Repairing without Retraining:
Avoiding Disparate Impact with Counterfactual Distributions

Hao Wang, Berk Ustun, Flavio P. Calmon

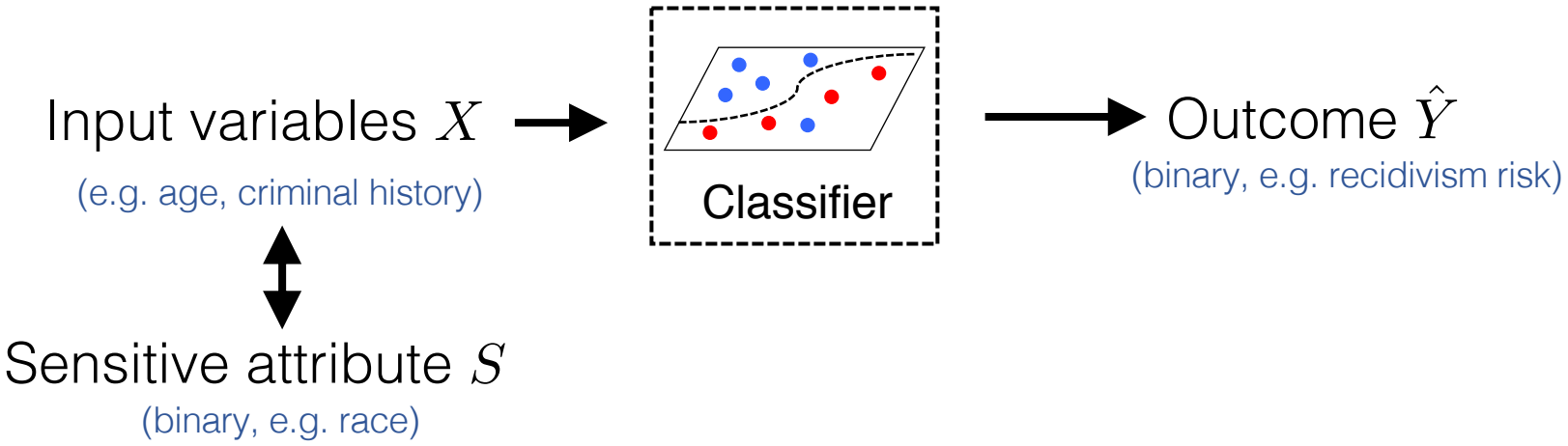
hao_wang@g.harvard.edu, {berk,Flavio}@seas.harvard.edu

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Outline

- Use cases
  - A bank enters a new market and discovers its credit score underperforms on customers over 60 years of age
  - A rural clinic purchases a classification model to detect lung cancer and discovers that patients in a certain subgroup have high FPR
- Framework and methodology
  - “Counterfactual distribution”
  - Local perturbation and influence function
- Model repair

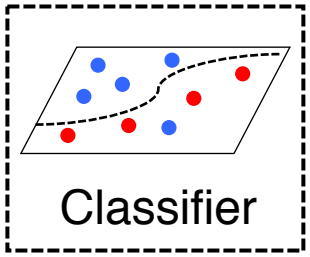
Disparate Impact



Disparate Impact

Input variables X

(e.g. age, criminal history)

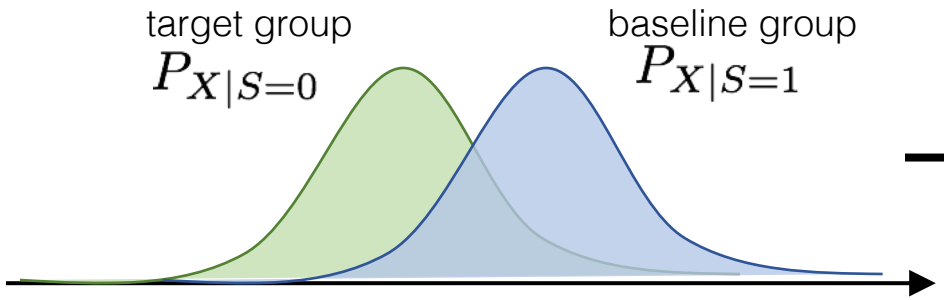


Outcome $\hat{Y}$

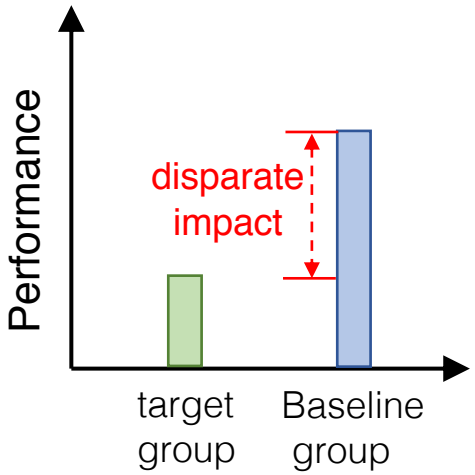
(binary, e.g. recidivism risk)

Sensitive attribute S

(binary, e.g. race)



Changes in input distribution…



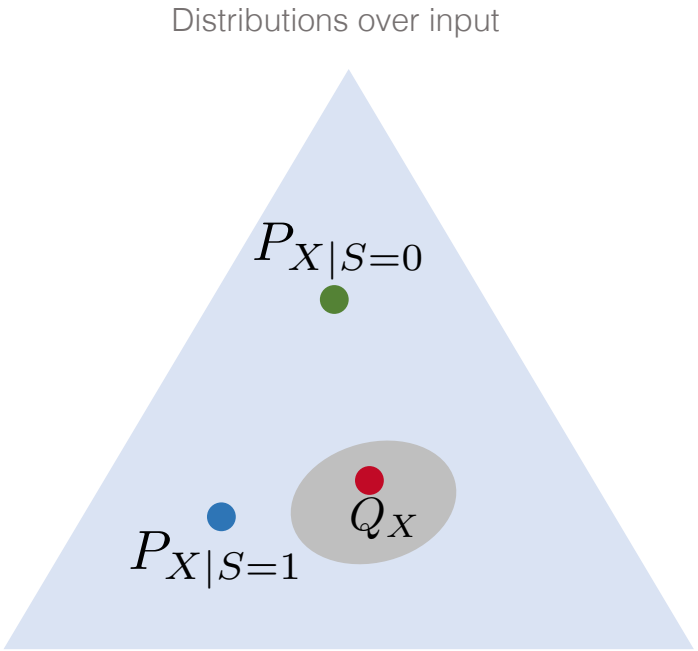
Can lead to different performance.

Counterfactual Distribution

Definition. For a given disparity metric $\mathsf{M}(\cdot)$, a counterfactual distribution is a distribution of input variables over the target group such that:

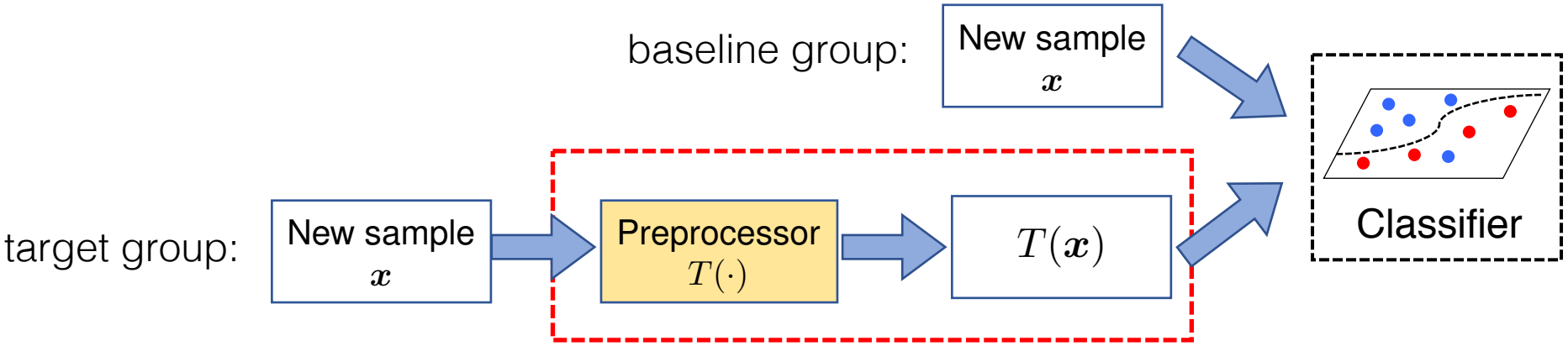
$$Q_X \in \underset{Q'_X \in \mathcal{P}}{\operatorname{argmin}} \left|\mathsf{M}(Q'_X)\right|,$$

where $\mathcal{P}$ is the set of probability distributions over $\mathcal{X}$.

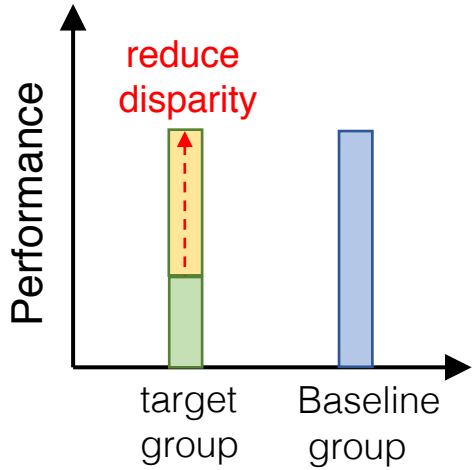


	Observed		Counterfactual		
	Female	Male	SP Female	FNR Female	FPR Male
Married	18%	63%	39%	23%	54%
Immigrant	10%	11%	11%	11%	12%
HighestDegree_is_HS	32%	32%	24%	28%	37%
HighestDegree_is_AS	7%	8%	9%	9%	6%
HighestDegree_is_BS	15%	18%	21%	17%	13%
HighestDegree_is_MSorPhD	6%	7%	13%	8%	5%
AnyCapitalLoss	3%	5%	8%	5%	4%
Age $\leq$ 30	39%	29%	29%	38%	35%
WorkHrsPerWeek<40	38%	17%	33%	37%	19%
JobType_is_WhiteCollar	34%	19%	36%	35%	15%
JobType_is_BlueCollar	5%	34%	4%	5%	39%
JobType_is_Specialized	23%	21%	29%	23%	20%
JobType_is_ArmedOrProtective	1%	2%	1%	1%	3%
Industry_is_Private	73%	69%	64%	69%	70%
Industry_is_Government	15%	12%	22%	17%	12%
Industry_is_SelfEmployed	5%	15%	8%	6%	13%

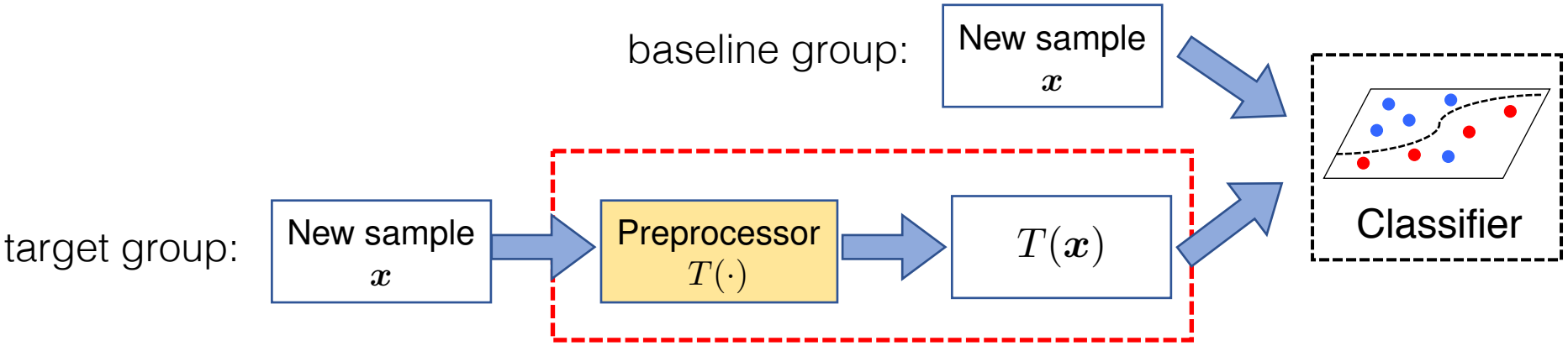
Goal: Model Repair



Goal: repair a classifier that has disparate impact by preprocessing the data

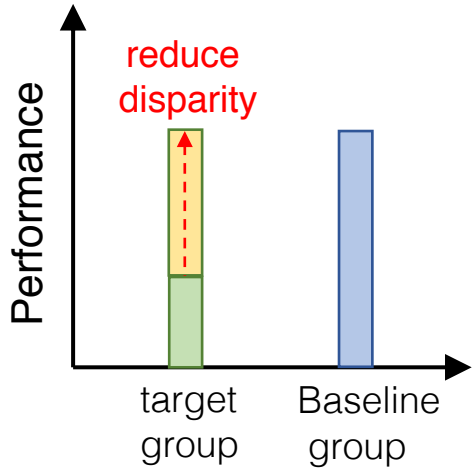


Goal: Model Repair



We build the pre-processor in two steps:

1) Compute a counterfactual distribution that minimizes disparate impact.
2) Solve an optimal transport problem between the distribution of the target population and the counterfactual distribution.



Numerical Experiments: COMPAS and UCI Adult

			Original Model			Repaired Model		Target Group AUC	
Dataset	Metric	Target Group	Baseline Group	Target Group	**Disc. Gap**	Target Group	**Disc. Gap**	Before Repair	After Repair
`adult`	SP	Female	0.696	0.874	**0.178**	0.688	**-0.007**	0.895	0.758
`adult`	FNR	Female	0.478	0.639	**0.161**	0.483	**0.004**	0.895	0.880
`adult`	FPR	Male	0.021	0.119	**0.098**	0.023	**0.002**	0.829	0.714
`compas`	SP	White	0.514	0.594	**0.079**	0.533	**0.018**	0.704	0.667
`compas`	FNR	White	0.350	0.487	**0.137**	0.439	**0.088**	0.704	0.699
`compas`	FPR	Non-white	0.190	0.278	**0.087**	0.160	**-0.029**	0.732	0.680

[Bache and Lichman, 2013], [Angwin et al., 2016]

Repairing without Retraining:
Avoiding Disparate Impact with Counterfactual Distributions

Poster Session:

Thursday 06:30 -- 09:00 PM

Pacific Ballroom

http://github.com/ustunb/ctfdist