Foundation of Data Science
Lecture 11, Module 2
Fall 2022

Rumi Chunara, PhD

Slides from KDD 2016 Tutorial given by Sara Hajian, Francesco Bonchi, Carlos Castillo,
ICML 2019 Tutorial by Silvia Chiappa & Jan Leike
and slides on Algorithmic Fairness from Julia Stoyanovich incorporated.

Addressing AI Biases

DS-GA 3001.009: Special Topics in Data Science: Responsible Data Science

CS 294: Fairness in Machine Learning

UC Berkeley, Fall 2017

Catalog › Computer Science Courses

Data Science Ethics

Learn how to think through the ethics surrounding privacy, data sharing, and algorithmic decision-making.

UNIVERSITY OF MICHIGAN

Fair ML for Health

NeurIPS 2019 Workshop, Vancouver Convention Center, Canada

Definitions of “Bias”

Statistical bias: a model is biased if it doesn’t summarize the data correctly

Societal bias: a dataset or a model is biased if it does not represent the world “correctly”, e.g., data is not representative, there is measurement error, or the **world is “incorrect”**

The world as it is or as it should be?

More on Statistical Bias

Is statistical **bias** sufficient to address?

- A common view: “The model summarizes the data correctly. If the data is biased - it’s not the algorithm’s fault”

However:

- statistical bias says nothing about error distribution
- data biases are inevitable - training data is not identical between groups - we must account for them

Reframing: focus on designing systems that support human values.

Sometimes we may decide to introduce statistical bias to correct for societal bias!

Bias in Algorithms

- Algorithms can **treat features differently** in inference (Pruning and quantizing deep neural networks amplifies algorithmic bias. (e.g. Hooker, Sara, et al. "Characterising bias in compressed models." *arXiv preprint arXiv:2010.03058* (2020), Hooker, Sara, et al. "What Do Compressed Deep Neural Networks Forget?." *arXiv preprint arXiv:1911.05248* (2019).)

- Work on memorization and variance of gradients (VoG) shows that **hard examples are learnt later in training**, and that **learning rates impact what is learnt** (Agarwal, Chirag, and Sara Hooker. "Estimating Example Difficulty using Variance of Gradients." *arXiv preprint arXiv:2008.11600* (2020)., Jiang, Ziheng, et al. "Characterizing Structural Regularities of Labeled Data in Overparameterized Models." *arXiv e-prints* (2020): arXiv-2002.)

- Models which are guaranteed to be differentially private introduce **disparate impact on model accuracy** (Bagdasaryan, Eugene, Omid Poursaeed, and Vitaly Shmatikov. "Differential privacy has disparate impact on model accuracy." *Advances in Neural Information Processing Systems* 32 (2019): 15479-15488.)

Legal constructs of discrimination

Disparate treatment is the illegal practice of treating an entity, such as a creditor or employee, differently based on a **protected characteristic** such as race, gender, age, religion, sexual orientation, or national origin.

Disparate impact is the result of systematic disparate treatment, where disproportionate **adverse impact** is observed on members of a **protected class**.

http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

Protected class: A **protected class** is a group of people sharing a common trait who are legally **protected** from being discriminated against on the basis of that trait. Examples of **protected** traits include race, gender, age, disability, and veteran status.

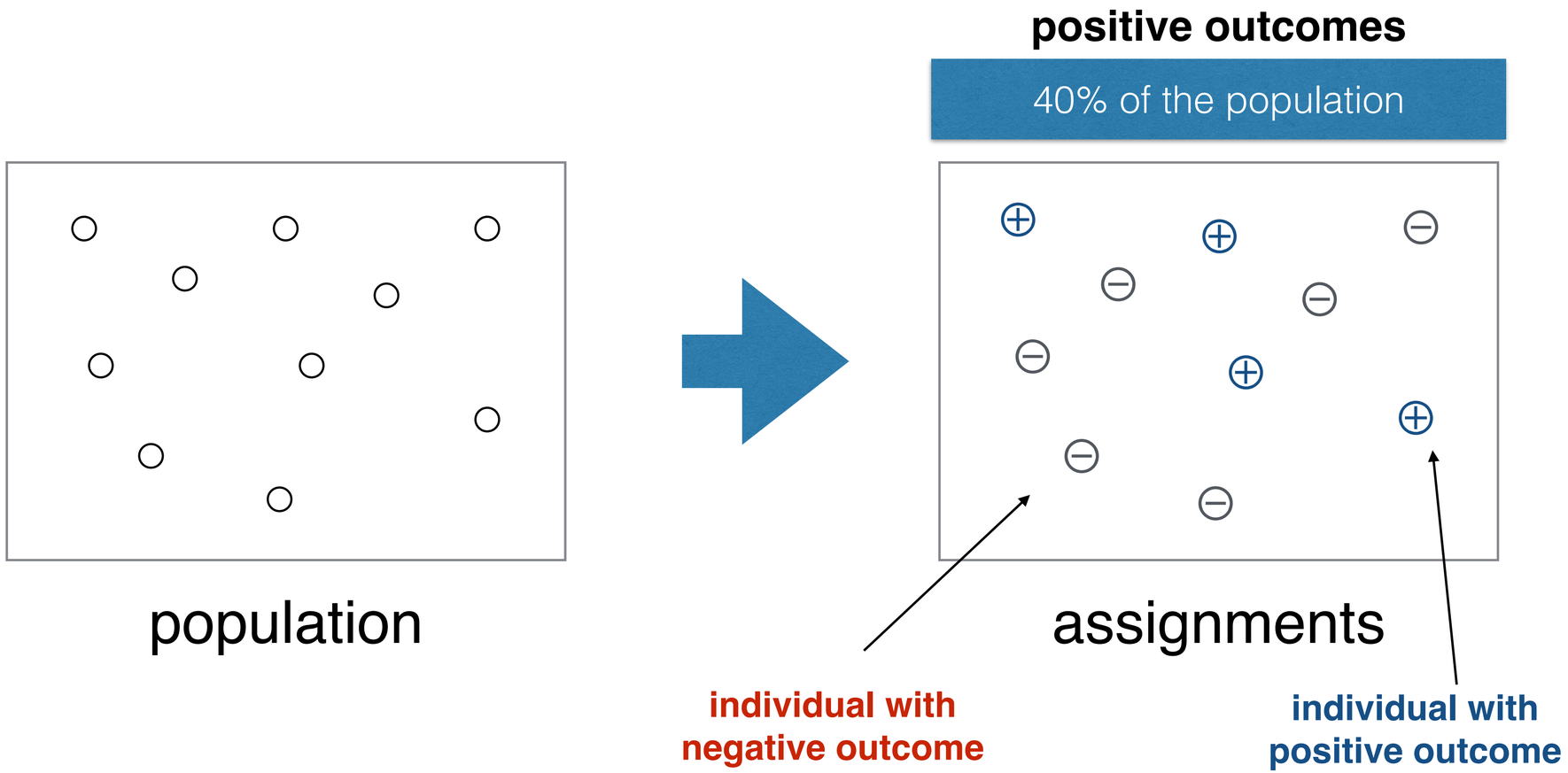
Example Algorithms

Consider the following algorithms in practice for which positive or negative **outcomes** are assigned to individuals.

Positive Outcomes	Negative Outcomes
offered employment	denied employment
accepted to school	rejected from school
offered a loan	denied a loan
offered a discount	not offered a discount

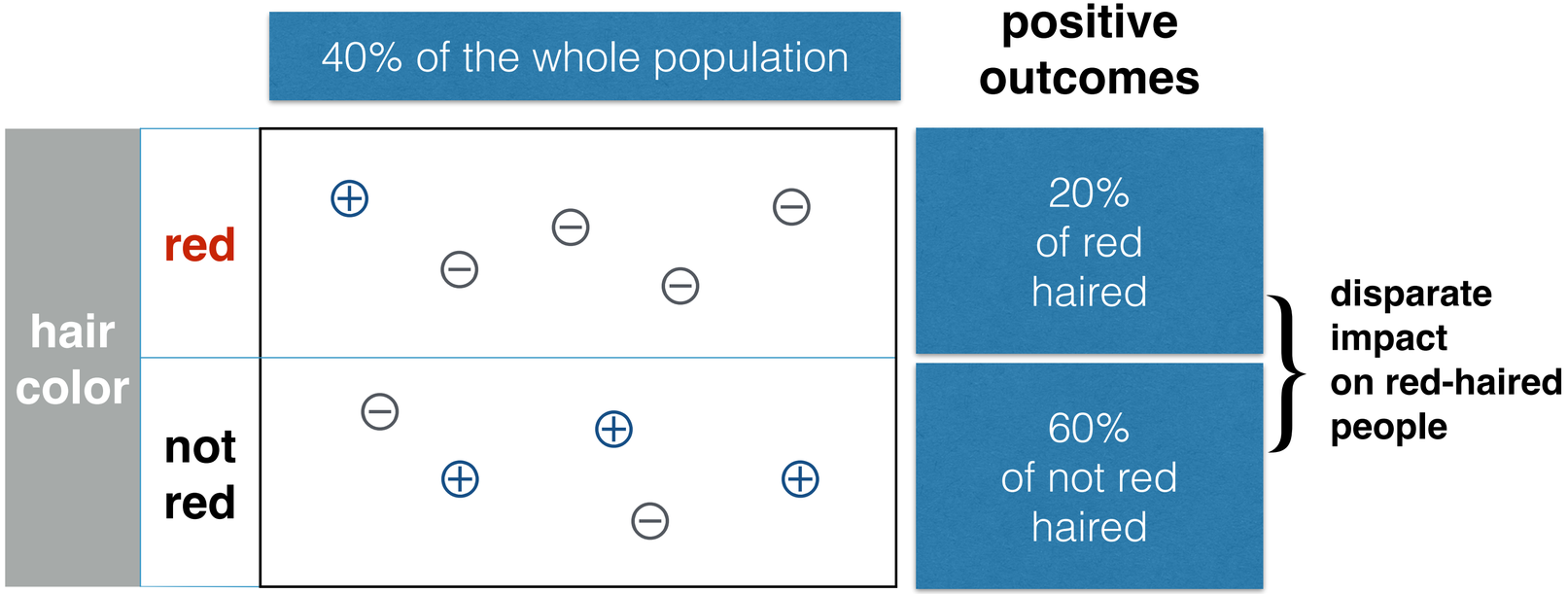
Assigning outcomes to populations

Algorithmic Fairness is concerned with how outcomes are assigned to a population

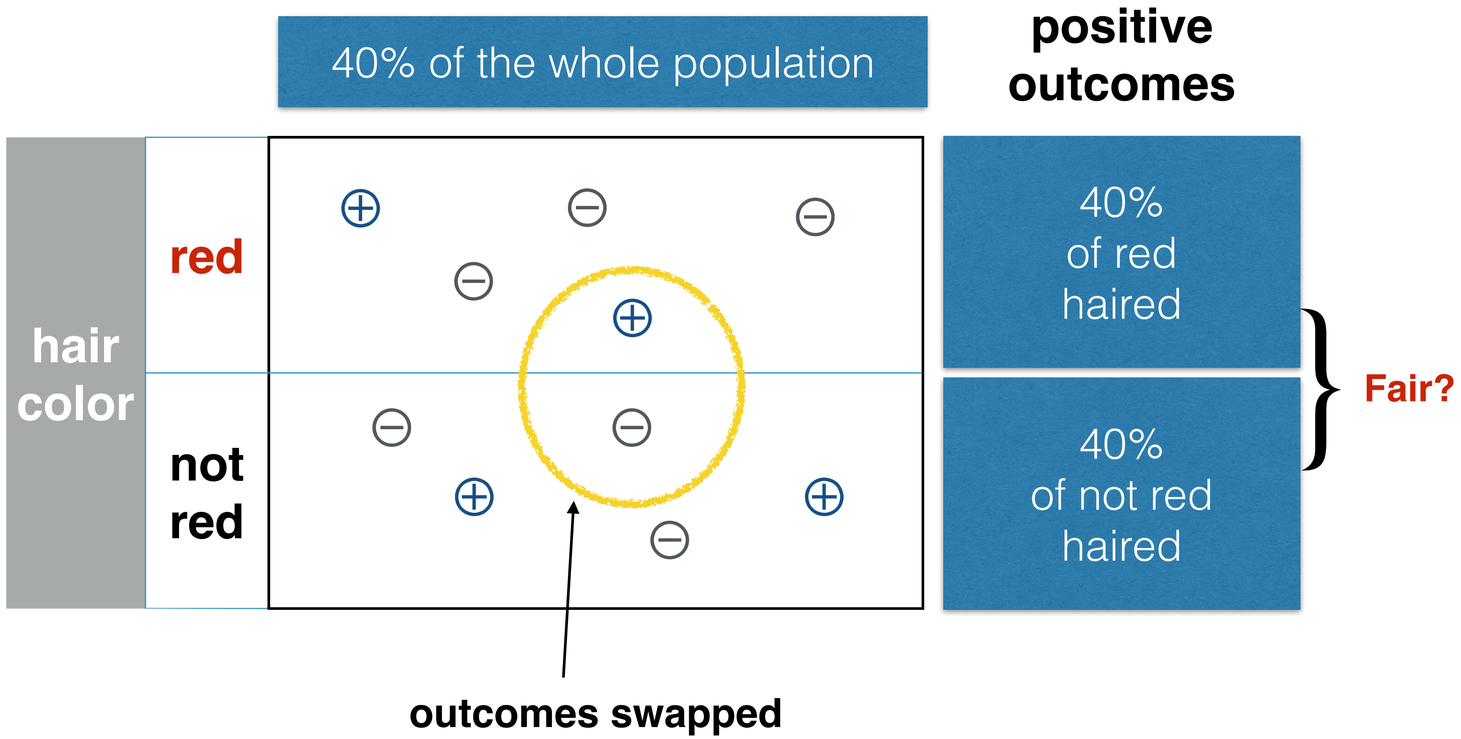


Sub-populations may be treated

Sub-population: those with red hair
(under the same assignment of outcomes)

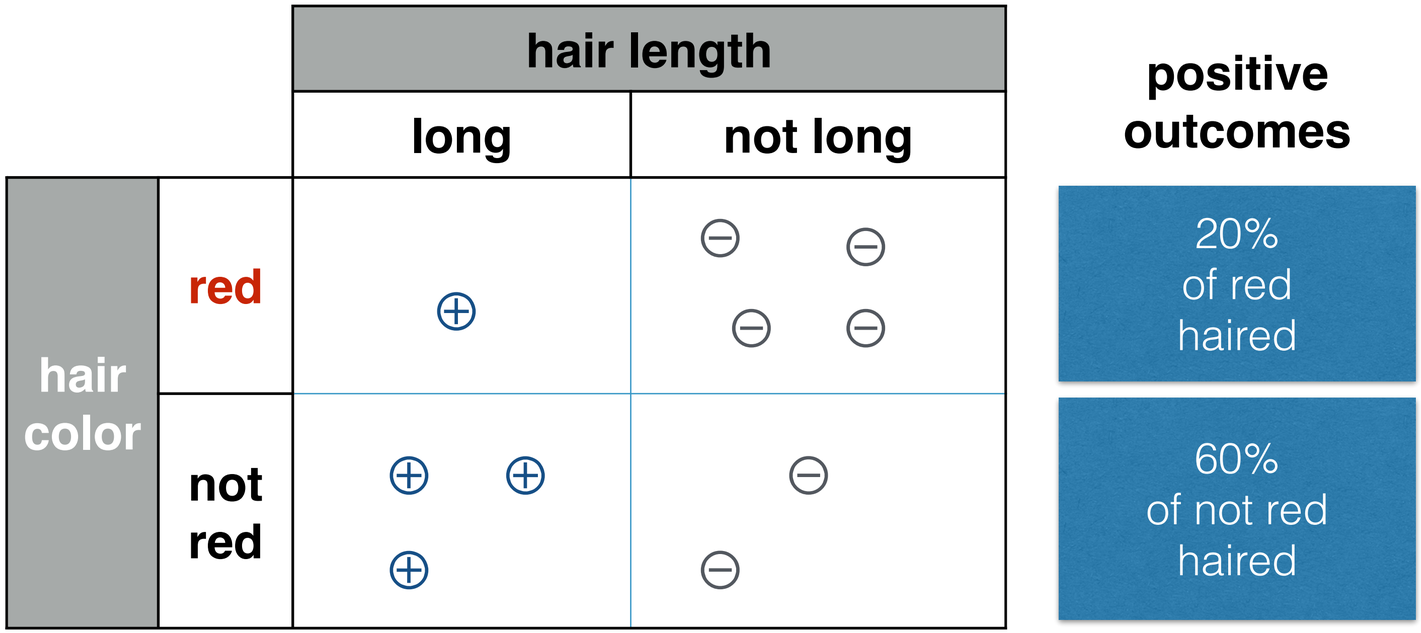


Outcome fairness



Redundant encoding

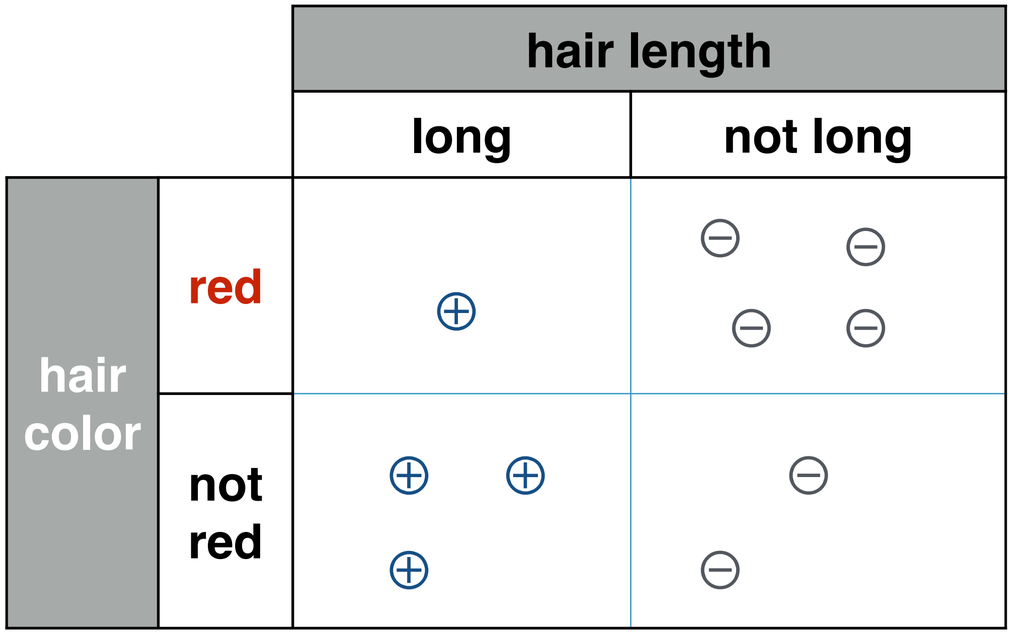
Now consider the assignments under both
hair color (protected) and **hair length** (innocuous)



Deniability
The algorithm has adversely impacted red-haired people, but claims that outcomes are assigned according to hair length.

Blinding is not an excuse

Removing **hair color** from the algorithm's assignment process does not prevent discrimination!



positive outcomes

20% of red haired

60% of not red haired

Assessing disparate impact

Discrimination is assessed by the effect on the protected sub-population, not by the input or by the process that lead to the effect.

Redundant encoding

Let's replace hair color with **race** (protected),
hair length with **zip code** (innocuous)

		zip code	
		10025	**10027**
race	**black**	⊕	⊖ ⊖ ⊖ ⊖
	white	⊕ ⊕ ⊕	⊖ ⊖

positive outcomes

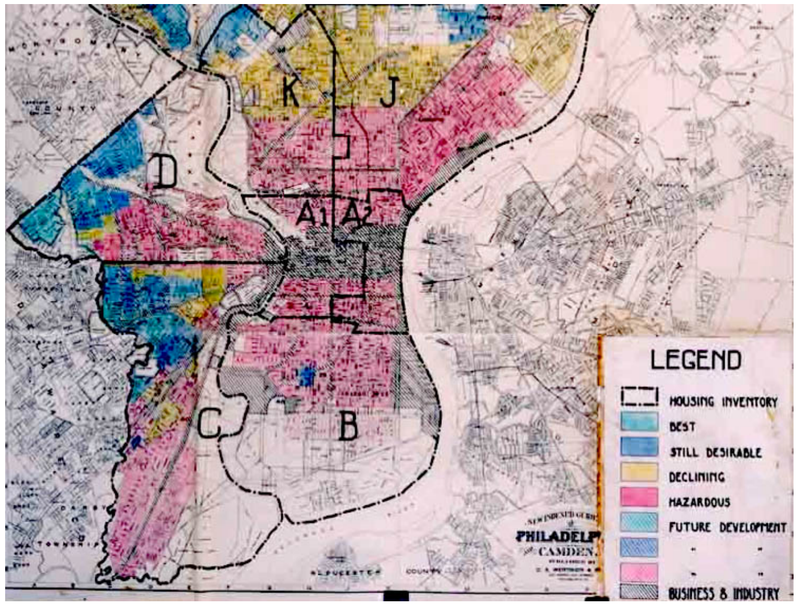
20% of black

60% of white

Redlining

Redlining is the practice of arbitrarily denying or limiting financial services to specific neighborhoods, generally because its residents are people of color or are poor.

Philadelphia, 1936



wikipedia

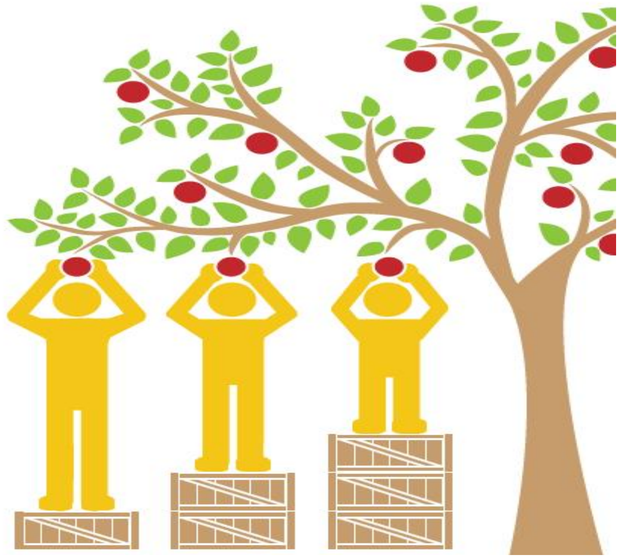
Households and businesses in the red zones could not get mortgages or business loans.

Two notions of fairness

individual fairness

equality

group fairness



equity

Effect on sub-populations

Simpson's paradox
disparate impact at the full population level disappears or reverses when looking at sub-populations!

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

favored group

women
women
men
women
men
women

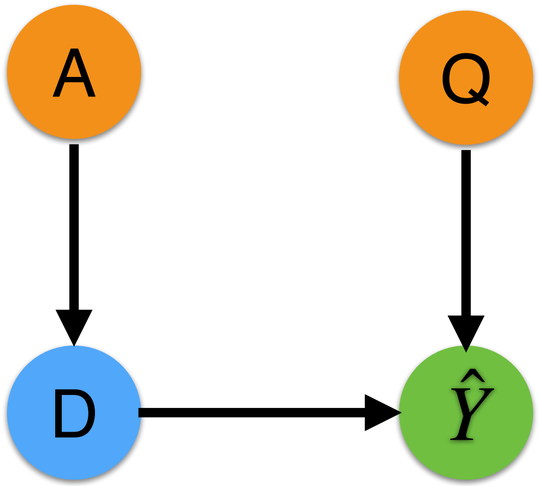
whole population

35%
of women

44%
of men

UC Berkeley 1973: women applied to more competitive departments, with low rates of admission among qualified applicants.

Bayesian Networks for Causal Fairness Analysis

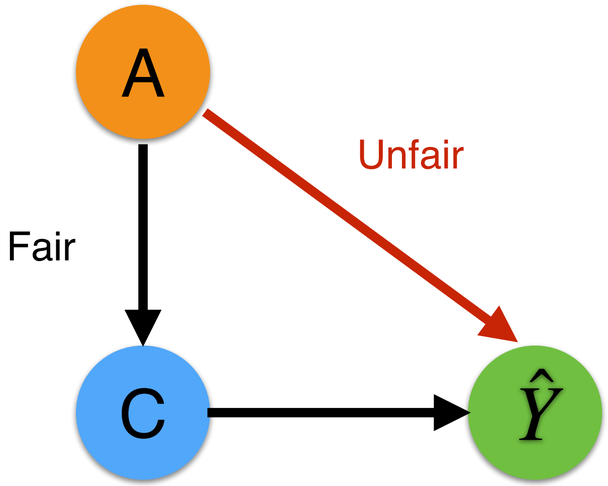


Nodes represent random variables:

- A = gender
- Q = Qualifications
- D= Department choice
- Y = admission decision

Links express (directed) causal influence.

Unfair Causal Paths



Understand whether there is a direct influence of A on $\hat{Y}$, namely a direct path A → $\hat{Y}$, by checking whether

$$p(Y = 1 \mid \hat{Y} = 1, A = 0) = p(Y = 1 \mid \hat{Y} = 1, A = 1)$$

Algorithmic Fairness Metrics and Approaches

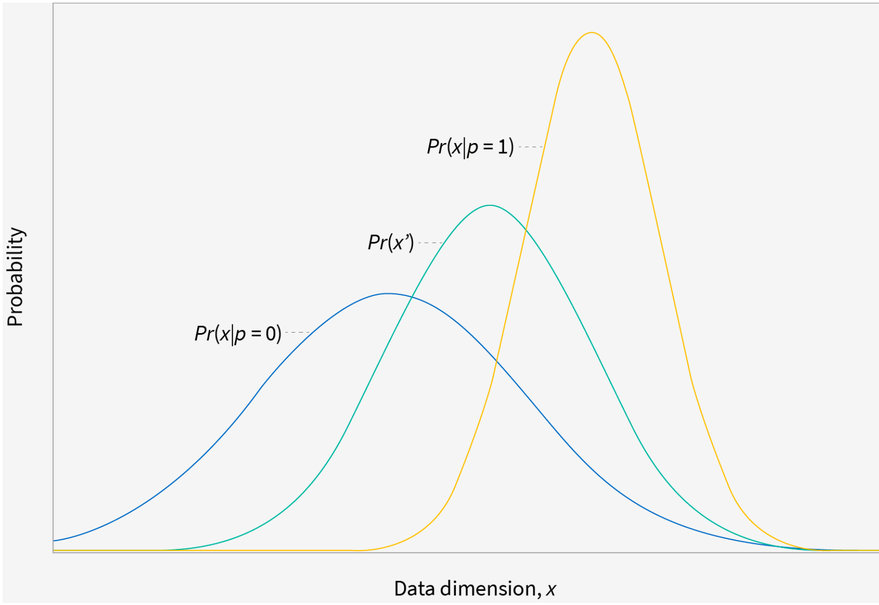
Attention to Fairness through the Data Lifecycle

Post-Processing	In-Processing	Pre-Processing	Data Collection
• Change thresholds • Trade off accuracy for fairness	• Adversarial training • Regularize for fairness • Constrain to be fair	• Modify labels • Modify input data • Modify label/data pairs • Weight label/data pairs	• Identify lack of examples or variates and collect

Pre-processing to manually remove discrimination from the dataset

Manipulate labels: Change some of the training labels, classify, for decisions close to boundary swap labels so positive outcome more likely for disadvantaged group, repeat.

Manipulate observed data: Manipulate individual data dimensions *x* in a way that depends on the protected attribute *p,* i.e. align cumulative distributions for feature *x* when *p* is 0 and 1 to a median distribution (right).



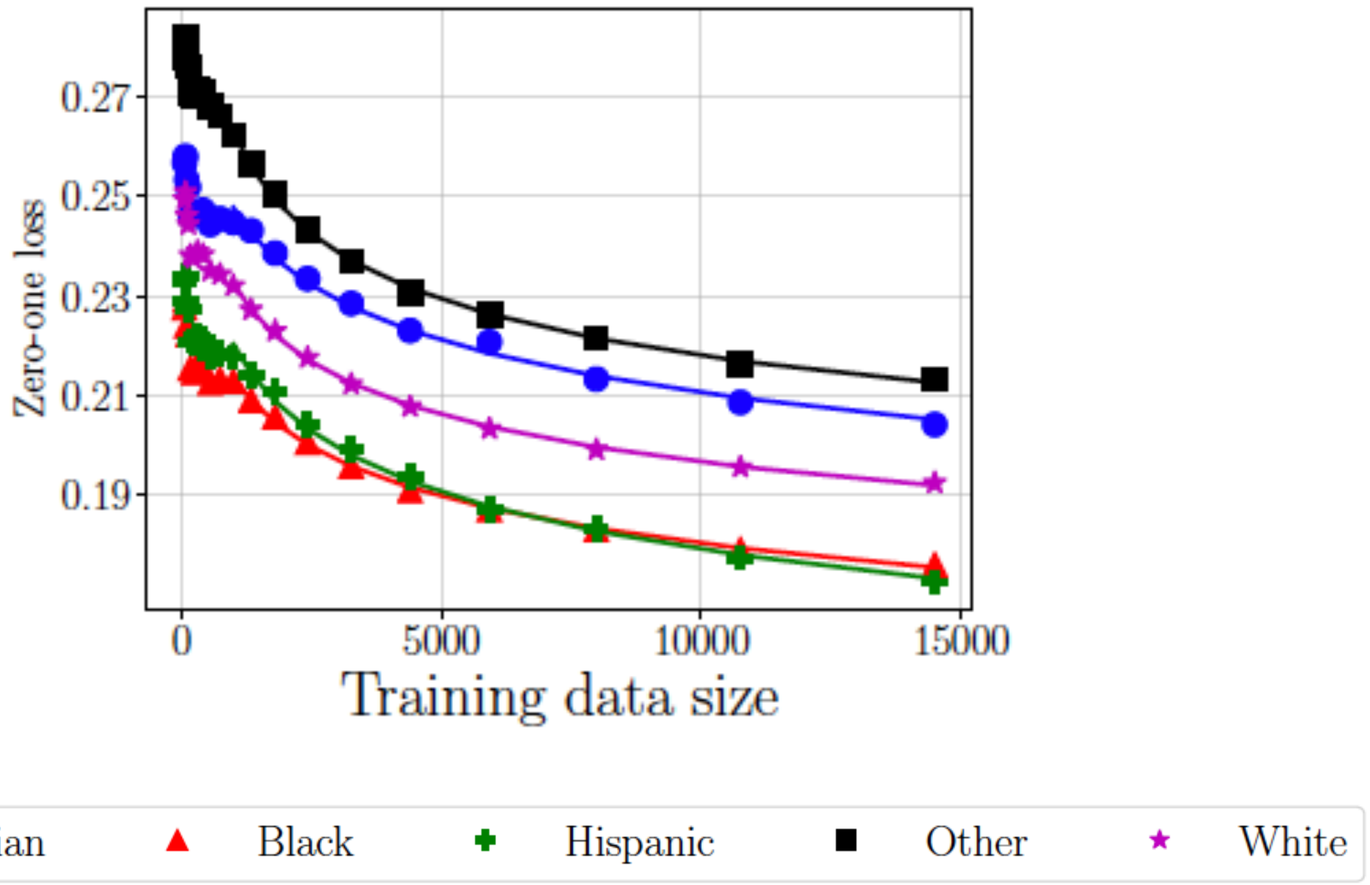
Manipulate labels and data: Learn a randomized transformation $Pr(\mathrm{x}', y' \mid \mathrm{x}, y, p)$ that transforms data pairs $\{\mathrm{x}, y\}$ to new data values $\{\mathrm{x}', y'\}$ in a way that depends on the protected attribute *p* (optimization subject to limits on prejudice/distortion of original values).

Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." *Knowledge and Information Systems* 33.1 (2012): 1-33.

Feldman, Michael, et al. "Certifying and removing disparate impact." *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015.

Pre-processing

Data collection: collection of additional samples in minority groups decreases loss.



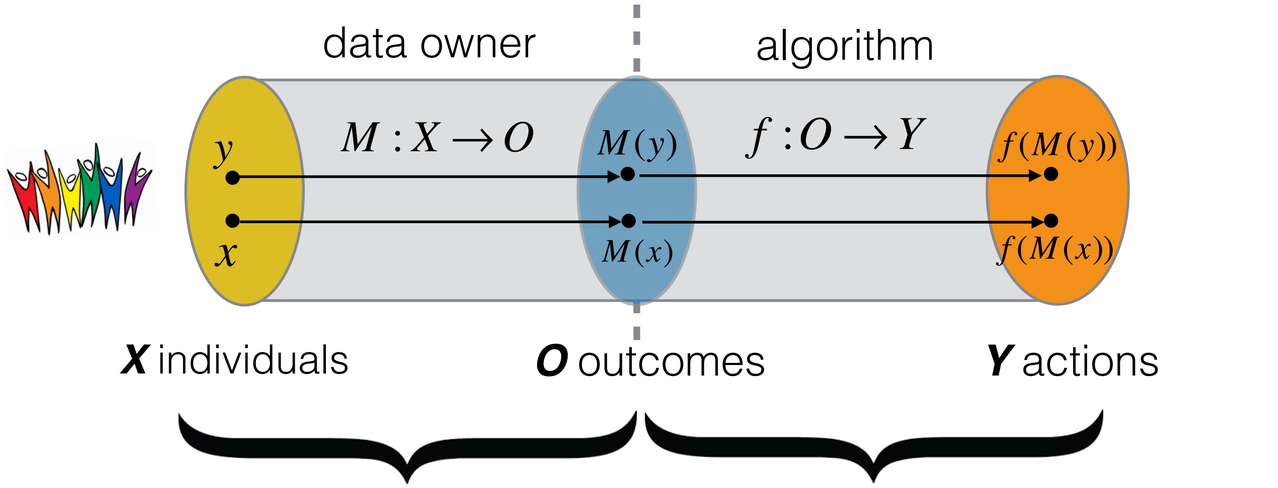
Chen, Irene Y., Fredrik D. Johansson, and David Sontag. "Why is my classifier discriminatory?." *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018.

APA

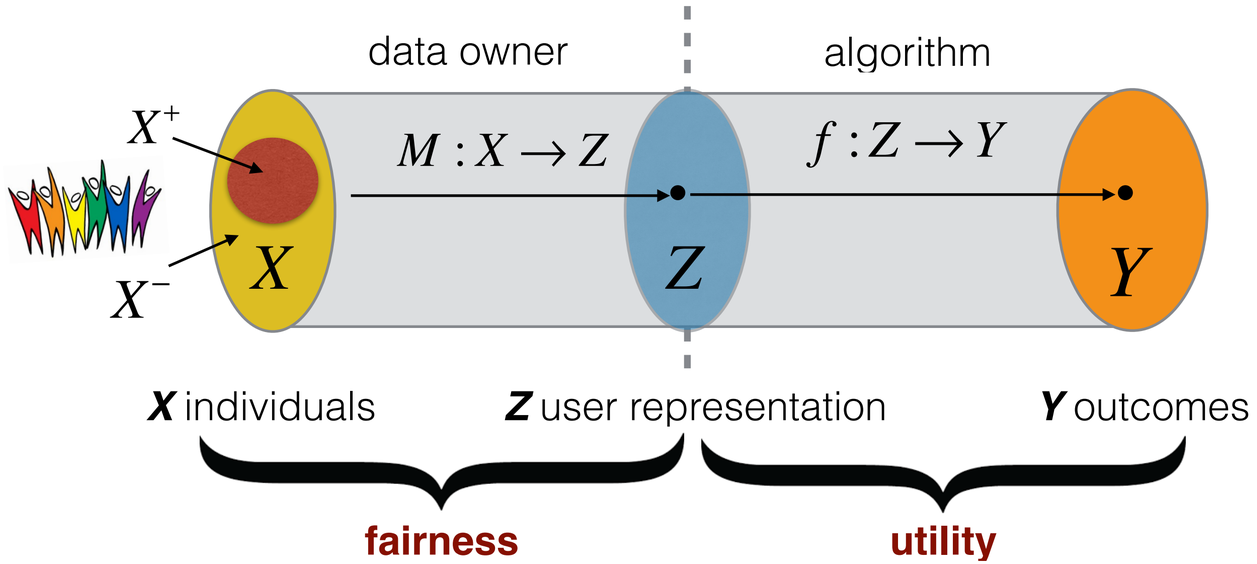
Pre/In-process

Fairness through awareness: Individuals who are similar for the purpose of classification task should be treated similarly.

Learning fair representations: Idea: remove reliance on a "fair" similarity measure, instead learn representations of individuals, distances



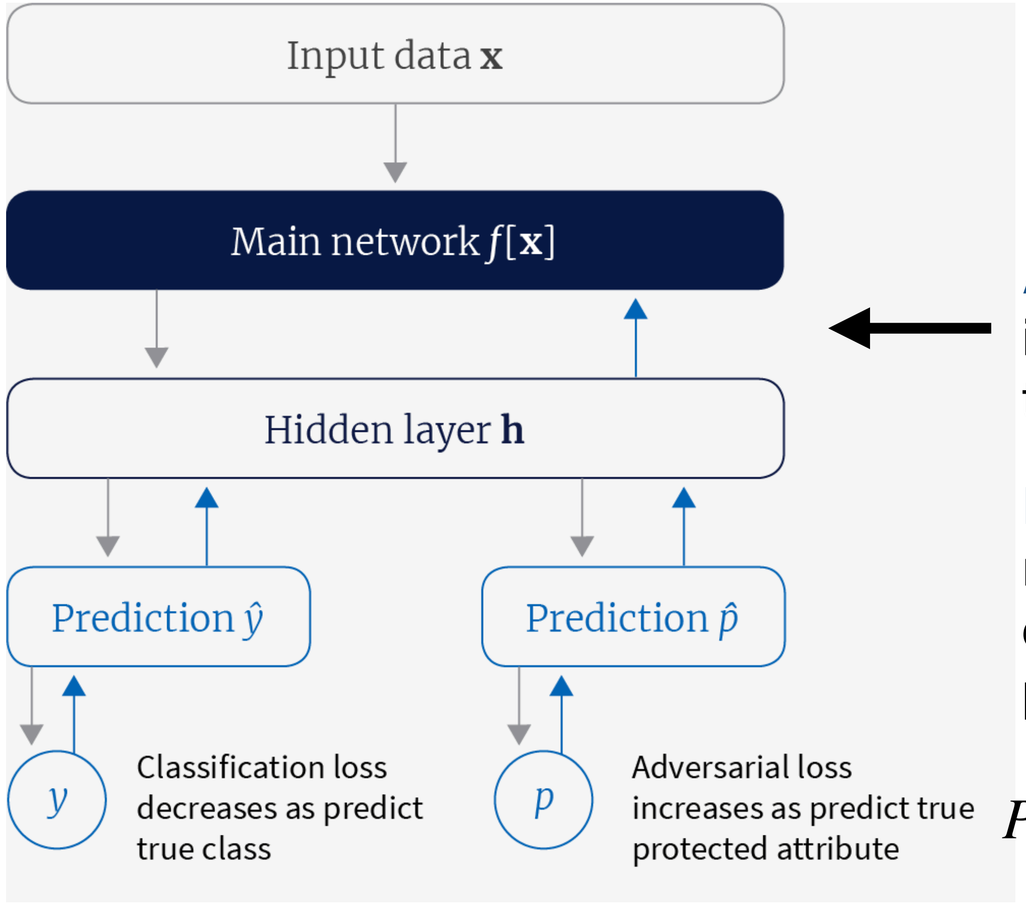
$$\forall x, y \in X \quad \left|\left|M(x), M(y)\right|\right| \leq d(x, y)$$



Zemel, Rich, et al. "Learning fair representations." *International conference on machine learning*. PMLR, 2013.

Dwork, Cynthia, et al. "Fairness through awareness." *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012.

In-processing Algorithms



Adversarial de-biasing: reduces evidence of protected attributes in predictions by trying to simultaneously fool a second classifier that tries to guess the protected attribute p (right).

Prejudice removal by regularization: Add an extra regularization condition to the output of logistic regression classifier that tries to minimize the mutual information between protected attribute *p* and prediction, *y*.

$$PI = \sum_{y,p} Pr(y \mid \mathbf{x}, p) \log \frac{Pr(y,p)}{Pr(y)Pr(p)} = \sum_{y,p} Pr(y \mid \mathbf{x}, p) \log \frac{Pr(y \mid p)}{Pr(y)}$$

$$L_{reg} = \sum_{i} \sum_{y,p} Pr(\hat{y}_i \mid \mathbf{x}_i, p_i) \log \frac{Pr(\hat{y}_i \mid p)}{Pr(y)}$$

Beutel, Alex, et al. "Data decisions and theoretical implications when adversarially learning fair representations." *arXiv preprint arXiv:1707.00075* (2017).

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning." Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018.

Kamishima, Toshihiro, Shotaro Akaho, and Jun Sakuma. "Fairness-aware learning through regularization approach." *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011.

Fairness Constraints in Optimization

Minimize loss with fairness constraints for a distribution *D:*

$$\min_{\theta \in \Theta} \mathit{Loss}(\theta, D)$$

$$\mathit{s.t.}\ \mathit{FairViol}(\theta, D) < \epsilon$$

Singh, H., Singh, R., Mhasawade, V. and Chunara, R., 2021, March. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 3-13).

Post-processing

Group Statistical Fairness Metrics

Statistical parity [Dwork et al. 2012]: equality of acceptance rates across groups

Demographic parity or statistical parity suggests that a predictor is unbiased if the prediction $\hat{y}$ is independent of the protected attribute p: $Pr(\hat{y} \mid p) = Pr(\hat{y})$

Deviations from statistical parity (or *disparate impact* which is the ratio of the two terms*:*

$$SPD = Pr(\hat{y} = 1, p = 1) - Pr(\hat{y} = 1, p = 0)$$

Equalized odds [Hardt et al. 2016]: equality of false positive, false negative rates across groups

Equality of odds is satisfied if the prediction $\hat{y}$ is conditionally independent to the protected attribute p, given the true value y:

$$Pr(\hat{y} \mid y, p) = Pr(\hat{y} \mid y)$$

Equality of opportunity has the same mathematical formulation as equality of odds, but is focused on one particular label y=1y=1 of the true value so that:

$$Pr(\hat{y} \mid y = 1, p) = Pr(\hat{y} \mid y = 1)$$

Calibration [Kleinberg et al. 2016]: equality of positive predictive values (PPV) across groups

$$Pr(y = 1 \mid p = 1) = Pr(y = 1 \mid p = 0)$$

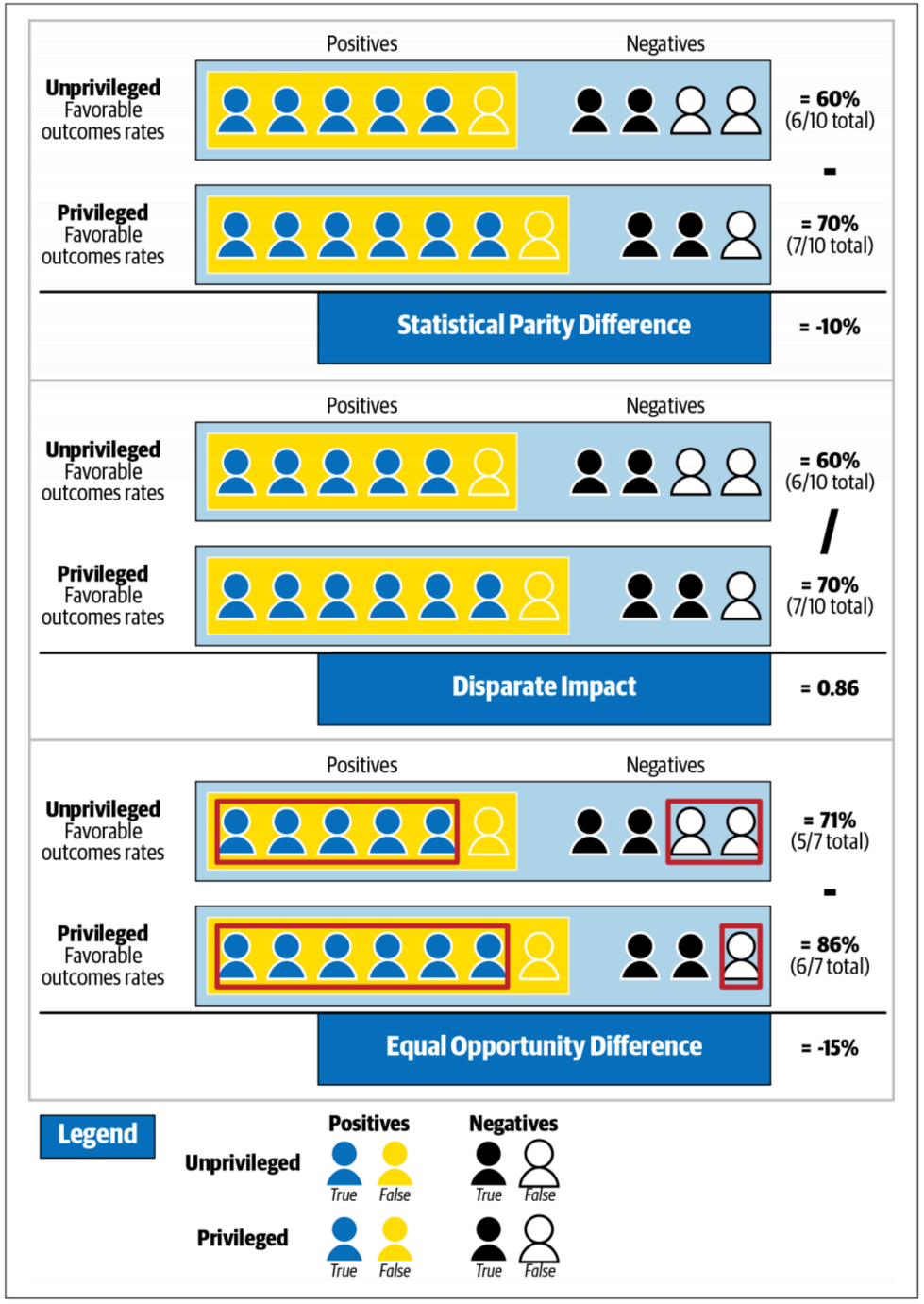


Figure 1-2. How bias is measured

How do decide which metrics?

In **assistive cases** like receiving extra care, separation (equalized odds) is the preferred fairness metric because it relates to recall (true positive rate), which is of primary concern in these settings.

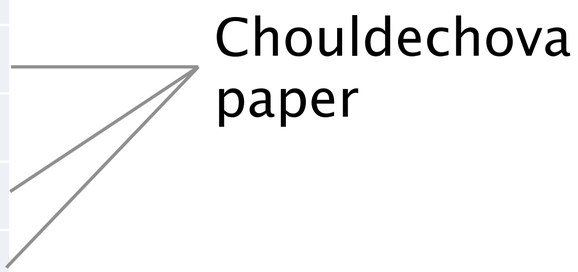
If receiving care management had been a **non-punitive act**, then sufficiency (calibration) would have been the preferred fairness metric because precision is of primary concern in non-punitive settings. (Precision is equivalent to positive predictive value, which is one of the two components of the average predictive value difference.)

Varshney, K.R., 2022. Trustworthy Machine Learning. *Chappaqua, NY, USA: Independently Published*.

Impossibility theorem

based on a slide by Arvind Narayanan

Metric	Equalized under
Selection probability	Demographic parity
Pos. predictive value	Predictive parity
Neg. predictive value	
False positive rate	Error rate balance
False negative rate	Error rate balance
Accuracy	Accuracy equity



All these metrics can be expressed in terms of FP, FN, TP, TN

If these metrics are equal for 2 groups, some trivial algebra shows that the prevalence (in the COMPAS example, of recidivism, as measured by re-arrest) is also the same for 2 groups

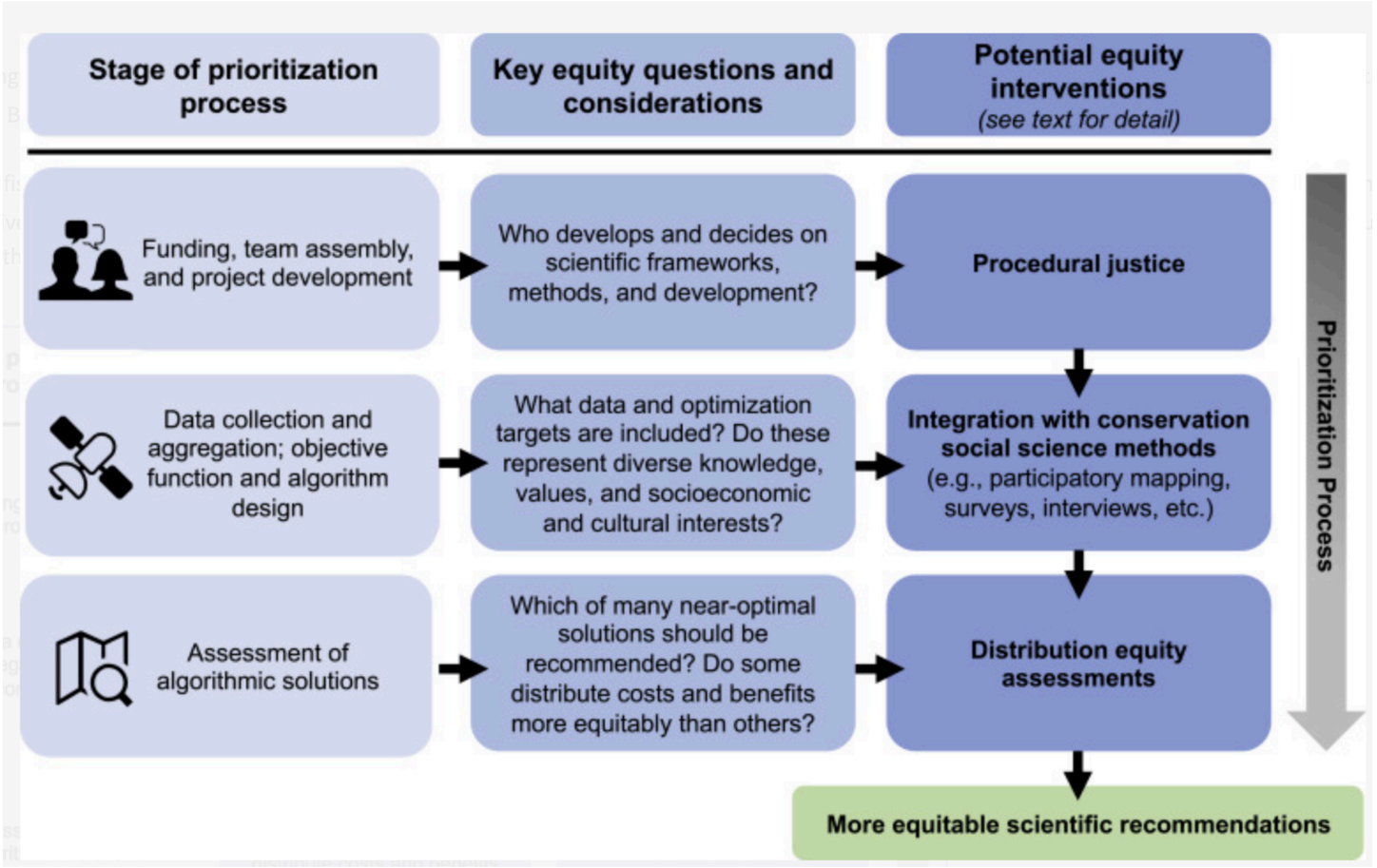
Ways to evaluate binary classifiers

based on a slide by Arvind Narayanan

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

364 impossibility theorems

Algorithmic Fairness is Not Enough



Chapman, Melissa S., et al. "Promoting equity in the use of algorithms for high-seas conservation." *One Earth* 4.6 (2021): 790-794.