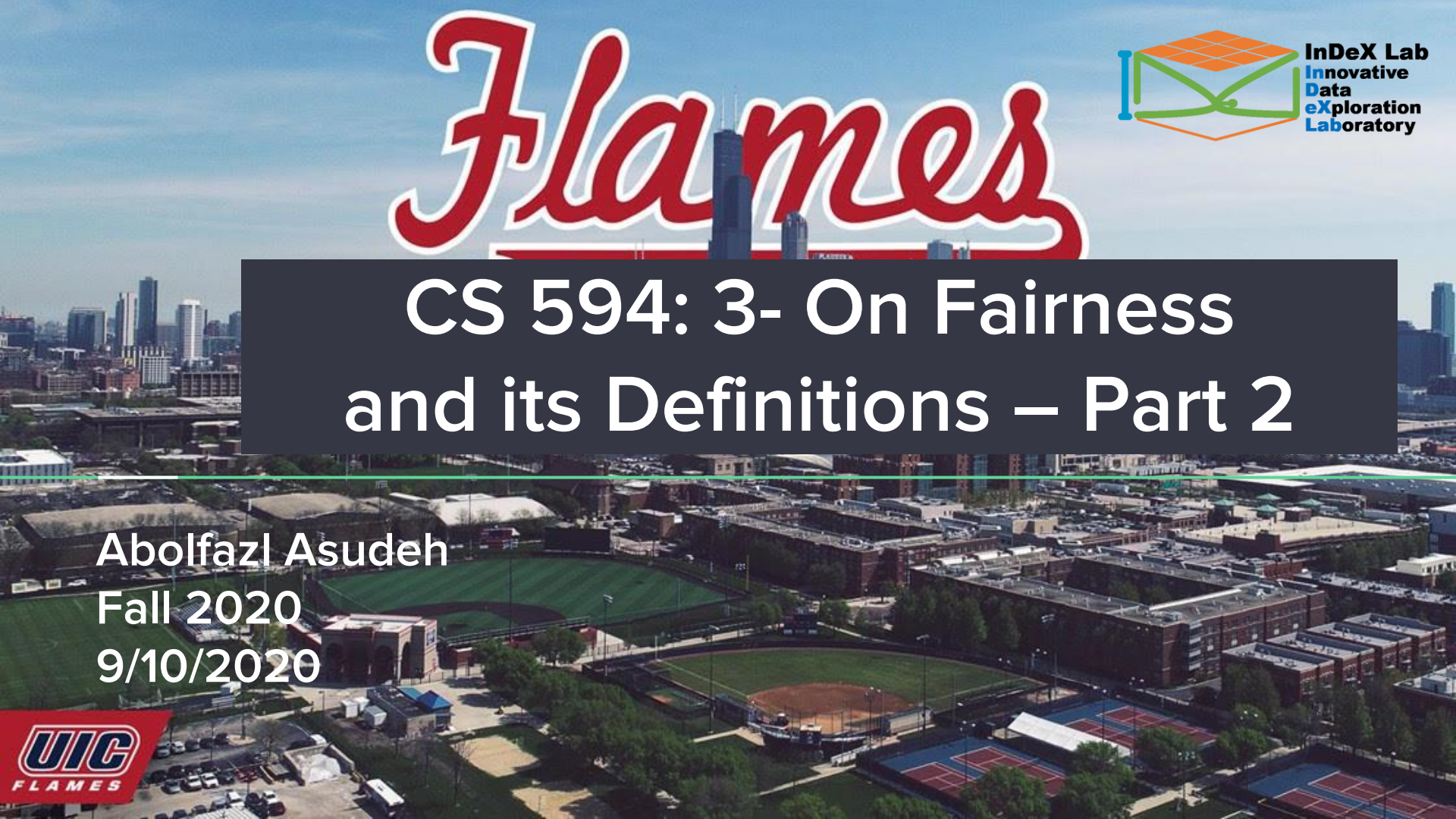


CS 594: 3- On Fairness and its Definitions – Part 2

Abolfazl Asudeh
Fall 2020
9/10/2020

References:

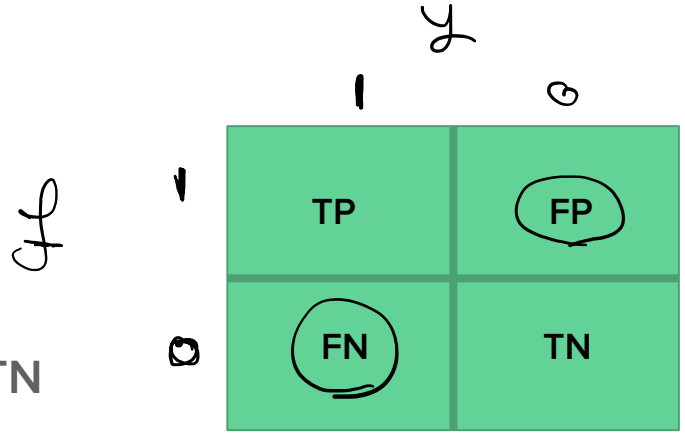
1. Solon Barocas, Moritz Hardt, Arvind Narayanan. “Fairness in Machine Learning, Limitations and Opportunities” (book)
2. (Tutorial) Arvind Narayanan. "21 fairness definitions and their politics" In FAT*, 2018.
3. Ashudeep Singh and Thorsten Joachims. "Fairness of exposure in rankings" In KDD. 2018
4. Ke Yang and Julia Stoyanovich. "Measuring fairness in ranked outputs." In SSDBM. 2017
5. Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im) possibility of fairness." arXiv:1609.07236 (2016).
6. Cynthia Dwork, et al. "Fairness through awareness." In ITCS. 2012.
7. (Tutorial) A. Asudeh, H. V. Jagadish. Fairly Evaluating and Scoring Items in a Data Set. PVLDB, 13(12): 3445-3448, 2020

Group Fairness

Fairness in (binary) classification

One very simple set up is to assume a classification task with a known correct answer, at least after the fact. Every classification algorithm is likely to have some error.

So, we define a matrix as in this figure:



An ideal classifier has only TP and FP.

But a real classifier has non-zero FN and TN

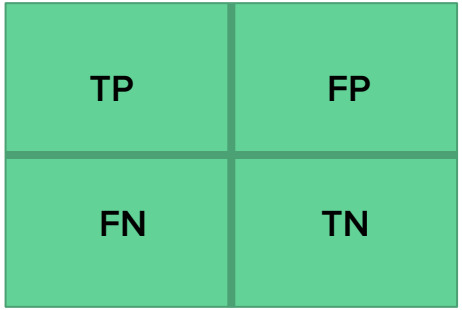
From Whose Perspective?

- Decision Maker
- Defendant

Decision maker:

Of those labeled positive, how many are truly positive

Positive Predictive Value: TP/(TP+FP)

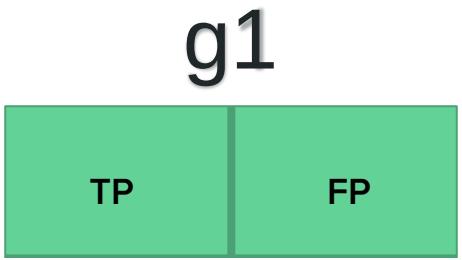


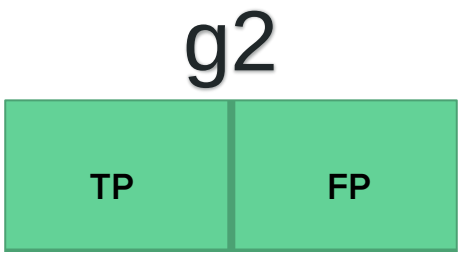
Decision maker:

Of those labeled positive, how many are truly positive

Positive Predictive Value: TP/(TP+FP)

PP1 = PP2



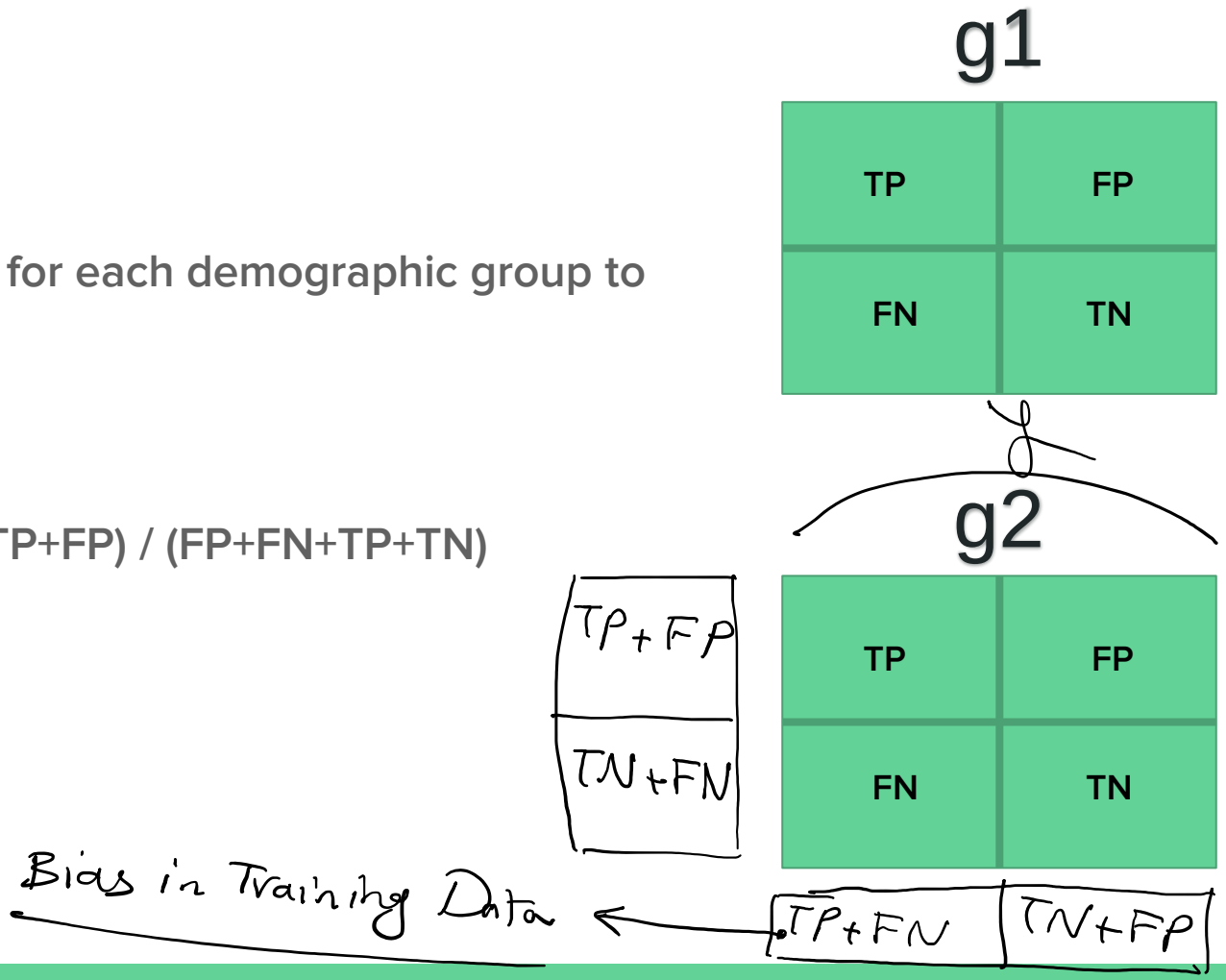


Defendant:

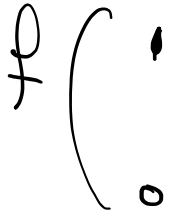
What is the likelyhood for each demographic group to be labeled as positive

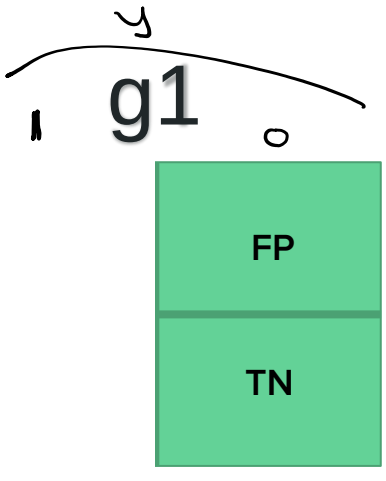
Demographic Parity: (TP+FP) / (FP+FN+TP+TN)

DP1 = DP2



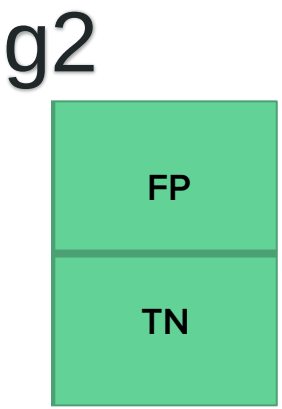
Decision maker:





How likely is an individual to be mistakenly labeled positive.

Of those that are truly negative, how many are labeled positive



False Positive Rate: FP/(FP+TN)

FP1 = FP2

Similarly:

False negative rate, True positive rate, True Negative rate

Decision maker:

How frequently does the system produce the wrong label

Error Rate: (FP+FN)/(FP+FN+TP+TN)

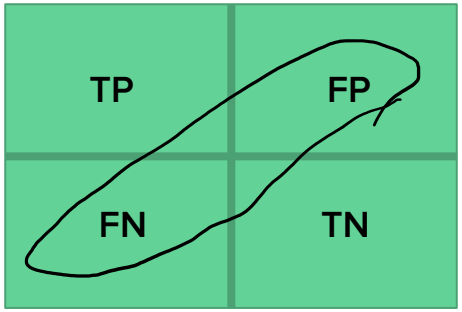
ER1 = ER2

How frequently does the system produce the correct label

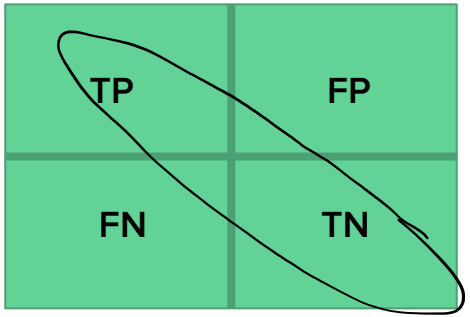
Accuracy: (TP+TN)/(FP+FN+TP+TN)

AR1 = AR2

g1



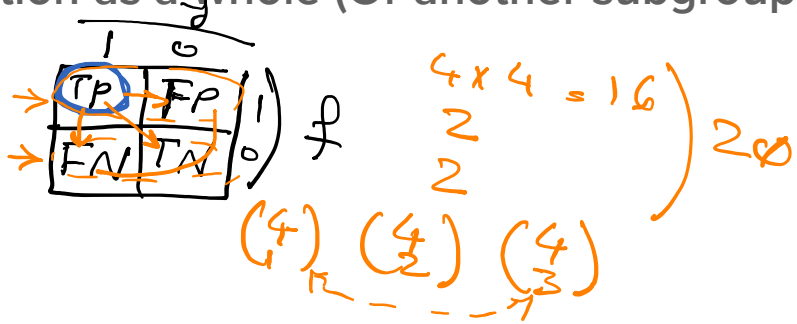
g2



21 definitions of fairness (Narayanan)

For every protected group, compute metric of choice, and compare against the same metric for the population as a whole (Or another subgroup).

How many can you count?

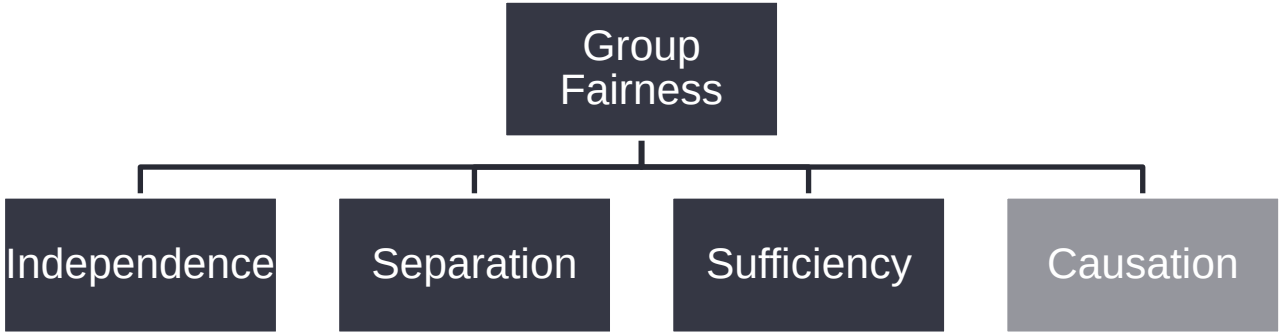


We will later discuss:

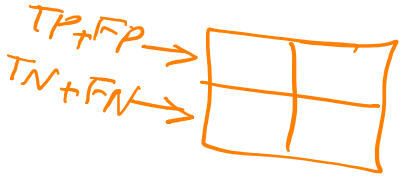
Not all definitions can be satisfied simultaneously, in general.

Even 3 can be impossible (Chouldechova, and several follow on papers).

Group Fairness categories



Independence



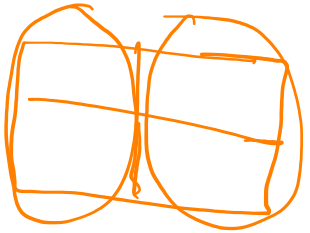
- A model satisfies independence if $f_\theta(X) \perp S$. That is, the outcome of the model is independent from the sensitive attribute(s)
- In a binary classification setting: $\mathrm{P}(f_\theta(X) = 1|S = a) = \mathrm{P}(f_\theta(X) = 1|S = b)$
- Fairness notions such as **demographic parity** or statistical parity follow the independence model
- (TP+FP) / (FP+FN+TP+TN)

$P(f(x)=1|S=a) = P(f(x)=1)$

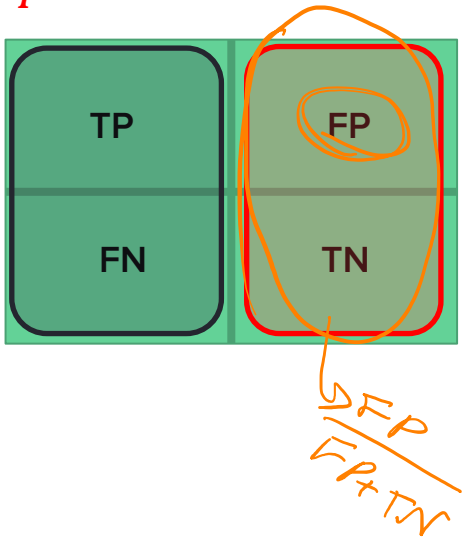
$|P(f(x)=1|S=a) - P(f(x)=1)| \leq \varepsilon$

TP	FP
FN	TN

Separation



- f_θ satisfies separation, if its outcome is independent from the sensitive attribute(s) conditional on the target variable: $f_\theta(X) \perp S \mid Y$
- Here we are looking at the columns of our matrix



- E.g.: Equalized Odds, False Positive Rate, ...

Equalized Odds

Both True Positive Rate and False Positive Rate should be equal for protected subgroup and others.

- for all demographic groups *a*, *b* the two constraints

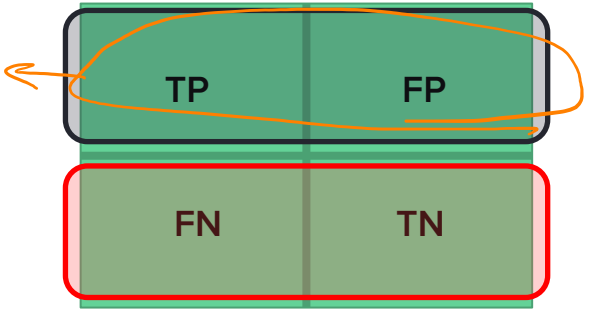
$$\mathrm{P}(f_\theta(X) = 1 | Y = 1, S = a) = \mathrm{P}(f_\theta(X) = 1 | Y = 1, S = b)$$
$$\mathrm{P}(f_\theta(X) = 1 | Y = 0, S = a) = \mathrm{P}(f_\theta(X) = 1 | Y = 0, S = b)$$

Sufficiency

- f_θ satisfies sufficiency, under the same model outcomes, sensitive attribute(s) and the true outcome are independent: $Y \perp S \mid f_\theta(X)$.

- Here we are looking at the rows of our matrix

$$\frac{TP}{TP + FP}$$



- E.g.: Predictive Parity.

Predictive Parity

- Equal positive predictive value for all demographic groups *a, b*

$$\mathrm{P}(Y = 1 | f_\theta(X) = 1, S = a) = \mathrm{P}(Y = 1 | f_\theta(X) = 1, S = b)$$

- Equal negative predictive value for all demographic groups *a, b*

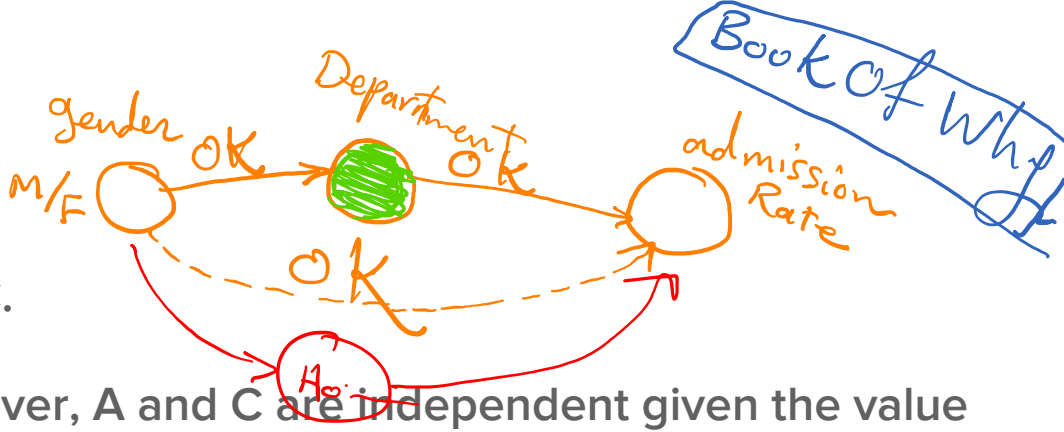
$$\mathrm{P}(Y = 0 | f_\theta(X) = 0, S = a) = \mathrm{P}(Y = 0 | f_\theta(X) = 0, S = b)$$

Other possibilities

$$\mathrm{P}(Y = 1 | f_\theta(X) = 0, S = a) = \mathrm{P}(Y = 1 | f_\theta(X) = 0, S = b)$$
$$\mathrm{P}(Y = 0 | f_\theta(X) = 1, S = a) = \mathrm{P}(Y = 0 | f_\theta(X) = 1, S = b)$$

Need Domain Expert!?

Causality



If A affects B, which in turn affects C.

Then transitively, A affects C. However, A and C are independent given the value of B.

For example, race can affect socioeconomic status, which in turn can affect whether one is hired. If there is no other way that race affects hiring, then we can get hiring separated from race given socioeconomic status.

Dep. is admissible variable

$$DP: P(f(x)=1 \mid Dep, S=a) = P(f(x)=1 \mid Dep)$$

How to bound disparities?

- Equal can never mean exactly equal in practice. A threshold is used to declare unfairness. Usually set at 80%.
- Disparity on Difference
- Disparity on Ratio