CSCI 699: Trustworthy ML (from an optimization lens)

Vatsal Sharan

Fall 2025

Lecture 6, Oct 1

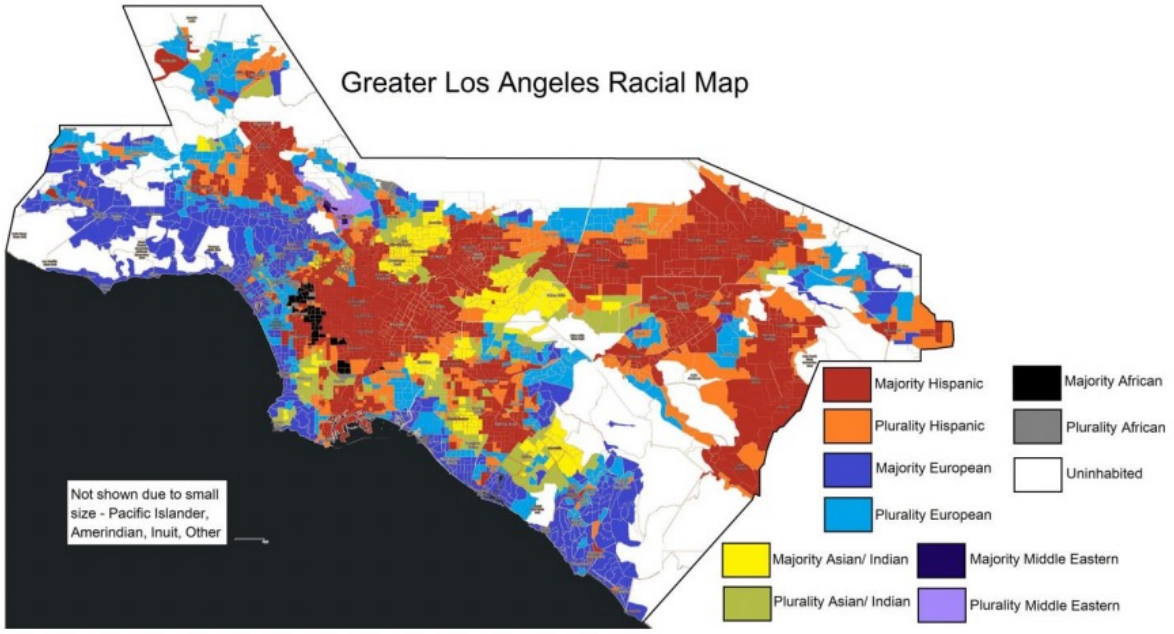
USC University of Southern California


Bennett
The Christian Science Monitor
JUSTICE
DEPARTMENT

Recap

How to obtain fair classifiers?

Observation: No fairness by just excluding sensitive attributes
Why? Sensitive attribute can often be reconstructed from other features



Zip code has a lot of information about race

Ensuring fairness in classification: **Group & Individual fairness notions**

Two broad classes of fairness notions in classification:

Individual fairness: Algorithm treats **similar individuals similarly**

Group fairness: Algorithm is **"unbiased" on protected groups** (such as race, gender etc.)

Individual fairness

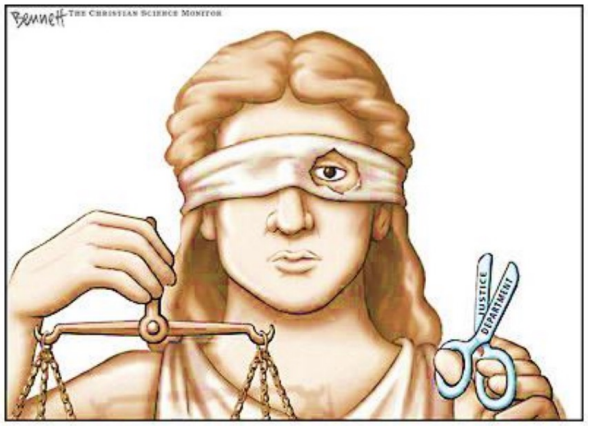
Define a ***metric*** $d(x, x')$ for the similarity between any two individuals x and x'.

e.g.: $d(x, x') = \| x - x' \|_2$

If classifier predicts $p(x)$ as the probability of label being one for x, if

$$|p(x) - p(x')| \leq \mu \, d(x, x'),$$

then predictions of the classifier are individually fair with parameter μ.



If these two individuals are similar, then their risk scores should be similar.

Fairness Through Awareness. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, Richard Zemel. 2011

Group fairness

Group fairness notions require that the models predictions obey certain properties over protected groups (e.g. by race, gender).

Many different notions have been proposed

- Statistical parity
- Equalized odds
- Calibration across groups

Statistical parity/Demographic parity

Binary classification setup (e.g. admitting a student to a degree program)

- Classifier f
- Datapoint (x, y)
- Sensitive attribute $a \in \{0,1\}$

Statistical parity (also known as demographic parity): $\Pr_x[f(x) = 1 \mid a = 1] = \Pr_x[f(x) = 1 \mid a = 0]$

In words: **Predictions are independent of sensitive attribute**

E.g., admit equal fraction of men or women into program

Can be too strong if labels and sensitive attribute are not independent.

E.g. if one demographic is more likely to be qualified than the other

Equalized odds & Equality of opportunity

Same binary classification setup (e.g. admitting student to degree program)

- Classifier f
- Datapoint (x, y)
- Sensitive attribute $a \in \{0,1\}$

Equalized odds: Following 2 constraints are satisfied

1. $\Pr_{x,y}[f(x) = 1 \mid a = 1, y = 1] = \Pr_{x,y}[f(x) = 1 \mid a = 0, y = 1]$
2. $\Pr_{x,y}[f(x) = 0 \mid a = 1, y = 0] = \Pr_{x,y}[f(x) = 0 \mid a = 0, y = 0]$

Just having 1st constraint (for $y = 1$) gives **equality of opportunity**

Equivalently:
Recall for $y = 1$ is the same for both groups (1st condition)
False positive rate (**FPR**) is the same for both groups (2nd condition)

Recall for class 1 = $\Pr_{x,y}[f(x) = 1 \mid y = 1]$

FPR = $\Pr_{x,y}[f(x) = 1 \mid y = 0]$

Also equivalent to saying: Conditioned on label, prediction is independent of sensitive attribute

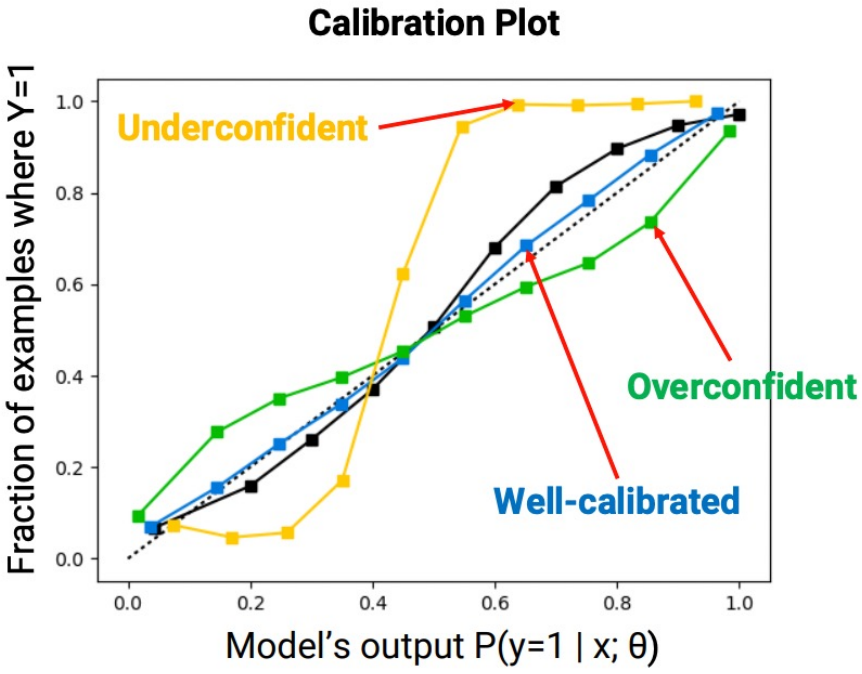
From *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

Calibration

Calibration: A model f for binary classification is calibrated if

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha] = \alpha$$

Informally, this says that “predictions mean what they should”



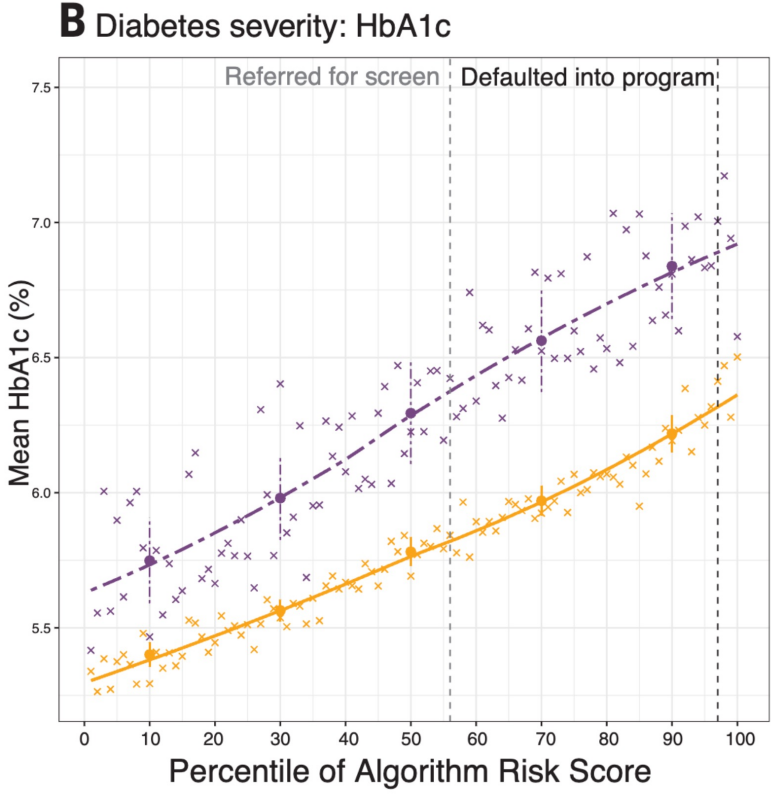
This is known as a ***reliability diagram***

Calibration across groups

A model f for binary classification is calibrated for groups defined by sensitive attribute a if

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 1] = \alpha\,,$$
$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 0] = \alpha.$$

Informally, this says that “predictions mean what they should for each group”



Achieving these notions: Post-processing for statistical parity

Consider binary classification setup (e.g. admitting a student to a degree program)

- Predictor p which predicts a score in $[0,1]$ (higher score => higher probability of label 1)
- Datapoint (x, y)
- Sensitive attribute $a \in \{0,1\}$

We want use p to get classifier f which maximizes accuracy but obeys statistical parity:

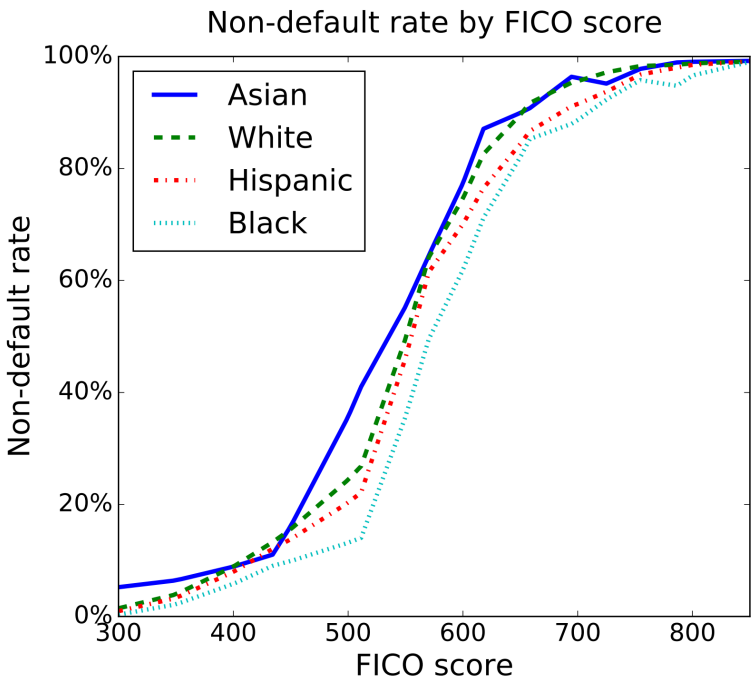
$$\Pr_x[f(x) = 1 \mid a = 1] = \Pr_x[f(x) = 1 \mid a = 0]$$

- What would f be if we only wanted to maximize accuracy? (Suppose $p(x) = \Pr[y = 1|x]$)
  Threshold the predictions at 0.5!

- How to ensure statistical parity? ***Threshold the predictions appropriately!***

The input data: FICO scores and their distribution by race

- FICO scores based on 300k TransUnion scores from 2003
- Range from 300-850, aim to predict risk of defaulting on a loan
- 620 is common threshold for good loan rates. Corresponds to 82% non-default rate in the data



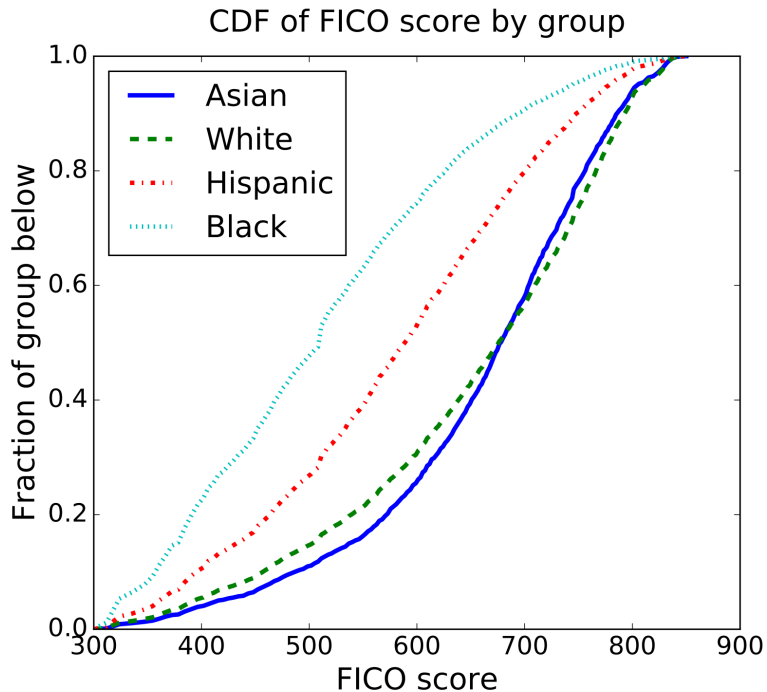
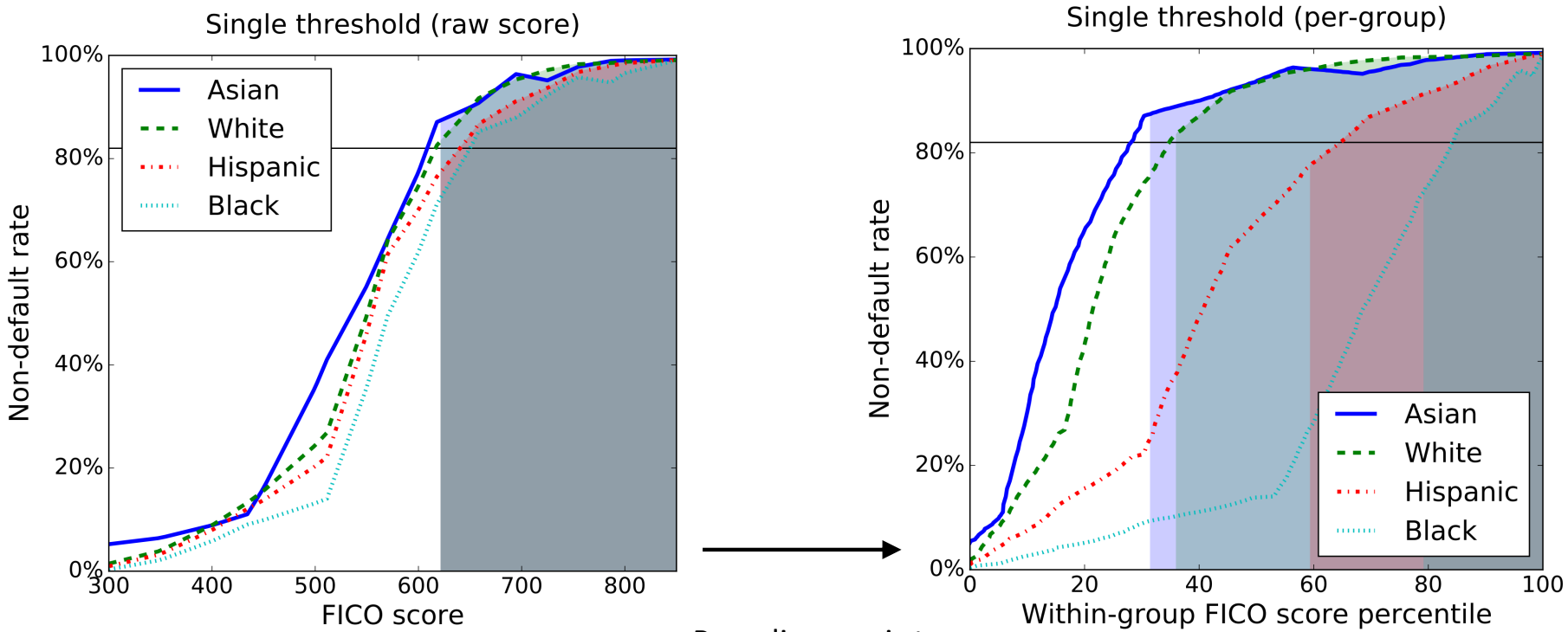


Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

What happens when we pick a single threshold?

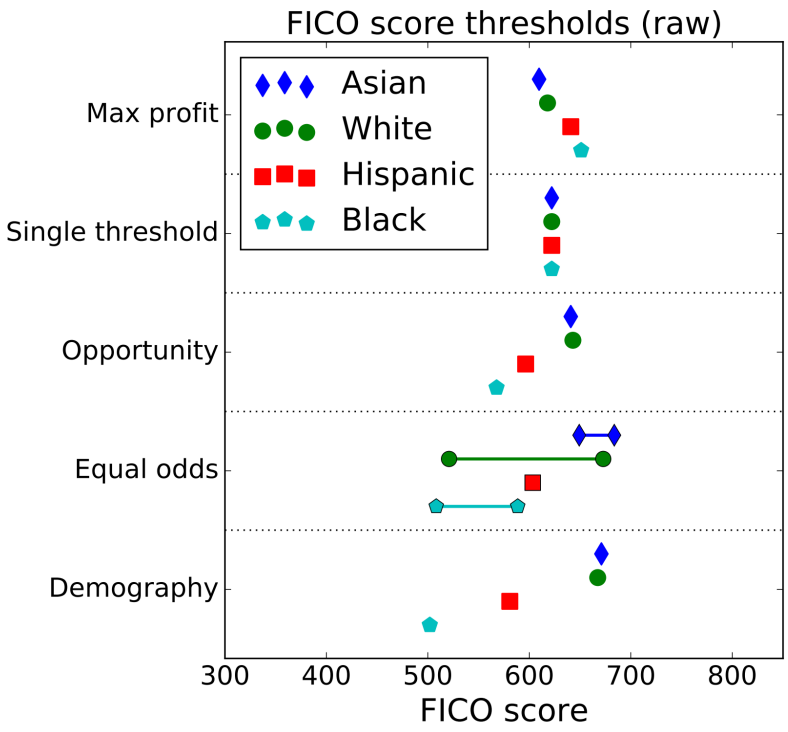


Rescaling x-axis to represent within group thresholds

Claim: $\Pr[f(x) = 1 \mid y = 1, a]$ is fraction of area under the curve that is shaded

Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

What are thresholds for different rules?



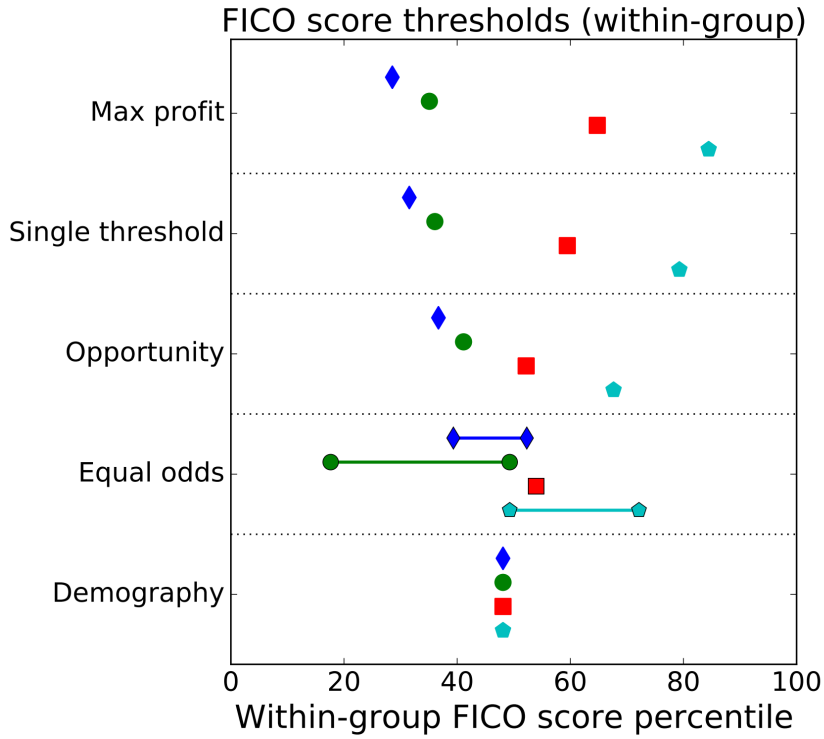
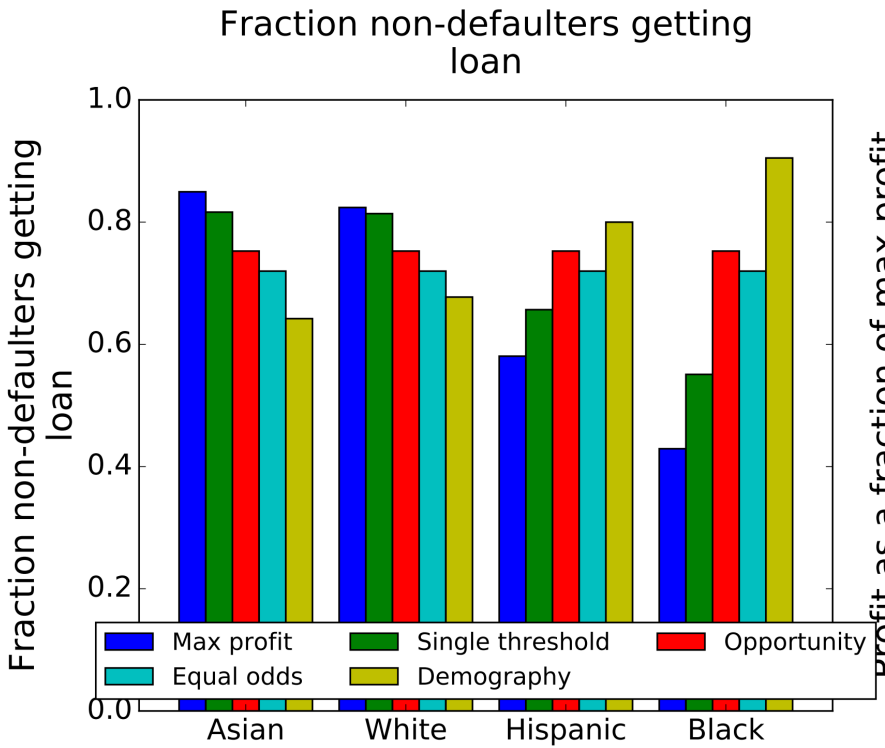


Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

What fairness and utility do different rules obtain?



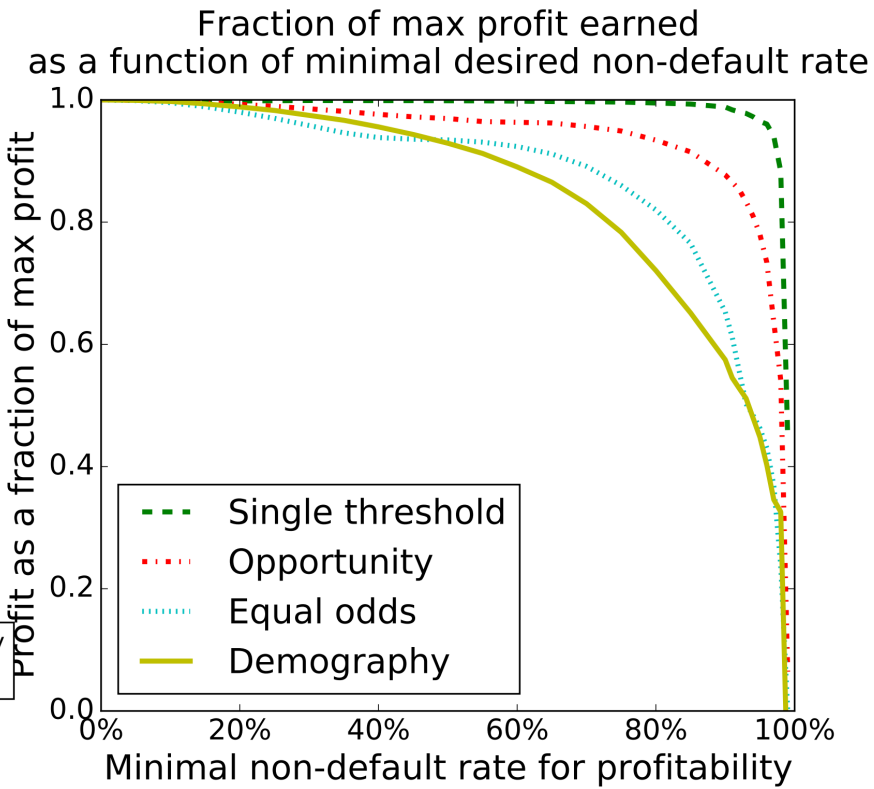
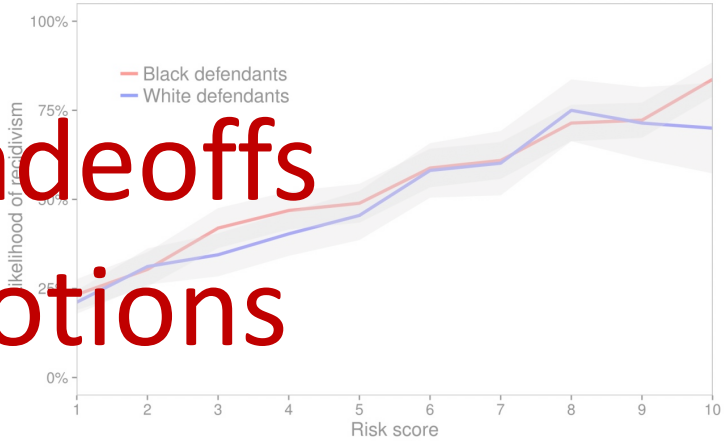


Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

Inherent tradeoffs between notions



Consider three basic fairness notions

- **Calibration for groups:**

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, x \in g_0] = \alpha\,,$$
$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, x \in g_1] = \alpha.$$

- **Balance for positive class:** The average score received by individuals with $y = 1$ is the same for the two groups.

$$\frac{\sum_{(x,y):\, x \in g_0,\, y=1} f(x)}{|(x,y): x \in g_0,\ y = 1|} = \frac{\sum_{(x,y):\, x \in g_1, y=1} f(x)}{|(x,y): x \in g_1,\ y = 1|}$$

- **Balance for negative class:** The average score received by individuals with $y = 0$ is the same for the two groups.

$$\frac{\sum_{(x,y):\, x \in g_0, y=0} f(x)}{|(x,y): x \in g_0,\ y = 0|} = \frac{\sum_{(x,y):\, x \in g_1, y=0} f(x)}{|(x,y): x \in g_1,\ y = 0|}$$

Except in special cases, impossible to achieve all 3!!

Let the base rate for group $g_i = \Pr[y = 1 \mid \ x \in g_i\ ]$.

Theorem (Kleinberg-Mullainathan-Raghavan '16): Consider any predictor f which satisfies (1) calibration across groups, (2) balance for positive class, and (3) balance for negative class. Then,

- either $f(x)$ is a perfect predictor, with $f(x) \in \{0,1\}$,
- or the two groups have equal base rates.

Similar result is also true for classifiers which satisfy approximate versions of the 3 fairness conditions.

Inherent tradeoff in fairness, proof sketch

Step 2: Relate to balance conditions

N_t = # people in group t
M_t = # qualified people in group t
$\Big\}$ $\frac{M_t}{N_t} = \alpha_t$

Let x_t be avg. score given to -ve class in group t

Let y_t " " +ve class "

$(N_t - M_t)\, x_t + M_t y_t = M_t$ (under calibration)

$\left(1 - \frac{M_t}{N_t}\right) x_t + \left(\frac{M_t}{N_t}\right) y_t = \frac{M_t}{N_t}$

$(1-\alpha_t)\, x_t + \alpha_t y_t = \alpha_t$

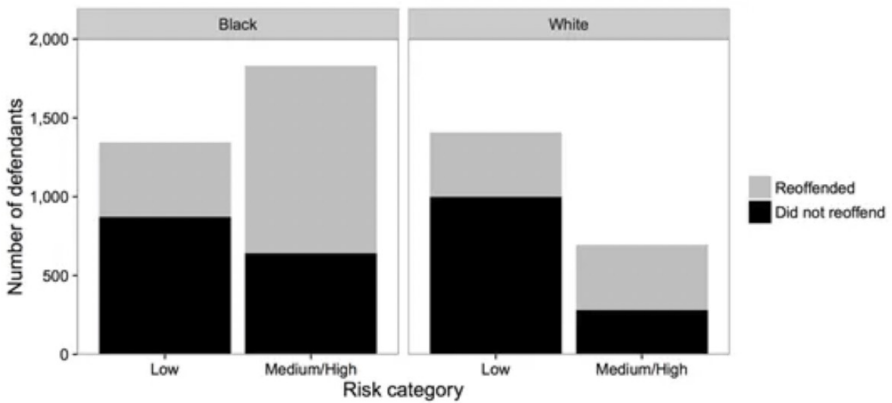
$\Big\}$ $(1-\alpha_1)x + \alpha_1 y = \alpha_1$
$(1-\alpha_2)x + \alpha_2 y = \alpha_2$

By ② & ③, $x_t = x \;\forall t$ & $y_t = y \;\forall t$

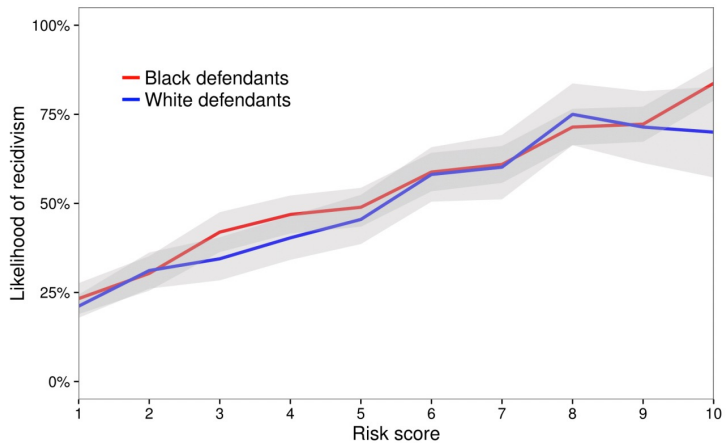
What does this say about COMPAS?

- The recidivism rate among black defendants in the data is 51%, compared to 39% for White defendants.
- Classifier is far from perfect (~60% accuracy)

COMPAS: Unfair because black defendants who did not recommit crime are assigned higher score (i.e. does not obey balance)



COMPAS: Fair because probability of recommitting crime is similar for a given risk score, for both groups (i.e. is calibrated)



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

Also see *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,* Chouldechova et al. '17

Individual fairness via uncertainty quantification, applied to rankings

Individual fairness defines fairness via a metric

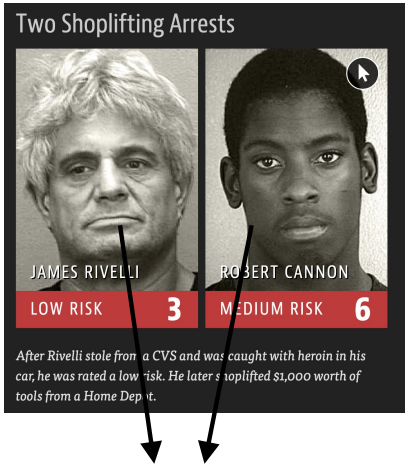
Define a ***metric*** $d(x, x')$ for the similarity between any two individuals x and x'.

e.g.: $d(x, x') = \| x - x' \|_2$

If classifier predicts $p(x)$ as the probability of label being one for x, if

$$|p(x) - p(x')| \leq \mu \, d(x, x'),$$

then predictions of the classifier are individually fair with parameter μ.



Individual fairness via a metric: If these two individuals are similar, then their risk scores should be similar.

However, it can be difficult to get access to this metric.

A solution:

1. Define the metric to be the distance between ground truth distribution
2. Now instead of attempting to measure similarity between individuals, obtain uncertainty in prediction for each individual
3. Use randomization to make decisions (so that discontinuities are not introduced due to thresholding)

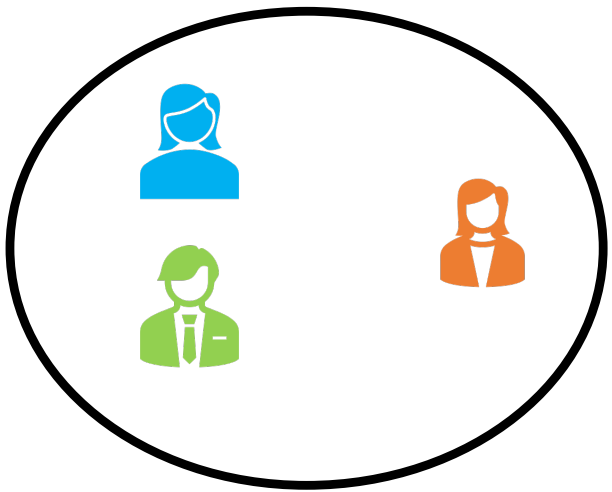
Individual fairness in rankings

Consider the simpler setting where we are trying to select one of these 3 candidates for a single job.

Candidate	GPA	Interview score	Work history
	3.5	excellent	2
	3.3	excellent	3
	3.8	good	1

Individual fairness:

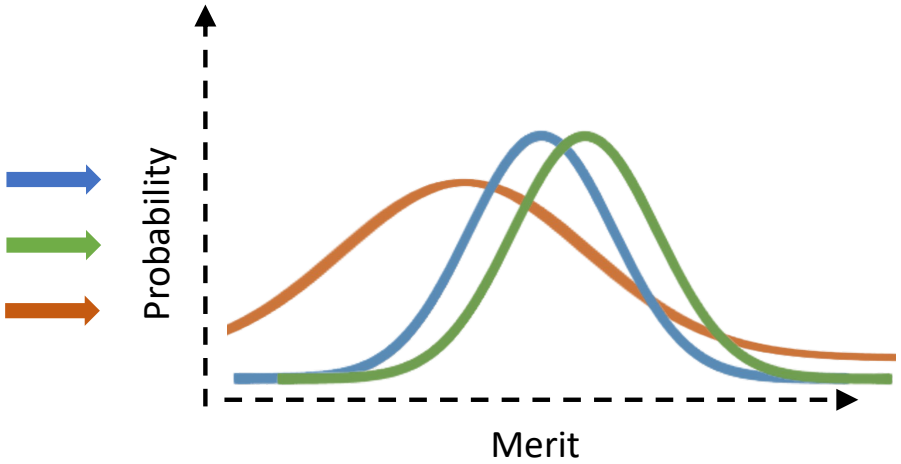
Observable features enforce a Lipschitz continuity condition on model via some metric



An alternate view on individual fairness

Singh-Kempe-Joachims'21: Observable features induce a posterior merit distribution of the individuals for the job

Candidate	GPA	Interview score	Work history
	3.5	excellent	2
	3.3	excellent	3
	3.8	good	1

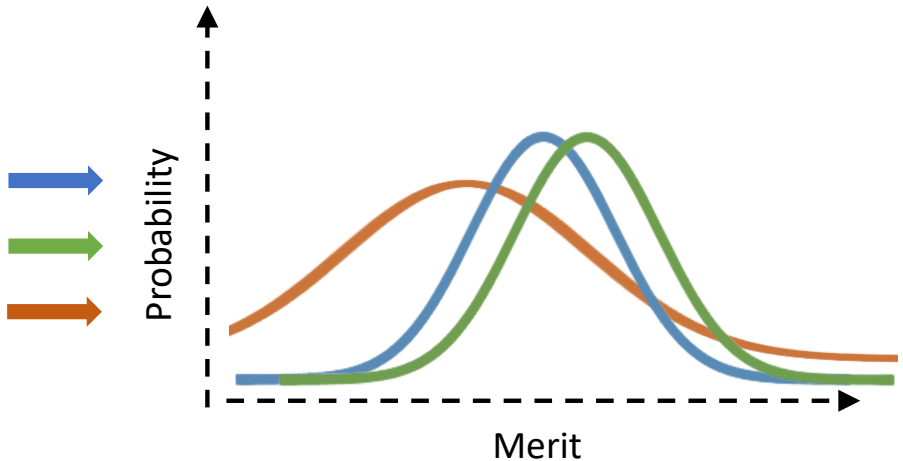


ML models can often give these distributions over merits

An alternate view on individual fairness

Singh-Kempe-Joachims'21: Observable features induce a posterior merit distribution of the individuals for the job

Candidate	GPA	Interview score	Work history
	3.5	excellent	2
	3.3	excellent	3
	3.8	good	1



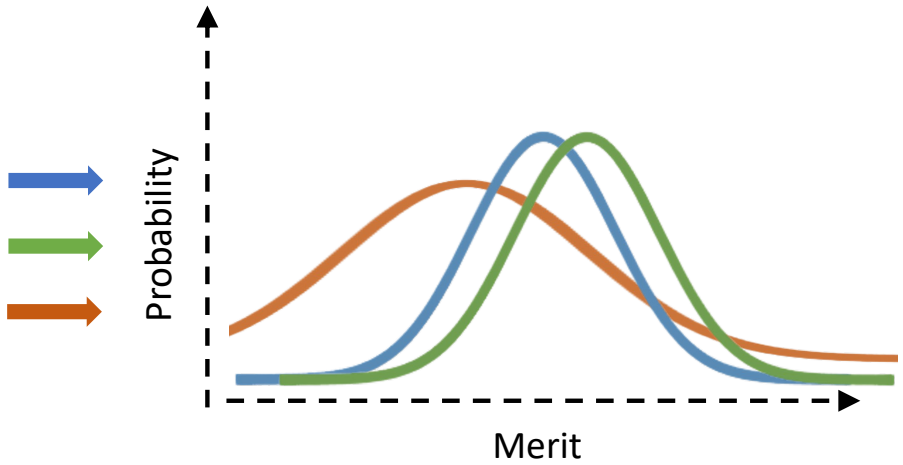
Suppose goal is to select one candidate.
Define utility as the expected merit of selected candidate.

How will a utility maximizing algorithm select candidates here?
What would individual fairness require?

An alternate view on individual fairness

Singh-Kempe-Joachims'21: Observable features induce a posterior merit distribution of the individuals for the job

Candidate	GPA	Interview score	Work history
	3.5	excellent	2
	3.3	excellent	3
	3.8	good	1



Uncertainty as the keystone of fairness:

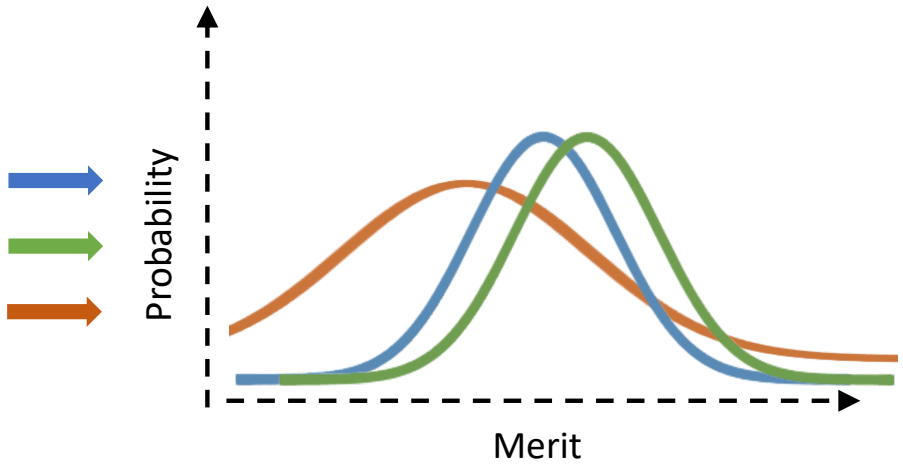
similar individuals should be treated similarly because similar observable features give rise to similar posterior distributions over merits.

If a candidate is the top candidate with prob. p, then they should get job with prob. p

An alternate view on individual fairness

Singh-Kempe-Joachims'21: Observable features induce a posterior merit distribution of the individuals for the job

Candidate	GPA	Interview score	Work history
	3.5	excellent	2
	3.3	excellent	3
	3.8	good	1

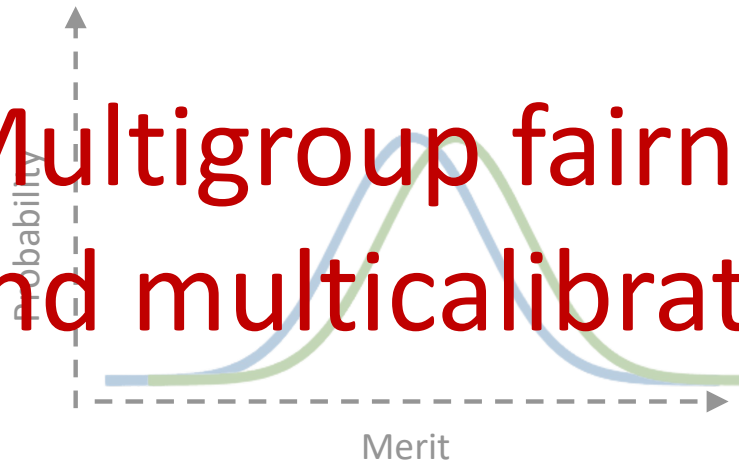


Proposed solution (for selecting one candidate):

Sample merits, measure probability of each candidate being top. If a candidate is top candidate with prob. p, they are selected with prob. p

Can also relax fairness requirement to tradeoff utility

Multigroup fairness and multicalibratoin

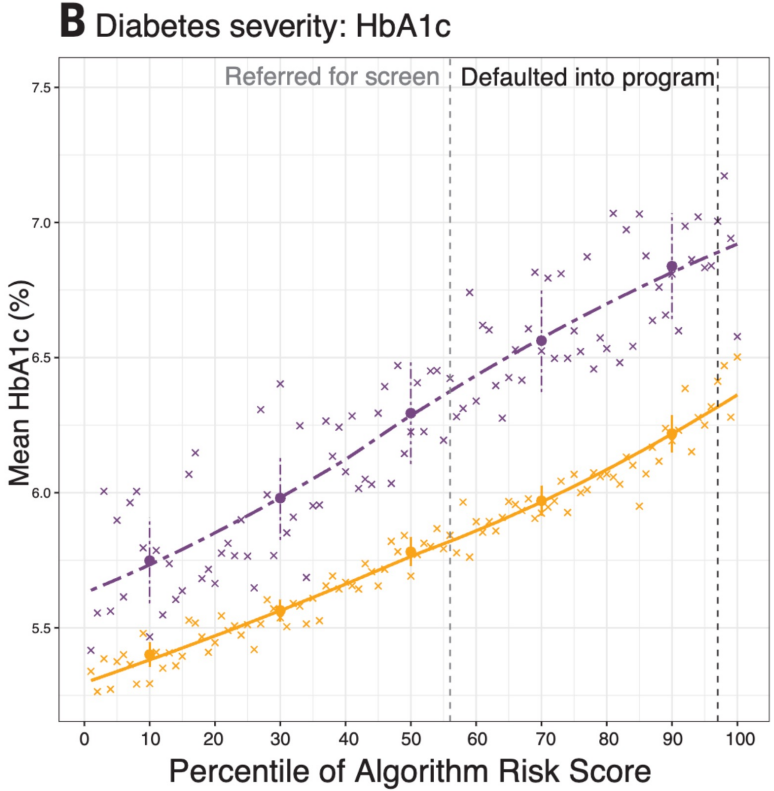


Earlier: Calibration across groups

A model f for binary classification is calibrated for groups defined by sensitive attribute a if

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 1] = \alpha\,,$$
$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 0] = \alpha.$$

Informally, this says that predictions mean what they should for each group.



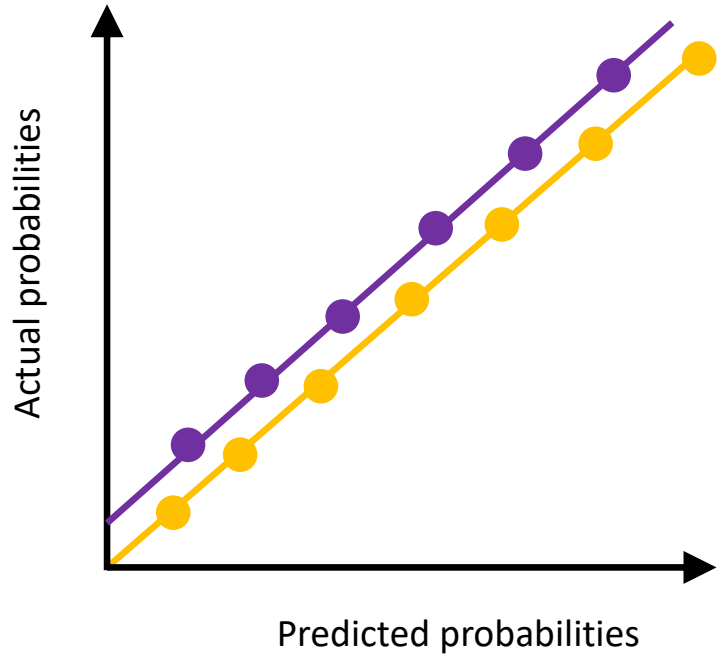
Existing medical risk predictors are miscalibrated across groups

Is calibration across groups enough?

A model f for binary classification is calibrated for groups defined by sensitive attribute a if

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 1] = \alpha\,,$$
$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 0] = \alpha.$$

Informally, this says that predictions mean what they should for each group.



Ideal scenario: Calibrated, meaningful predictions for both yellow and purple groups.

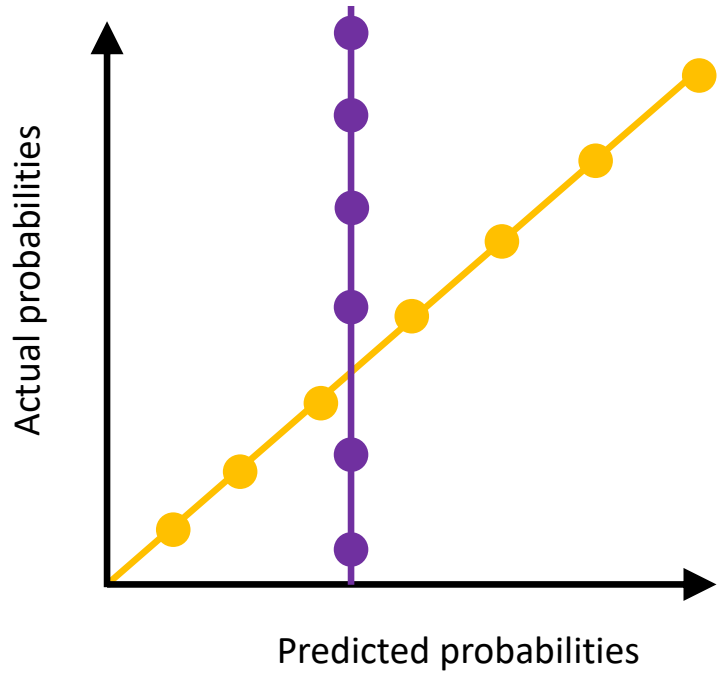
Fig. based on talk by Michael Kim, Multicalibration and Outcome Indistinguishability, Simons Institute

Is calibration across groups enough?

A model f for binary classification is calibrated for groups defined by sensitive attribute a if

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 1] = \alpha\,,$$
$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 0] = \alpha.$$

Informally, this says that predictions mean what they should for each group***, but they may not say much!***



Issue (**Algorithmic stereotyping**): Can achieve calibration on purple group by just predicting average on the group

Fig. based on talk by Michael Kim, Multicalibration and Outcome Indistinguishability, Simons Institute

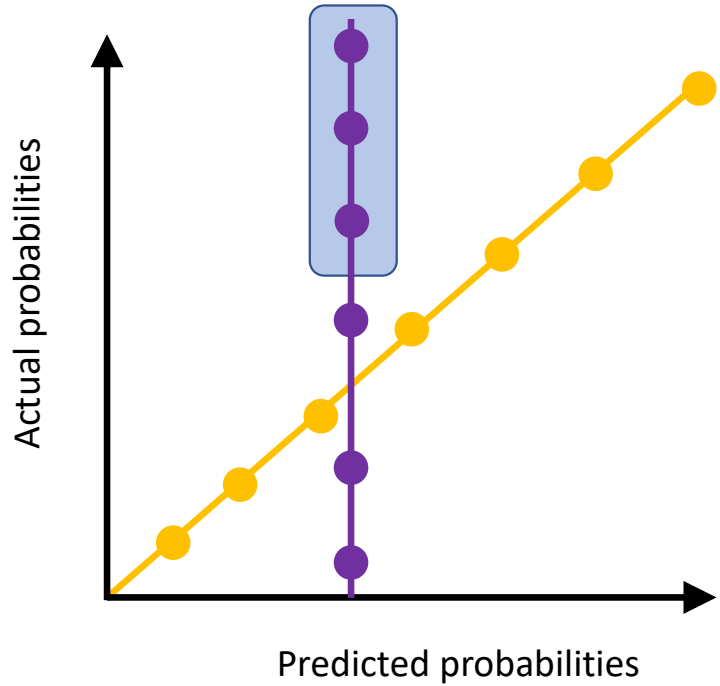
Calibration across rich set of groups

Idea: Let's try to get calibration on ***all*** groups in the data.

Can we get calibration on groups which correspond to single datapoints in the population?

No, since it would require estimating the ground truth probabilities accurately for each datapoint.

Instead: Can we capture some large, meaningful set of groups?



The model here is not calibrated on the individuals in the group in the blue box .

Fig. based on talk by Michael Kim, Multicalibration and Outcome Indistinguishability, Simons Institute

Multicalibration: Calibration for identifiable groups

Definition (Multicalibration, Hebert-Johnson et al. '18): A predictor f is multicalibrated with respect to a collection of groups C, if f is calibrated for every $c \in C$, i.e.

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, x \in c] = \alpha \ \forall \alpha \ .$$

Group calibration

- $C =$ small number of groups
- **Easy to achieve**
- **Weak, allows algorithmic stereotyping**

Individual fairness

- $C = \{\{x\} : x \in \text{domain}\}$
- **Cannot be efficiently achieved**
- **Strong individual-level fairness guarantee**

Multicalibration: Calibration for identifiable groups

Definition (Multicalibration, Hebert-Johnson et al. '18): A predictor f is multicalibrated with respect to a collection of groups C, if f is calibrated for every $c \in C$, i.e.

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, x \in c] = \alpha \ \forall \alpha \ .$$

Group calibration

- C = small number of groups
- **Easy to achieve**
- **Weak, allows algorithmic stereotyping**

Individual fairness

- $C = \{\{x\} : x \in \text{domain}\}$
- **Cannot be efficiently achieved**
- **Strong individual-level fairness guarantee**

Multi-calibration

- C = potentially infinite
- **Efficiently achievable (under certain conditions)**
- **Much stronger fairness protection than group calibration**

Multicalibration: Some nice properties

Definition (Multicalibration, Hebert-Johnson et al. '18): A predictor f is multicalibrated with respect to a collection of groups C, if f is calibrated for every $c \in C$, i.e.

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, x \in c] = \alpha \ \forall \alpha \ .$$

- **(Efficiently achievable)** If C is defined by a collection of groups such that it is easy to check for violation of calibration for $c \in C$, then can efficiently postprocess a predictor to be multicalibrated with respect to C (without hurting its accuracy).
- **(Gives predictors optimal for multiple downstream loss functions)** If f is multicalibrated, then is f is optimal with respect to multiple loss functions at the same time (up to postprocessing). This is called *omniprediction* (Gopalan et al. 2022).
- **(Gives predictors robust to distribution shifts)** If f is multicalibrated with respect to some groups, then can get group robustness for those groups (Kim et al. 2022).