Fairness in ML 2: Equal opportunity and odds

Privacy & Fairness in Data Science

CS848 Fall 2019

Slides adapted from https://fairmlclass.github.io/4.html

Outline

- Recap: Disparity impact
  - Issues with Disparate Impact
- Observational measure of fairness
  - Equal opportunity and Equalized odds
  - Predictive Value Parity
  - Tradeoff
- Achieving Equalized Odds
  - Binary Classifier

Recap: Disparate Impact

- Let $D=(X, Y, C)$ be a labeled data set, where $X = 0$ means protected, $C = 1$ is the positive class (e.g., admitted), and Y is everything else.
- We say that a classifier f has **disparate impact (DI) of $\boldsymbol{\tau}$** ($0 < \tau < 1$) if:

$$\frac{\Pr(f(Y) = 1 \mid X = 0)}{\Pr(f(Y) = 1 \mid X = 1)} \leq \tau$$

that is, if the protected class is positively classified less than τ times as often as the unprotected class. (legally, $\tau = 0.8$ is common).

Recap: Disparate Impact

Y (features) — *X (protected attribute)* — *f(Y) (prediction)*

X1	…	…	…	…	Race	Bail
0	…	0	1	…	1	1 (Y)
1	…	1	0	…	1	0 (N)
1	…	1	0	…	0	0 (N)
..	…	…	…	…	…	…

protected group (Race = 0)

$$P_{X=0}[E] = \Pr[E|X = 0] \quad P_{X=1}[E] = \Pr[E|X = 1]$$

Recap: Disparate Impact

Y (features) — *X (protected attribute)* — *f(Y) (prediction)*

X1	…	…	…	…	Race	Bail
0	…	0	1	…	1	1 (Y)
1	…	1	0	…	1	0 (N)
1	…	1	0	…	0	0 (N)
..	…	…	…	…	…	…

protected group (Race = 0)

Classifier f has DI of τ:

$$\frac{P_{X=0}[f(Y)=1]}{P_{X=1}[f(Y)=1]} \leq \tau$$

Demographic parity (or the reverse of disparate impact)

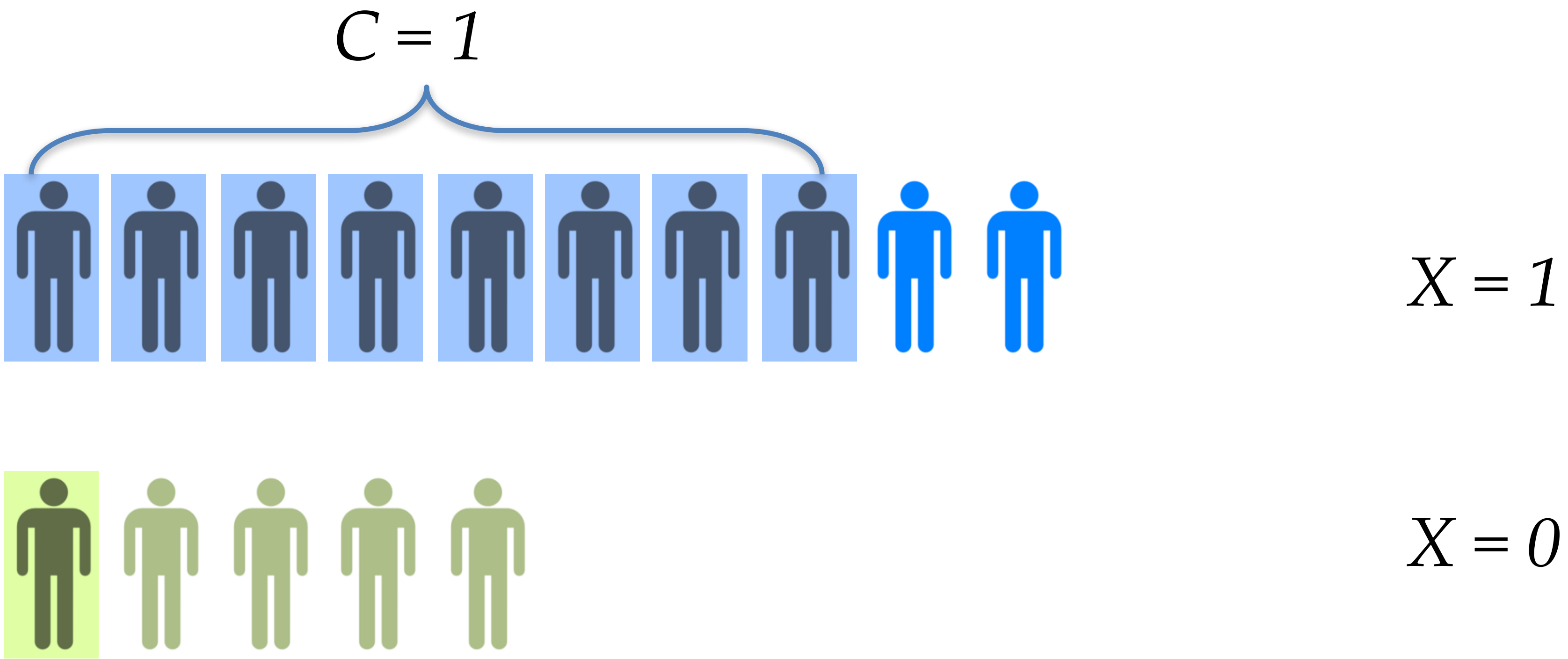
- Definition. Classifier f satisfies **demographic parity** if f is independent of X

- When f is binary 0/1-variables, this means, for all groups x and x',

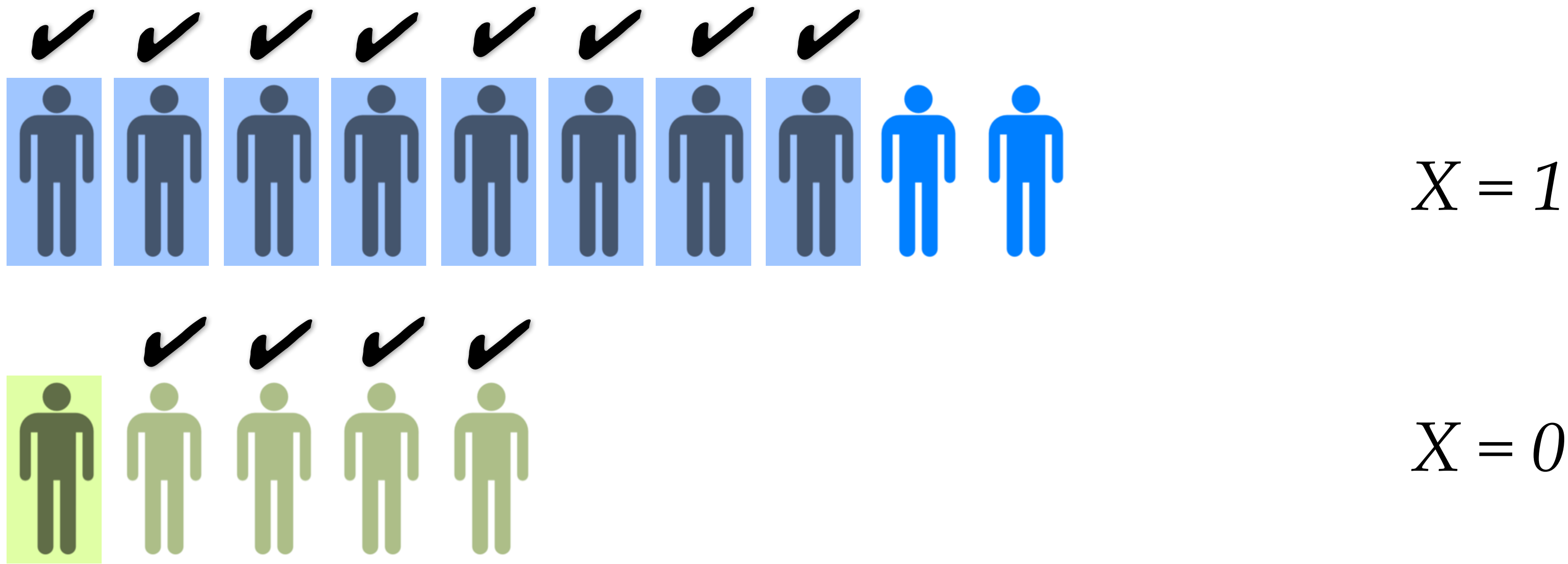
$$P_{X=x}[f(Y) = 1] = P_{X=x'}[f(Y) = 1]$$

- Approximate versions:
  - $\frac{P_{X=x}[f(Y)=1]}{P_{X=x'}[f(Y)=1]} \geq 1 - \epsilon$
  - $\left|P_{X=x}[f(Y) = 1] - P_{X=x'}[f(Y) = 1]\right| \leq \epsilon$

Demographic parity Issues



Demographic parity Issues



- Does not seem “fair” to allow random performance on X = 0
- Perfect classification is impossible

Outline

- ~~Recap: Disparity impact~~
  - ~~Issues with Disparate Impact~~

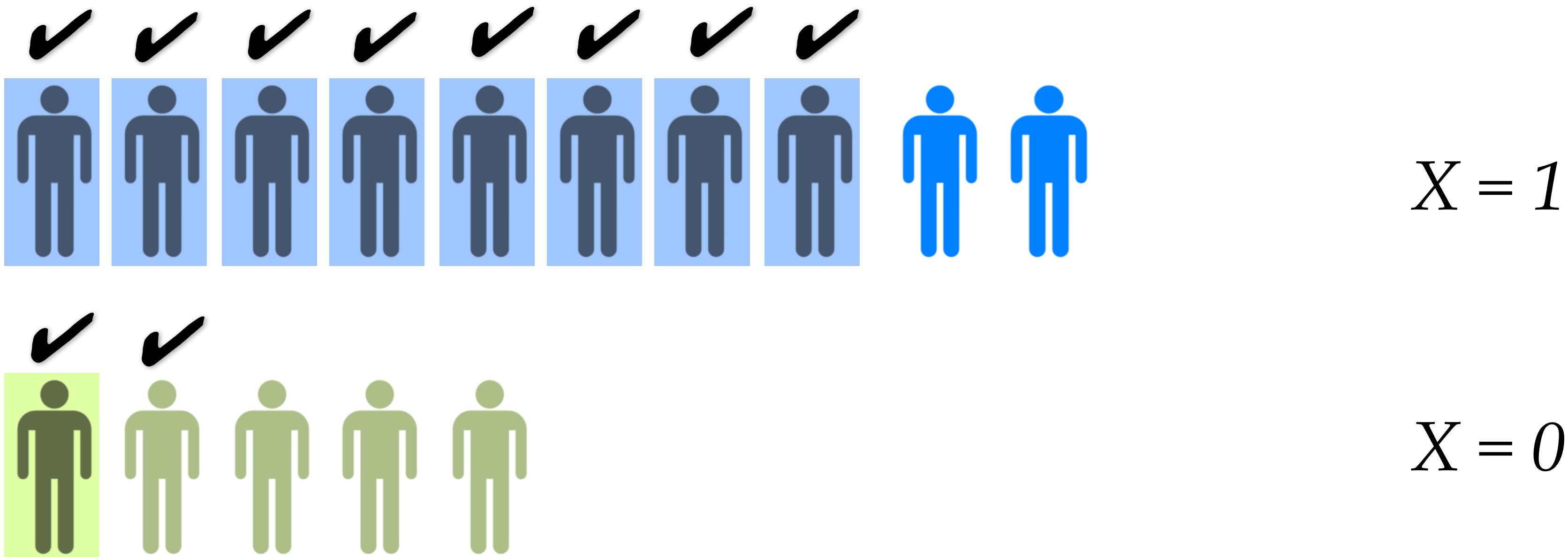
- Observational measure of fairness
  - Equal opportunity and Equalized odds
  - Predictive Value Parity
  - Tradeoff

- Achieving Equalized Odds
  - Binary Classifier

True Positive Parity (TPP)
(or equal opportunity)

- Assume classifier *f* and label *C* are binary 0/1-variables
- Definition. Classifier *f* satisfies **true positive parity** if for all groups x and x',
  $$P_{X=x}[f(Y) = 1 | C = 1] = P_{X=x'}[f(Y) = 1 | C = 1]$$
- When positive outcome (1) is desirable
- Equivalently, primary harm is due to false negatives
  - Deny bail when person will not recidivate

TPP



- Forces similar performance on C = 1

False Positive Parity (FPP)

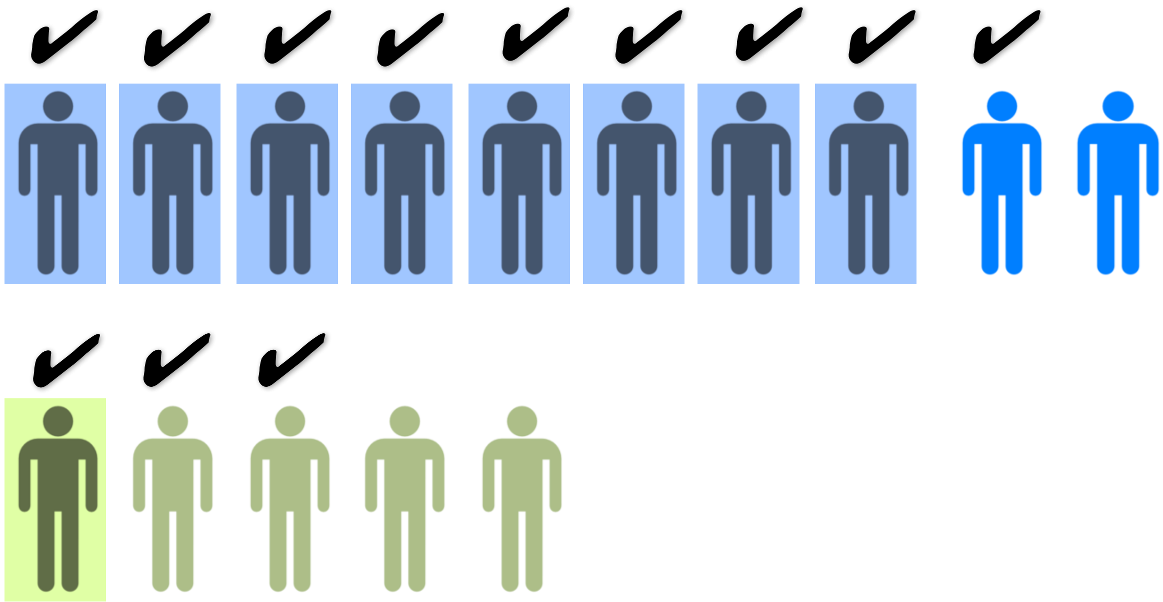
- Assume classifier *f* and label *C* are binary 0/1-variables

- Definition. Classifier *f* satisfies **false positive parity** if for all groups x and x',
$$P_{X=x}[f(Y) = 1|C = 0] = P_{X=x'}[f(Y) = 1|C = 0]$$

- TPP & FPP: **Equalized Odds**, or **Positive Rate Parity**

f satisfies equalized odds if
f is conditionally independent of ***X*** *given* ***C****.*

Positive Rate Parity



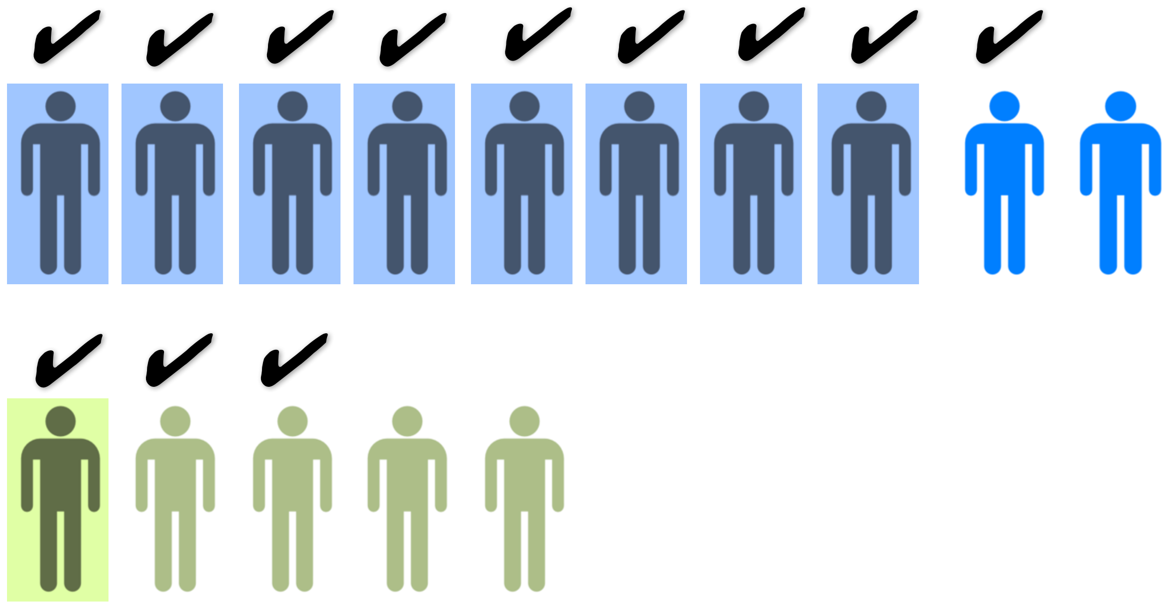
$X = 1$

$X = 0$

$P_{X=1}[f(Y) = 1 \mid C = 1] =?$ $P_{X=1}[f(Y) = 1 \mid C = 0] =?$

$P_{X=0}[f(Y) = 1 \mid C = 1] =?$ $P_{X=0}[f(Y) = 1 \mid C = 0] =?$

Positive Rate Parity



$X = 1$

$X = 0$

$$P_{X=1}[f(Y) = 1 \mid C = 1] = 1 \quad P_{X=1}[f(Y) = 1 \mid C = 0] = 1/2$$

$$P_{X=0}[f(Y) = 1 \mid C = 1] = 1 \quad P_{X=0}[f(Y) = 1 \mid C = 0] = 1/2$$

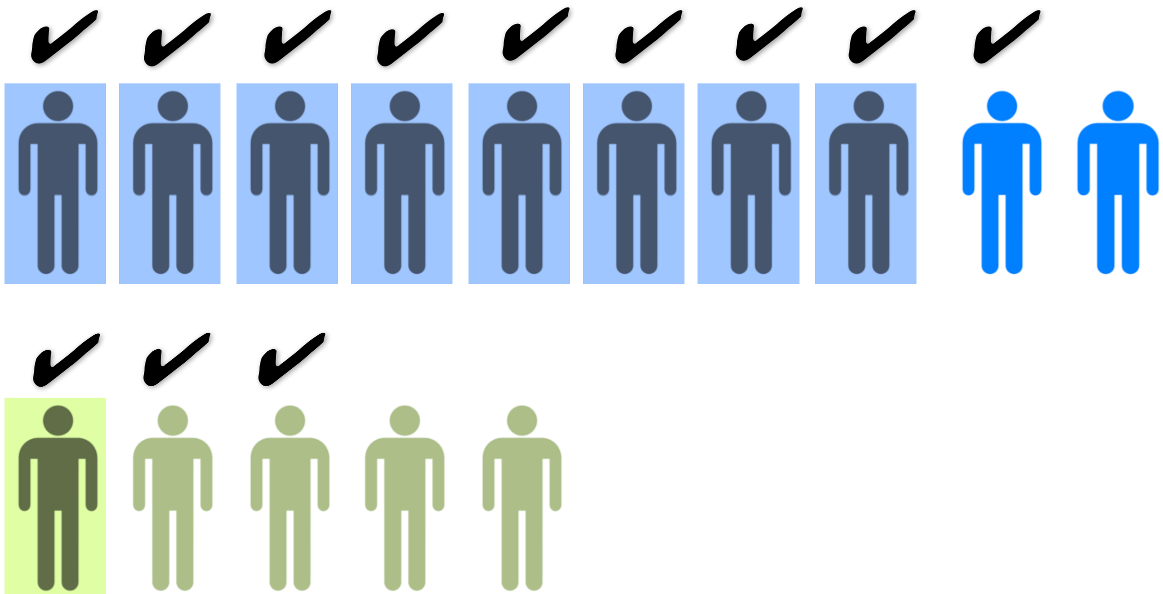
Outline

- ~~Recap: Disparity impact~~
  - ~~Issues with Disparate Impact~~
- Observational measure of fairness
  - ~~Equal opportunity and Equalized odds~~
  - Predictive Value Parity
  - Tradeoff
- Achieving Equalized Odds
  - Binary Classifier

Predictive Value Parity

- Assume classifier f and label C are binary 0/1-variables
- Definition. Classifier f satisfies
  - **positive predictive value parity if** if for all groups x and x',
    $P_{X=x}[C = 1|f(Y) = 1] = P_{X=x'}[C = 1|f(Y) = 1]$
  - **negative predictive value parity if** if for all groups x and x',
    $P_{X=x}[C = 1|f(Y) = 0] = P_{X=x'}[C = 1|f(Y) = 0]$
  - predictive value parity if satisfies both of the above.
- Equalized chance of success given acceptance.

Predictive Value Parity



$X = 1$

$X = 0$

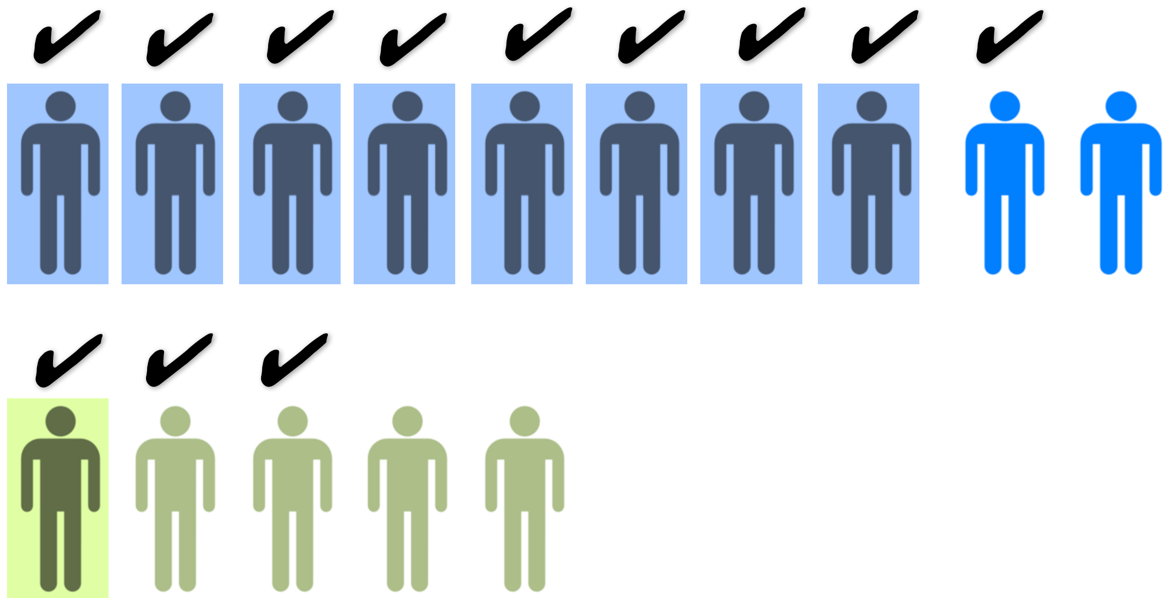
$P_{X=1}[C = 1 \mid f(Y) = 1] =$

$P_{X=1}[C = 1 \mid f(Y) = 0] =$

$P_{X=0}[C = 1 \mid f(Y) = 1] =$

$P_{X=0}[C = 1 \mid f(Y) = 0] =$

Predictive Value Parity



$X = 1$

$X = 0$

$P_{X=1}[C = 1 \mid f(Y) = 1] = 8/9$ $\quad$ $P_{X=1}[C = 1 \mid f(Y) = 0] = 0$

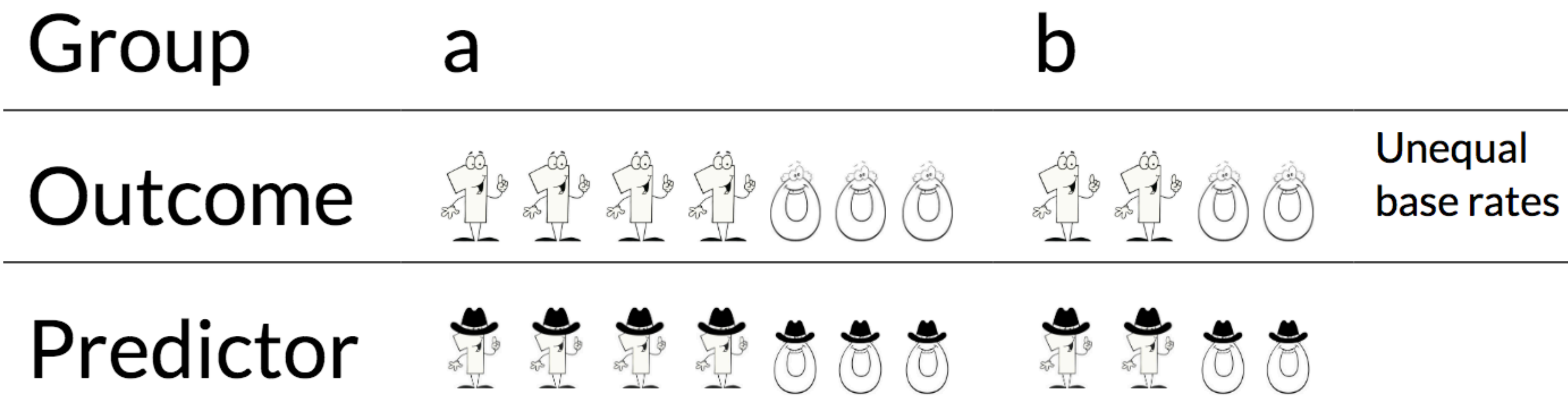
$P_{X=0}[C = 1 \mid f(Y) = 1] = 1/3$ $\quad$ $P_{X=0}[C = 1 \mid f(Y) = 0] = 0$

Trade-off

- Proposition. Assume differing base rates and an imperfect classifier $f \neq C$. Then either
  - Positive rate parity fails, or
  - Predictive value parity fails.

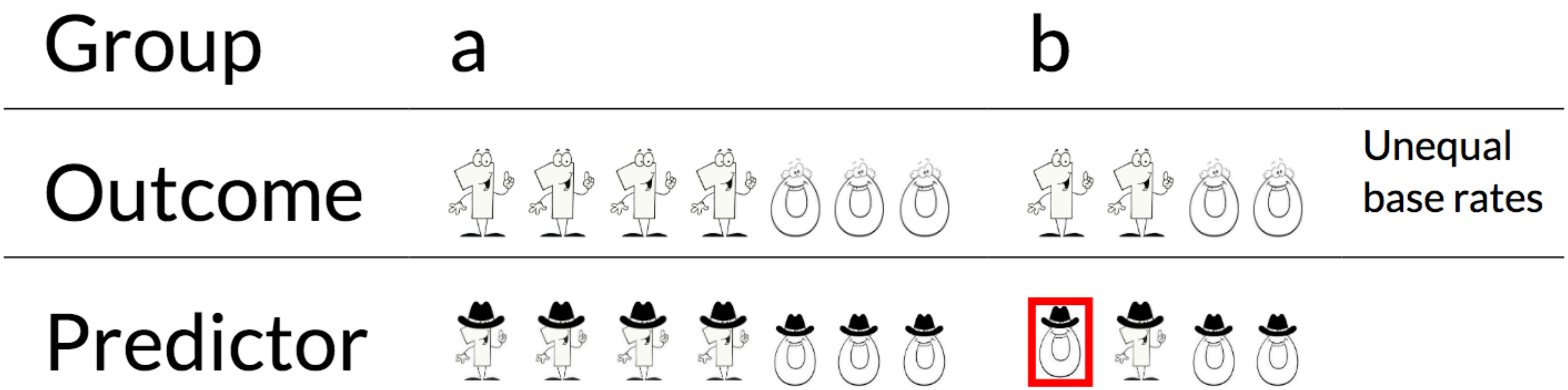
- We will look at a similar result later in the course due to Kleinberg, Mullainathan and Raghavan (2016)

Intuition



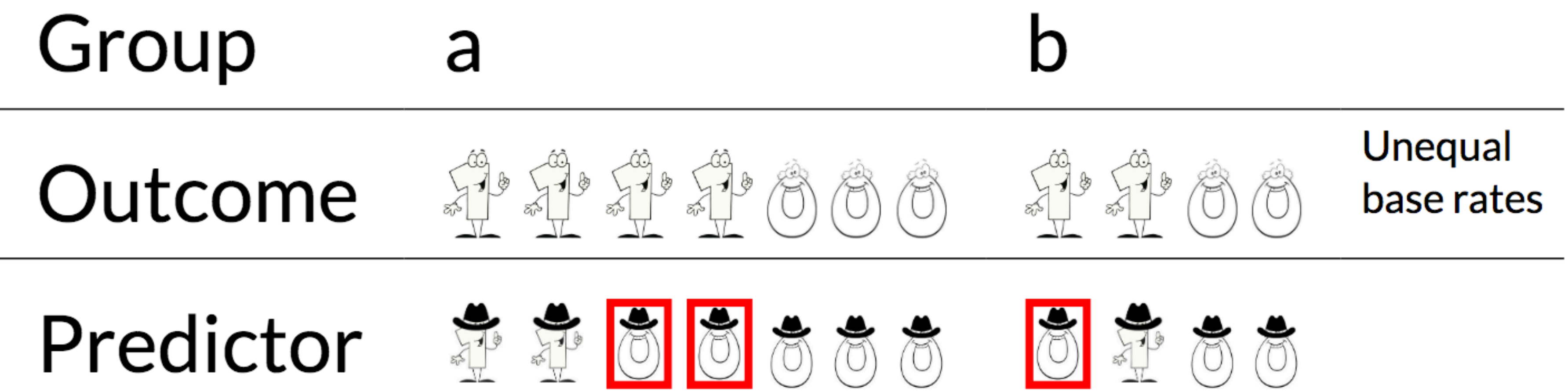
- So far, predictor is perfect.
- Let's introduce an error.

Intuition



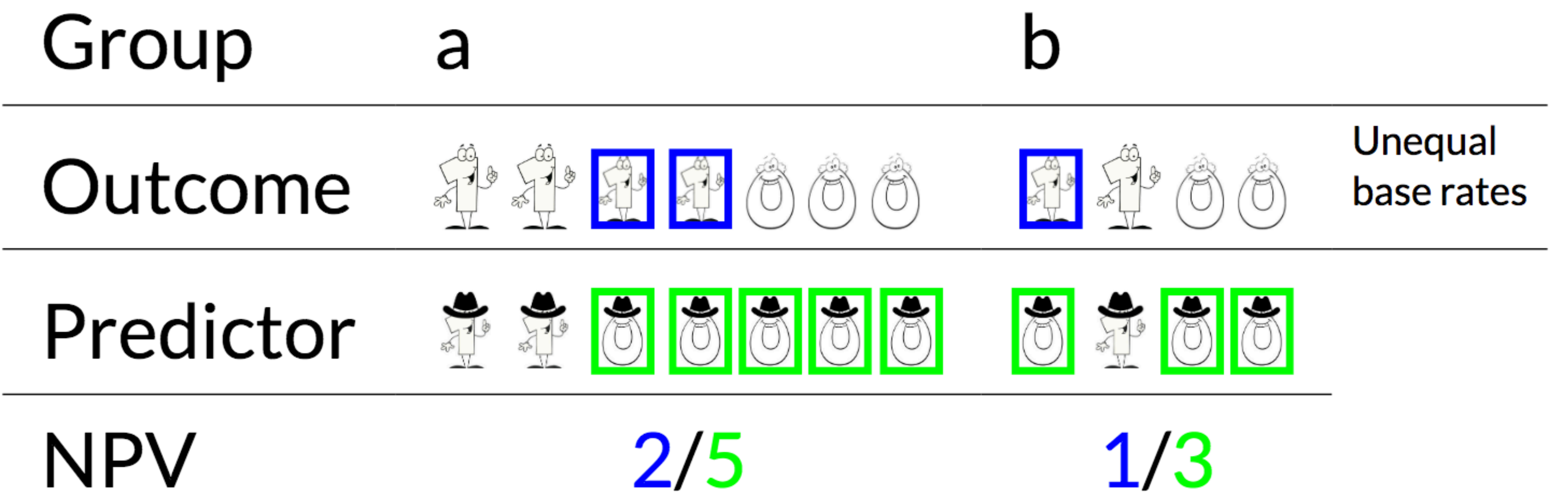
- But this doesn't satisfy positive rate parity!
- Let's fix that!

Intuition



- Satisfies positive rate parity!

Intuition



- Does not satisfy predictive value parity!

Proof. Assume unequal base rates $p_a, a \in \{0, 1\}$, imperfect classifier $C \neq Y$, and positive rate parity. W.l.o.g., $p_0 > 0$ (since $p_0 = p_1 = 0$ is trivial)
Show that predictive value parity fails.

Proof by googling the first Wiki entry on this:

$$\mathrm{PPV_a} = \frac{\mathrm{TPR}p_a}{\mathrm{TPR}p_a + \mathrm{FPR}(1-p_a)} \qquad \mathrm{NPV_a} = \frac{(1-\mathrm{FPR})(1-p_a)}{(1-\mathrm{TPR})p_a + (1-\mathrm{FPR})(1-p_a)}$$

Hence, $\mathrm{PPV}_0 = \mathrm{PPV}_1$ implies either $\mathrm{TPR} = 0$ or $\mathrm{FPR} = 0$. (But not both, since $C \neq Y$)

In either case, $\mathrm{NPV}_0 \neq \mathrm{NPV}_1$. Hence predictive value parity fails.

■

Outline

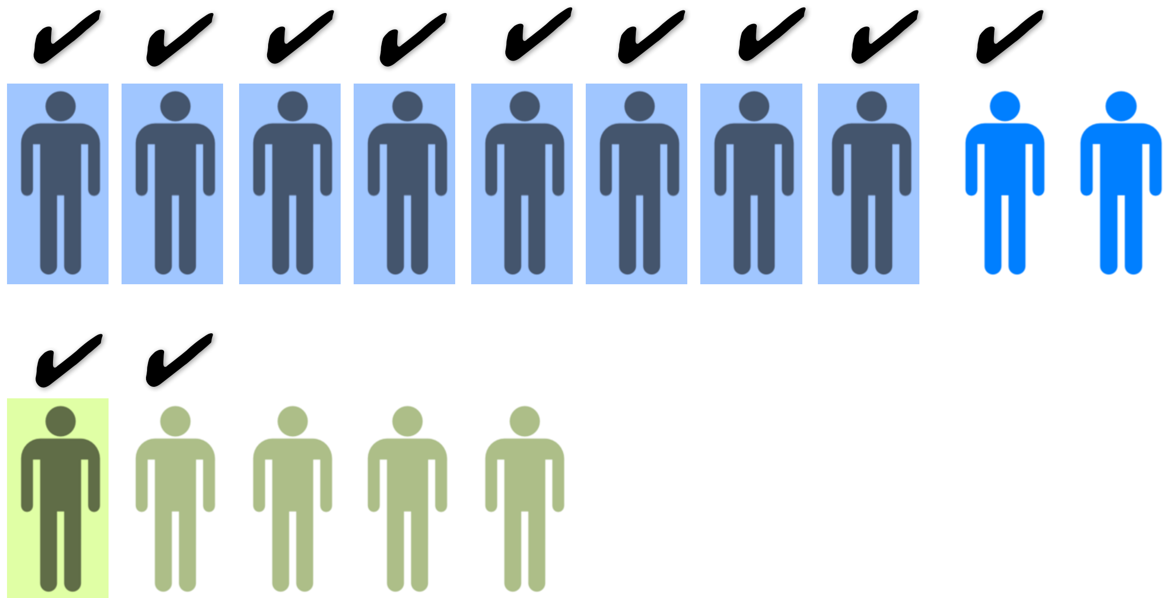
- ~~Recap: Disparity impact~~
  - ~~Issues with Disparate Impact~~
- ~~Observational measure of fairness~~
  - ~~Equal opportunity and Equalized odds~~
  - ~~Predictive Value Parity~~
  - ~~Tradeoff~~
- Achieving Equalized Odds
  - Binary Classifier

Equalized Odds

f satisfies equalized odds if
f is conditionally independent of protected X
given outcome C.

- Let $\hat{f}$ be any classifier out of the existing training pipeline for the problem at hand that fails to satisfy equalized odds

Classifier $\hat{f}$ that does not satisfy equalized odds



$X = 1$

$X = 0$

$$P_{X=1}[\hat{f}(Y) = 1 \mid C = 0] \neq P_{X=0}[\hat{f}(Y) = 1 \mid C = 0]$$

Derived Classifier

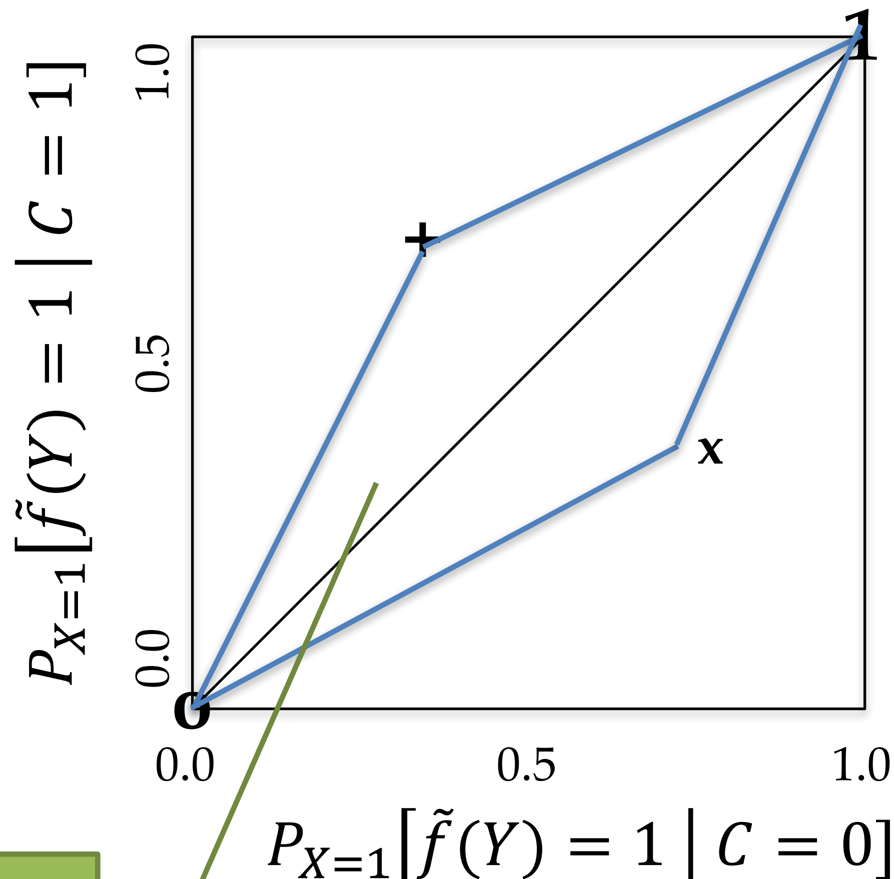
- A new classifier $\tilde{f}$ is **derived from $\hat{f}$ and the protected attribute X**
  - $\tilde{f}$ is independent of features Y conditional on $(\hat{f}, X)$
  - $P_{X=1}[\tilde{f}(Y) = c | C = 1]$ is $\sum_{c' \in \{0,1\}} P[c | \hat{f}(Y) = c', X = 1] \cdot P_{X=1}[\hat{f}(Y) = c' | C = 1]$
  - $P_{X=1}[\tilde{f}(Y) = c | C = 0]$ is $\sum_{c' \in \{0,1\}} P[c | \hat{f}(Y) = c', X = 1] \cdot P_{X=1}[\hat{f}(Y) = c' | C = 0]$
  - $P_{X=0}[\tilde{f}(Y) = c | C = 1]$
  - $P_{X=0}[\tilde{f}(Y) = c | C = 0]$

X=1	c'=0	c′=1
c=0	p0	p1
c=1	1-p0	1-p1

X=0	c′=0	c′=1
c=0	p2	p3
c=1	1-p2	1-p3

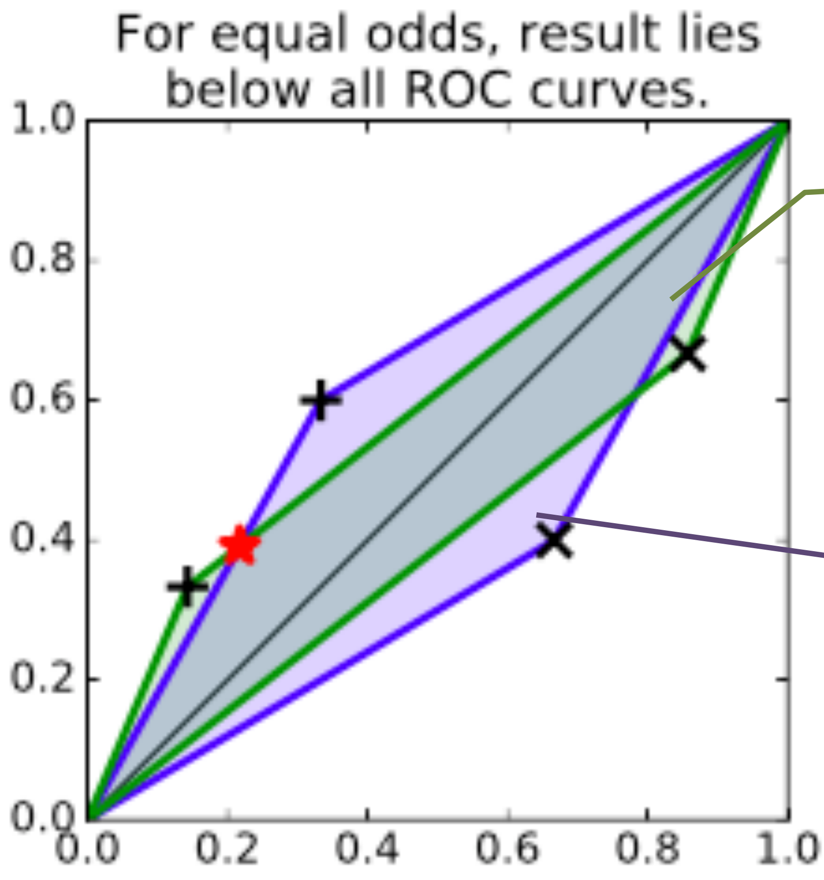
Derived Classifier

- Options for $\tilde{f}$:
  - $\tilde{f} = \hat{f}$ (+)
  - $\tilde{f} = 1 - \hat{f}$ (x)
  - $\tilde{f} = (1,1)$
  - $\tilde{f} = (0,0)$
  - Or some randomized combination of these



$\tilde{C}$ is in the enclosed region

Derived Classifier



$P_X[\tilde{f}(Y) = 1 \mid C = 1]$

$P_X[\tilde{f}(Y) = 1 \mid C = 0]$

$\tilde{f}$ is in this region for X = 0

$\tilde{f}$ is in this region for X = 1

Derived Classifier

- Loss minimization: $l: \{0,1\}^2 \rightarrow R$
  - Indicate the loss of predicting $\tilde{f}(Y) = c$ when the correct label is c''

- Minimize the expected loss E $[l(\tilde{f}(Y), C)]$ s.t.
  - $\tilde{f}$ is derived
  - $\tilde{f}$ satisfies equalized odds
    - $P_{X=1}[\tilde{f}(Y) = 1 | C = 1] = P_{X=0}[\tilde{f}(Y) = 1 | C = 1]$
    - $P_{X=1}[\tilde{f}(Y) = 1 | C = 0] = P_{X=0}[\tilde{f}(Y) = 1 | C = 0]$

Derived Classifier

- $E\left[l\left(\tilde{f}(Y), C\right)\right] = \sum_{c,c'' \in \{0,1\}} l(c, c'') \Pr[\tilde{f}(Y) = c, C = c'']$

- $\Pr[\tilde{f} = c, C = c'']$

$$= \Pr\left[\tilde{f} = c, C = c'' \middle| \tilde{f} = \hat{f}\right] \Pr\left[\tilde{f} = \hat{f}\right]$$
$$+ \Pr\left[\tilde{f} = c, C = c'' \middle| \tilde{f} \neq \hat{f}\right] \Pr\left[\tilde{f} \neq \hat{f}\right]$$
$$= \Pr\left[\hat{f} = c, C = c''\right] \Pr\left[\tilde{f} = \hat{f}\right]$$
$$+ \Pr\left[\hat{f} = 1 - c, C = c''\right] \Pr\left[\tilde{f} \neq \hat{f}\right]$$

Based on the joint distribution

$\hat{f}$

$\tilde{f}$

X=1	c'=0	c′=1
c=0	p0	p1
c=1	1-p0	1-p1

X=0	c'=0	c′=1
c=0	p2	p3
c=1	1-p2	1-p3

Summary: Multiple fairness measures

- Demographic parity or disparate impact
  - Pro: Used in the law
  - Con: Perfect classification is impossible
  - Achieved by modifying data

- Equal odds/ opportunity
  - Pro: Perfect classification is possible
  - Con: Different groups can get different rates of positive prediction
  - Achieved by post processing the classifier

Summary: Multiple fairness measures

- Equal odds/opportunity
  - Different groups may be treated unequally
  - Maybe due to the problem
  - Maybe due to bias in the dataset

- *While demographic parity seems like a good fairness goal for the society, …*
  Equal odds/opportunity seems to be measuring whether an algorithm is fair (independent of other factors like input data).

Summary: Multiple fairness measures

- Fairness through Awareness:
  - Need to define a distance function $d(x,x')$
  - A guarantee at the individual level (rather than on groups)
  - How does this connect to other notions of fairness?