What is Fairness?

A model is fair if its Performance is equally good for all!

all : group memberships

~ ~ different groups.

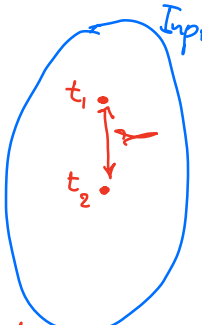
$Perf^A$: Performance of Alg. A

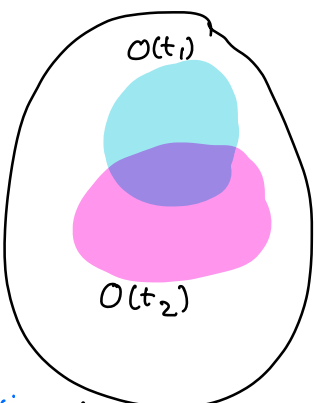
$Perf^A \mid g_i \simeq Perf^A$

all
- group level
- Subgroup level
- Indiv. level

Individual Fairness

Similar individuals should be treated Similarly





Normalization factor

output distance Metric

$$d(t_1, t_2) \geq \varepsilon \Delta(O(t_1), O(t_2))$$

[Fairness through Awareness, 2012]

* No deterministic Alg. can achieve Indv. Fairness.

Group Fairness

- We assume the existence of at least one Sensitive Attribute S (gender, race, Education, Income, ...)
  the value of S specify the groups

GPA	SAT	gender (S)	race (S)
—	—	M	W
—	—	F	B
—	—	F	B
—	—	M	W
—	—	...	...

{M, F, W, B, ...} ⇐ Non-Intersectional groups

{WM, WF, BM, BF, ...} ⇐ Intersectional

{WM, WM-with College Degree, ...}

Fully Specified Subgroup

Indivual-level — fully-specified Sup groups ..... Partially-specified Subgroups — groups {M, F, ...}

$|M| \geq |WM| \geq |WM\text{-}CD|$

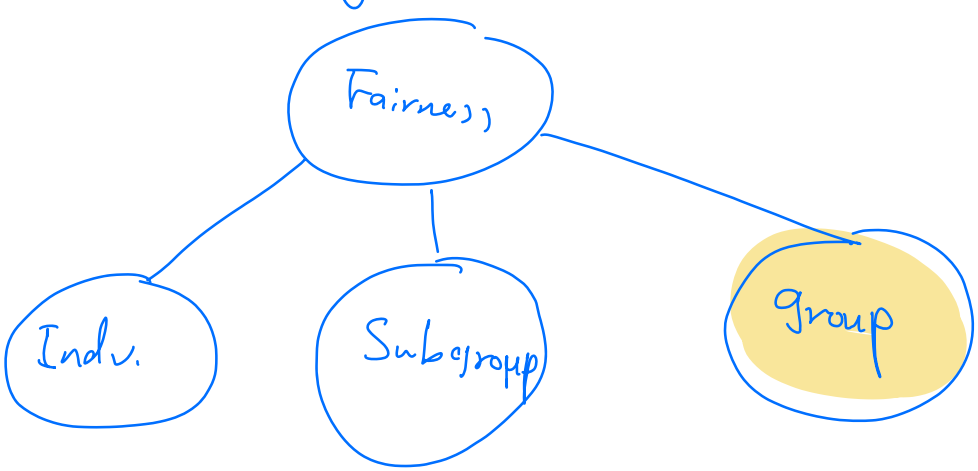
↳ and College degree

Subgroup Fairness Challenge:

↳ we may not have enough samples from all subgroups to draw Statistical Conclusions.

demographic groups
- Binary groups:
  - M vs F
  - B vs W
  - Majority vs Minority
  - ...
- Non_binary
  - over lapping { F, H, B, M, ... }
    - F → Female
    - H → Hispanic
  - Non_over lapping { B, W, H, ... }

We Assume Non_overlapping Binary groups for defining Fairness Metrics



Group Fairness (Binary Groups / Binary classification)

$$P(\alpha \mid \beta, g_1) = P(\alpha \mid \beta, g_2)$$

↓ ? Performance

$$P(\alpha \mid \beta, g_1) = P(\alpha \mid \beta)$$

Decision Maker: Out of the ones Labed as Positive (+1), how many are really Positive

→ Performance Metric

Prediction $f_\theta(x)$ \ True Label y	+	−
+	TP	FP
−	FN	TN

$$\left[\frac{TP}{TP+FP}\right]_{g_2} = \left[\frac{TP}{TP+FP}\right]_{g_1}$$

$$P(\underbrace{y=1}_{\alpha} \mid \underbrace{f=1}_{\beta}, g_1) = P(y=1 \mid f=1, g_2)$$

Positive Predictive Parity (Aka Calibration)

- $P(y=0 \mid f=0, g_1) = P(y=0 \mid f=0, g_2)$
  ↳ Negative Predictive Parity

- $P(y=0 \mid f=1, g_1)$ ....
- $P(y=1 \mid f=0, g_1)$ ....

PP Rate Parity AND NP Rate → Equal Opportunity

Decision Maker: How Likely Can an Indv. be falsely Labeled as positive.

↳ eg., If Someone is not dangerous but mistakenly labeled as one.

Prediction $f_\theta(x)$ \ True Label y	+	−
+	TP	FP
−	FN	TN

$$\left[\frac{FP}{FP+TN}\right]_{g_1} = \left[\frac{FP}{FP+TN}\right]_{g_2}$$

$$-P(f=1 \mid y=0, g_1) = P(f=1 \mid y=0, g_2)$$

False Positive Rate Parity)

- False Neg. rate Parity
- True " " "
- " Positive " "

Both FPRate Parity AND FNRate Parity → Equalized ODDs

Defendant: The likelyhood of Positive Prediction should be equal for all groups

→ R-Scores are not fair because black is more likely of being Predicted as Positive (Dangerous)

f	+	−
+	TP	FP
−	FN	TN

$$\left[\frac{TP+FP}{\underbrace{(TP+FP)+(TN+FN)}_{all}}\right]_{g_1} = \left[\frac{TP+FP}{all}\right]_{g_2}$$

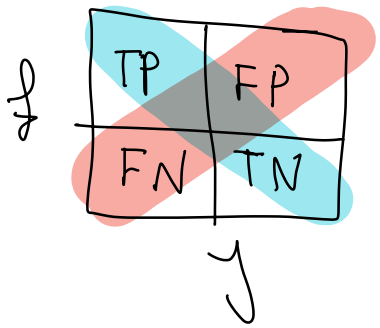
$$P(\underbrace{f=1}_{\alpha} \mid g_1) = P(\underbrace{f=1}_{\alpha} \mid g_2)$$

β: Null

→
- Demographic Parity
- Statistical "
- Disparate Impact

$$P(f=0 \mid g_1) = P(f=0 \mid g_2)$$

Decision Maker: { Accuracy, misclassification rate } is equal for the two groups



Accuracy Parity

$$\left[\frac{TP+TN}{all}\right]_{g_1} = \left[\frac{TP+TN}{all}\right]_{g_2}$$

$$P(y=f \mid g_1) = P(y=f \mid g_2)$$

Misclassification Rate (Error) Parity

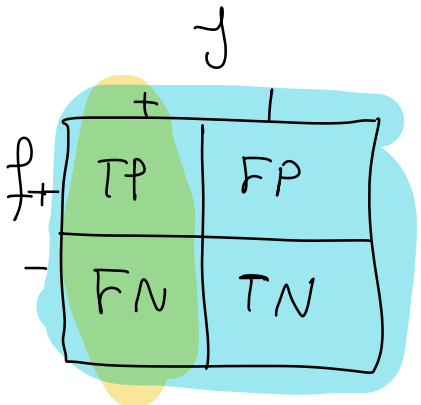
$$\left[\frac{FP+FN}{all}\right]_{g_1} = \left[\frac{FP+FN}{all}\right]_{g_2}$$

$$P(y \neq f \mid g_1) = P(y \neq f \mid g_2)$$

$4 + 4 + 2 + 2 + 2 + 4 + 4 = 22 > 21$

- 4 → rows
- 4 → Col
- 2 → Dem. Parity
- 2 → EO, EO
- 2 → Acc, Miss
- 4 → 1 cell / all
- 4 → Diameters

→ 21 definitions of fairness



Assumption $P(y=1 \mid g_1) \neq P(y=1 \mid g_2)$

↳ Bias in Data.

$y \not\perp s$

↳ The random variable y is not independent from the rand. variable s.

Simpson Paradox.

admitted : $y=1$
not admitted : $y=0$

$S=1$ M
$S=0$ F

UC Berkeley Admission Process is ~~sexist~~ because

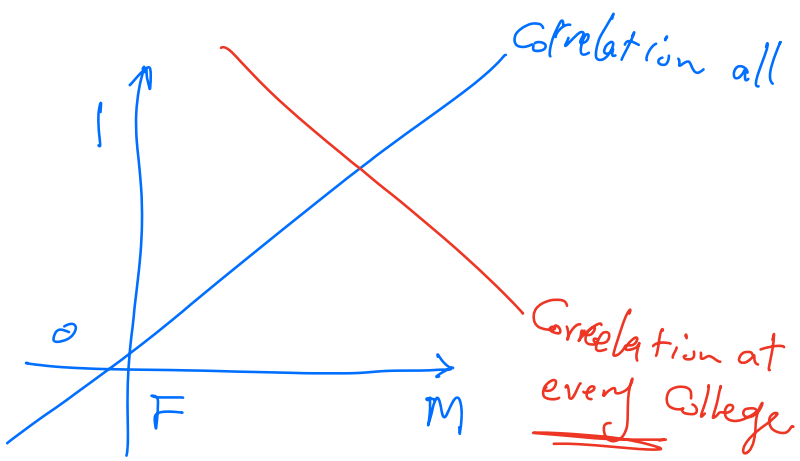
Overall : $P(f=1 \mid M) > P(f=1 \mid F)$

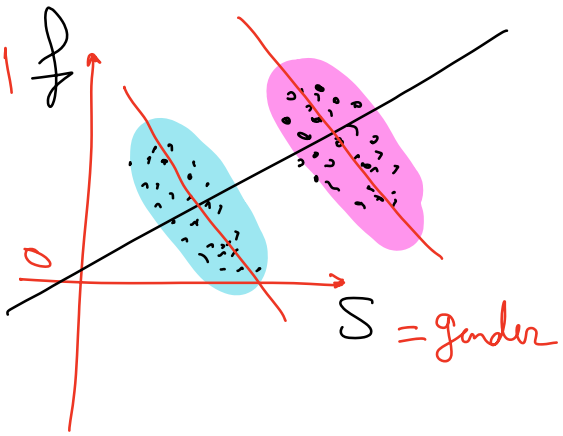
→ Demographic Parity

$f \not\perp S \Rightarrow Corr(f, S) \neq 0$

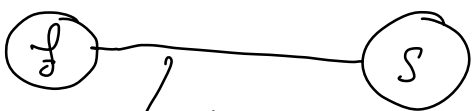
College-level : ~~almost always~~ Assume $\forall$ college

$P(f=1 \mid M) < P(f=1 \mid F)$

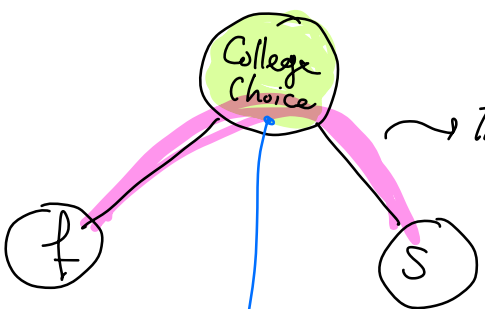




The admission process is fair because there is an admissible explanation for the overall disparity



→ An edge is added if the two variables are not independent

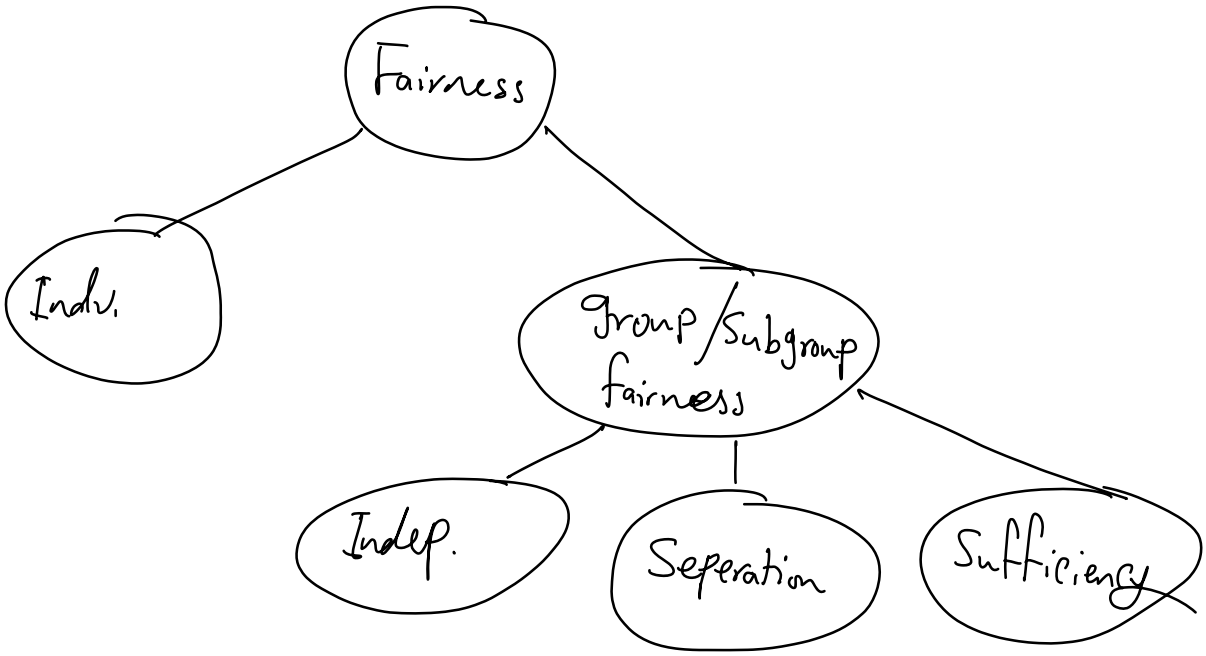


→ the disparity is because of the Path f-C-S

→ Admissible Variable

A disparity is unfairness iff it is not through admissible Variables.

A Categorization of Fairness Def.



Independence:

M satisfies independence

$S \perp f$

↳ e.g., demographic Parity

Indep. ⇒ demo. Parity

Sufficiency:

S should be indep from f Conditioned on y

$S \perp f \mid y$

↳ TPRP
FPRP
TNRP
FNRP
Equalized odds (EO)

Sufficiency ⇒ TPRP
...
EO

Seperation:

S should be indep. from Y Cond. on f

$S \perp Y \mid f$

↳ Pos. Predictive Parity (Calibration)
Neg. " "
Equal opportunity

Seperation ⇒ PPPR
...

If data is biased: $Y \not\perp S$, ← Assumption

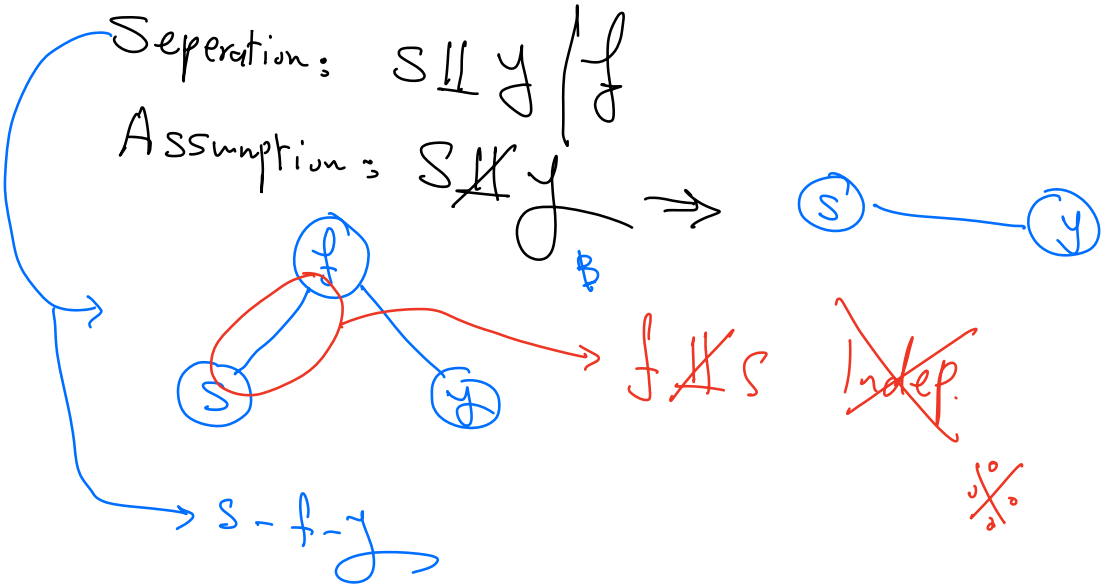
for any choice of two fairness Categories, It is IMOSSIBLE to satisfy BOTH

Thm: Indep. and Seperation Cannot be Satisfied at the Same time.

Indep. $S \perp f$

Seperation: $S \perp Y \mid f$

Assumption: $S \not\perp Y$

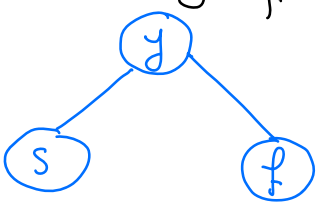


Independence & Sufficiency are Impossible mutually

Sufficiency $S \perp\!\!\!\perp f \mid y$

Indendence $S \perp\!\!\!\perp f$

if the Path $S-y-f$ exists $\Rightarrow S \not\perp\!\!\!\perp f$



$\Rightarrow$ The Path $S-y-f$ does not exist.

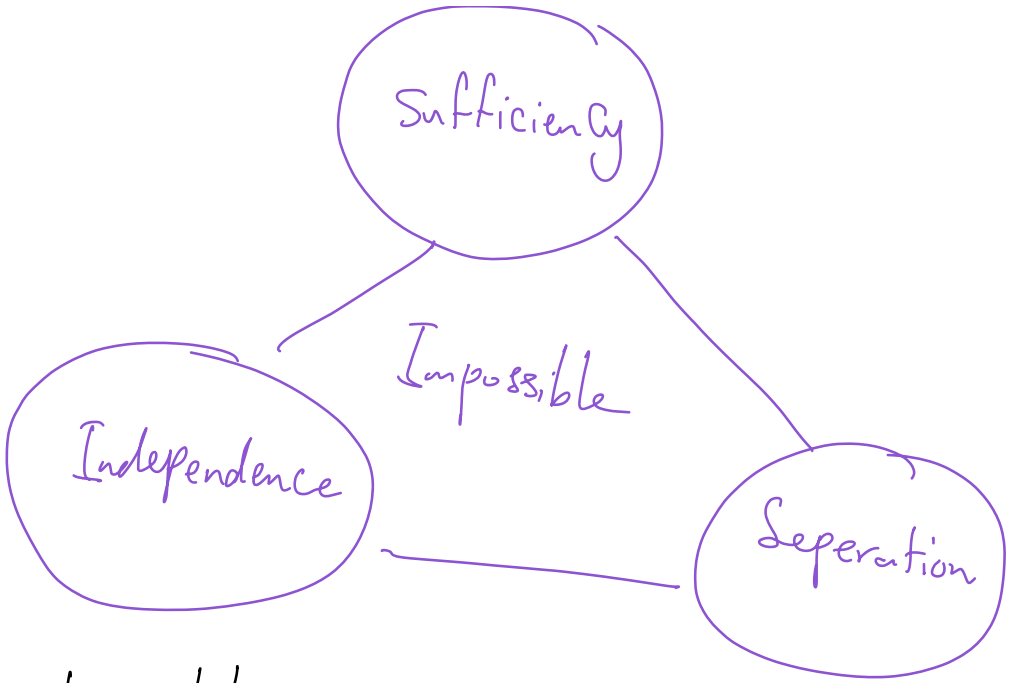
- $\Rightarrow y \perp\!\!\!\perp f$ $\Leftarrow$ meaningless (No Prediction Power)
- $\Rightarrow S \perp\!\!\!\perp y$ $\not\!\!X$. $\Leftarrow$ Bias Assumption

- Seperation & Sufficiency are mutually Impossible

- $S \perp\!\!\!\perp y \mid f$ Seperation
- $S \perp\!\!\!\perp f \mid y$ Sufficiency

$S \perp\!\!\!\perp (y, f)$

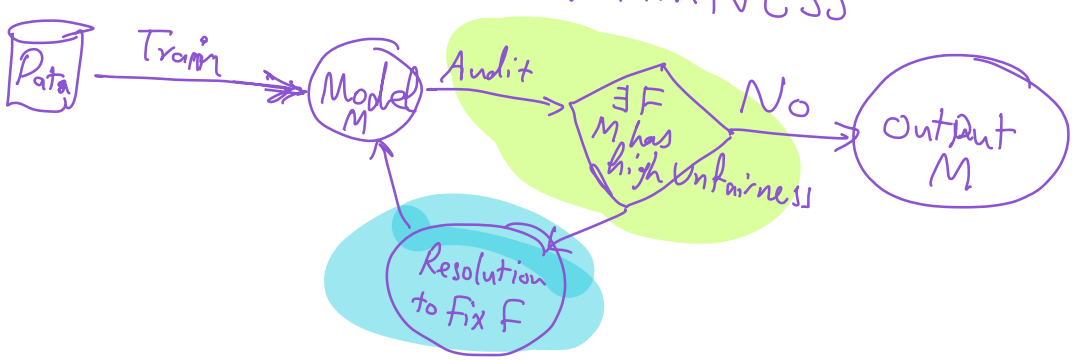
$\Rightarrow S \perp\!\!\!\perp y$ $\not\!\!X$. Bias Assumption



⇒ Impossibility theorems only prove that exact equality is impossible for mutually exclusive definitions

↳ we can still satisfy ALMOST EQUAL for all definitions.

↳ the goal should be to MIN UNFAIRNESS



Measuring Unfairness

- If Alg. A has perf. P_1 for g_1 & P_2 for g_2, how should we measure the unfairness?

~~X~~ $- |P_2 - P_1|$

e.g., - Unfairness, $F_1 = 0.3$

Case2 - $F_2 = 23$

Case3 - $F_3 = 0.0003$

Different Scenarios

⇒ Case 2: $P_1 = 236723$, $P_2 = 236700$ Seems fair

⇒ Case 3: $P_1 = 0.0004$, $P_2 = 0.0001$ Seems Unfair

— Subtraction (Normalized)

$$F = \left| \frac{P_1 - P_2}{P_1 + P_2} \right| \rightarrow F = 0.7$$

- $F = \dfrac{\min(P_1, P_2)}{\max(P_1, P_2)}$

$$\frac{1}{1+\varepsilon} < \frac{P_1}{P_2} < 1+\varepsilon$$