



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

Omri Ben-Dov¹, Samira Samadi¹, Amartya Sanyal², Alexandru Tifrea³

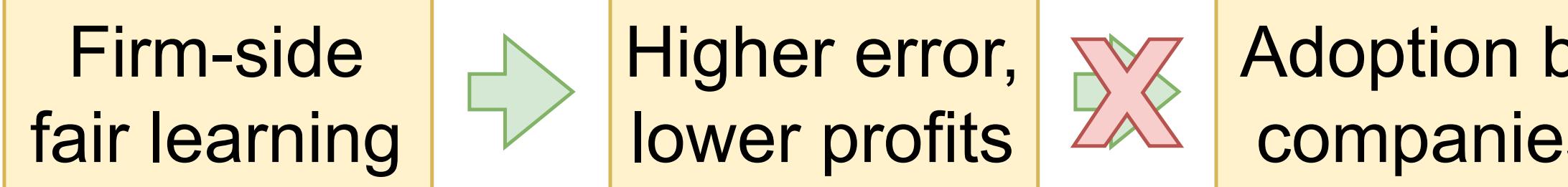
¹Max Planck Institute for Intelligent Systems, Tübingen AI Center, Tübingen, Germany

²Department of Computer Science, University of Copenhagen

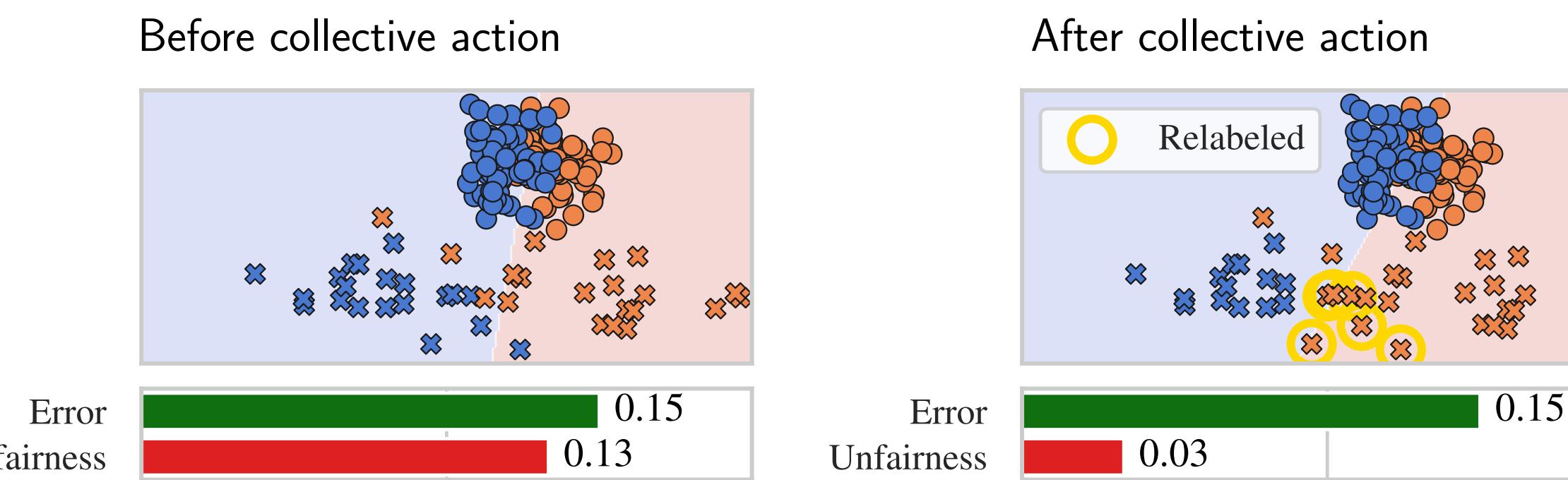
³ETH Zurich

Motivation

A profit-maximizing firm may ignore fair ML.



If a company has no incentive to be fair, but trains on user data, what can a minority do?



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

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

the collective can apply a relabeling strategy [1]

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

Setting the signal as a group counterfactual

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

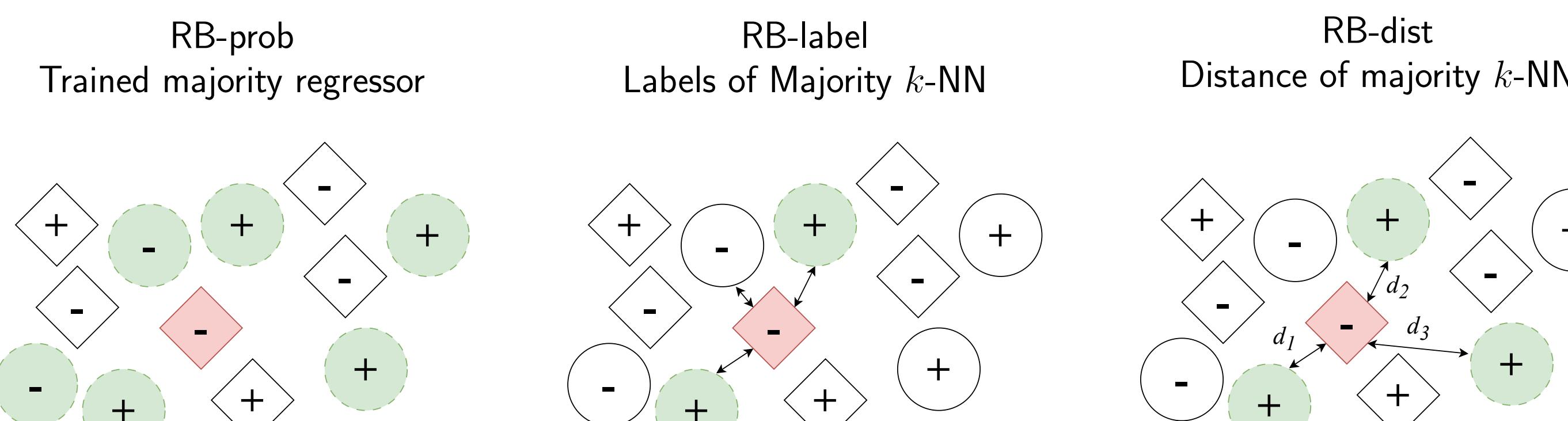
leads towards counterfactual fairness, in some cases promoting other forms of group fairness [2].

References

- [1] Moritz Hardt, Eric Mazumdar, Celestine Mender-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 Antithis 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

Counterfactuals are generally unknown. We propose methods to estimate the likelihood s of a positive counterfactual label.



$$s_i = f(x_i)$$

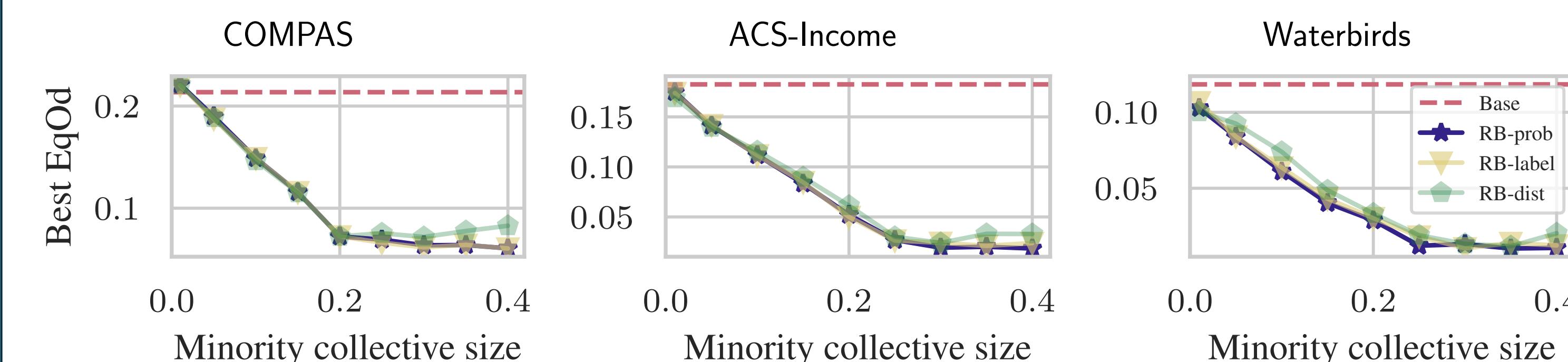
$$s_i = \sum_{j \in K_i} \mathbf{1}\{y_j = 1\}$$

$$s_i = -\frac{1}{k} \sum_{j \in K_i} \|x_i - x_j\|_2$$

The collective sorts the negative-labeled members by s and the top M flip their labels.

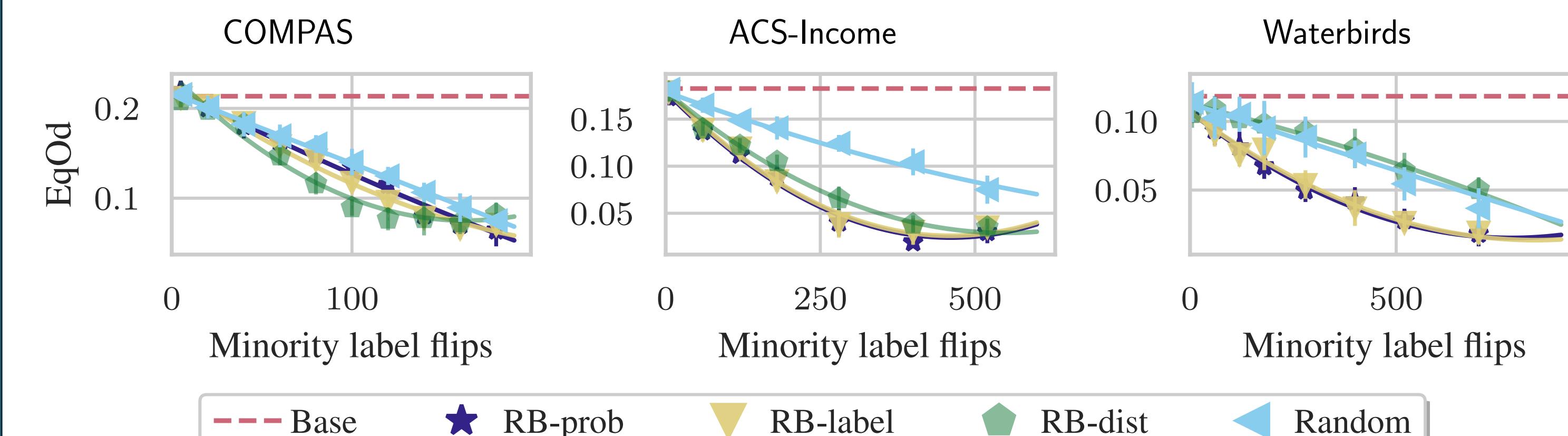
Importance of Collective Size

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



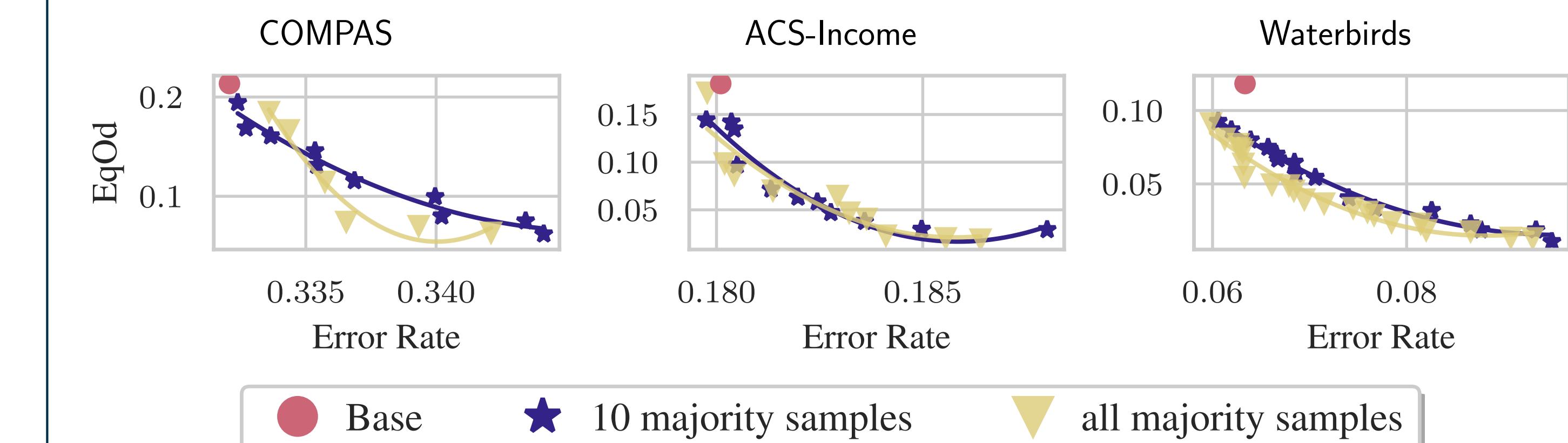
Required Number of Label Flips

Our methods require fewer flips than other relabeling baselines.



Restricting Access to Data

Limiting the collective's access to only few samples from the majority data has small effect on the Pareto frontiers.



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.

