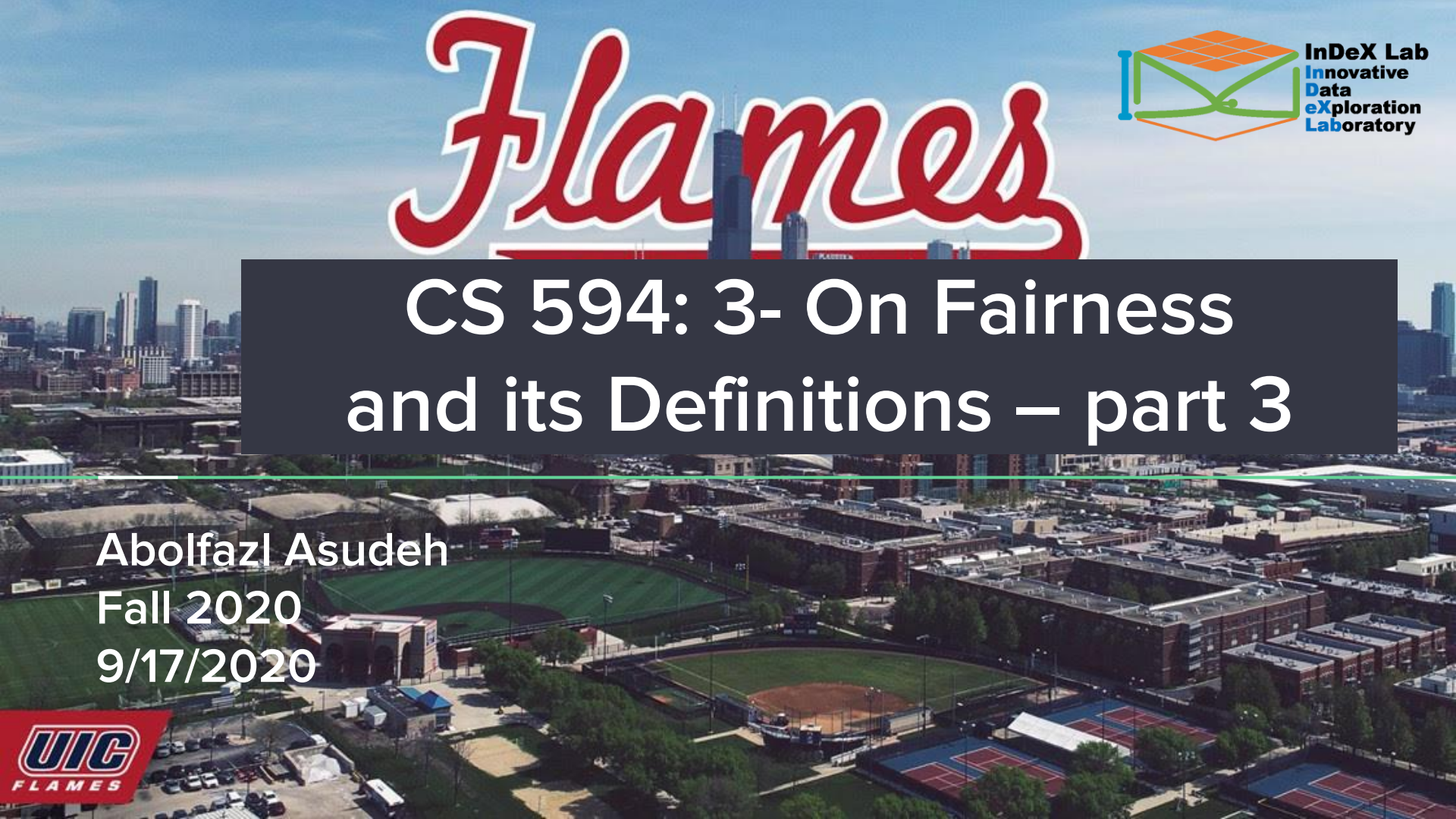


CS 594: 3- On Fairness and its Definitions – part 3

Abolfazl Asudeh

Fall 2020

9/17/2020

InDeX Lab
Innovative
Data
eXploration
Laboratory

UIC
FLAMES

Trade-offs and Impossibility Theorem

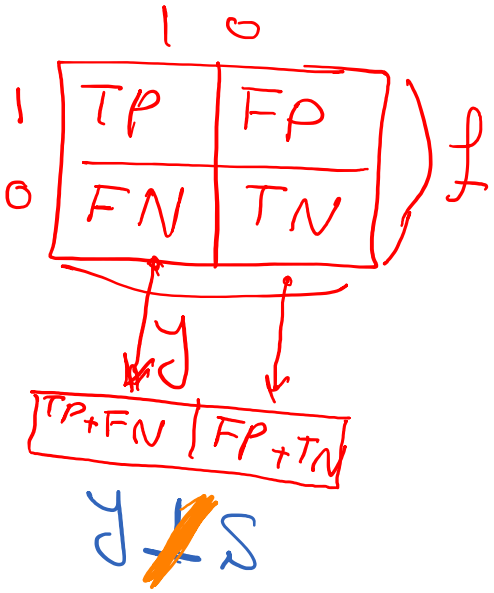
Slides prepared using:

A Tutorial on Fairness in Machine Learning

Ziyuan Zhong

Assumption

- Data is biased:
  - Y is not independent from S



e.g. Demo Parity

e.g. Predictive Parity

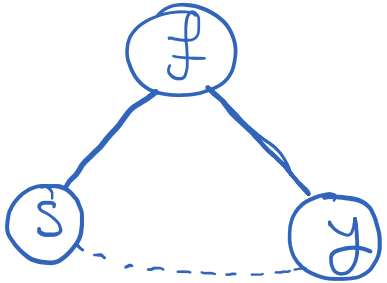
Independence v.s. Sufficiency

- If S and Y are not independent, then either independence holds or sufficiency, but not both

Bias: $Y \not\perp S$ (1)

Indep: $f_\theta \perp S$ (2)

Suff.: $S \perp Y \mid f$ (3)



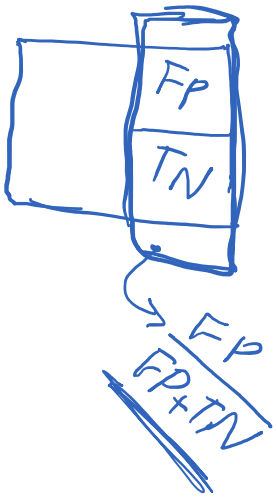
$Y \not\perp S$ (1)

But $S \perp Y \mid f$ (3) (Path S — f — y)

$f \not\perp y$ and $S \not\perp f$ (2)

Independence v.s. Separation

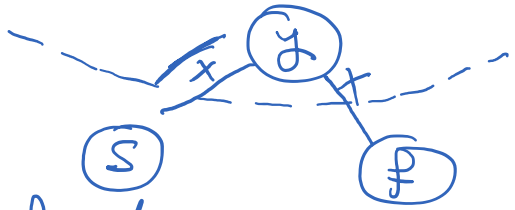
- If S and Y are not independent, then either independence holds or separation, but not both



① Bias: $Y \not\perp S$

③ Indep: $f \perp S$

③ Sep: $f \perp S \mid Y$ > (path S-Y-f does not exist)

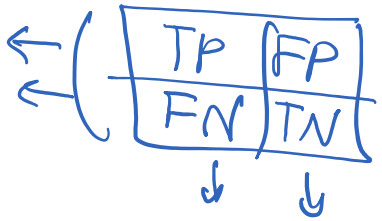


either | ~~$Y \perp S$~~ : ~~X~~ · ① (Bias)

or

$f \perp Y$: (random assignment) ← useless

Sufficiency v.s. Separation



- If S and Y are not independent, then either sufficiency holds or separation, but not both (under some assumptions...)

① Bias: $S \perp Y$

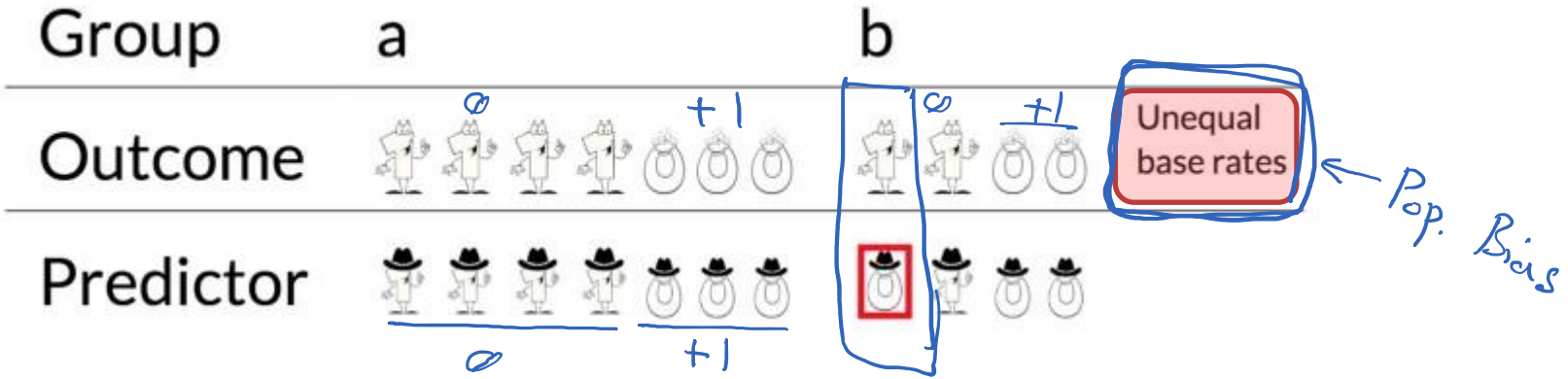
② Suf.: $S \perp f \mid y$ → S is indep. of f|y (marginal on f|y)

③ Sep.: $S \perp f \mid y$ → S is ~ ~ f|y (~ ~ cal)

⇒ $S \perp (f, y)$ ⟹ $S \perp y$ ✗ ①

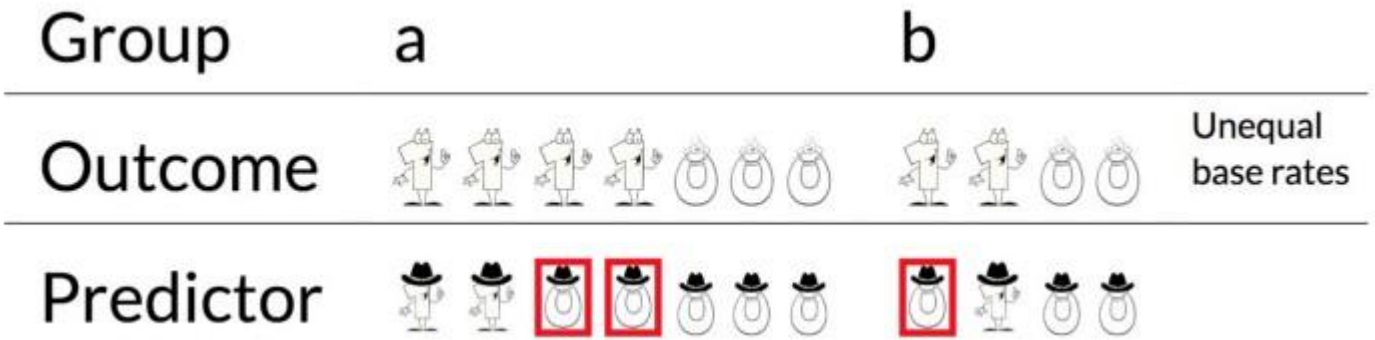
Sufficiency v.s. Separation

- Original classifier

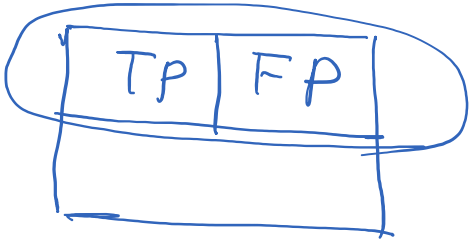


Sufficiency v.s. Separation

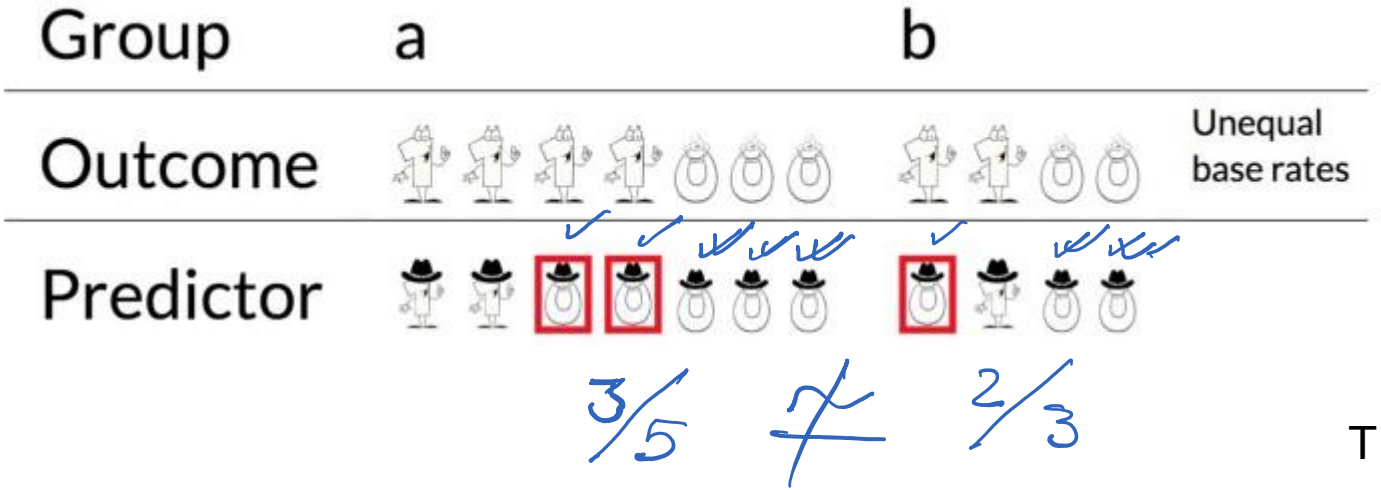
- Fixing False positive rate disparity



Sufficiency v.s. Separation



- Introduces disparity on PP



Accuracy v.s. Fairness

- (Assumption) original classifier:

$$\min L(\theta)$$

- Any intervention on the classifier will increase the loss

References:

1. Alexandra Chouldechova. “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.”, arxiv.org/abs/1610.07524v1 (2016)

2. Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." arXiv:1609.05807 (2016).

3. Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im) possibility of fairness." arXiv:1609.07236 (2016).

* link to related videos for [1,2] are posted in the course page

Individual Fairness and Subgroup Fairness

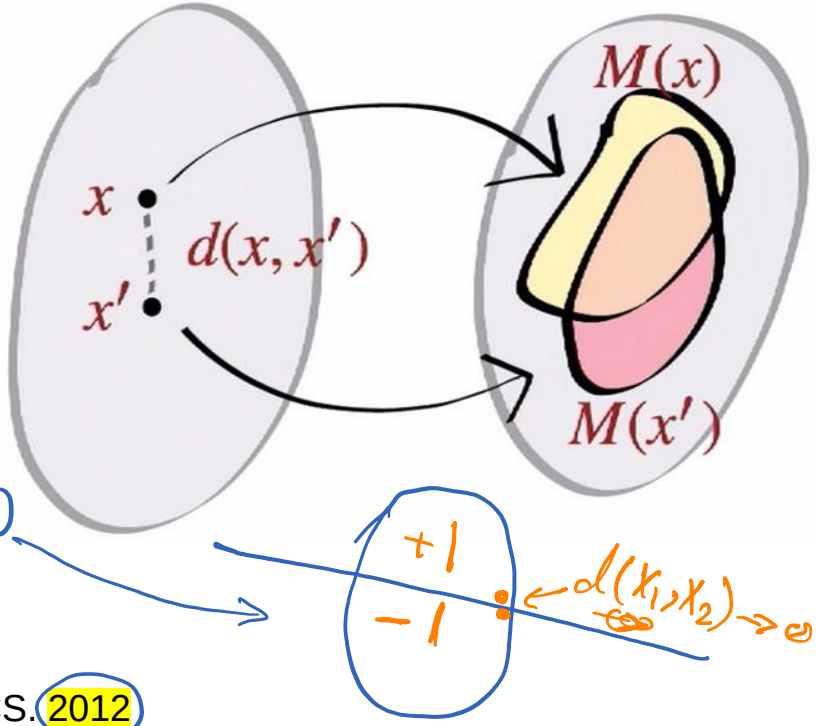
Individual Fairness

- Similar individual should have similar outcomes

$$\Delta\big(f_\theta(X_1), f_\theta(X_2)\big) \leq \varepsilon(d(X_1, X_2))$$

Montonic to d

- Question: Can any deterministic (binary) classifier achieve individual fairness? NO



* Cynthia Dwork, et al. "Fairness through awareness." In ITCS. 2012

Subgroup fairness

- Here, fairness (parity) is defined over the intersection of demographic groups:
  - E.g.: {Hispanic, Female}
  - Hispanic Female
  - Hispanic Female under the age of 20
  - .....
- s/t between indiv. fairness and group fairness (subgroups are small)
- Challenge: Combinatorial Space ← #P problems

Fairness in Rankings

Fairness in Ranking

Ranking is a much more complex output than a binary class label.

Defining fairness is correspondingly more involved.

Toy Example

Sale -- Normalized

Customer Satisfaction -- Normalized

				1 + 1
C		0.63	0.71	1.34
A2		0.72	0.65	1.37
C		0.58	0.78	1.36
A2		0.7	0.68	1.38
C		0.53	0.82	1.35
A2		0.61	0.79	1.4

Fairness at Top-k

The simplest measures consider the top-k, and then address it as if it were a classification task -- items ranked in the top-k have one label and the rest have another. Now we can use the entire fairness framework for classification.

This, of course, begs the question of choosing k.

Answers obtained could be very different depending on the value chosen.

Exposure-based assessment

In many ranking situations, such as in information retrieval or recommender systems, higher ranked items get more attention than lower ranked ones.

There often is not a hard cut-off. But it is possible to define a monotonically decreasing function, such as inverse rank, that quantifies how much attention an item gets.

Now we can aggregate the attention received per protected group, and make that into a criterion against which we assess fairness.

Probability-Based Assessment

If ordered lists were created separately for each protected group, and the lists were then merged at random, how likely is it that we will observe the ordering reported?

In a random merge of these sorted lists, it should be the case that for any pair of groups, if we consider all pairs of items in the list from these groups, there should be approximately as many pairs items with the group A item ranked higher as the number with group B items ranked higher. The difference in this number is a measure of bias, related to the probability measure discussed above.