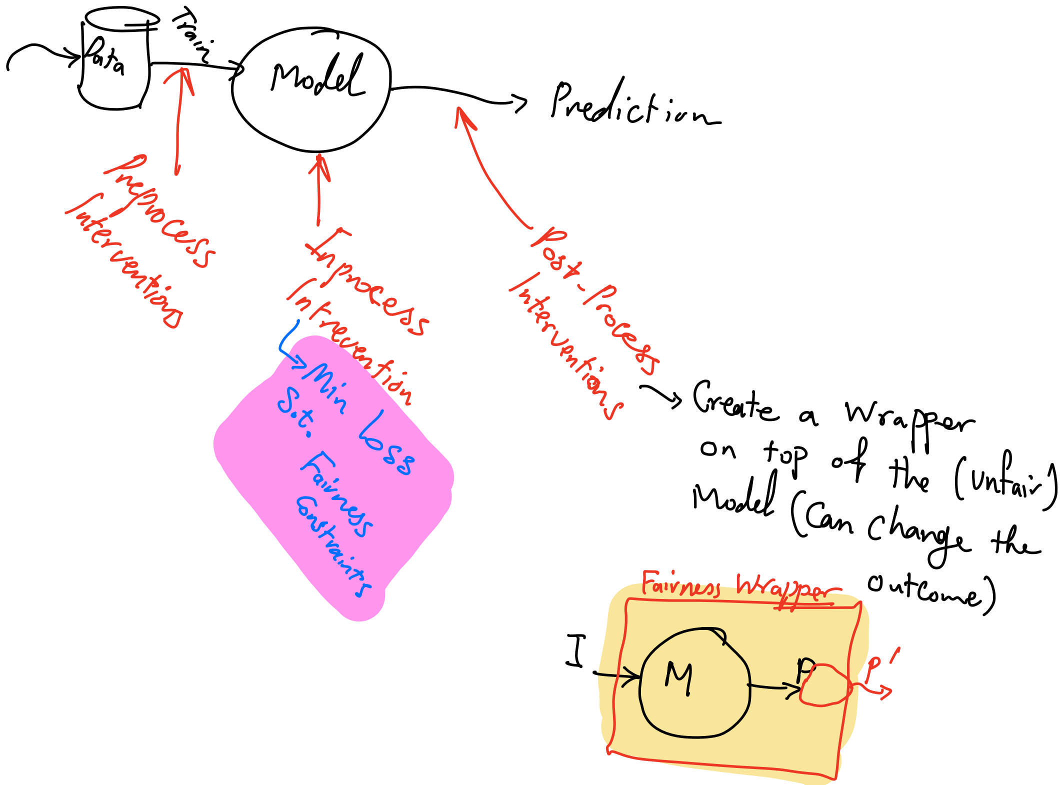
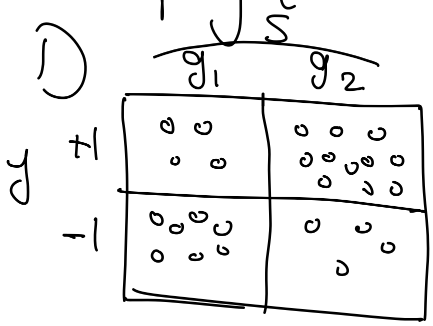
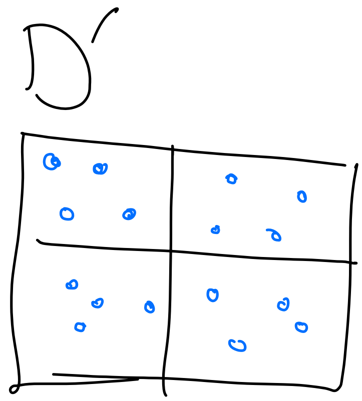


We are looking for a dataset that

1. has enough Samples from all groups.
2. Corr(X, y) is Maximized
3. Corr(X, S) is Minimized

- [Faisal Kamiran, 2015]
  - Sampling (with replacement)





- Data Augmentation:
  - Use additional real data?
  - Generate Synthetic Data (GenAI)

- Reweighting: Assigns a weight to every sample

	X_1	...	X_d	S	y	
t_1				M	+1	w_1
t_2				M	+1	w_2
t_3				F	+1	w_3
				⋮	⋮	⋮
t_n				M	-1	w_n

$$S \perp\!\!\!\perp y$$

$$P(S \wedge y) = P(S)\,P(y)$$

$$\forall\, t_k \in (s_i \wedge y_j)$$

$$w_k = \frac{P(s_i)\,P(y_j)}{P(s_i \wedge y_j)}$$

$$P(y) = \begin{cases} 0.6 & y=+1 \\ 0.4 & y=-1 \end{cases} \qquad P(S) = \begin{cases} 0.5 & M \\ 0.5 & F \end{cases}$$

$$P(S \wedge y) = \begin{cases} 0.4 & +, M \\ 0.2 & +, F \\ 0.1 & -, M \\ 0.3 & -, F \end{cases} \qquad W(S,y) = \begin{cases} \dfrac{0.6 \times 0.5}{0.4} & M, + \\ \dots \end{cases}$$

Preprocess Int.

- Sampling
- Reweighting
- Transformation
- Data Repair

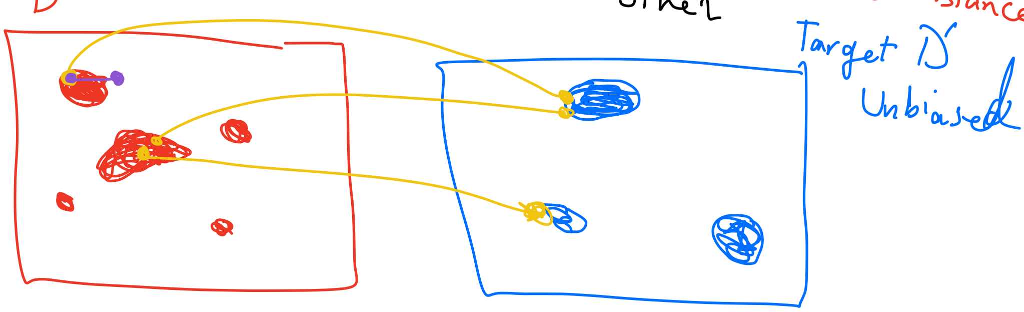
- Transformation

(A) Moving Earth Distance

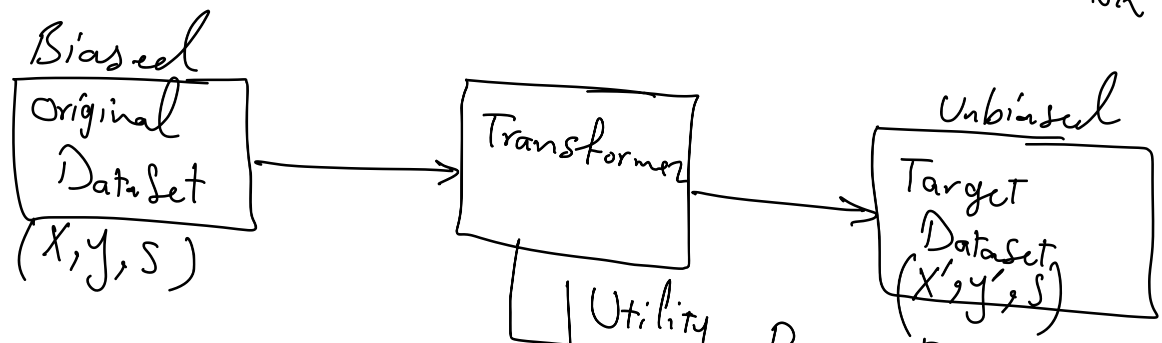
↳ If there are two data distributions, how can we minimally change one to become the other

ME-Distance

Biased D



(B) Opt. Preprocess for Discrimination Prevention



Utility $P_{X,y} \sim P_{X',y'}$

Individual Distortion $(x_i, y_i) \approx (x_i', y_i')$

$S \perp\!\!\!\perp y$

- Certifying & Removing Disparate Impacts

To prevent Disparate Impact you should make sure that X cannot predict S.

Disparate Impact Rule:

$$0.8 = t \leq \frac{P(f(x) = 1 \mid S=0)}{P(f(x)=1 \mid S=1)} \leq \frac{1}{t}$$

If X cannot predict S, NO classifier trained on X, f(x) can have disparate impact.

Balanced Error Rate (BER): for any predictor $g(\cdot)$ that predicts the (Binary) Sensitive attribute S,

$$BER(g(X)) = \frac{P(g=1 \mid S=0) + P(g=0 \mid S=1)}{2}$$

- S is ε-predictable by X, if $\exists\, g: X \to \{0,1\}$ where $BER(g) \leq \varepsilon$

- Theorem, for any target value t for disparate impact, there exists a value ε, such that if S is not ε-predicatable by X, any classifier trained on X satisfies the disparate impact requirement

Proof sketch:

$$BER(g) = \frac{P(g=1 \mid S=0) + P(g=0 \mid S=1)}{2}$$

$$= \frac{P(g=1 \mid S=0) + (1 - P(g=1 \mid S=1))}{2}$$

disparate impact requirement:

$$\frac{P(g=1 \mid S=0)}{P(g=1 \mid S=1)} \geq t \Rightarrow P(g=1 \mid S=1) \leq \frac{P(g=1 \mid S=0)}{t}$$

$$\Rightarrow BER(g) \leq \frac{P(g=1 \mid S=0) + 1 - \frac{P(g=1 \mid S=0)}{t}}{2}$$

$$= \frac{1}{2} + P(g=1 \mid S=0)\left(1 - \frac{1}{t}\right) = \varepsilon$$