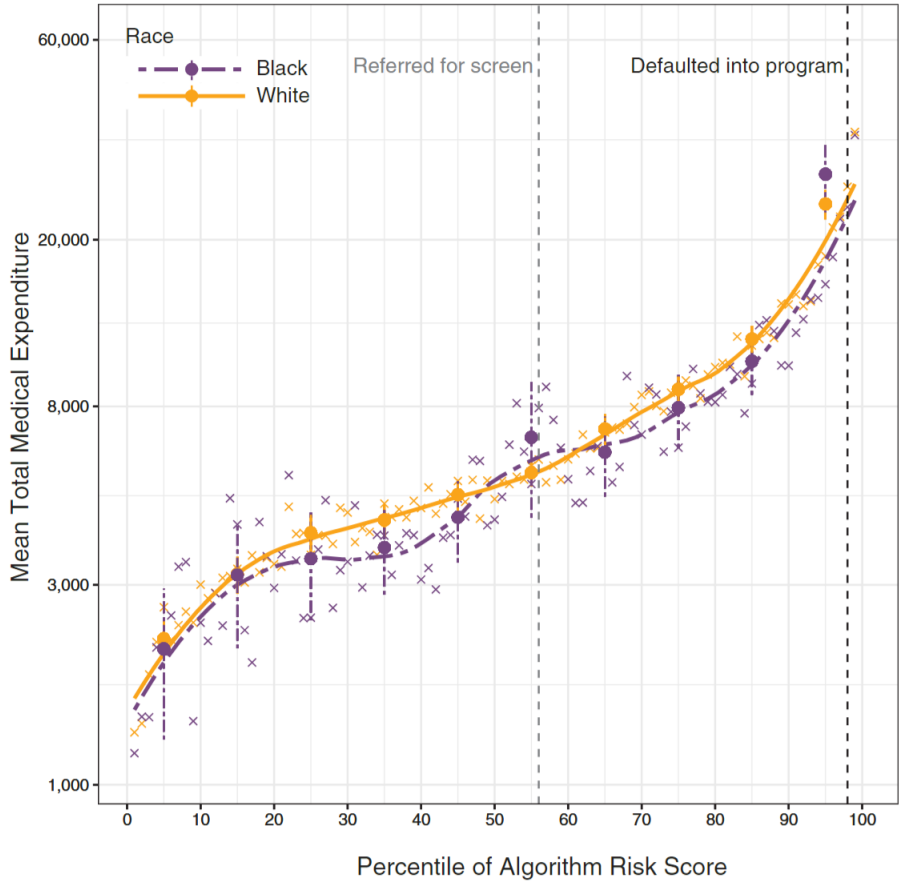


- Zoom in on most common chronic conditions
- Again, black patients are sicker than white patients at same risk score

Obermeyer et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 2019.

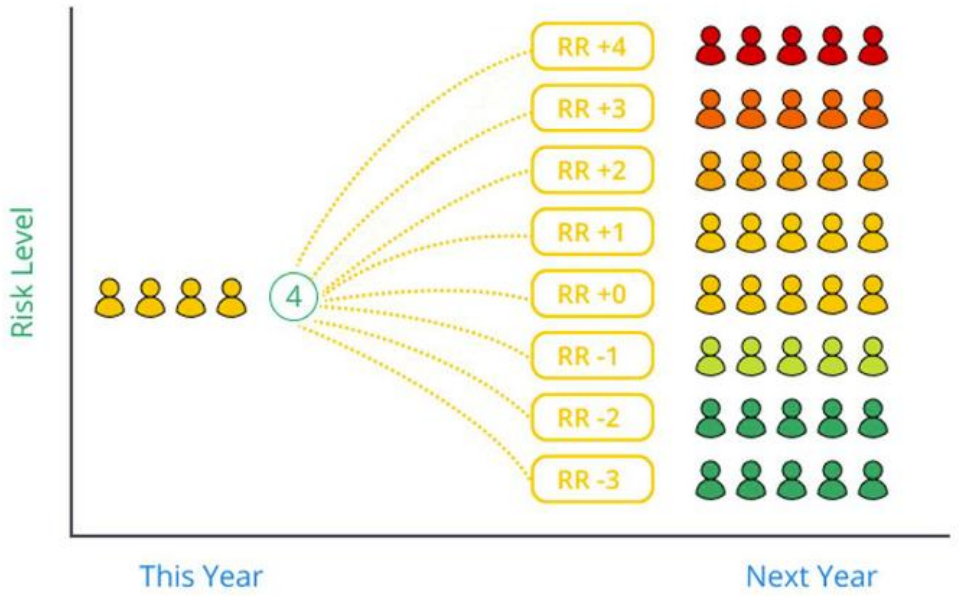
Why Might These Models be Biased?

- Model inputs: Patient age, sex, current health conditions, medicines
- Model was trained to predict **total medical costs in next year**
- Problem: Future medical cost is not same as need for medical care
  - Poor patients face more barriers to getting care
  - Lower health spending by black patients in general, possibly due to higher mistrust of medical system
- Risk score is actually **not** biased w.r.t. costs
  - Model *correctly learns from the data* that black patients with same medical conditions spend less money on average on healthcare
- Feedback loop: Underserved populations remain underserved
- Fix: Use other proxy besides cost (e.g., future health complications)



Obermeyer et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 2019.

Risk Models still Predict Cost



RISING RISK

MARA's powerful Rising Risk models

MARA Rising Risk models differentiate, classify, and allow users to assemble cohorts of members by whether their cost will stay the same, decrease, or rise. MARA results offer greater precision for classifying populations, identifying risk drivers and enabling better decisions for care programs, staffing needs to achieve results, and measuring success.

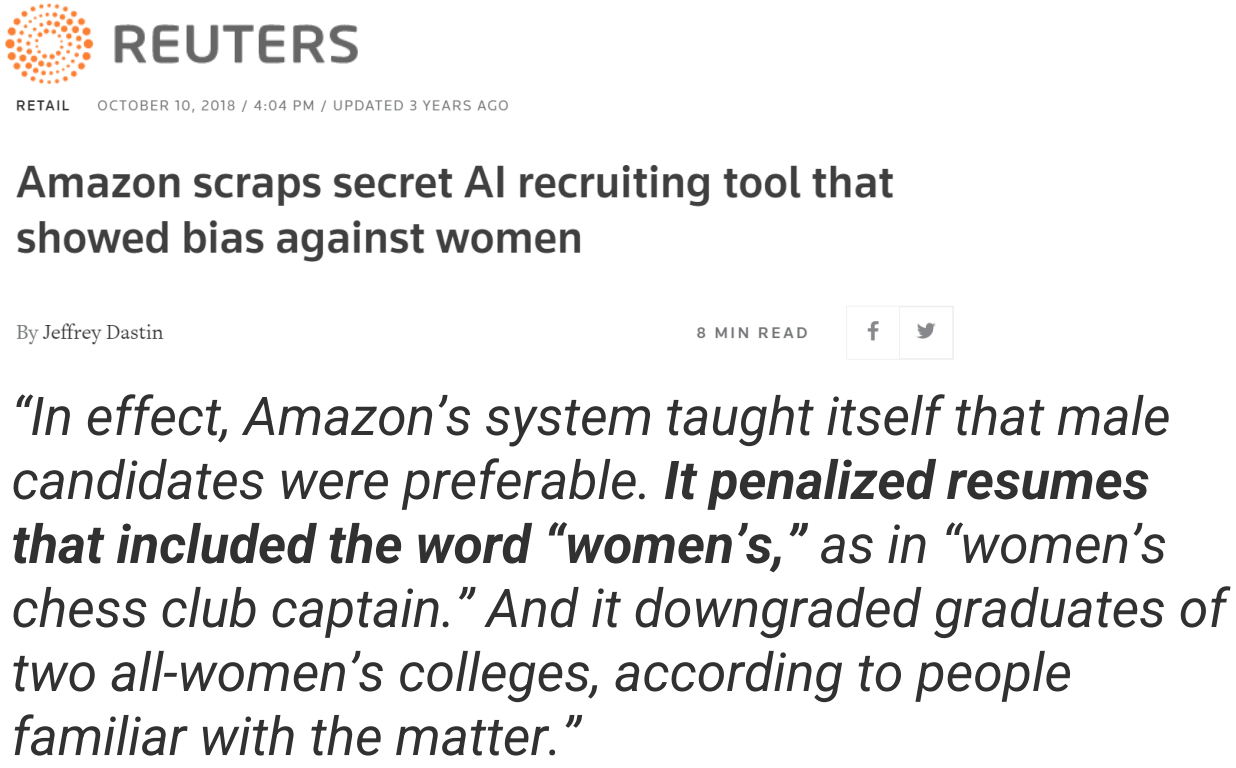
https://www.milliman.com/en/products/mara

Fairness Problems

- Allocative harms
- Unequal accuracy
- Representational harms

Allocation problems

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



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."*

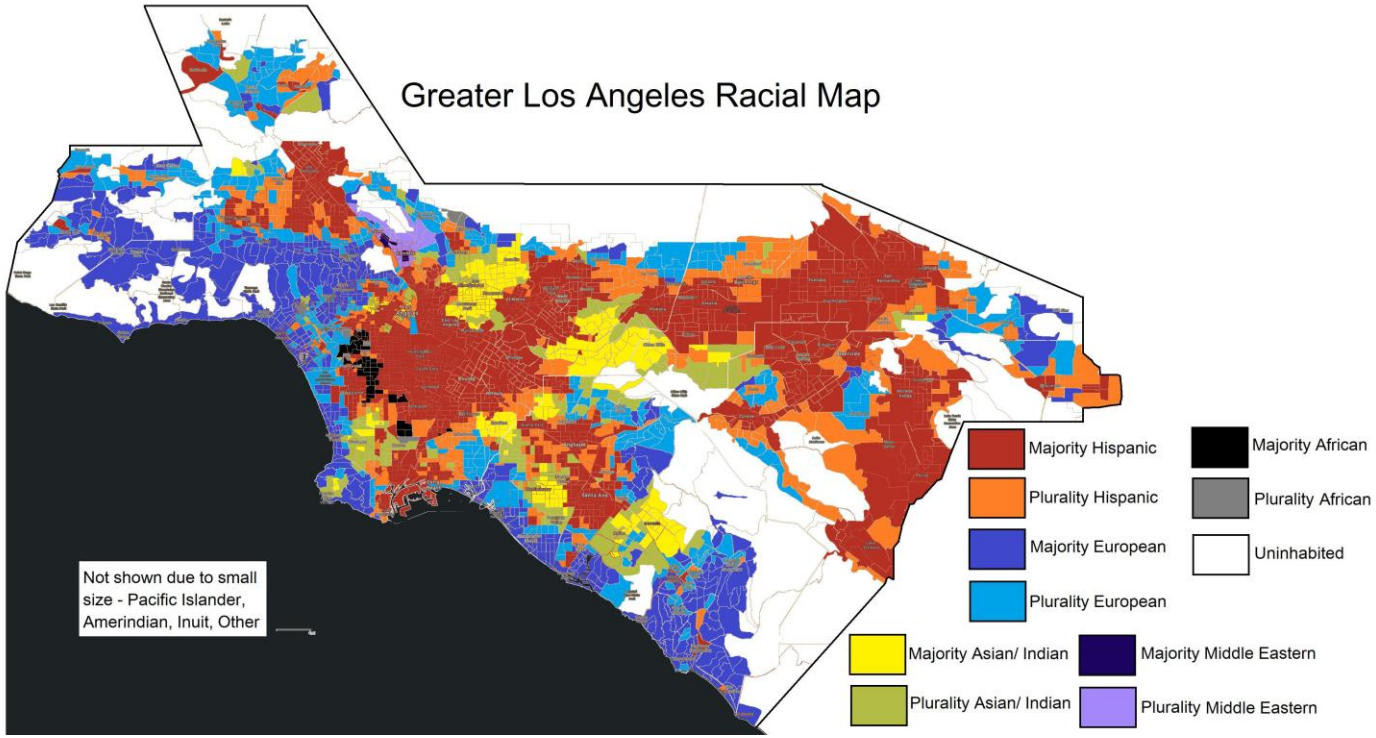
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

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
  - Example: Insurance risk model from before did **not** use race as a feature



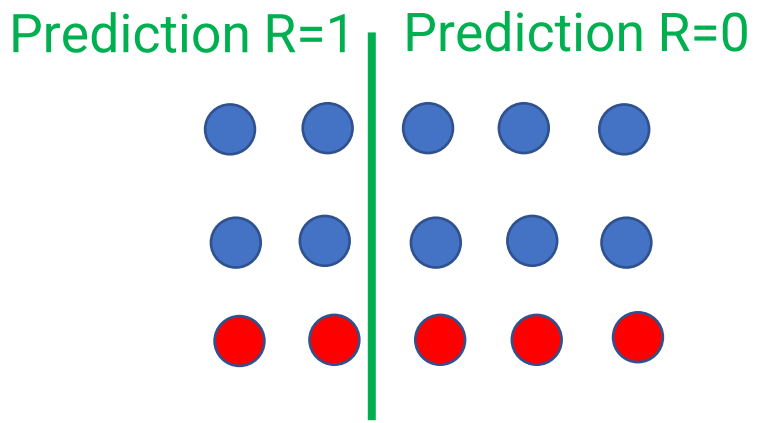
Barocas, Hardt, and Narayanan. Fairness and Machine Learning: Limitations and Opportunities.

How can we measure (un)fairness?

- As computer scientists, we would like to quantify what we mean by “fairness”
- But “fairness” is a complicated word to define…
  1. Independence (statistical parity)
  2. Separation (equalized odds)
  3. Sufficiency (calibration within groups)

1. Independence

- Independence: $R \perp A$
  - Equivalently for binary predictor:
    $P(R = 1 \mid A = a) = P(R = 1 \mid A = b) \forall a, b$
  - E.g., if you label 3% of white patients as high risk, must also label 3% of black patients as high risk
  - 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 $Y \not\perp A$
    - E.g., "Base rate" between white & black populations may be different for many reasons



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

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

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

2. Separation / Equalized odds

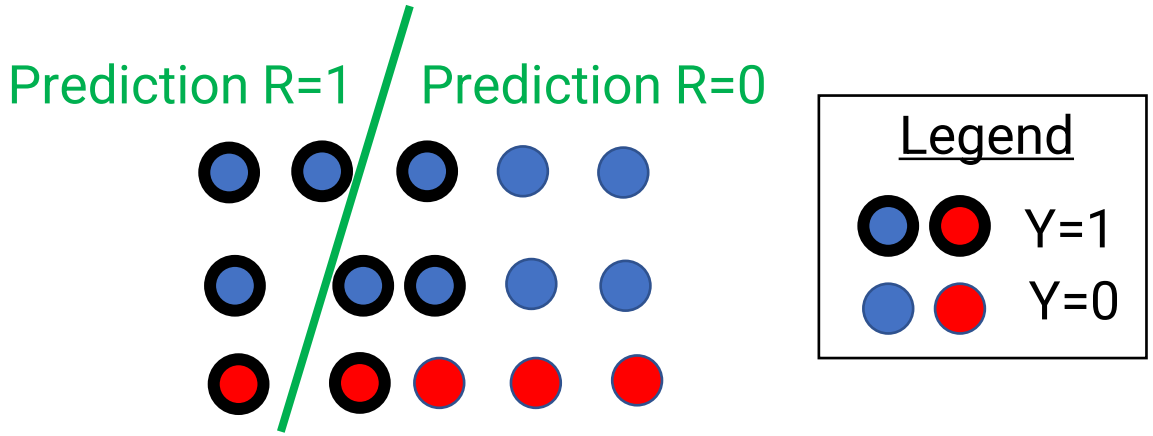
- Separation: $R \perp A \mid Y$
  - Equivalently for binary predictor:

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

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

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

  - Alternatively: Split dataset into Y=1 and Y=0 subsets, equalize the accuracy on each subset



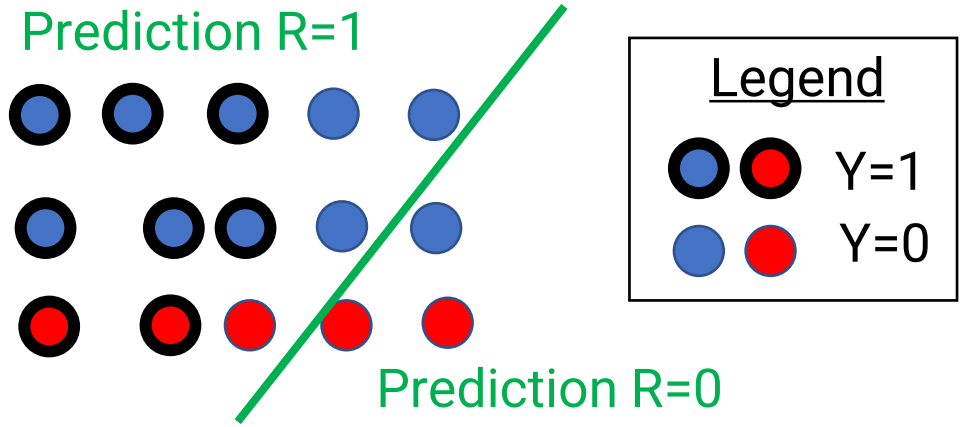
$$P(R = 1 \mid A = \bullet, Y = 1) = 3/6 = 1/2$$
$$P(R = 1 \mid A = \bullet, Y = 1) = 1/2$$
$$P(R = 0 \mid A = \bullet, Y = 0) = 4/4 = 1$$
$$P(R = 0 \mid A = \bullet, Y = 0) = 3/3 = 1$$

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

3. Sufficiency / Calibration within groups

- Sufficiency: $Y \perp A \mid R$
  - 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

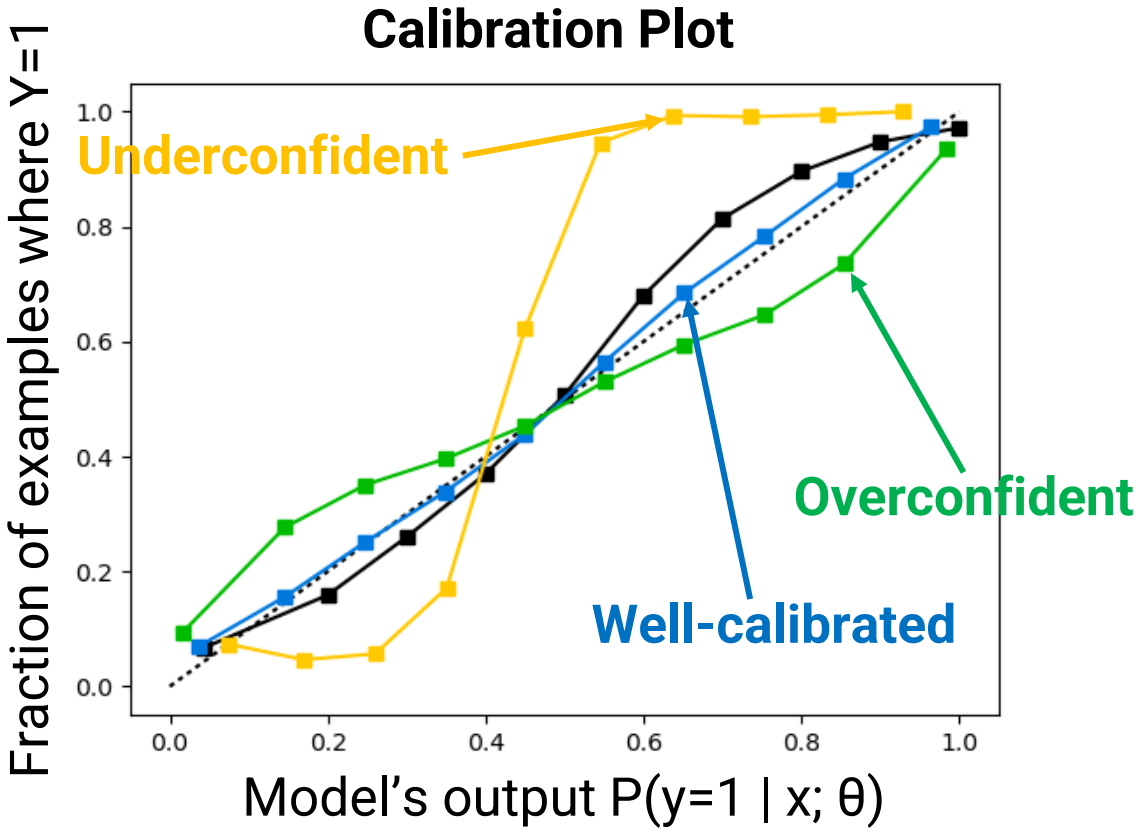
$$\frac{\text{Positives found by classifier}}{\text{Things predicted as positive}}$$

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

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**

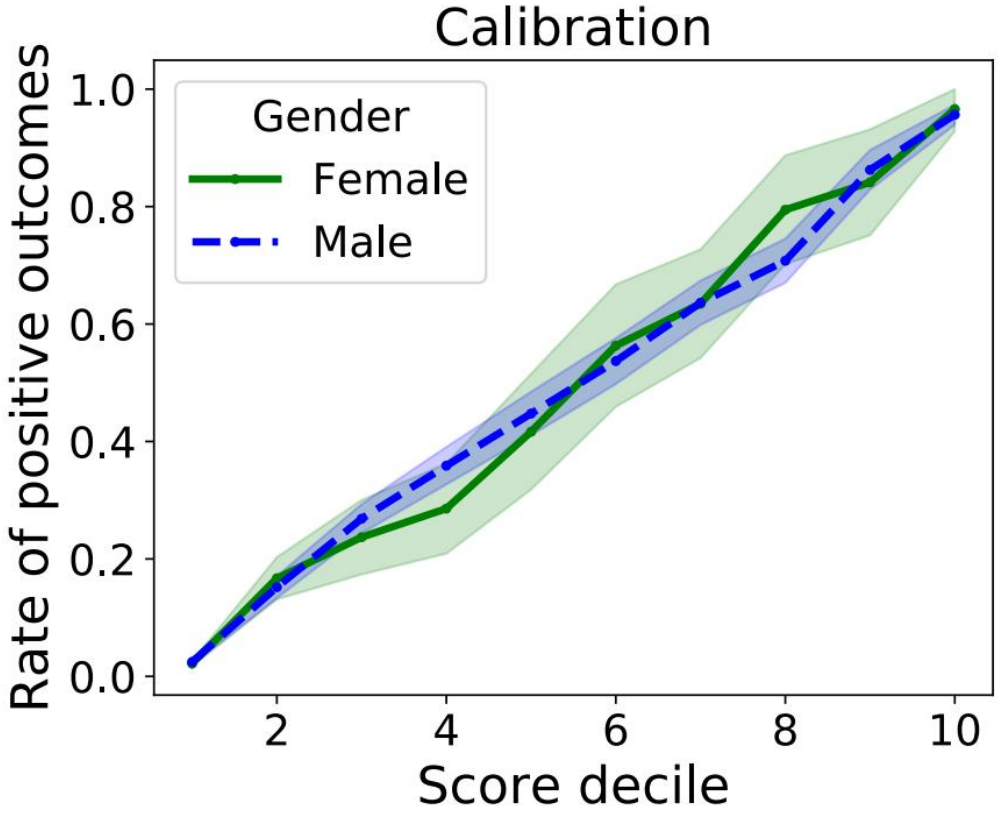


Sufficiency and Calibration

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

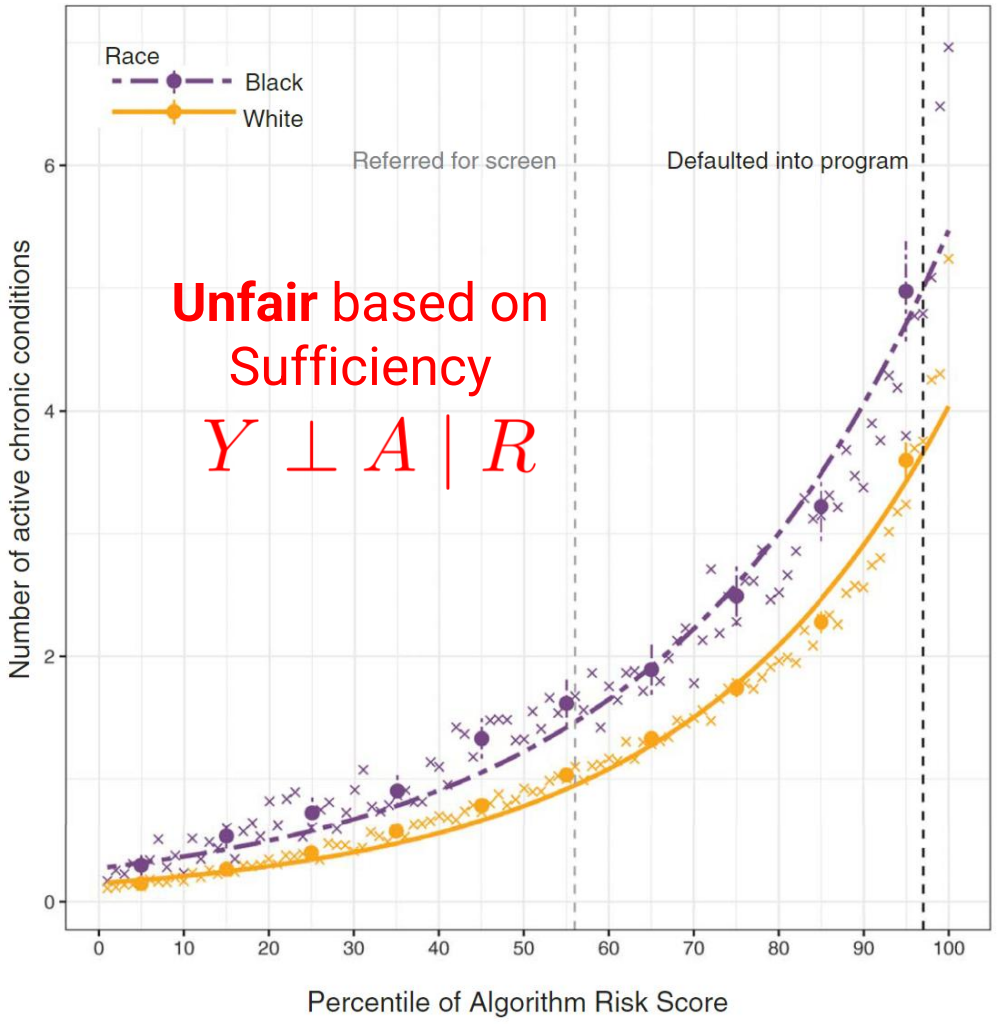
$$P(Y = 1 \mid A = a, R = r) =$$
$$P(Y = 1 \mid A = b, R = r) \forall r$$

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



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

Returning to Insurance



- Insurance risk models fail the test of sufficiency

Great, now we can make things fair…?

- Surprisingly, 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 $Y \perp A \mid R$

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

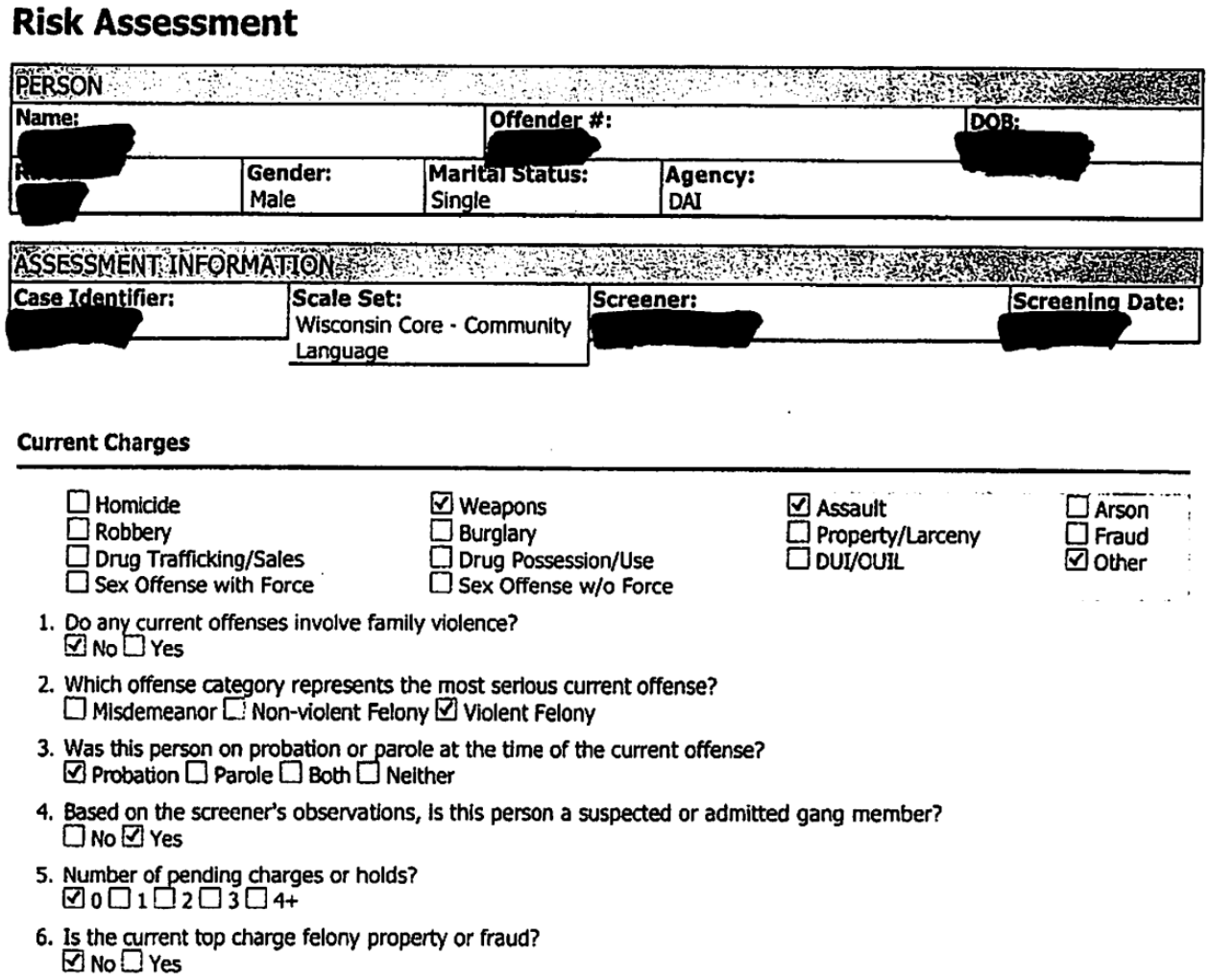
$$= P(Y \mid A = b)$$

Base rate of Y in population b

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

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]
[redacted]	Gender: Male	Marital Status: Single
	Agency: DAI	

ASSESSMENT INFORMATION

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

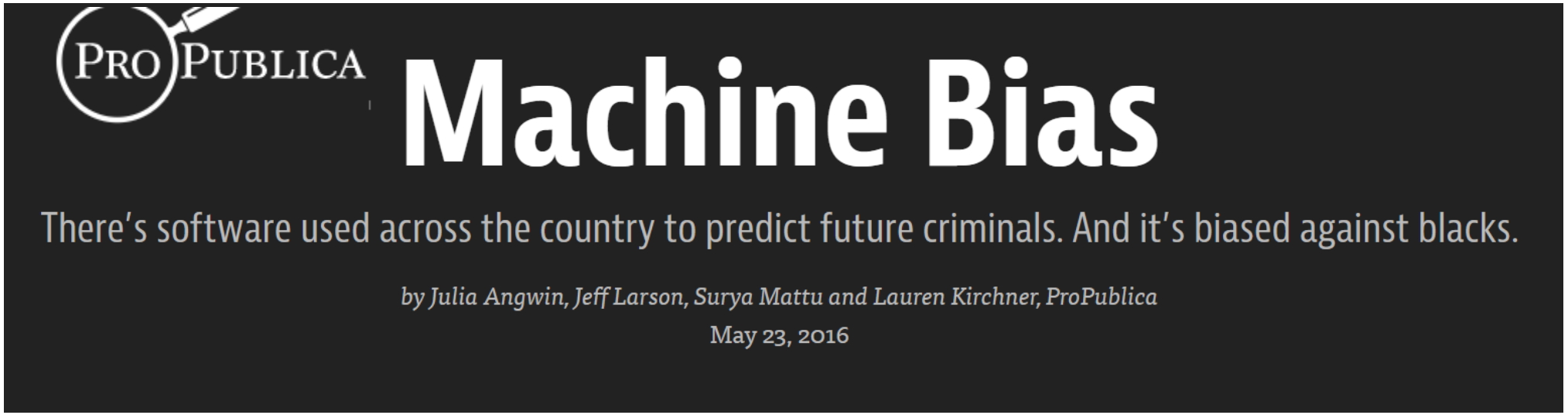
Current Charges

☐ 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

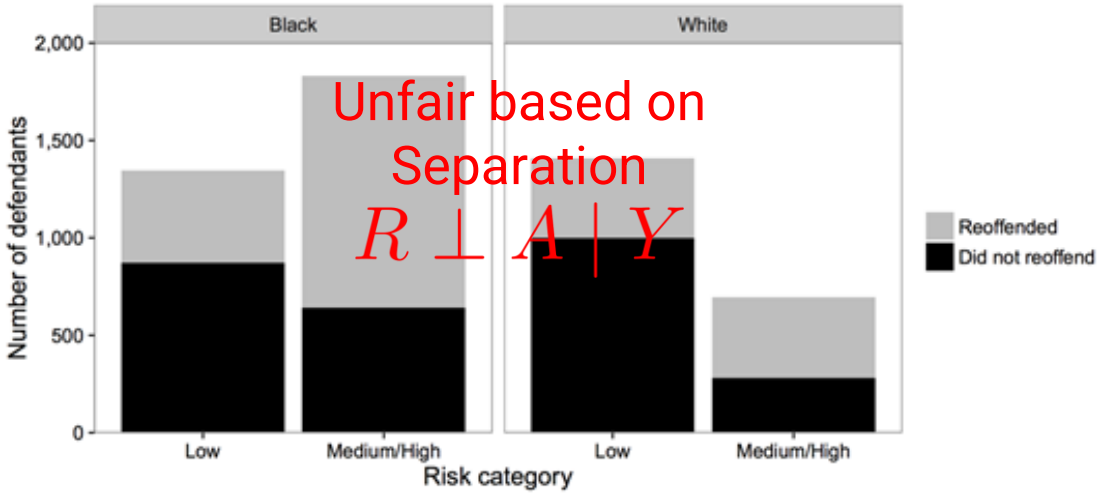


- "The formula was **particularly likely to falsely flag black defendants as future criminals**, wrongly labeling them this way at almost twice the rate as white defendants."
- "White defendants were mislabeled as low risk more often than black defendants."

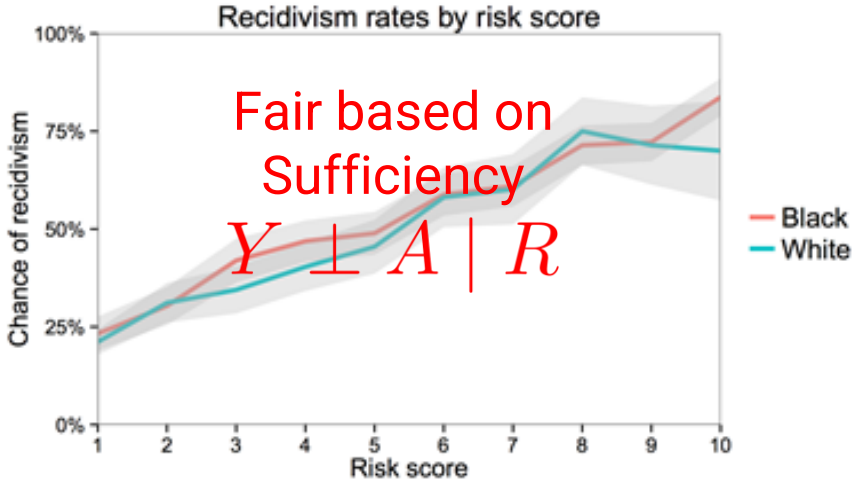
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

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

Where do we go from here?

- There is a fundamental trade-off between different natural notions of fairness
- Some systems may lie in a “gray area” where they appear fair in one way, but unfair in another
- Other systems may be more clearly unfair
- Auditing systems requires thinking deeply about what notion of fairness matters for the task at hand



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

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%



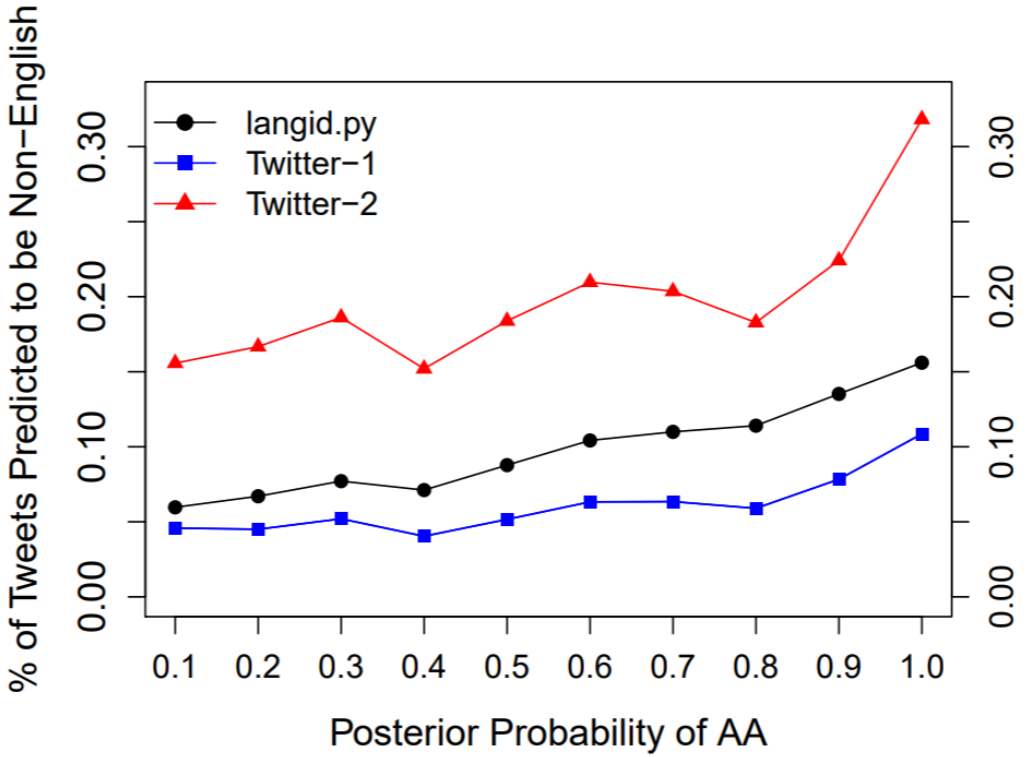
- 2018 study: Commercial facial recognition systems much less accurate on darker-skinned females than other groups

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

Announcements

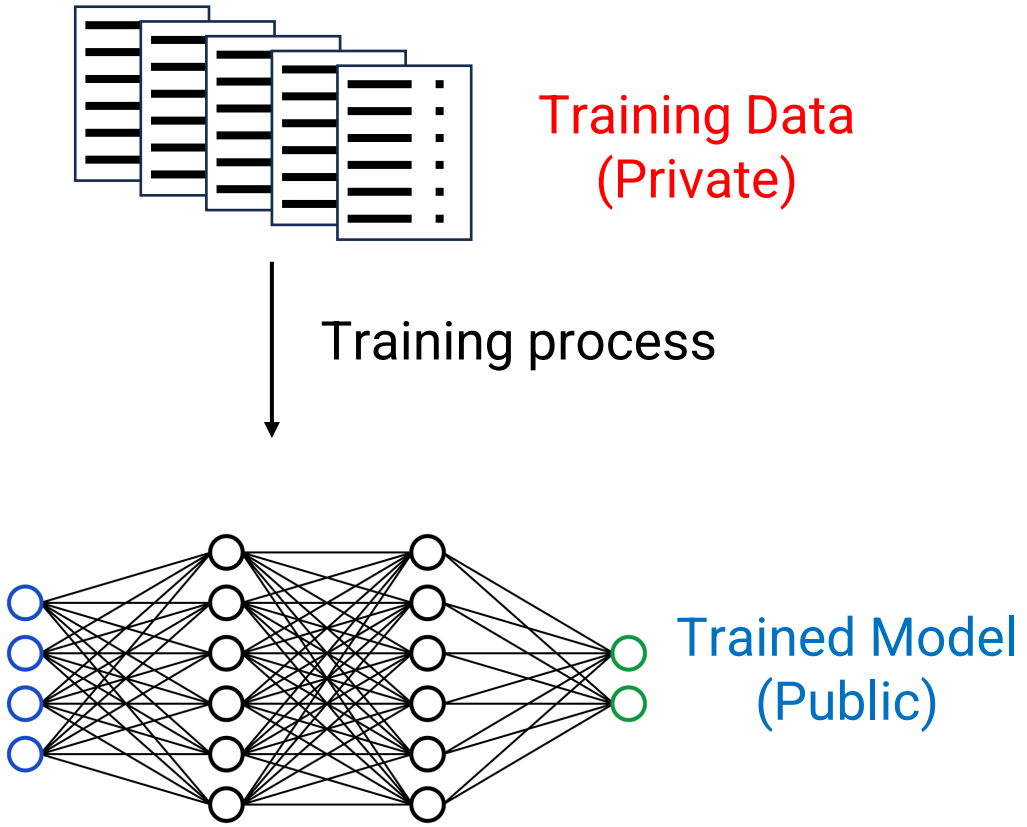
- Homework 4 out
  - Due Monday, November 30
- Final Project Report
  - Due Monday, December 7
  - 5-6 pages, use same LaTeX template as before
  - Show model improvements relative to midterm report
  - Submit code & Readme
  - See website for details

Announcements

- Final Exam Logistics
  - Friday, December 11 from 2-4pm
  - Room arrangement TBD
  - Allowed 2 (double-sided) 8.5”x11” sheets of paper
  - Exam is cumulative, more emphasis on post-midterm material
  - Practice exams & lecture videos will be posted tonight

Use of Private Training Data

- Machine learning models are often trained on private data
  - Social media companies: Train on user preferences, actions, etc.
  - Medical applications: Train on medical records of patients
- While training data is kept private, the models (or their predictions) are often publicly visible
- **Does this leak private information?**



Threats to Privacy

Generation

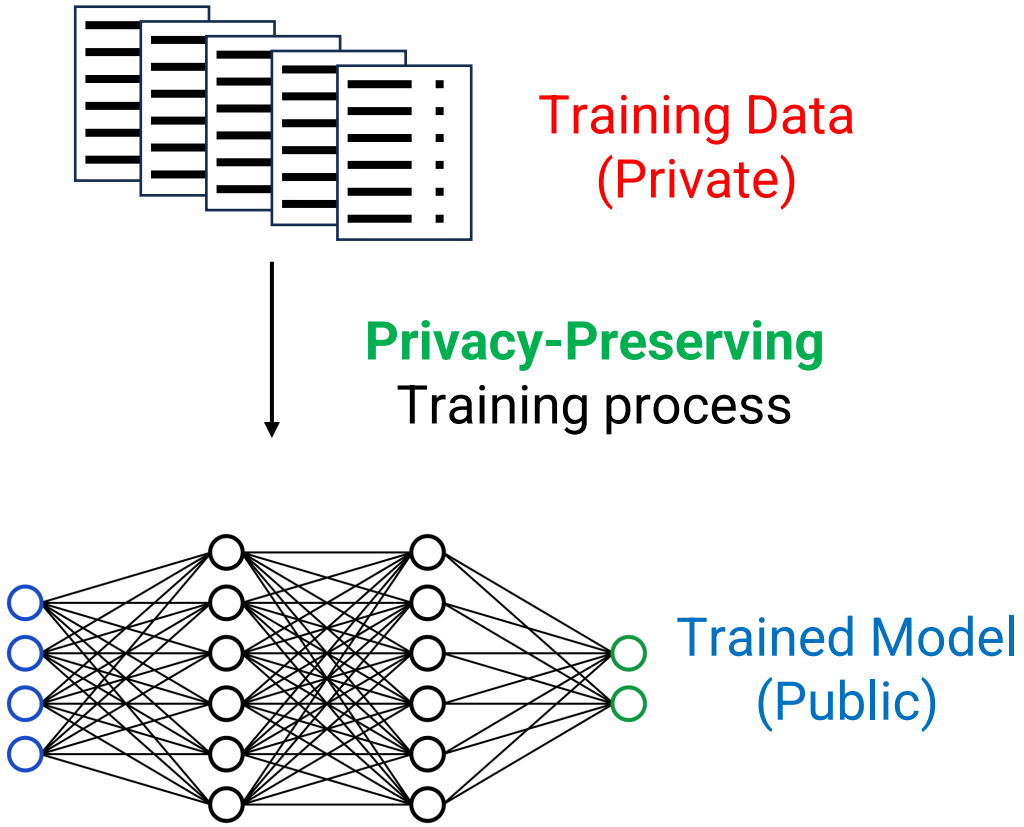
- Target: Language model
- Input: *"John Smith's social security number is"*
- Output: LM will fill in most likely answer based on its (internet-wide) training data
- If personal information leaked online, LLM might reveal it!

Membership Inference

- Target: Classifier for risk of heart disease given medical record
- Input: *[Medical record that states that John Smith has hypertension]*
- Output: Classifier outputs some probability of patient developing heart disease
- Membership Inference Attack: Infer whether input was part of training data
  - E.g., if model is highly confident, can be a sign it overfit to this input => it must have been part of the training data
- Attacker can thus infer John Smith's medical history!

Privacy-Preserving Algorithms

- Goal: Modify the training process to **guarantee** that the model cannot leak private information in the dataset?
- Intuitive guarantee: If we change the training data by a small amount, *the model will be very similar*
  - Corollary: If one person's private data is a small % of total data, *the resulting model will be very similar to the model that did not train on that data*
  - And the latter cannot leak private information!



Differential Privacy

For scalars ε and δ, an algorithm A is ***(ε, δ) differentially private*** if for any two neighboring datasets D and D', and for any S ⊆ Range(A):

E.g., differ by exactly 1 example

"Set of models with property X"

$$P(A(D) \in S) \leq e^{\varepsilon} * P(A(D') \in S) + \delta$$

"Probability that the model trained on **D** has property X"

"Probability that the model trained on **D'** has property X"

"Fudge factors": The two probabilities are within δ (additively) and e^{ε} (multiplicatively) of each other

Punchline: **Hard to distinguish** models trained on D and D'—they have similar probability distributions!

Differentially Private SGD

Recall Original SGD:

for t = 1, …, T: Each t (i.e., each pass over the dataset) is called one “epoch”

Randomly partition training examples into batches B_1, …, B_k

for i = 1, …, k:

$$\theta \leftarrow \theta - \eta \cdot \frac{1}{|B_i|} \sum_{(x,y) \in B_i} \nabla_\theta \ell\left(y, g(x)\right)$$

Update based on sample mean within current batch

Differentially Private SGD

DP-SGD:

for t = 1, …, T: Each t (i.e., each pass over the dataset) is called one "epoch"

Randomly partition training examples into batches B_1, …, B_k

for i = 1, …, k:

$$\theta \leftarrow \theta - \eta \cdot \frac{1}{|B_i|} \sum_{(x,y)\in B_i} \text{CLIP}\left(\left(\nabla_\theta \ell\left(y, g(x)\right)\right) + \text{NOISE}\right)$$

Clip each gradient to some maximum L_2 norm

Add Gaussian noise

These two steps can ensure DP property!

- ε and δ determined by scale of clipping and noise
  - Clipping prevents any single example from having large influence
  - Noise creates uncertainty, attacker's job is harder
- Trade-off: Model will often be of worse quality, due to noisy updates

Conclusion

- ML systems learn patterns in the data, including ones we may not intend
- Fairness: Breadth of potential harms
  - To individuals being evaluated
  - To users attempting to use tools
  - To broader society due to shifts in perception
- Privacy: Can leak info about private training data
  - Differential privacy is one way to define “privacy-preservation”
  - DP-SGD gives us way to ensure that ML models preserve privacy