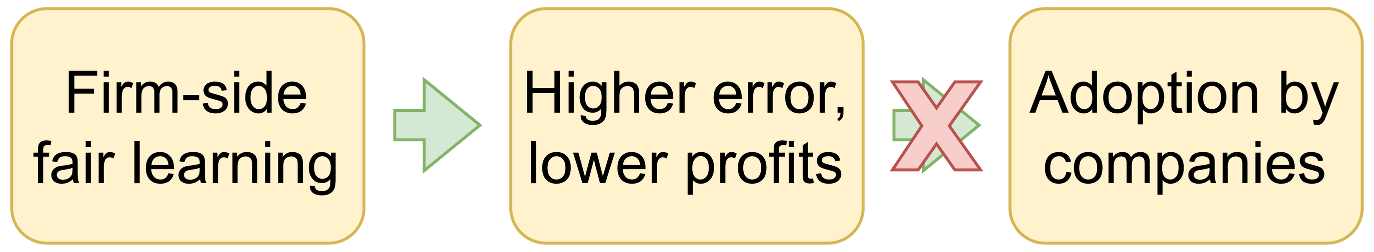
Fairness for the People, by the People: Minority Collective Action

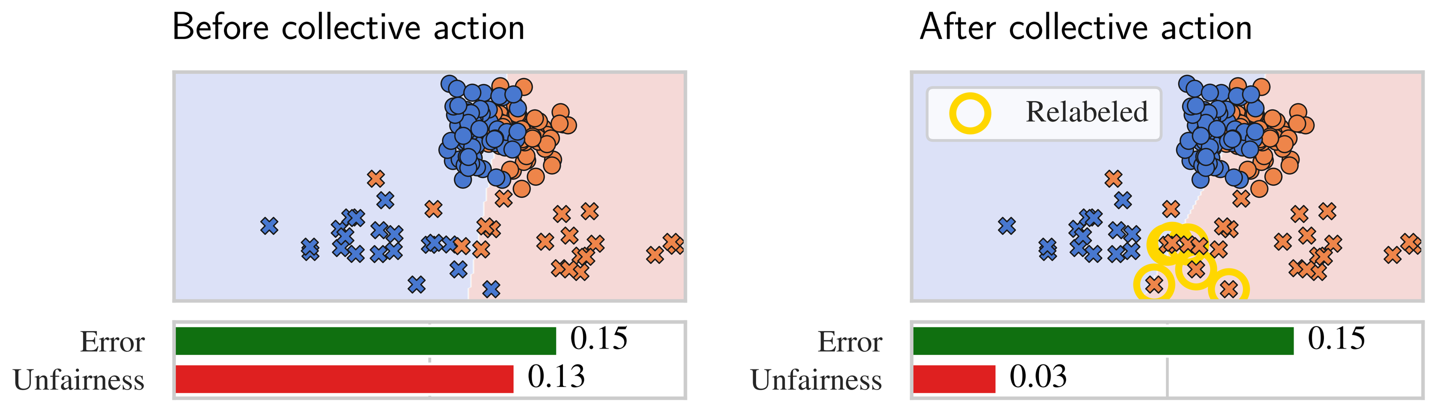
Omri Ben-Dov[1], Samira Samadi[1], Amartya Sanyal[2], Alexandru Ţifrea[3]

[1]Max Planck Institute for Intelligent Systems, Tübingen AI Center, Tübingen, Germany [2]Department of Computer Science, University of Copenhagen [3]ETH Zurich

Motivation



When a company learns from user-data, can a minority induce fairness?



Algorithmic Collective Action

A firm trains a classifier h on user-data and a α-sized group of users collaborate to modify their data.

To make a classifier ignore a signal g

$$S(\alpha) = \mathbb{P}_0 \left[h(g(x)) = h(x) \right],$$

the collective can apply a relabeling strategy [1]

$$y \to \operatorname*{argmax}_{y' \in \{0,1\}} \mathbb{P}_0 \left(y' | g(x) \right).$$

Setting the signal as a group counterfactual

$$g(x) = x_{A \leftarrow 0}$$

leads towards counterfactual fairness, in some cases promoting other forms of group fairness [2].

References

[1] Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. Algorithmic Collective Action in Machine Learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 12570–12586, 2023.

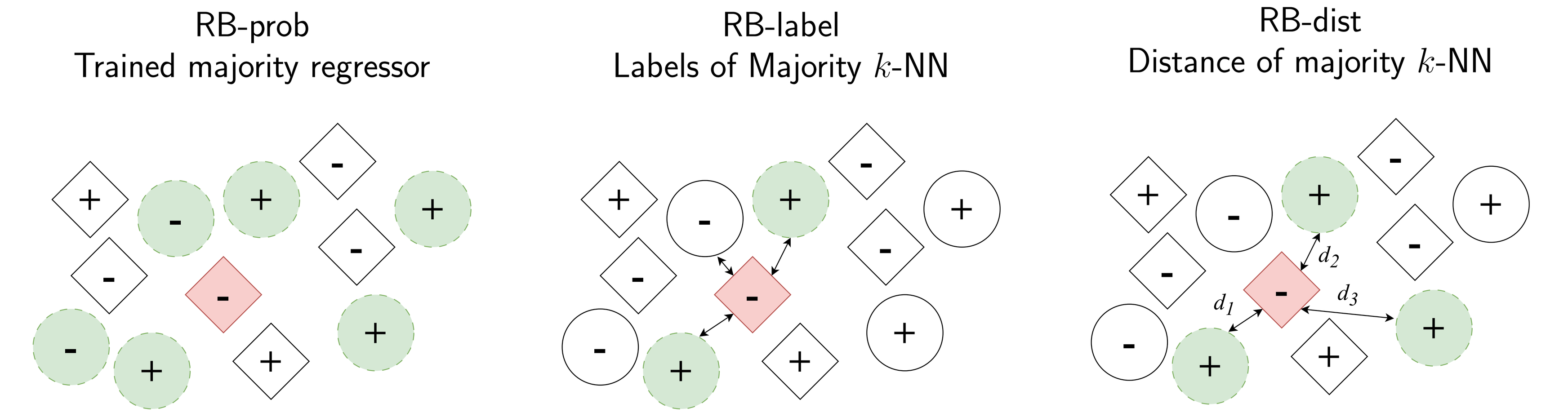
[2] Jacy Anthis and Victor Veitch. Causal context connects counterfactual fairness to robust prediction and group fairness. In *Advances in Neural Information Processing Systems*, volume 36, pages 34122–34138. Curran Associates, Inc., 2023.

[3] Nikola Jovanović, Mislav Balunovic, Dimitar Iliev Dimitrov, and Martin Vechev. FARE: Provably Fair Representation Learning with Practical Certificates. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15401–15420. PMLR, 2023.

[4] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Approximating the Counterfactuals

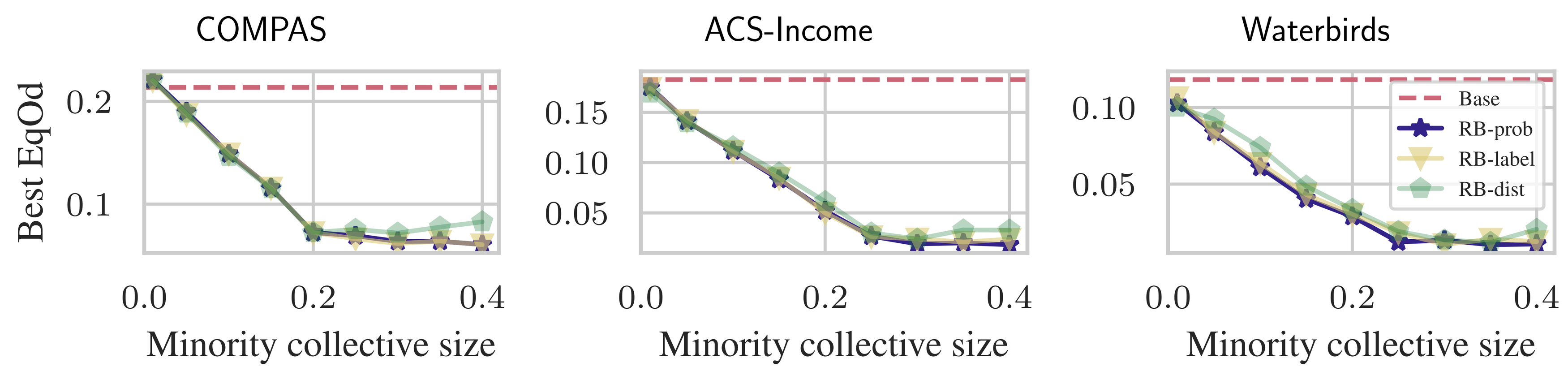
Relabel the M negative candidates with highest score s.



$s_i = f(x_i)$ $\quad$ $s_i = \sum_{j \in K_i} \mathbf{1}\{y_j = 1\}$ $\quad$ $s_i = -\frac{1}{k} \sum_{j \in K_i} \|x_i - x_j\|_2$

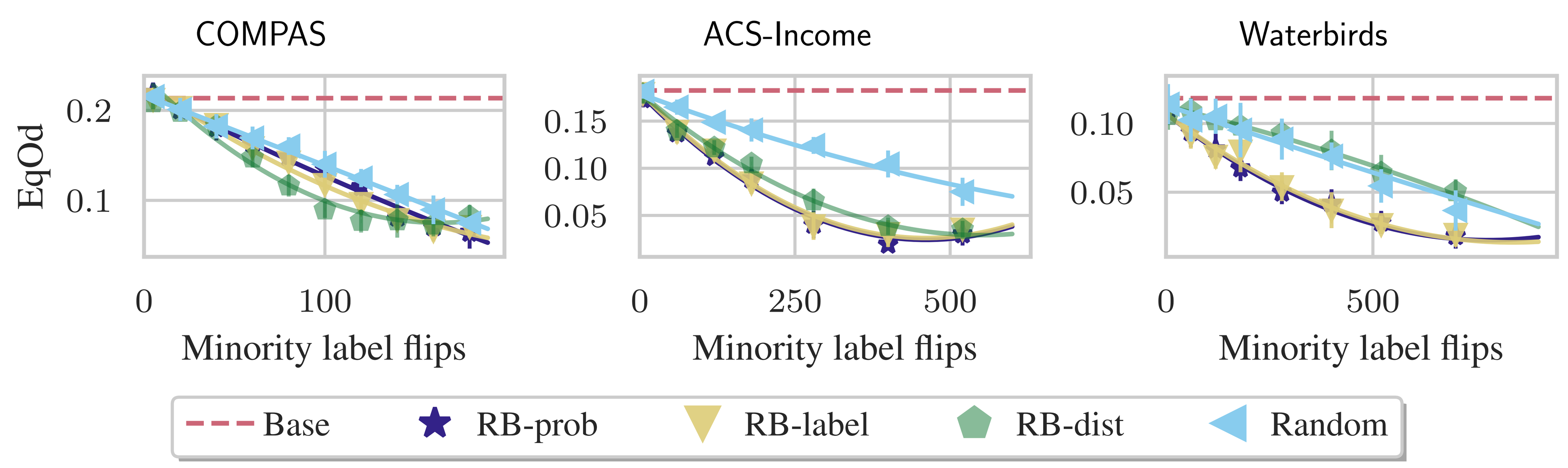
Importance of Collective Size

20–30% of the minority attains the least fairness violation.



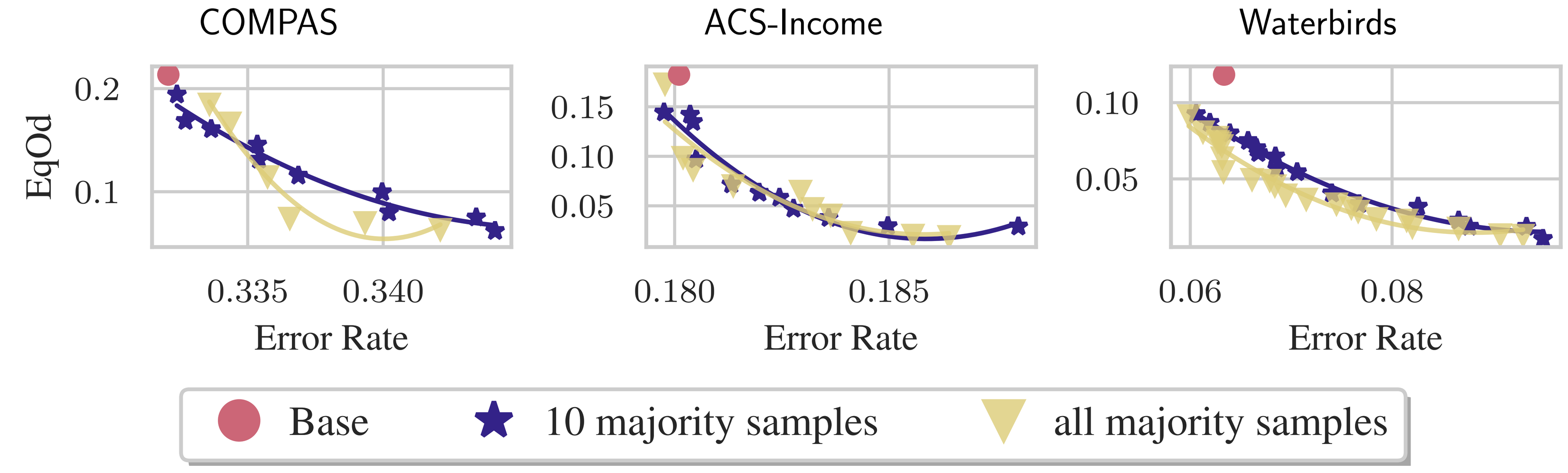
Relabeling Efficiency

Our methods are more efficient than relabeling baselines.



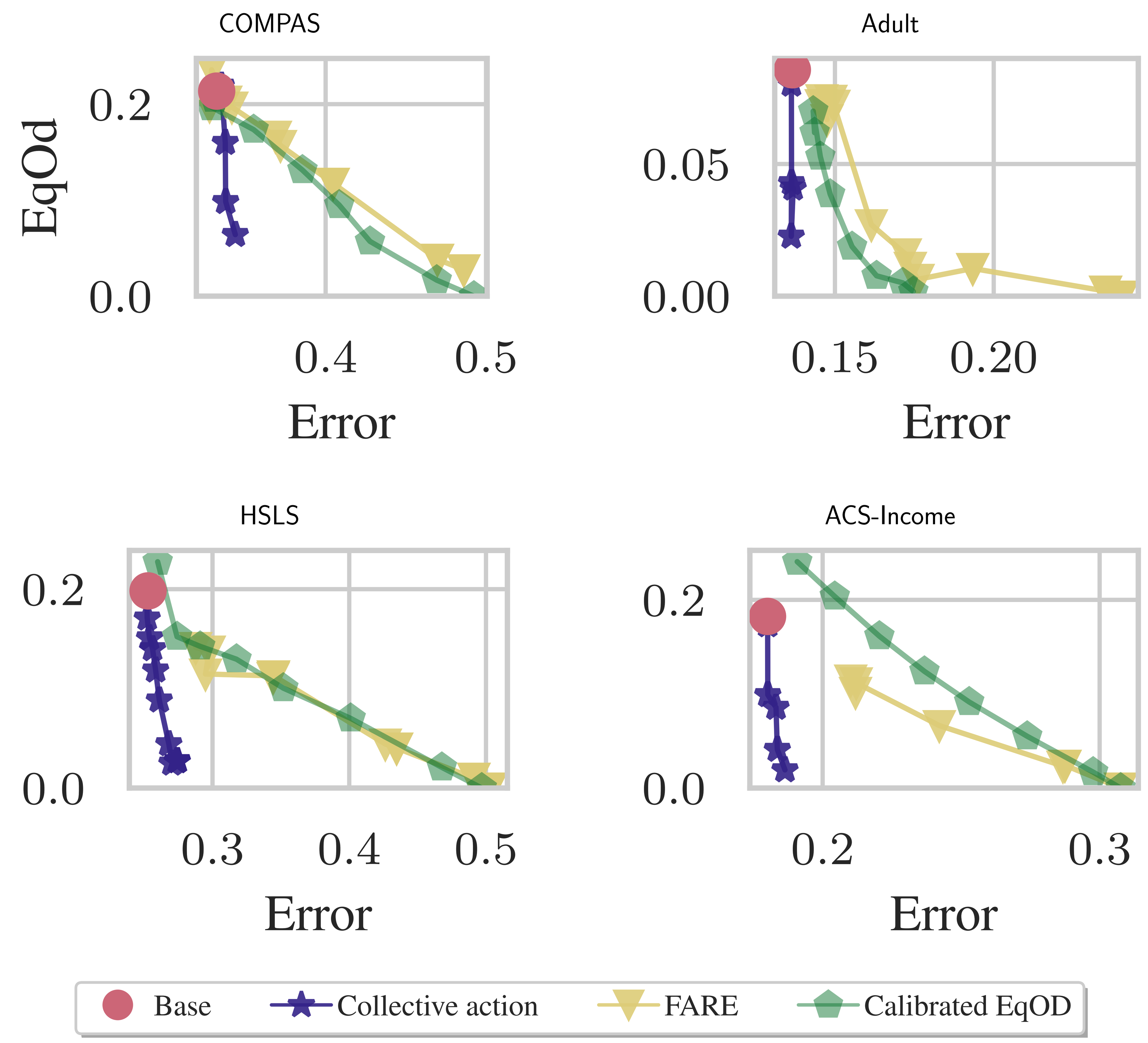
Varying Knowledge

k-NN methods are effective even with a few majority points.



Comparison With Firm-Side Methods

Unlike firm-side FARE [3] and calibrated equalized odds [4], a minority cannot get perfect fairness, but adds smaller error.



arXiv: https://arxiv.org/abs/2508.15374