# Quantifying fairness and discrimination in predictive models

Arthur Charpentier

16th Annual Conference of Thailand Econometric Society TES'2023

# Motivation

NEWSLETTERS
Sign up to read our regular email newsletters

NewScientist

SUBSCRIBE AND SAVE 69%

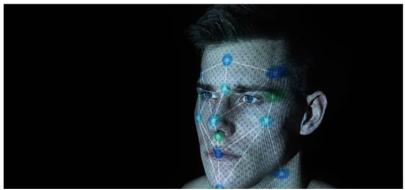
News Podcasts Video Technology Space Physics Health More Shop Courses Events Tours Jobs

## Discriminating algorithms: 5 times AI showed prejudice

Artificial intelligence is supposed to make life easier for us all – but it is also prone to amplify sexist and racist biases from the real world

TECHNOLOGY 12 April 2018 , updated 27 April 2018

By Daniel Cossins



TRENDING LATEST VIDEO FREE

1. Dingo genome suggests Australian icon not descended from domestic dogs
2. How *Minecraft* is helping children with autism make new friends
3. A third of people aged over 70 are sexually active, survey reveals
4. Harmful air pollution now affects 99 per cent of everyone on Earth
5. Breaking the News exhibition shows Edward Snowden's smashed drives

PROPUBLICA

Donate

## How We Analyzed the COMPAS Recidivism Algorithm

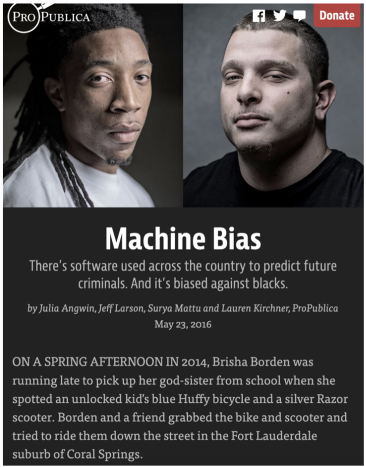
*by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin*
May 23, 2016

← Read the story

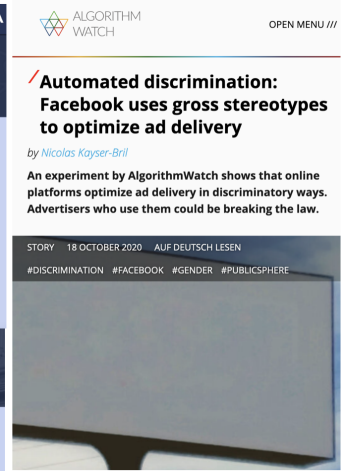
Across the nation, judges, probation and parole officers are increasingly using algorithms to assess a criminal defendant's likelihood of becoming a recidivist – a term used to describe criminals who re-offend. There are dozens of these risk assessment algorithms in use. Many states have built their own assessments, and several academics have written tools. There are also two leading nationwide tools offered by commercial vendors.

We set out to assess one of the commercial tools made by Northpointe, Inc. to discover the underlying accuracy of their recidivism algorithm and to test whether the algorithm was biased against certain groups.

# Motivation


PRO PUBLICA
Donate
Machine Bias
There's software used across the country to predict future criminals. And it's biased against blacks.
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016
ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.


IABA
PUBLIC CRITICISM OF INSURANCE PRICING PRACTICES
A 2015 Study by the Consumer Federation of America (CFA) claimed that "on average, a good driver in a predominantly African American Community will pay considerably more for state-mandated auto insurance coverage than a similarly situated driver in a predominantly White community."
An analysis by ProPublica and Consumer Reports in 2017 drew a similar conclusion, stating that "this disparity may amount to a subtler form of redlining, a term that traditionally refers to denial of services or products to minority areas." While there were methodological flaws in the analysis, the article raised a number of questions about whether insurance rates were biased against minorities.
RACE AND INSURANCE
CONGRESS TARGETS DISCRIMINATION IN AUTO INSURANCE
H.R. 1756: Preventing Credit Score Discrimination in Auto Insurance Act
This bill intended to amend the Fair Credit Reporting Act to prohibit the use of consumer credit information in auto insurance decision-making but did not receive a vote in the 116th Congress.
H.R. 3693: Prohibit Auto Insurance Discrimination Act


ALGORITHM WATCH
OPEN MENU ///
Automated discrimination: Facebook uses gross stereotypes to optimize ad delivery
by Nicolas Kayser-Bril
An experiment by AlgorithmWatch shows that online platforms optimize ad delivery in discriminatory ways. Advertisers who use them could be breaking the law.
STORY 18 OCTOBER 2020 AUF DEUTSCH LESEN
#DISCRIMINATION #FACEBOOK #GENDER #PUBLICSPHERE

# Motivation

Fairness Through Awareness

Cynthia Dwork* Moritz Hardt† Toniann Pitassi‡ Omer Reingold§ Richard Zemel¶

November 30, 2011

**Algorithmic decision making and the cost of fairness**

Sam Corbett-Davies, Stanford University, scorbett@stanford.edu
Emma Pierson, Stanford University, emmap1@stanford.edu
Avi Feller, Univ. of California, Berkeley, afeller@berkeley.edu
Sharad Goel, Stanford University, scgoel@stanford.edu
Aziz Huq, University of Chicago, huq@uchicago.edu

Equality of Opportunity in Supervised Learning

Moritz Hardt Eric Price Nathan Srebro

October 11, 2016

Fairness in Criminal Justice Risk Assessments: The State of the Art

Richard Berk[a,b], Hoda Heidari[c], Shahin Jabbari[c], Michael Kearns[c], Aaron Roth[c]

Fair prediction with disparate impact:
A study of bias in recidivism prediction instruments

Alexandra Chouldechova *

The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*

Sam Corbett-Davies, Stanford University
Sharad Goel, Stanford University

August 14, 2018

HUMAN DECISIONS AND MACHINE PREDICTIONS*

Jon Kleinberg
Himabindu Lakkaraju
Jure Leskovec
Jens Ludwig
Sendhil Mullainathan

The Frontiers of Fairness in Machine Learning

Alexandra Chouldechova* Aaron Roth†

October 23, 2018

DOI:10.1145/3376898

**A group of industry, academic, and government experts convene in Philadelphia to explore the roots of algorithmic bias.**

BY ALEXANDRA CHOULDECHOVA AND AARON ROTH

**A Snapshot of the Frontiers of Fairness in Machine Learning**

**A Survey on Bias and Fairness in Machine Learning**

NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI

FAIRNESS IN MACHINE LEARNING: A SURVEY

A PREPRINT

**Simon Caton**
University College Dublin
Dublin, Ireland
simon.caton@ucd.ie

**Christian Haas**
University of Nebraska at Omaha
Omaha, US
christianhaas@unomaha.edu

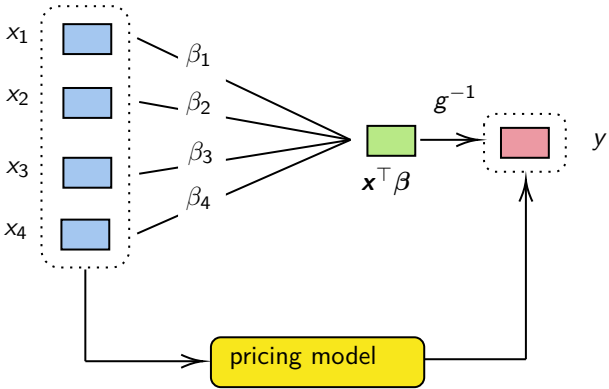
# Motivation

Very important topic in many industries,
e.g. insurance and (retail) banking

- "*Technology is neither good nor bad; nor is it neutral* ", Kranzberg (1986)
- "*Machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitely ask for* ", Kearns and Roth (2019)

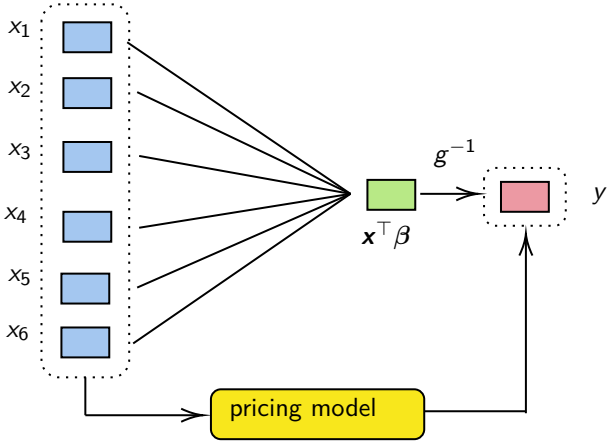
Quick overview of the topic today...

# GLM, ML & Big Data



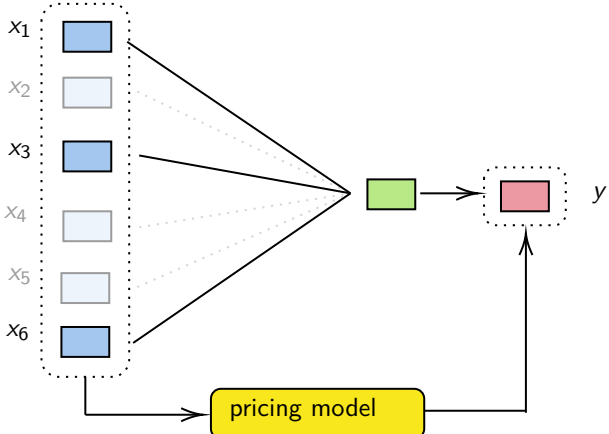

GLM: $\min \sum_{i=1}^{n} \ell(y_i, g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}))$ where $\boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}$,

# GLM, ML & Big Data



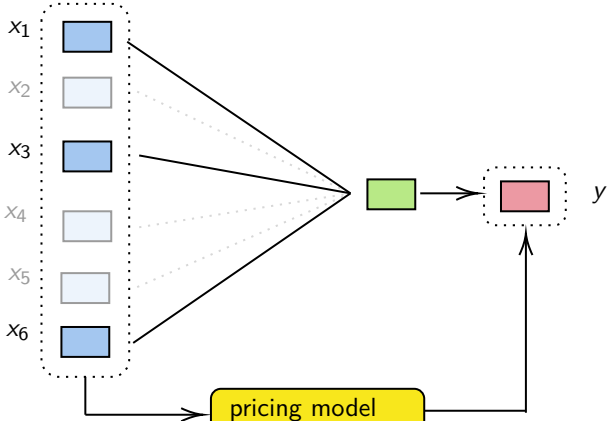

Data enrichment: $\min \sum_{i=1}^{n} \ell(y_i, g^{-1}(\boldsymbol{x}_i^\top\boldsymbol{\beta}))$ where $\boldsymbol{x}_i^\top\boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$,

# GLM, ML & Big Data



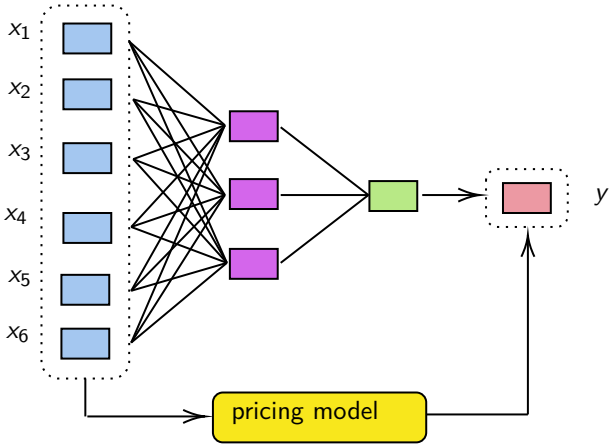

LASSO and penalty: $\min \left\{ \sum_{i=1}^{n} \ell(y_i, g^{-1}(\boldsymbol{x}_i^\top \beta)) \right\}$ subject to $\dim(\beta) \leq s$

# GLM, ML & Big Data



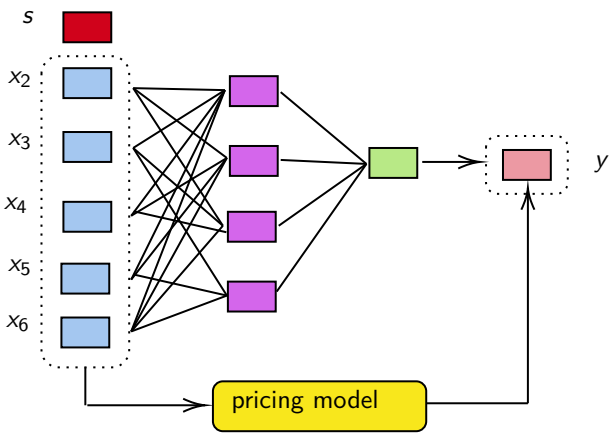

LASSO and penalty: $\min \left\{ \sum_{i=1}^{n} \ell(y_i, g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})) + \lambda \cdot \dim(\boldsymbol{\beta}) \right\}$

# GLM, ML & Big Data



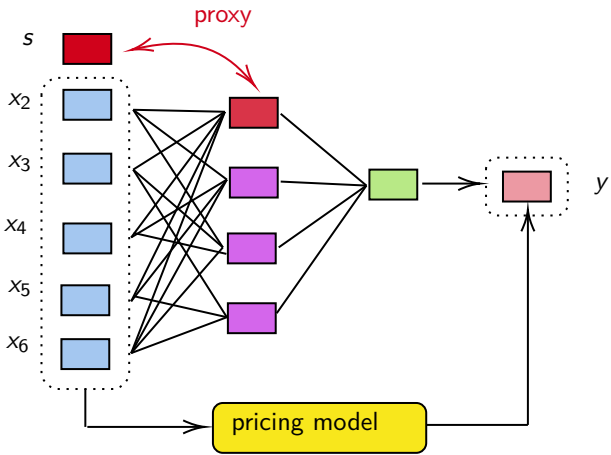

Neural Nets: $\min \sum_{i=1}^{n} \ell(y_i, g^{-1}(\omega_1 z_{1i} + \omega_2 z_{2i} + \omega_3 z_{3i}))$ where $z_{ji} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j$.

# GLM, ML & Big Data



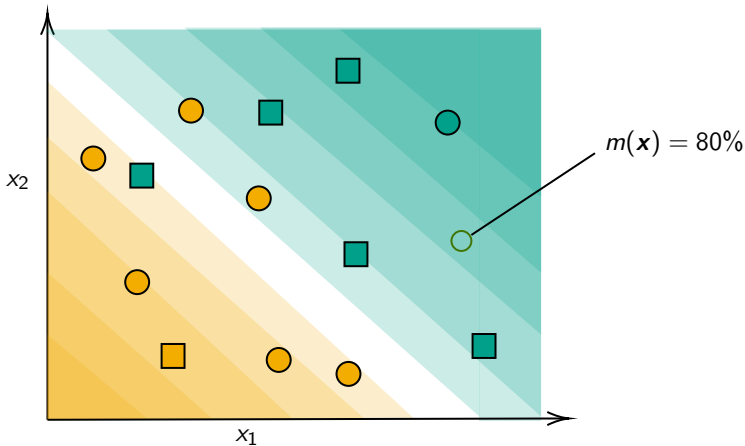

Fairness by unawarness (remove the sensitive variable $s$)

# GLM, ML & Big Data



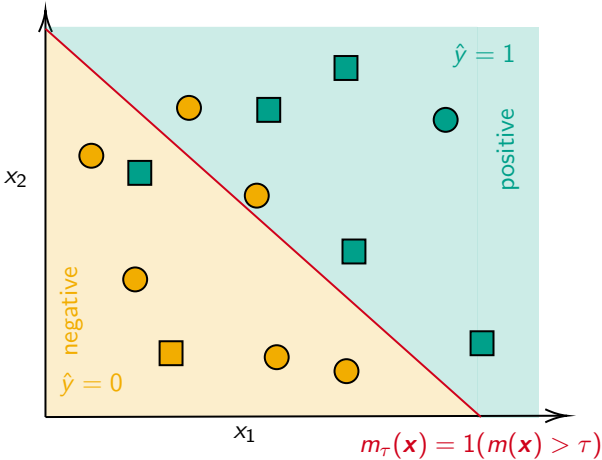

Possible statistical discrimination if $z_1$ and $s$ are highly correlated (demographic parity)

# Notations



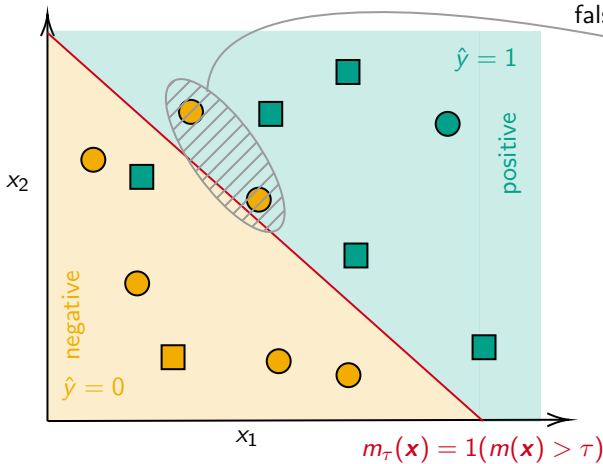

Dataset $\{(y_i, \boldsymbol{x}_i)\}$, $y_i \in \{0, 1\}$, score $m(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = g^{-1}(\boldsymbol{x}^\top \beta)$

# Notations




For $\tau \in (0, 1)$, define classifier $m_\tau(\boldsymbol{x}) = \mathbf{1}(m(\boldsymbol{x}) > \tau) \in \{0, 1\}$

# Notations




false positive

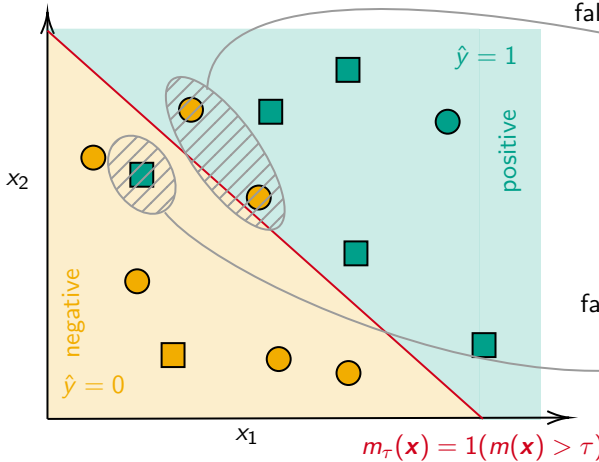
| | $\hat{y} = 0$ | $\hat{y} = 1$ |
|---|---|---|
| $y = 0$ | TN | FP |
| $y = 1$ | FN | TP |

positive

P

From $m_\tau$ consider the associated confusion matrix (negative/positive, true/false)

# Notations



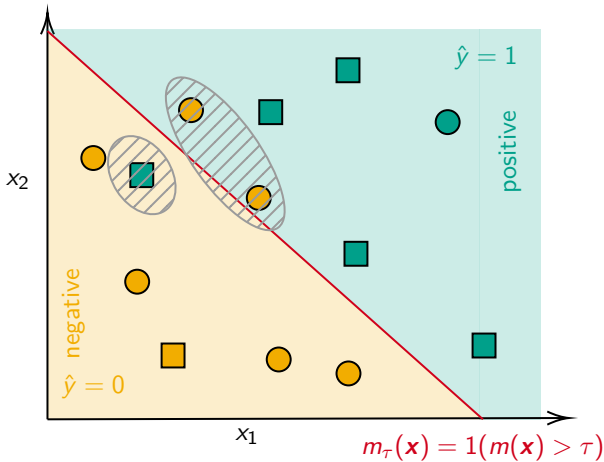

false positive

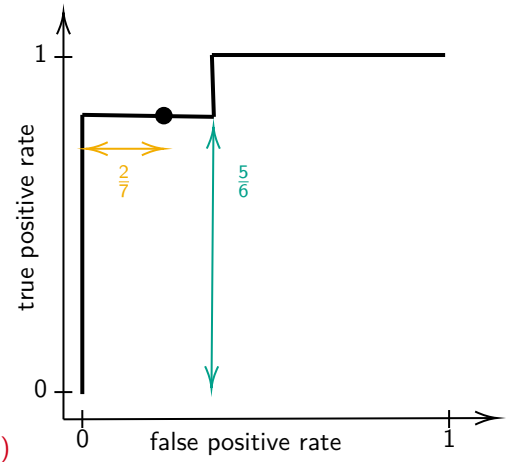
| | $\hat{y} = 0$ | $\hat{y} = 1$ |
|---|---|---|
| $y = 0$ | TN | FP |
| $y = 1$ | FN | TP |

positive
P

false negative

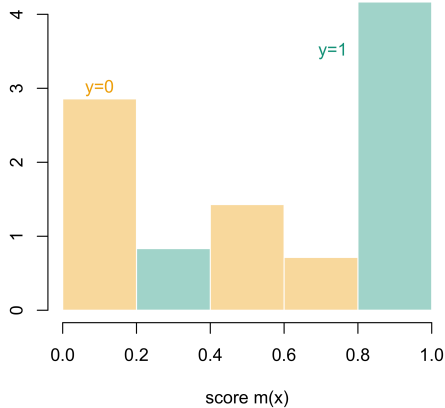
# Notations

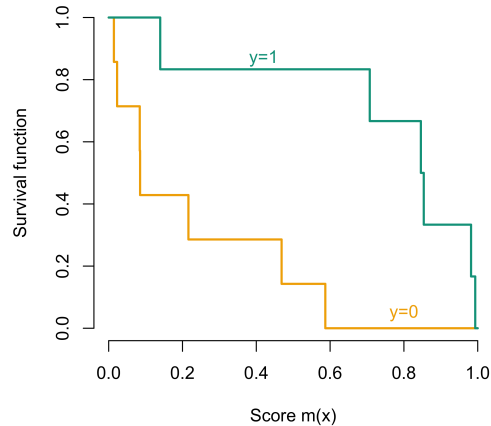



Changing $\tau$ lead to the ROC curve (TPR against FPR)

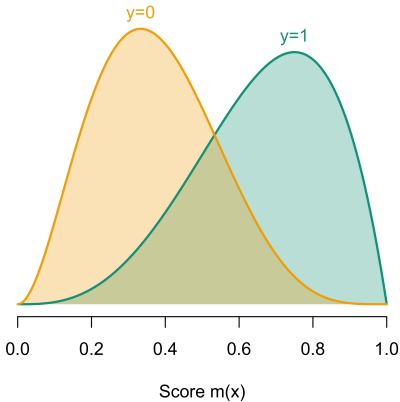
# Notations

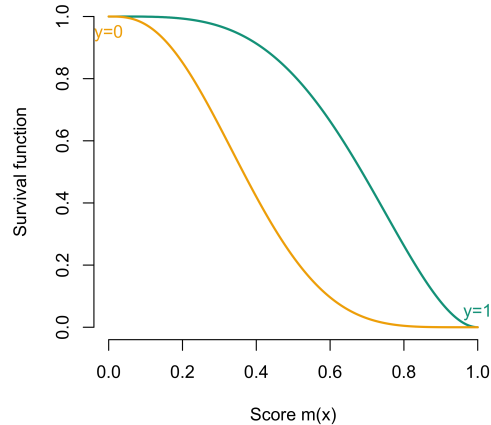



Consider the distribution of scores $m(\boldsymbol{x}_i)$ when $y_i = 0$ and $y_i = 1$

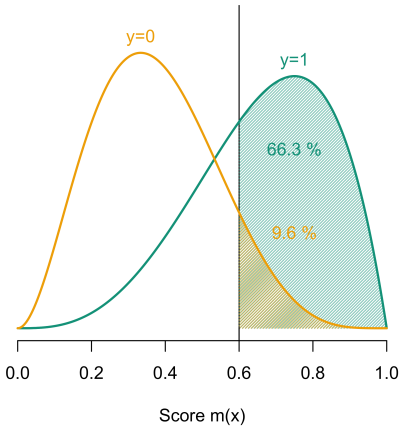
# Notations

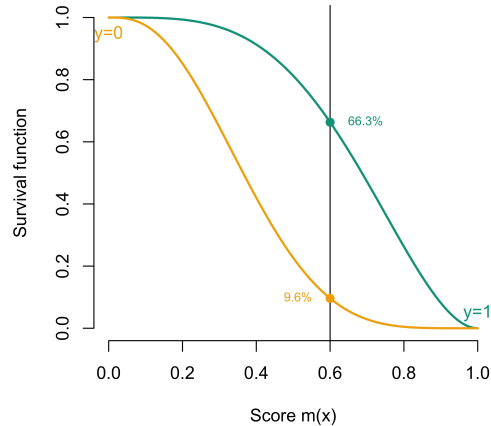



Consider the distribution of scores $m(\boldsymbol{x}_i)$ when $y_i = 0$ and $y_i = 1$ (continuous version)

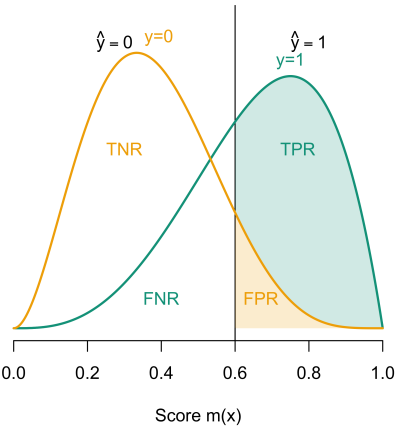
# Notations

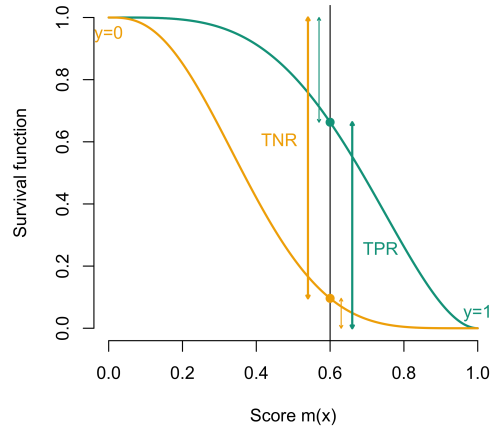



$\tau = 60\%$, FPR$\sim$ 9.6% and TPR$\sim$ 66.3% (continuous version)

# Notations




$\tau = 60\%$, Given $\tau$, one visualize FPR, TNR, TPR and FNR (continuous version)

# Notations

$$\begin{cases} y \in \{0,1\} & \text{variable of interest (binary)} \\ s \in \{0,1\} & \text{protected variable (sensitive)} \\ \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d & \text{'explanatory' variables} \\ m : \mathcal{X} \to [0,1] & \text{score } m(\boldsymbol{x}) = g^{-1}(\boldsymbol{x}^\top \beta), \text{ e.g. } \text{logit}(\boldsymbol{x}^\top \beta) \\ & m(\boldsymbol{x}) = \mathbb{P}[Y = 1 | \boldsymbol{X} = \boldsymbol{x}] \\ \tau \in (0,1) & \text{threshold} \\ \widehat{y} \in \{0,1\} & \text{classifier } \widehat{y} = m_\tau(\boldsymbol{x}) \mathbf{1}(m(\boldsymbol{x}) > \tau) \end{cases}$$

**Remark** for people who prefer regression over classification:
instead of $\mathbb{P}[\widehat{Y} = 1 | \cdots]$ or $\mathbb{P}[Y = 1 | \cdots]$, read

$$\begin{cases} \mathbb{E}[\widehat{Y} | \cdots] \text{ or } \mathbb{E}[Y | \cdots] & \text{weak version} \\ \mathbb{P}[\widehat{Y} \in A | \cdots] \text{ or } \mathbb{P}[Y \in A | \cdots] \ \forall A \subset \mathcal{Y} & \text{strong version} \end{cases}$$

# Group fairness

At least 21 definitions of "fairness", Narayanan (2018).
"*focus on equality of treatment among groups of people from criteria requiring equality of treatment among couples of similar individuals*", Castelnovo et al. (2022)

**Fairness Through Unawareness** Kusner et al. (2017) We will speak of fairness through unawareness if the sensitive attribute $s$ is not explicitly used in the decision function $\widehat{y}$, i.e. neither in the construction of the score $m$, nor in the choice of the threshold level $\tau$, allowing to pass from $m$ to $\widehat{y}$.

... obviously not sufficient.

# Demographic Parity

"*independence is equivalent to requiring the same positive prediction ratio across groups identified by the sensitive features. This form of independence is usually known as Demographic Parity (DP), statistical parity, or sometimes as group fairness*", Castelnovo et al. (2022)

**Demographic Parity** Corbett-Davies et al. (2017), Agarwal (2021) A decision function $\widehat{y}$ satisfies demographic parity if $\widehat{Y} \perp\!\!\!\perp S$, i.e.

$$\mathbb{P}[\widehat{Y} = y|S = 0] = \mathbb{P}[\widehat{Y} = y|S = 1], \ \forall y \in \{0, 1\}$$

**Disparate impact** Feldman et al. (2015)] A decision function $\widehat{Y}$ has a disparate impact, for a given threshold $d$, if,

$$\min\{\frac{\mathbb{P}[\widehat{Y} = 1|S = 0]}{\mathbb{P}[\widehat{Y} = 1|S = 1]}, \frac{\mathbb{P}[\widehat{Y} = 1|S = 1]}{\mathbb{P}[\widehat{Y} = 1|S = 0]}\} < d \text{ (usually 80\%)}.$$

# Equalized Odds

Separation critera: independence when $Y = 0$ or $Y = 1$

**True positive equality, Equalized Odds** Hardt et al. (2016) We will speak of equality of opportunity, or parity of true positives, if

$$\mathbb{P}[\widehat{Y} = 1 | S = 0, Y = 1] = \mathbb{P}[\widehat{Y} = 1 | S = 1, Y = 1]$$

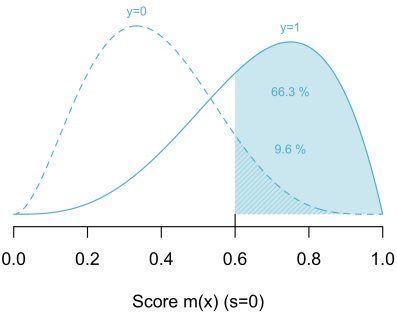
or equivalently $\text{TPR}_0 = \dfrac{\text{TP}_0}{\text{FN}_0 + \text{TP}_0} = \dfrac{\text{TP}_1}{\text{FN}_1 + \text{TP}_1} = \text{TPR}_1$.

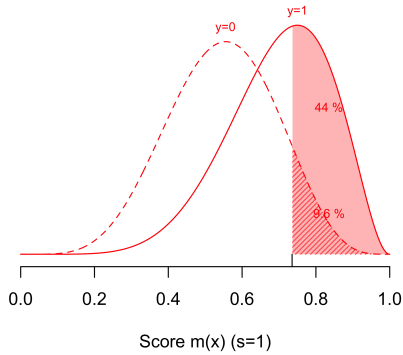
**False positive equality** Hardt et al. (2016) We will speak of equality of false positives if

$$\mathbb{P}[\widehat{Y} = 1 | S = 0, Y = 0] = \mathbb{P}[\widehat{Y} = 1 | S = 1, Y = 0],$$

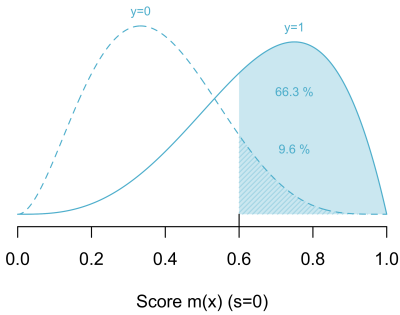
or equivalently $\text{FPR}_0 = \dfrac{\text{FP}_0}{\text{TN}_0 + \text{FP}_0} = \dfrac{\text{FP}_1}{\text{TN}_1 + \text{FP}_1} = \text{FPR}_1$.

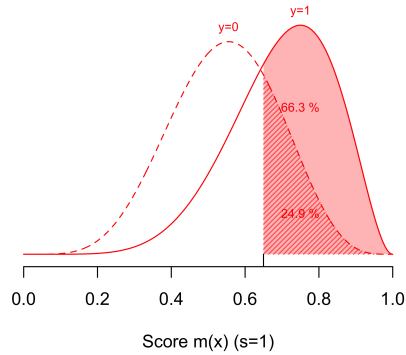
# Equalized Odds



# Equalized Odds



# Equalized Odds

**Equalized of opportunity** Hardt et al. (2016) The parity of false positives and true positives is called equality of opportunity,

$$\mathbb{P}[\widehat{Y} = 1|S = 0, Y = y] = \mathbb{P}[\widehat{Y} = 1|S = 1, Y = y], \ \forall y \in \{0, 1\}$$

in other words, $\widehat{Y} \perp\!\!\!\perp S$ conditionally on $Y$.

On can also use any measure based on confusion matrices, such as $\phi$, introduced by Matthews (1975),

**$\phi$-fairness** Chicco and Jurman (2020) We will have $\phi$-fairness if $\phi_1 = \phi_0$, where $\phi_s$ denotes Matthews correlation coefficient for the $s$ group,

$$\phi_s = \frac{\mathrm{TP}_s \cdot \mathrm{TN}_s - \mathrm{FP}_s \cdot \mathrm{FN}_s}{\sqrt{(\mathrm{TP}_s + \mathrm{FP}_s)(\mathrm{TP}_s + \mathrm{FN}_s) \cdot (\mathrm{TN}_s + \mathrm{FP}_s)(\mathrm{TN}_s + \mathrm{FN}_s)}}$$

# Equalized Odds

All those measures are based on some choice of thresholds, but it is also possible to consider a global measures of calibration, such as the area under the curve,

**AUC fairness** Borkan et al. (2019) We will have AUC fairness if $\text{AUC}_1 = \text{AUC}_0$, where $\text{AUC}_s$ is the AUC for the $s$ group.

We find a similar idea in Beutel et al. (2019). The problem with the AUC is that we can have identical AUCs, but very different underlying ROC curves. So, it can be interesting to consider a notion of fairness based on the ROC curves. As a reminder, we had defined the ROC curve as $t \mapsto \text{TPR} \circ \text{FPR}^{-1}(t)$.

**Equality of ROC curves** Vogel et al. (2021) Let $\text{FRP}_s(t) = \mathbb{P}[m(\boldsymbol{X}) > t | Y = 0, S = s]$ and $\text{TPR}_s(t) = \mathbb{P}[m(\boldsymbol{X}) > t | Y = 1, S = s]$. Set $\Delta_{TPR}(t) = \text{TPR}_1 \circ \text{TPR}_0^{-1}(t) - t$ et $\Delta_{FRP}(t) = \text{FPR}_1 \circ \text{FPR}_0^{-1}(t) - t$. We will have an fairness of ROC curves if $\|\Delta_{TPR}\|_\infty = \|\Delta_{FPR}\|_\infty = 0$.

# Equalized Odds

From Hardt et al. (2016)

$$\begin{cases} \text{True positive equality :} & \mathbb{P}[\widehat{Y}=1|S=0,Y=1]=\mathbb{P}[\widehat{Y}=1|S=1,Y=1] \\ \text{False positive equality :} & \mathbb{P}[\widehat{Y}=1|S=0,Y=0]=\mathbb{P}[\widehat{Y}=1|S=1,Y=0] \\ \text{False negative equality :} & \mathbb{P}[\widehat{Y}=0|S=0,Y=1]=\mathbb{P}[\widehat{Y}=0|S=1,Y=1] \end{cases}$$

but one can consider equality of ratios...

**Equal treatment** Berk et al. (2021) We have equality of treatment, the rate of false positives and false negatives are identical in the protected groups,

$$\frac{\mathbb{P}[\widehat{Y}=1|S=0,Y=0]}{\mathbb{P}[\widehat{Y}=0|S=0,Y=1]}=\frac{\mathbb{P}[\widehat{Y}=1|S=1,Y=0]}{\mathbb{P}[\widehat{Y}=0|S=1,Y=1]}$$

# Calibration

Instead of $\widehat{Y}$ focus on $m(\boldsymbol{X})$ (if possible)

**Class balance** Kleinberg et al. (2016) We will have class balance in the weak sense if

$$\mathbb{E}[m(\boldsymbol{X})|Y=y,S=0]=\mathbb{E}[m(\boldsymbol{X})|Y=y,S=1],\ \forall y\in\{0,1\}$$

or in the strong sense if

$$\mathbb{P}[m(\boldsymbol{X})\leq\mu|Y=y,S=0]=\mathbb{P}[m(\boldsymbol{X})\leq\mu|Y=y,S=1],\ \forall\mu\in[0,1],\ \forall y\in\{0,1\}.$$

**Calibration (or accuracy) parity** Kleinberg et al. (2016), Zafar et al. (2017)] We have calibration parity if

$$\mathbb{P}[Y=1|m(\boldsymbol{X})=\mu,S=0]=\mathbb{P}[Y=1|m(\boldsymbol{X})=\mu,S=1],\ \forall\mu\in[0,1].$$

A weaker version can be related to $Y\perp\!\!\!\perp S$ conditionally on $\widehat{Y}$

# Calibration

We can go further by asking not only for parity, but also for a good calibration

**Good calibration** Kleinberg et al. (2017) We have an fairness of good calibration if

$$\mathbb{P}[Y = 1|m(\boldsymbol{X}) = \mu, S = 0] = \mathbb{P}[Y = 1|m(\boldsymbol{X}) = \mu, S = 1] = \mu, \ \forall \mu \in [0, 1].$$

Nice property, but usually never satisfied by ML models (without fairness issue),

$$\mathbb{E}[Y|m(\boldsymbol{X}) = \mu] \neq \mu, \ \forall \mu$$

This "good calibration" property of the model $m$, also called "well-calibration" in Dawid (1982), and "autocalibration" in Van Calster et al. (2019), Krüger and Ziegel (2021) and Denuit et al. (2021) in the context of regression, i.e. $\mathbb{E}[Y|m(\boldsymbol{X}) = \mu] = \mu$, is a standard property in econometrics, in generalized linear models, but not in most machine learning algorithms.

# Varia

**Non-reconstruction of the protected attribute** Kim (2017) If we cannot tell from the result $(\boldsymbol{x}, m(\boldsymbol{x}), y$ and $\widehat{y})$ whether the subject was a member of a protected group or not, we will talk about fairness by non-reconstruction of the protected attribute

$$\mathbb{P}[S = 0|\boldsymbol{X}, m(\boldsymbol{X}), \widehat{Y}, Y] = \mathbb{P}[S = 1|\boldsymbol{X}, m(\boldsymbol{X}), \widehat{Y}, Y].$$

Used in Zhang et al. (2018) to suggest a adversarial approach to mitigate discrimination

# Wrap-Up on Group Fairness

| | | |
|---|---|---|
| *statistical parity*, Dwork et al. (2012) | $\mathbb{P}[\widehat{Y}=1\|S=s]=\text{cst},\ \forall s$ | independence |
| *conditional stat. parity*, Corbett-Davies et al. (2017) | $\mathbb{P}[\widehat{Y}=1\|S=s,X=x]=\text{cst}_x,\ \forall s,y$ | $\widehat{Y} \perp\!\!\!\perp P$ |
| *equalized odds*, Hardt et al. (2016) | $\mathbb{P}[\widehat{Y}=1\|S=s,Y=y]=\text{cst}_y,\ \forall s,y$ | separation |
| *equalized opportunity*, Hardt et al. (2016) | $\mathbb{P}[\widehat{Y}=1\|S=s,Y=1]=\text{cst},\ \forall s$ | |
| *predictive equality*, Corbett-Davies et al. (2017) | $\mathbb{P}[\widehat{Y}=1\|S=s,Y=0]=\text{cst},\ \forall s$ | $\widehat{Y} \perp\!\!\!\perp S \mid Y$ |
| *balance (positive)*, Kleinberg et al. (2017) | $\mathbb{E}[m(\boldsymbol{X})\|S=s,Y=1]=\text{cst},\ \forall s$ | $S \perp\!\!\!\perp P \mid Y$ |
| *balance (negative)*, Kleinberg et al. (2017) | $\mathbb{E}[m(\boldsymbol{X})\|S=s,Y=0]=\text{cst},\ \forall s$ | |
| *conditional accuracy equality*, Berk et al. (2017) | $\mathbb{P}[Y=y\|S=s,\widehat{Y}=y]=\text{cst}_y,\ \forall s,y$ | sufficiency |
| *predictive parity*, Chouldechova (2017) | $\mathbb{P}[Y=1\|S=s,\widehat{Y}=1]=\text{cst},\ \forall s$ | |
| *calibration*, Chouldechova (2017) | $\mathbb{P}[Y=1\|S=s,m(\boldsymbol{X})=m]=\text{cst}_s,\ \forall s,m$ | $Y \perp\!\!\!\perp S \mid \widehat{Y}$ |
| *well-calibration*, Chouldechova (2017) | $\mathbb{P}[Y=1\|S=s,m(\boldsymbol{X})=m]=s,\ \forall s,m$ | |
| *accuracy equality*, Berk et al. (2017) | $\mathbb{P}[\widehat{Y}=Y\|m(\boldsymbol{X})=m]=\text{cst},\ \forall m$ | |
| *treatment equality*, Berk et al. (2017) | $\frac{\text{FN}_s}{\text{FP}_s}=\text{cst}_s,\ \forall s$ | |

# Individual fairness

"*Individual fairness is embodied in the following principle: similar individuals should be given similar decisions. This principle deals with the comparison of single individuals rather than focusing on groups of people sharing some characteristics. On the other hand, group fairness starts from the idea that there are groups of people potentially suffering biases and unfair decisions, and thus tries to reach equality of treatment for groups instead of individuals*", Castelnovo et al. (2022)

**Lipschitz property** Duivesteijn and Feelders (2008), Luong et al. (2011) A decision function $\widehat{Y}$ satistfies the Lipschitz property if

$$d_y(\widehat{y}_i, \widehat{y}_j) \leq d_x(\boldsymbol{x}_i, \boldsymbol{x}_j), \ \forall i, j = 1, \cdots, n.$$

Two "close" individuals (in the sense of unprotected characteristics $\boldsymbol{x}$) must have the same forecast

# Conterfactual fairness

"what would have been the decision if that individual had a different gender?"

| | Sensitive $s$ | Outcome | | | Age | School | Height | Weight |
|---|---|---|---|---|---|---|---|---|
| | $s_i$ | $y_i$ | $y^\star_{i,S\leftarrow 1}$ | $y^\star_{i,S\leftarrow o}$ | $x_{1,i}$ | $x_{2,i}$ | $x_{3,i}$ | $x_{4,i}$ |
| 1 | 0 – M | 1 | 1 | ? | 37 | 14 | 160 | 56 |
| 2 | 1 – F | 1 | ? | 1 | 28 | 12 | 156 | 54 |
| 3 | 0 – M | 0 | 0 | ? | 53 | 11 | 190 | 87 |

**Counterfactual fairness** Kusner et al. (2017) If the prediction in the real world is the same as the prediction in the counterfactual world where the individual would have belonged to a different demographic group, we have counterfactual fairness, i.e.

$$\mathbb{P}[Y^\star_{S\leftarrow 1} = 1 | \boldsymbol{X} = \boldsymbol{x}] - \mathbb{P}[Y^\star_{S\leftarrow 0} = 1 | \boldsymbol{X} = \boldsymbol{x}] = 0, \ \forall \boldsymbol{x}.$$

# Conterfactual fairness

"what would have been the outcome if that individual had a different treatment?"

| | Treatment | Outcome | | | Age | School | Height | Weight |
|---|---|---|---|---|---|---|---|---|
| | $t_i$ | $y_i$ | $y^\star_{i,T\leftarrow 1}$ | $y^\star_{i,T\leftarrow o}$ | $x_{1,i}$ | $x_{2,i}$ | $x_{3,i}$ | $x_{4,i}$ |
| 1 | 1 | 121 | 121 | ? | 37 | 14 | 160 | 56 |
| 2 | 0 | 109 | ? | 109 | 28 | 12 | 156 | 54 |
| 3 | 1 | 162 | 162 | ? | 53 | 11 | 190 | 87 |

**Counterfactual fairness** Kusner et al. (2017) If the prediction in the real world is the same as the prediction in the counterfactual world where the individual would have belonged to a different demographic group, we have counterfactual fairness, i.e.

$$\mathbb{P}[Y^\star_{S\leftarrow 1} = 1 | \boldsymbol{X} = \boldsymbol{x}] - \mathbb{P}[Y^\star_{S\leftarrow 0} = 1 | \boldsymbol{X} = \boldsymbol{x}] = 0, \ \forall \boldsymbol{x}.$$

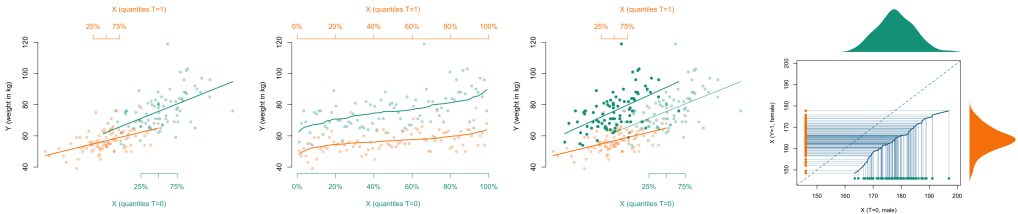
# Conterfactual fairness

Related to the concept of conditional average treatment,

**CATE** Hahn (1998), Heckman et al. (1998) Conditional ATE is CATE($x$)

$$\mathbb{E}\big[Y^*_{T\leftarrow 1} - Y^*_{T\leftarrow 0}\big|X = x\big]$$

Unfortunately, standard estimates are *ceteris paribus*, see Charpentier et al. (2023) at ECONVN2023 for a *mutandis mutatis* estimate, using optimal transport...



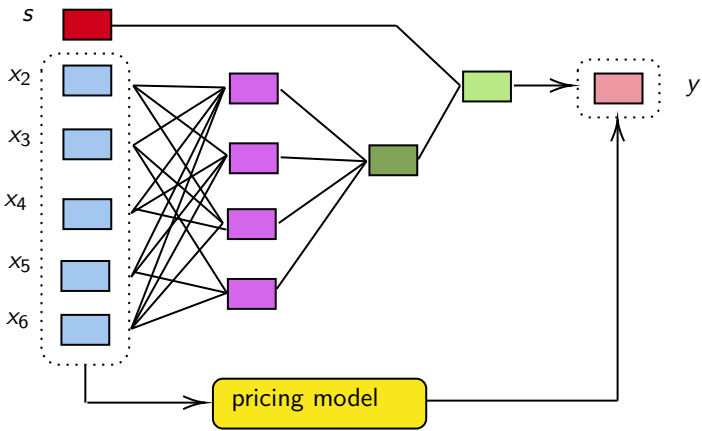
# Correction and mitigation

- Pre-processing changing the training data
  use weights or change $\boldsymbol{x}$'s into fair-$\boldsymbol{x}$'s
  Luong et al. (2011), Kamiran and Calders (2012), Zemel et al. (2016)
- In-processing adding a penalty term to balance accuracy with fairness
  Berk et al. (2017), Agarwal et al. (2018, 2019)
- Post-processing select separately thresholds for each group
  maximizes accuracy and minimizes demographic parity
  Corbett-Davies et al. (2017) and Menon and Williamson (2018)

# Pre-processing

$\boldsymbol{S}$ collection of $k$ sensitive variables, $\boldsymbol{S} = (\boldsymbol{s}_1 \ \cdots \ \boldsymbol{s}_k)$, i.e. a $n \times k$ matrix.
The orthogonal projection on $\mathcal{V}\{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_k\}$ is associate to matrix
$\Pi_S = \boldsymbol{S}(\boldsymbol{S}^\top \boldsymbol{S})^{-1}\boldsymbol{S}^\top$, while $\Pi_{S^\perp} = \mathbb{I} - \Pi_S$ (see Gram-Schmidt orthogonalization).
Let $\widetilde{\boldsymbol{S}}$ denote the collection of centered vectors
If $\boldsymbol{X} = (\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_p)$, for any $\boldsymbol{x}_j$, define $\boldsymbol{x}_{j^\perp} = \Pi_{\widetilde{S}^\perp}\boldsymbol{x}_j$.
$\boldsymbol{x}_{j^\perp}$ is orthogonal to any $\boldsymbol{s}_\ell$.
Then, train model $m$ on $\boldsymbol{x} = (x_{1^\perp}, \cdots, x_{p^\perp})$
And similarly the centered version of $\boldsymbol{x}_{j^\perp}$ is then also orthogonal to any $\boldsymbol{s}$.
See also Samadi et al. (2018) and Pelegrina et al. (2022) on fair-PCA, to reduce dimension, or Grari et al. (2022) for the use of fair-autoencoders

# In-processing

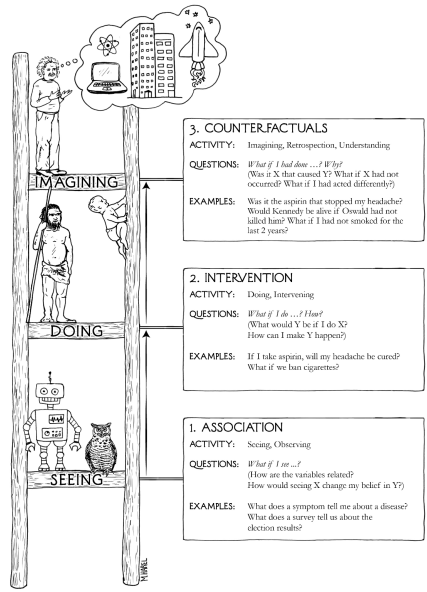



Penalizing discrimination $\min \left\{ \sum_{i=1}^{n} \ell(y_i, \widehat{y}_i)) + \lambda \cdot \text{cor}(\widehat{y}, s) \right\}$ (adversarial learning)

# Take-away

- quick survey on fairness and discrimination
- ongoing work on individual fairness and causality (see ECONVN2023)
- Insurance: biases, discrimination & fairness (to appear in 2023)

Source: Pearl and Mackenzie (2018)



# References


Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR.

Agarwal, A., Dudík, M., and Wu, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR.

Agarwal, S. (2021). Trade-offs between fairness and interpretability in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression. *arXiv*, 1706.02409.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., et al. (2019). Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2212–2220.

# References


Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):1–21.

Charpentier, A. (2023). *Insurance: biases, discrimination and fairness*. Springer Verlag.

Charpentier, A., Flachaire, E., and Gallic, E. (2023). Causal inference with optimal transport. In Thach, N. N., Kreinovich, V., Ha, D. T., and Trung, N. D., editors, *Optimal Transport Statistics for Economics and Related Topics*. Springer Verlag.

Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *arXiv*, 1701.08230.

Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

# References


Denuit, M., Charpentier, A., and Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics & Economics*.

Duivesteijn, W. and Feelders, A. (2008). Nearest neighbour classification with monotonicity constraints. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 301–316. Springer.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

Grari, V., Charpentier, A., Lamprier, S., and Detyniecki, M. (2022). A fair pricing model via adversarial learning. *ArXiv*, 2202.12008.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

# References


Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *The review of economic studies*, 65(2):261–294.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Kim, P. T. (2017). Auditing algorithms for discrimination. *University of Pennsylvania Law Review*, 166:189.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv*, 1609.05807.

Kranzberg, M. (1986). Technology and history:" kranzberg's laws". *Technology and culture*, 27(3):544–560.

Krüger, F. and Ziegel, J. F. (2021). Generic conditions for forecast dominance. *Journal of Business & Economic Statistics*, 39(4):972–983.

# References


Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. NIPS.

Luong, B. T., Ruggieri, S., and Turini, F. (2011). $k$-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Menon, A. K. and Williamson, R. C. (2018). The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR.

Narayanan, A. (2018). Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, volume 1170, page 3.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Pelegrina, G. D., Brotto, R. D., Duarte, L. T., Attux, R., and Romano, J. M. (2022). Analysis of trade-offs in fair principal component analysis based on multi-objective optimization. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

# References


Samadi, S., Tantipongpipat, U., Morgenstern, J. H., Singh, M., and Vempala, S. (2018). The price of fair pca: One extra dimension. *Advances in neural information processing systems*, 31.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7.

Vogel, R., Bellet, A., Clémen, S., et al. (2021). Learning fair scoring functions: Bipartite ranking under roc-based fairness constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 784–792. PMLR.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *arXiv*, 1507.05259.

Zemel, Y. L. M. W. R., Louizos, C., and Swersky, K. (2016). The variational fair autoencoder. In *Proceedings of the international conference on learning representations*.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *arXiv*, 1801.07593.