Algorithmic fairness

Gal Yona, PhD Candidate, Weizmann Institute

Data-driven algorithms are all around us

Data-driven algorithms are all around us

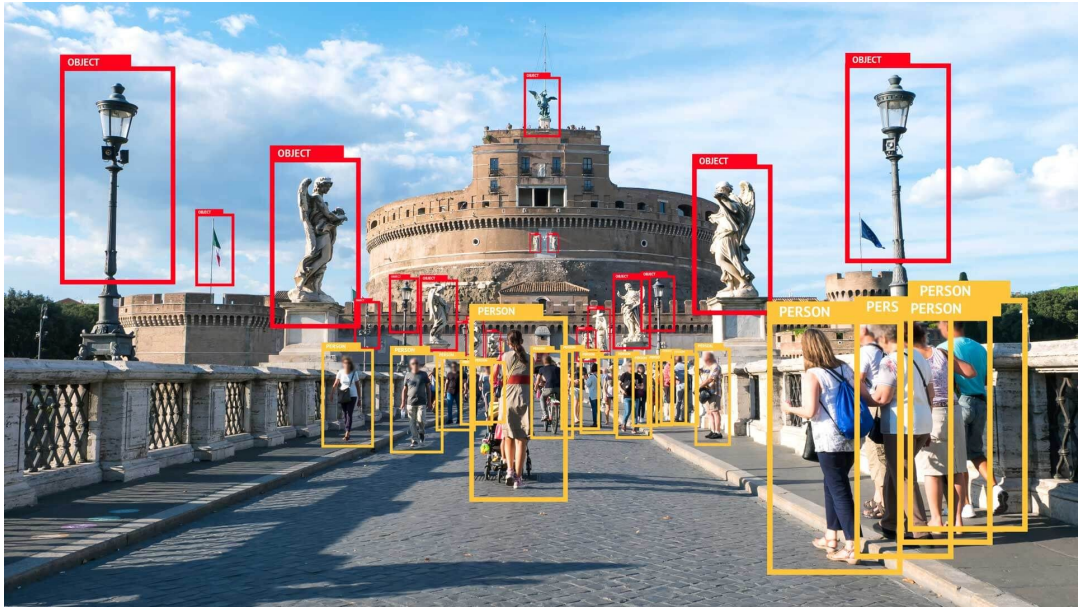


Image recognition

Data-driven algorithms are all around us

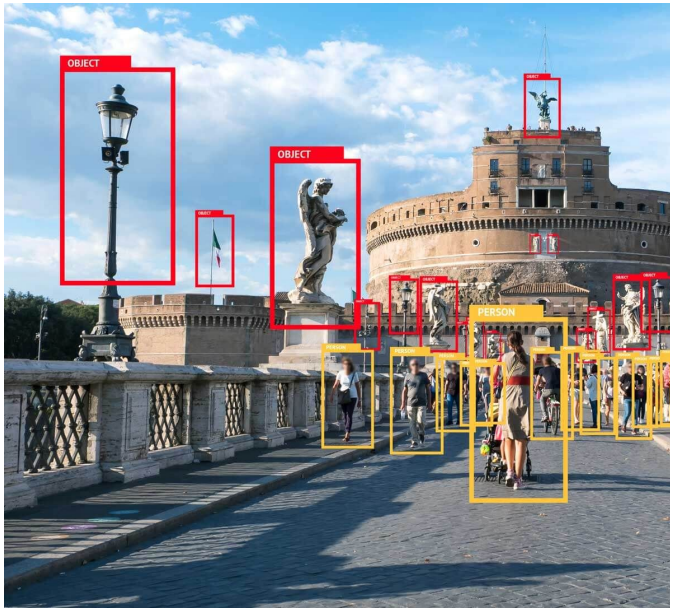
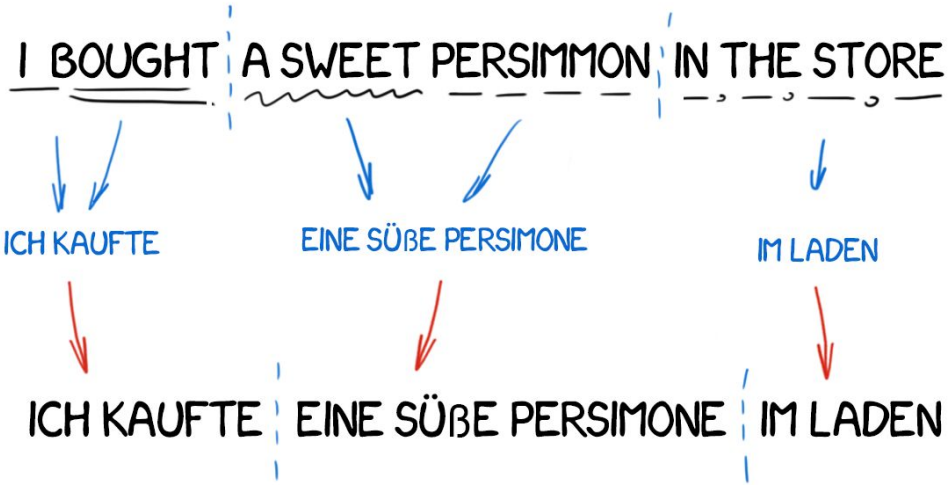


Image recognition

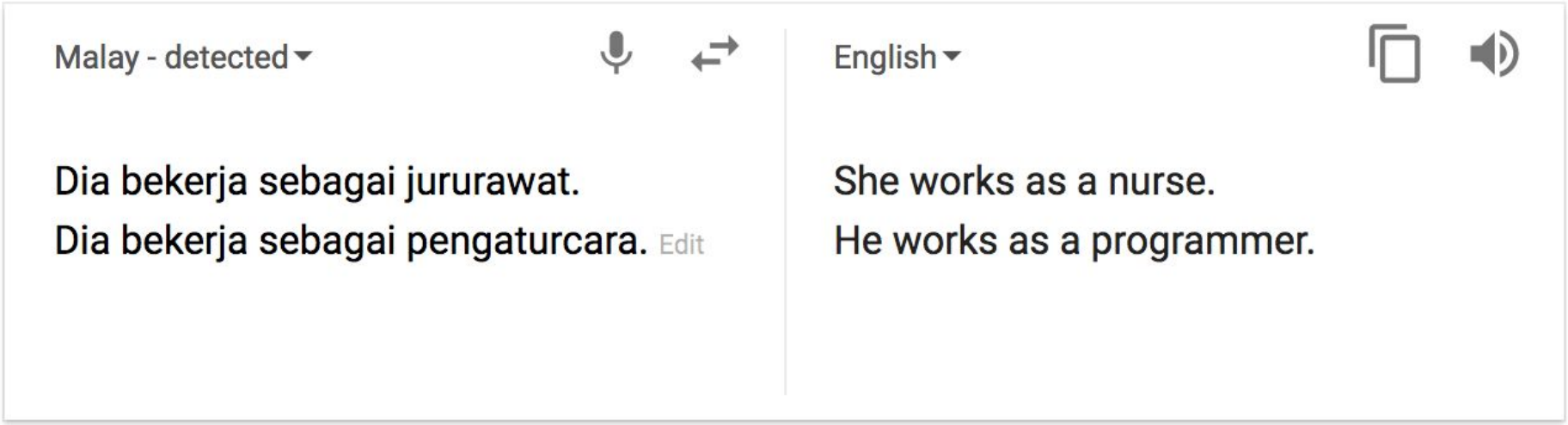


Machine translation

Why worry?

Why worry?

(1) …. are these tasks nearly solved?



Why worry?

(2) Modern applications now include automating human judgement and predicting (and informing) social outcomes

Why worry?

(2) Modern applications now include automating human judgement and predicting (and informing) social outcomes

Arvind Narayanan, "How to recognize AI snakeoil"
https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

Today's talk: algorithmic fairness

- "Algorithmic": focus on data-driven algorithms
- A topic of active and interdisciplinary research
- Today: a glimpse of the CS perspective
  - sources of unfairness
  - defining fairness: common notions, weaknesses, ways to strengthen them

Today's talk: algorithmic fairness

- "Algorithmic": focus on data-driven algorithms
- A topic of active and interdisciplinary research
- Today: a glimpse of the CS perspective
  - sources of unfairness
  - defining fairness: common notions, weaknesses, ways to strengthen them

Problem formulation → Gathering data → Learning

Problem formulation | Gathering data | Learning

What are you interested in predicting? (**Y**)

How are you planning to predict it? (**X**)

Problem formulation | Gathering data | Learning

What are you interested in predicting? (Y)

How are you planning to predict it? (X)

Obtaining a collection **S** of "labeled" pairs (x,y), representative of the target distribution

Problem formulation | Gathering data | Learning

What are you interested in predicting? (Y)

How are you planning to predict it? (X)

Obtaining a collection S of “labeled” pairs (x,y), representative of the target distribution

Search a candidate class **H** for the model that makes the “most accurate” predictions on **S**

Problem formulation | Gathering data | Learning

What are you interested in predicting? (**Y**)

How are you planning to predict it? (**X**)

Obtaining a collection **S** of "labeled"), representative of the target distribution

Search a candid for the model that makes the "most accurat" ons on **S**

Potential unfairness can enter at any stage.

Y true "credit-worthiness"

"credit-worthiness" not observed, resort to proxies

X *all* financial histories

X missing important features (not random)

S *representative* and *accurate* sample of pairs (x,y)

S not representative or biased, partial feedback

No issues in the learning procedure

Accuracy on (smaller, harder to predict) subpopulations might be compensated

This is *not* a toy example!

← **Thread**

DHH ✓
@dhh

The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

10:34 PM · Nov 7, 2019 · Twitter for iPhone

9.7K Retweets **29.3K** Likes

This is *not* a toy example!

Thread

DHH
@dhh

The @AppleCard is su
wife and I filed joint ta
property state, and ha
Yet Apple's black box
credit limit she does. I

10:34 PM · Nov 7, 2019 · Twitter fo

9.7K Retweets **29.3K** Likes

Carmine Granucci
@whoiscarmine

Replying to @dhh and @AppleCard

Just read this thread. My wife has a way better score than me, almost 850, has a higher salary and was given a credit limit 1/3 of mine. We had joked that maybe Apple is just sexist. Seems like it's not a joke. Beyond f'ed up.

5:08 PM · Nov 9, 2019 · Tweetbot for iOS

226 Retweets **2K** Likes

Official response

“we have not and never will make decisions based on factors like gender. In fact, we do not know your gender or marital status during the Apple card application process”

Official response

“we have not and never will make decisions based on factors like gender. In fact, we do not know your gender or marital status during the Apple card application process”

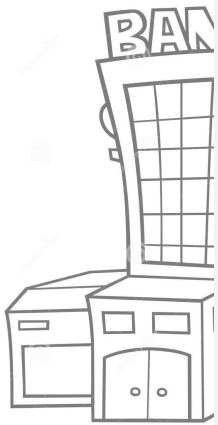
“some of our customers have told us they received lower credit lines than they expected. In many cases, this is because their existing cards are … under their spouse’s primary account – which may result in the applicant having limited personal credit history”.

Y true "credit-worthiness"

X *all* financial histories

S *representative* and *accurate* sample of pairs (x,y)

No issues in the learning procedure



Ideal world: perfect predictions possible, no fairness issues

Practice: sub-optimal choices, each carries potential for some individuals to be negatively affected more than others

Want a methodical way of finding these issues, and possibly addressing them

Defining fairness

For whom?

- Groups of individuals protected by law or ethics

What's fair?

- model performs "equally well" across groups

Group notions of fairness: general recipe

1. Define notion of “harm”

Group notions of fairness: general recipe

1. Define notion of "harm"

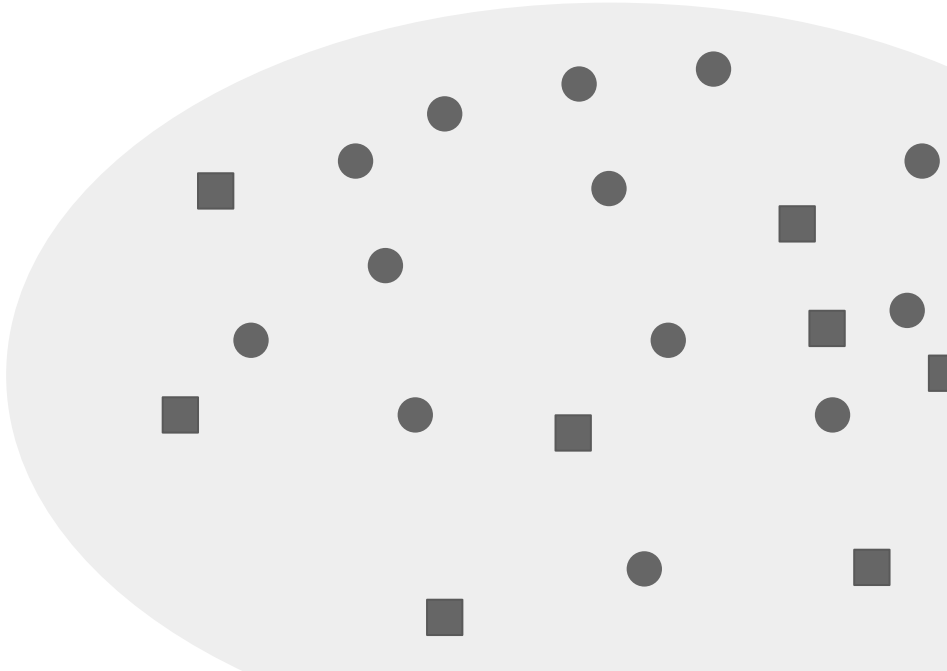
Classification (Y binary)

- $\tilde{Y} = 0$
  (not getting a loan)
- $Y = 1, \tilde{Y} = 0$
  (not getting a loan, though qualified)

Regression (Y in [0,1]):
Calibration

Group notions of fairness: general recipe

1. Define notion of “harm”
2. Partition universe across a protected attribute A (e.g. gender)



Group notions of fairness: general recipe

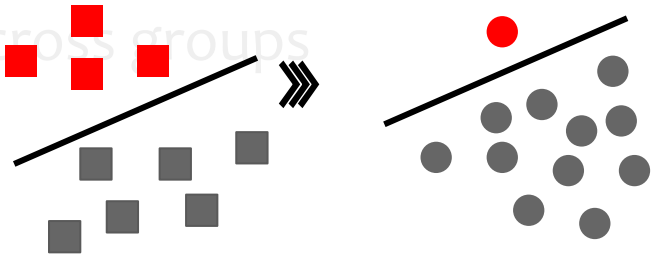
1. Define notion of “harm”
2. Partition universe across a protected attribute A (e.g. gender)

Group notions of fairness: general recipe

1. Define notion of “harm”
2. Partition universe across a protected attribute A (e.g. gender)
3. Model is unfair if the distribution of induced "harms" is different across groups

Group notions of fairness: general recipe

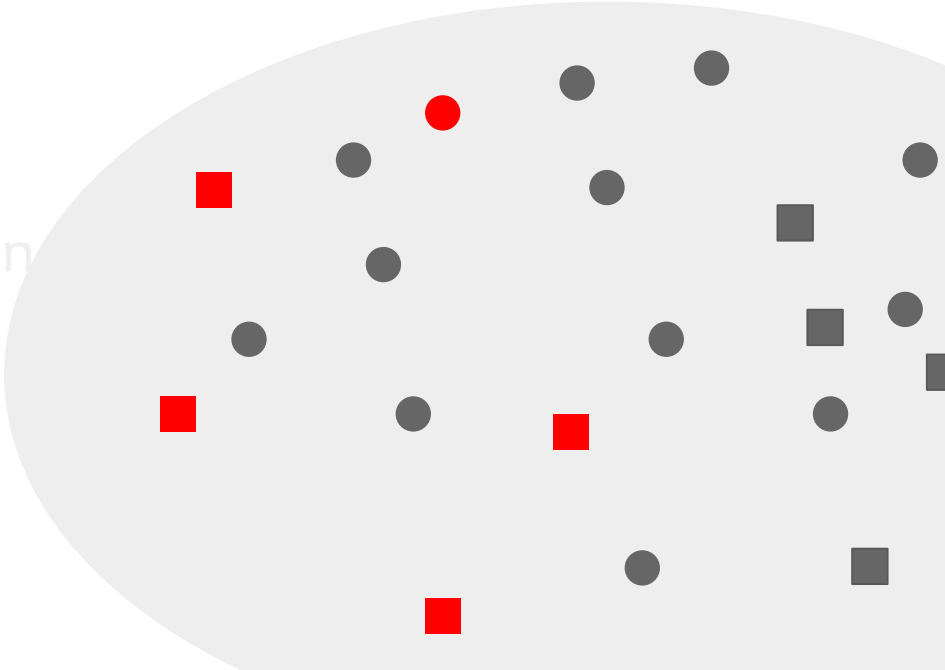
1. Define notion of "harm"
2. Partition universe across a protected attribute A (e.g. gender)
3. Model is unfair if the distribution of induced "harms" is different across groups



Men

Women

Harmed by model



Group notions of fairness

Violation of group-fairness → unfairness

Group notions of fairness

Violation of group-fairness → unfairness

Missing important features? Too little data from minority? Issues with the learning procedure?

Group notions of fairness

Violation of group-fairness → unfairness

No violations of group-fairness → fairness?

Group notions of fairness

Violation of group-fairness → unfairness

No violations of group-fairness → fairness?

NO!

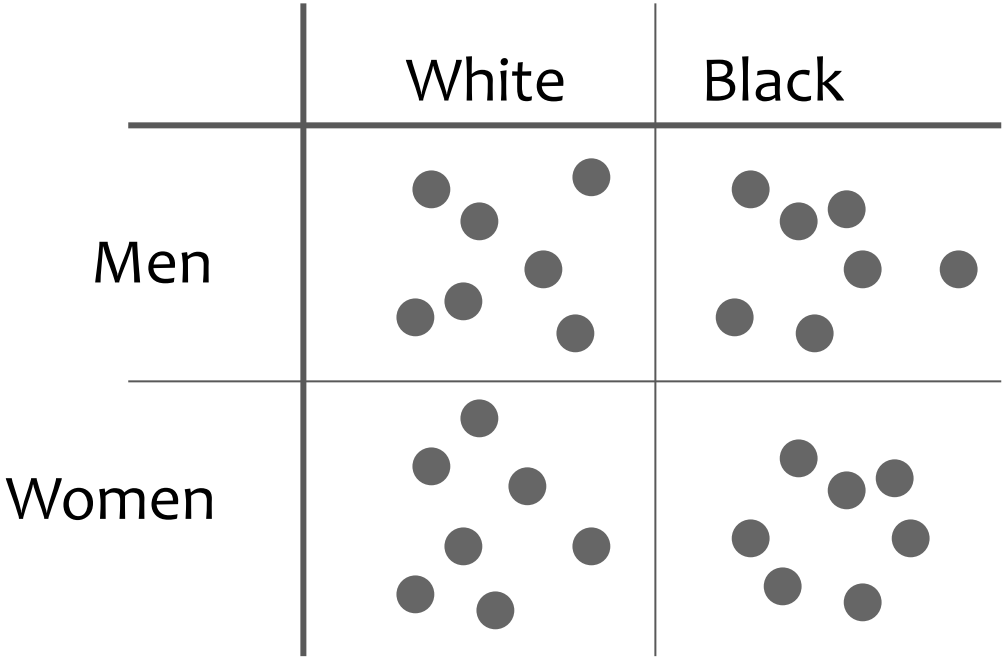
violations of group-fairness → fairness?

Intuition:

- Group fairness says something about the probability of harm for a random member of A = 0 versus a random member of A = 1
- …. But *you* are not a random member!

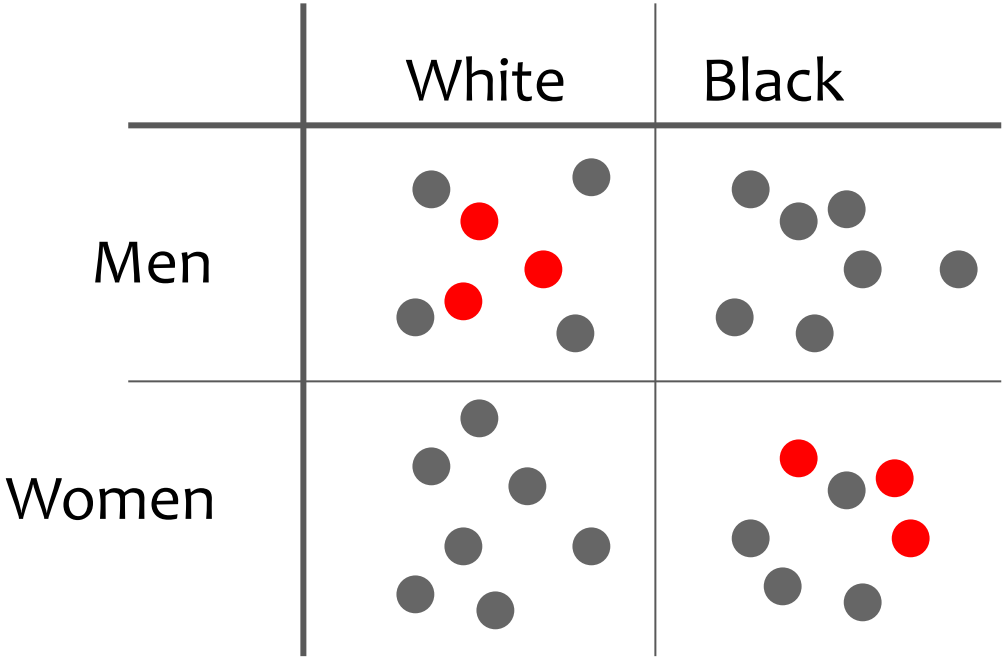
In fact, a "group-fair" classifier might be very unfair to **individuals** or more **structured subgroups**

Weaknesses of group fairness



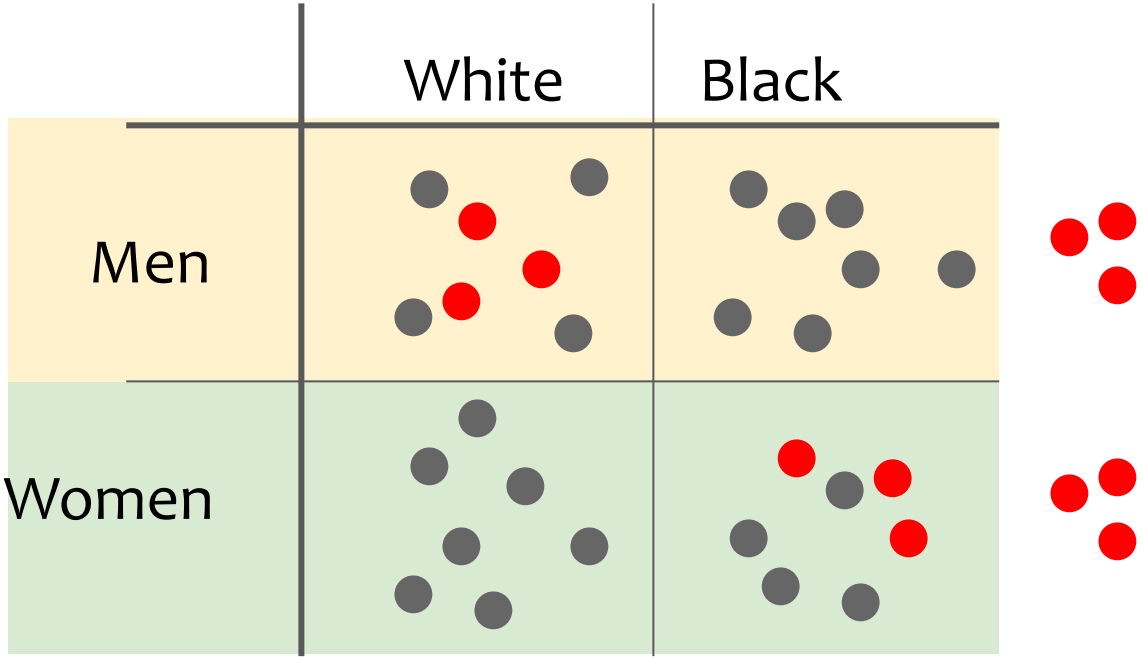
Kearns, M., Neel, S., Roth, A. and Wu, Z.S., 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*.

Weaknesses of group fairness



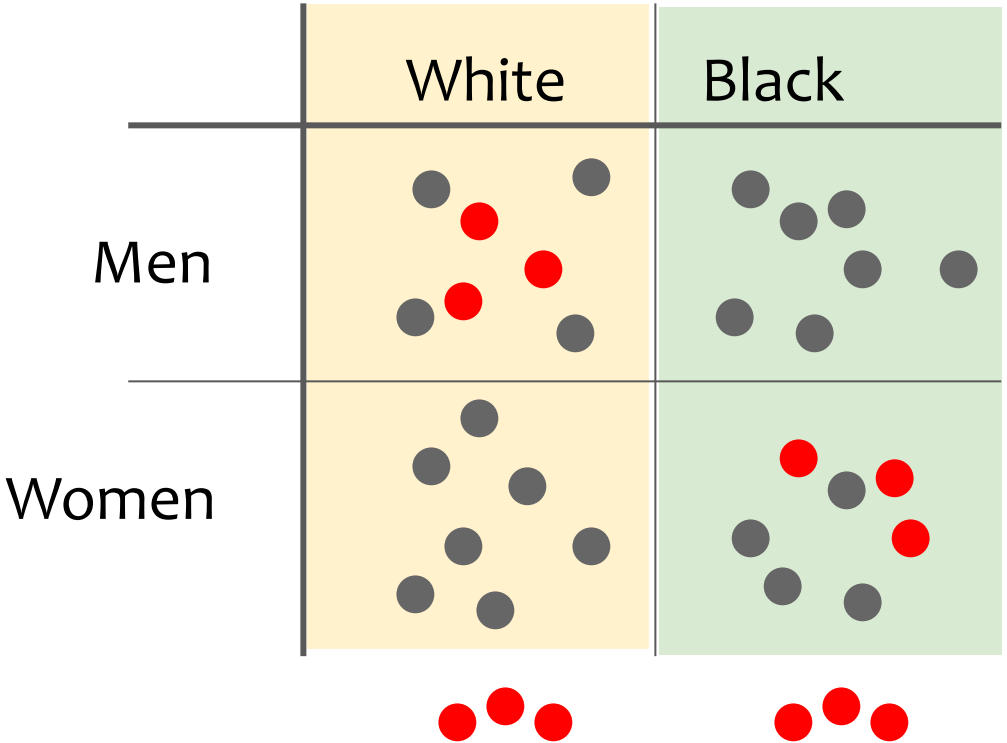
Kearns, M., Neel, S., Roth, A. and Wu, Z.S., 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*.

Weaknesses of group fairness



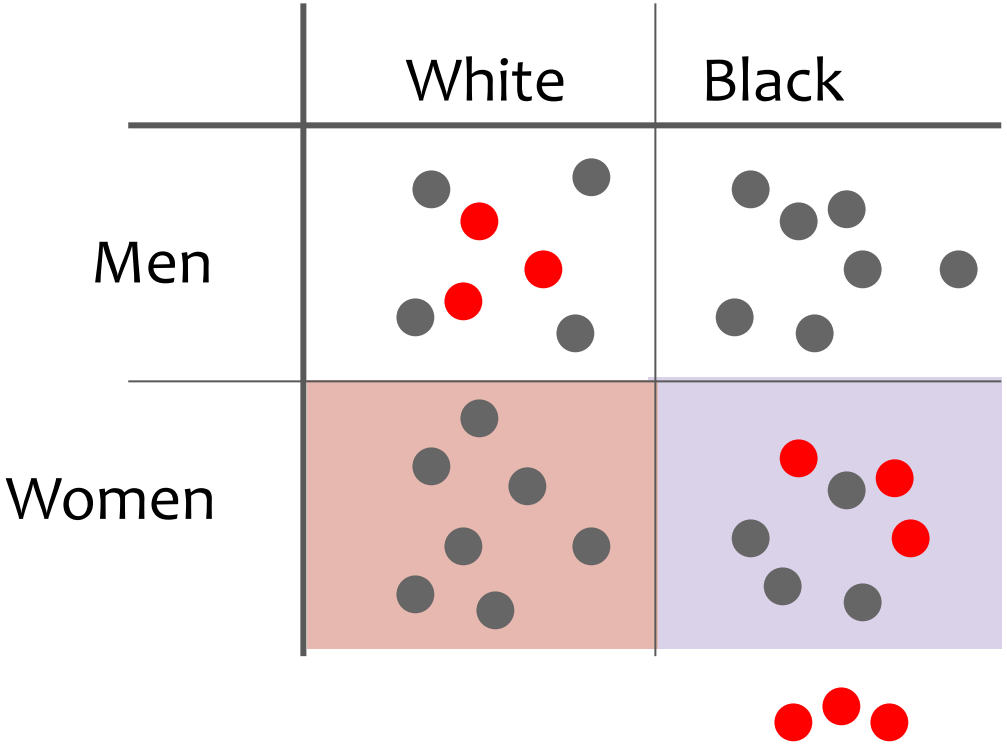
Kearns, M., Neel, S., Roth, A. and Wu, Z.S., 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*.

Weaknesses of group fairness



Kearns, M., Neel, S., Roth, A. and Wu, Z.S., 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*.

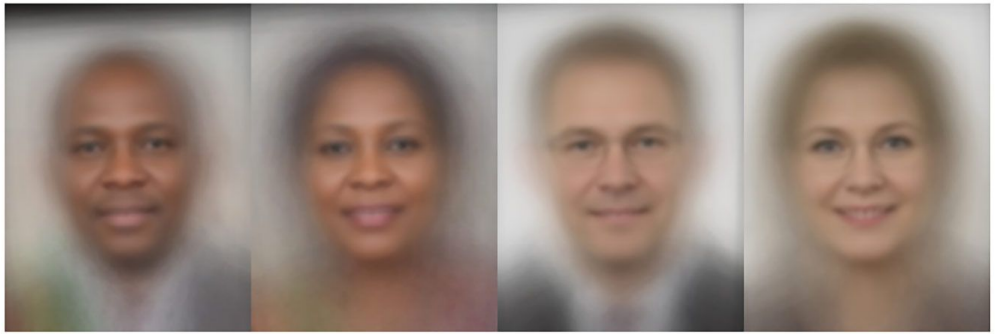
Weaknesses of group fairness



Kearns, M., Neel, S., Roth, A. and Wu, Z.S., 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*.

A problem in practice.

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%



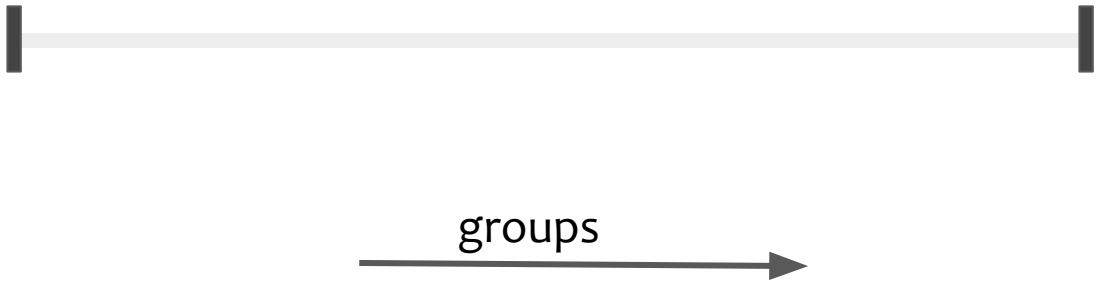
Gender shades: Intersectional Accuracy Disparities in Commercial Gender Classification Buolamwini, Gebru 2018

Not unique to machine learning.

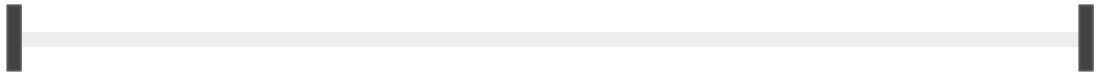
> “Crenshaw found that Black women in employment discrimination cases were unsuccessful, in part, because courts compared their claims against the experiences of similarly situated Black men (for racial discrimination cases) or white women (for gender discrimination)—both groups that, unlike Black women, enjoy systematic advantages along at least one historically-contingent dimension “

Hoffmann, A.L., 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society, 22(7), pp.900-915.

A spectrum of guarantees



A spectrum of guarantees

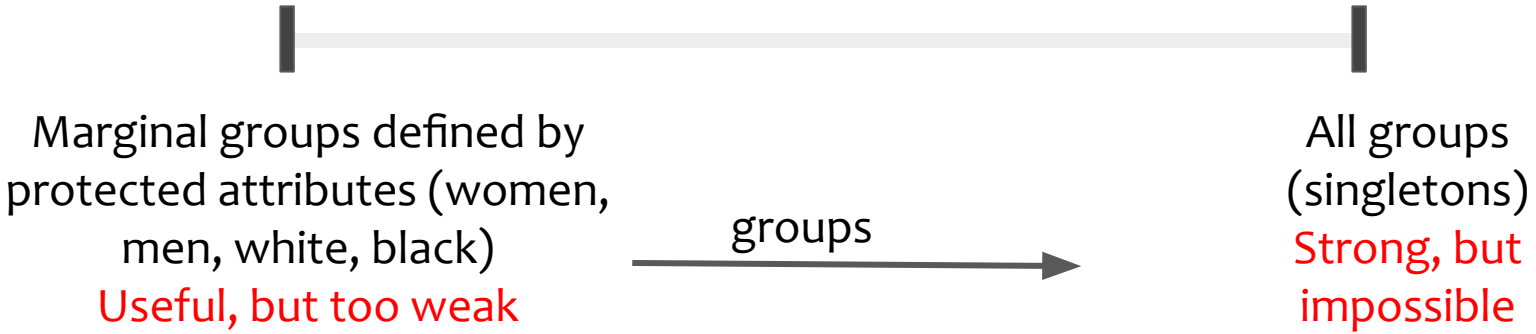


Marginal groups defined by protected attributes (women, men, white, black)

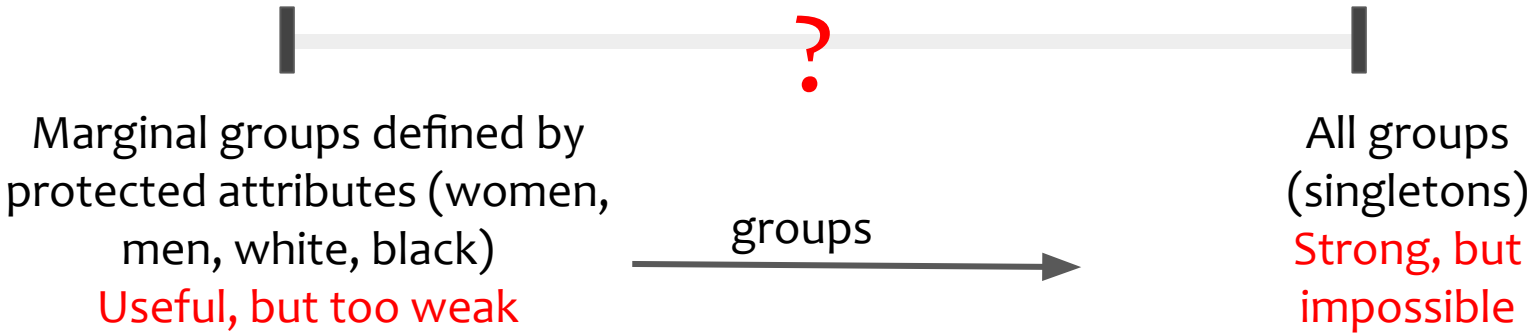
Useful, but too weak

groups

A spectrum of guarantees



What's in between?



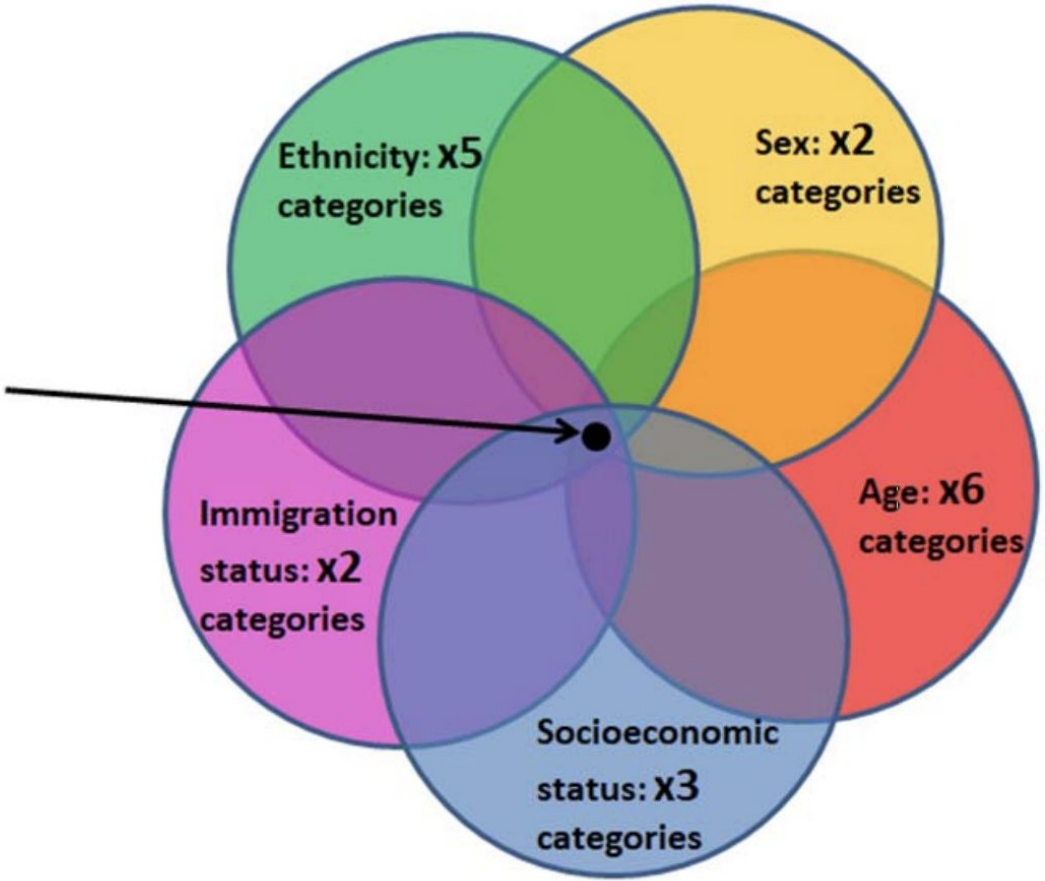
Between individual and group fairness

Clalit Research Institute

CVD prediction

Population	Predictors	Protected Variables
n = 1,021,041	Age	Age Group
Ages 40-79	Sex	Sex
No Previous CVD Event	LDL	Immigration Status
10 Year Follow-up	HDL	Socioeconomic Status
	Total Cholesterol	Ethnicity
	Systolic BP	
	Treated Hypertension	
	Smoking Status	
	Diabetes	

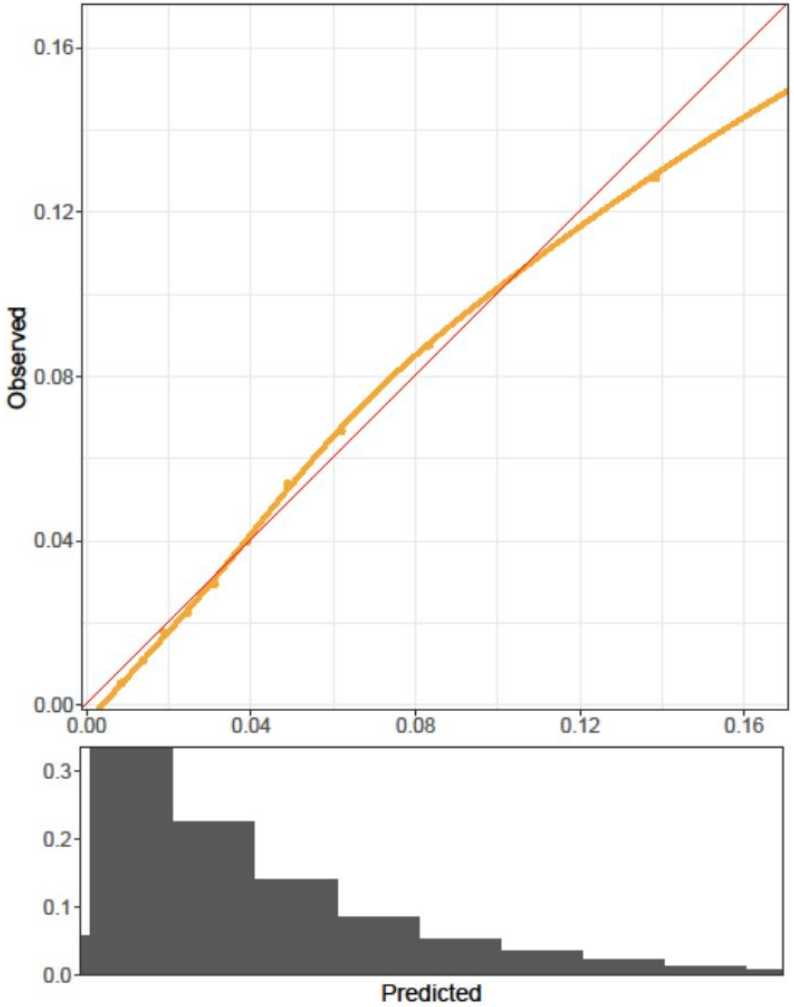
A subpopulation of Caucasian females, 50-59 years of age, of low socioeconomic status and a history of immigration



5·2·6·3·2 = 360 sub-groups!

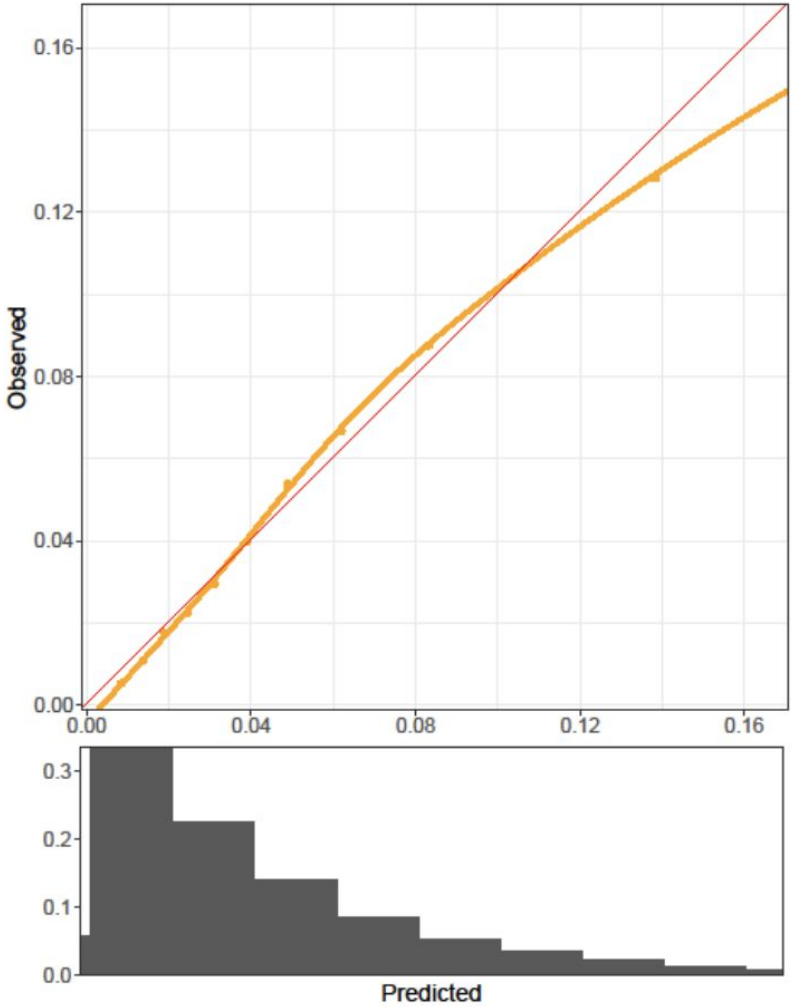
Harm: mis-calibration

"probabilities don't mean what they say they mean"

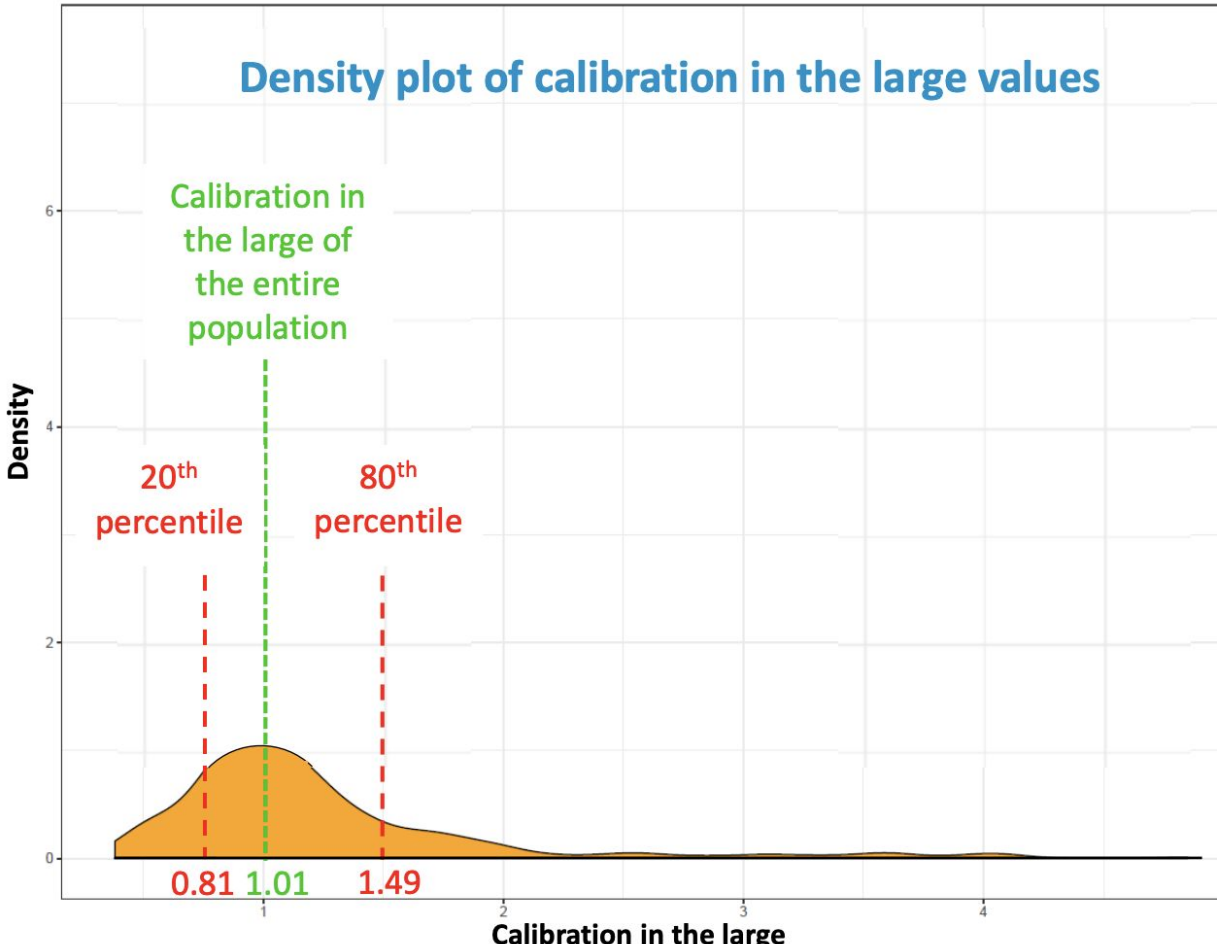


Harm: mis-calibration
"probabilities don't mean what they say they mean"

Baseline model is very well-calibrated, *globally*



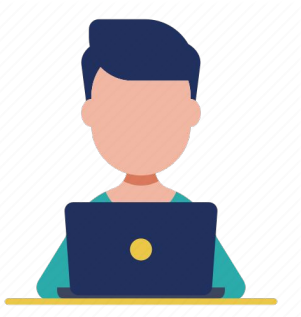
On a subgroup level, calibration is poor



Auditing versus learning

So far, we've only discussed auditing: *determining* whether a particular classifier is unfair; Now we're interested in the problem of finding a classifier that *isn't unfair*

Learner



Auditor

"Informal" theorem ([HKRR, KNRW]): two problems are computationally equivalent

Post-processing for subgroup calibration

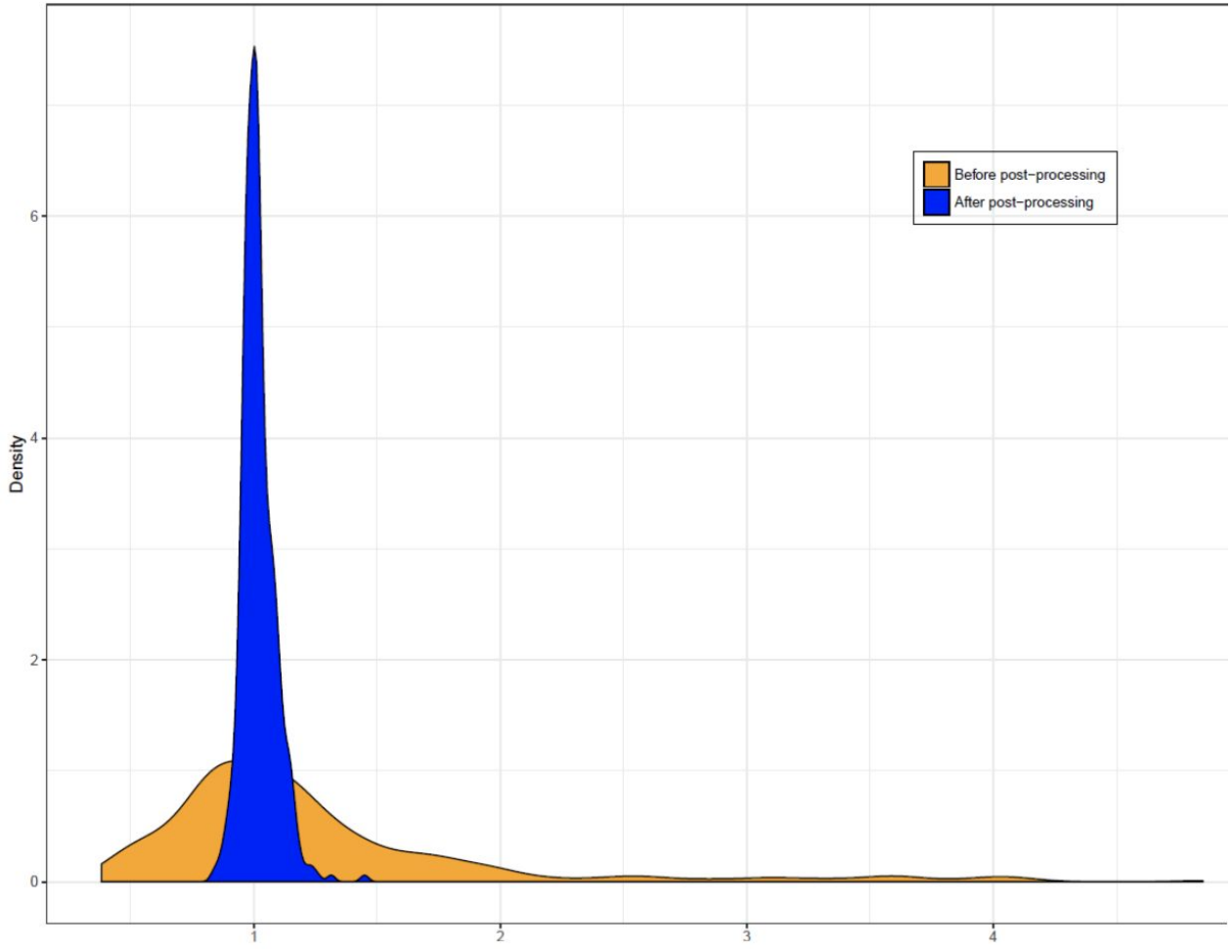
Input: An initial predictor p', a collection of subpopulations $\mathcal{C}$, a training set $D = \{(x_i, y_i)\}_{i=1}^{m}$ and a violation parameter $\alpha > 0$

Output: A post-processed predictor p satisfying $(\mathcal{C}, \alpha)$-multiaccuracy on D

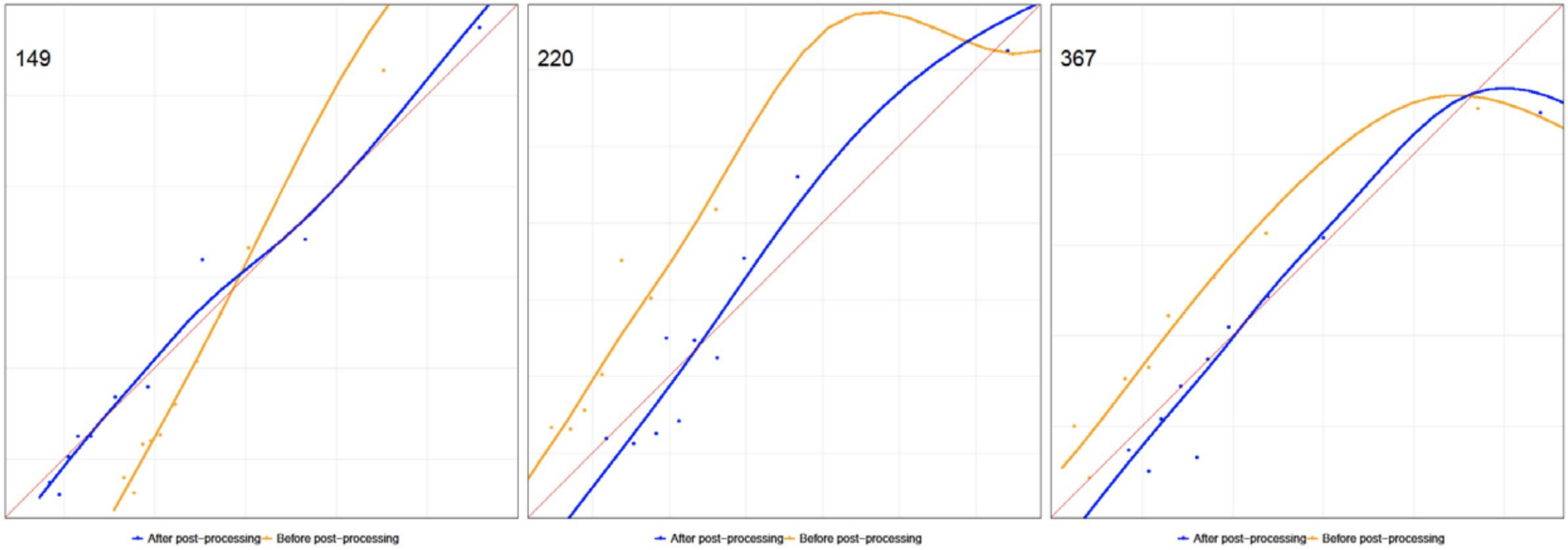
$p \leftarrow p'$
done $\leftarrow$ `False`

while $\neg$*done* **do**
  done $\leftarrow$ `True`
  foreach $S \in \mathcal{C}$ **do** // iterate over subpopulations
    $\Delta_S = \frac{1}{|S \cap D|} \cdot \left(\sum_{x_i \in S} y_i - \sum_{x_i \in S} p_i\right)$ // mag. of violation on S
    if $|\Delta_S| > \alpha$ **then**
      $p \leftarrow p + \Delta_S \cdot \mathbf{1}_S$ // update predictor by "nudging" S
      done $\leftarrow$ `False`
    end if
  end foreach
end while
return p

Post-processing procedure reduces variance by 99%



Post-processing improves subgroup calibration



Practical Limitations

- For many protected attributes, approach quickly become infeasible
  - Computational
  - Statistical

- What if protected attributes aren't explicitly present in my data?

Between individual and group fairness

A different approach: take **all groups with "short description"** (parameterized by **C**)

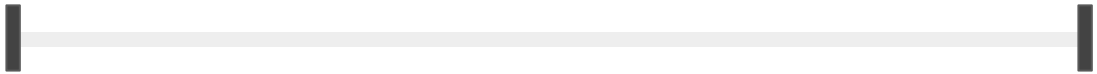
E.g.

a. Conjunctions, up to three attributes
b. Decision trees of up to depth 5

Description of groups is *implicit* rather than explicit

Between individual and group fairness

Group {me} *doesn't have* a short description



Marginal groups defined by protected attributes (women, men, white, black)
Useful, but too weak

groups

All groups (singletons)
Strong, but impossible

Revisiting the limitations

Revisiting the limitations

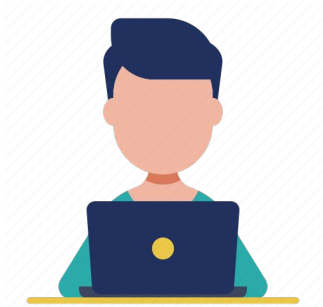
(1) For many protected attributes, approach quickly become infeasible

Revisiting the limitations

(1) For many protected attributes, approach quickly become infeasible

Before:
enumerating over exp. many groups
Now:
fit decision tree to residuals

Learner



Auditor

Revisiting the limitations

(2) What if protected attributes aren't explicitly present in my data?

If data is rich: group "women" might be effectively included even if not an explicit feature: it may be inferred from other features

useful when storing/collection private information is prohibited

Summary

1. Discrimination is one of several “societal concerns” in modern machine learning
2. CS perspective: what’s a good, strong definition?
    a. Group notions of fairness
    b. Ways they can be abused
    c. Subgroup fairness notions

Lots of data related issues...

In general, auditing and learning fairly require a lot of sensitive information

- Privacy?
- Incentives?

Subgroup fairness:

- Strength of guarantee depends on the richness of the data!
- how can we make sure C includes "important groups"?

Thanks!

References

1. Kearns, M., Neel, S., Roth, A. and Wu, Z.S., 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness.
2. Hébert-Johnson, U., Kim, M.P., Reingold, O. and Rothblum, G.N., 2017. Calibration for the (computationally-identifiable) masses.
3. Kim, M.P., Ghorbani, A. and Zou, J., 2019, January. Multiaccuracy: Black-box post-processing for fairness in classification.
4. Barda, N., Dagan., N., Addressing Fairness in Prediction Models by Improving Subpopulation Calibration"

Additional resources

- ethical algorithm book (privacy,)

- fairness in ml book (economic, causality)

- Aaron's survey (recent advances from an academic perspective)

- fat*? (Community that brings together interdisciplinary research)

- talk to me!