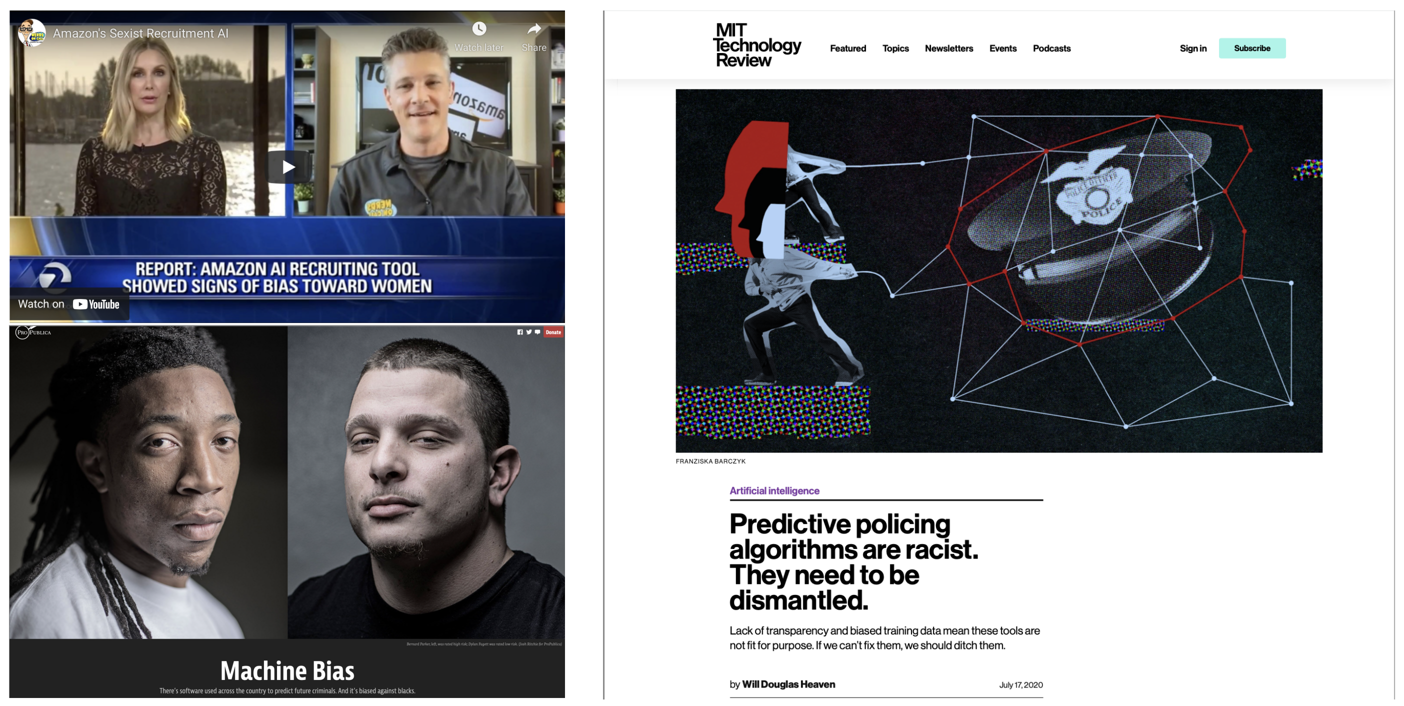
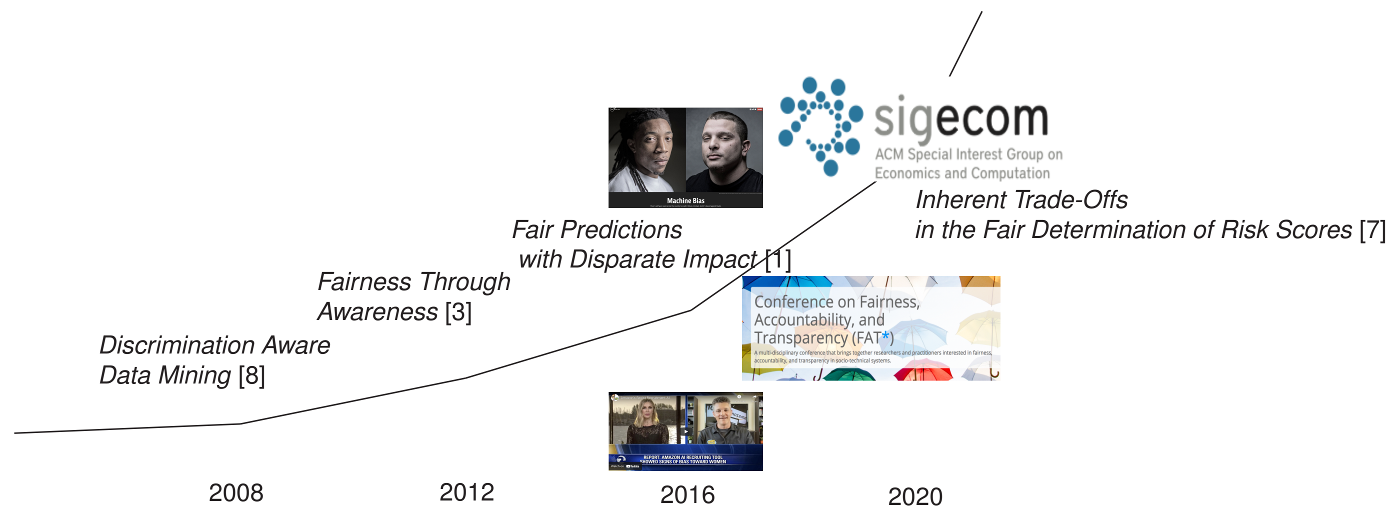
Multiaccurate Proxies for Downstream Fairness

Emily Diana (ediana@wharton.upenn.edu)
with Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, Saeed Sharifi-Malvajerdi
University of Pennsylvania

ALGORITHMIC FAIRNESS IN THE NEWS



ALGORITHMIC FAIRNESS IN THE LITERATURE



RESEARCH QUESTION

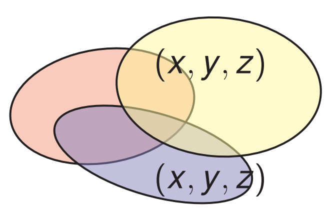
- Algorithmic fairness aims to understand and prevent bias in machine learning models.
- Often one wants to train a model that is fair with respect to a sensitive feature that has been redacted from training data?
- Could be for legal or policy reasons:
  - In the United States it is against the law to use race as an input to consumer lending models.
  - Many large consumer-facing organizations choose not to ask their customers for such information.

How do we make a model fair with respect to race if we don't have data about race?

FRAMEWORK

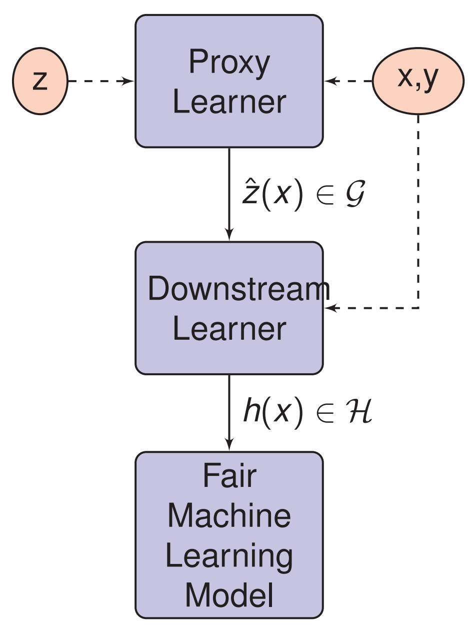
- Data domain Ω divided into K groups:

$$\Omega = \overbrace{\mathcal{X}}^{\text{non-sensitive features}} \times \underbrace{\mathcal{Y}}_{\text{label}} \times \overbrace{\mathcal{Z}}^{\text{sensitive feature}}$$



- Proxy model class $\mathcal{G} : \mathcal{X} \to \mathbb{R}^K$
- Proxy $\hat{z} \in \mathcal{G}$: vector of K real numbers $(\hat{z}_1, ..., \hat{z}_K)$
- Downstream model class $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$

Proxy Learner aims to find proxy $\hat{z}$ such that if a Downstream Learner trains a model h that is fair with respect to $\hat{z}$, h is also fair with respect to z.



KEY INSIGHT: PROXY CAN BE REAL VALUED

We can write fairness constraints, usually defined with respect to binary valued group membership using a real valued proxy

$$\begin{aligned}\Pr[h(x) \neq y | z_k = 1] &= \frac{\Pr[z_k = 1, h(x) \neq y]}{\Pr[z_k = 1]} \\ &= \frac{\mathbb{E}[1[z_k = 1]\, 1[h(x) \neq y]]}{\mathbb{E}[1[z_k = 1]]} \\ &= \frac{\mathbb{E}[z_k 1[h(x) \neq y]]}{\mathbb{E}[z_k]}\end{aligned}$$

KEY INSIGHT: REPLACE Z WITH $\hat{Z}$

If the following holds:

$$\frac{\mathbb{E}[z_k 1[h(x) \neq y]]}{\mathbb{E}[z_k]} = \frac{\mathbb{E}[\hat{z}_k(x) 1[h(x) \neq y]]}{\mathbb{E}[\hat{z}_k(x)]}$$

Then if a model is fair with respect to $\hat{z}$

$$\frac{\mathbb{E}[\hat{z}_{k_i}(x) 1[h(x) \neq y]]}{\mathbb{E}[\hat{z}_{k_i}(x)]} = \frac{\mathbb{E}[\hat{z}_{k_j}(x) 1[h(x) \neq y]]}{\mathbb{E}[\hat{z}_{k_j}(x)]}$$

it also satisfies fairness constraints with respect to the true attribute z.

MAIN RESULT: PROXY DEFINITION

We say $\hat{z}$ is an α-proxy for z if for all classifiers $h \in \mathcal{H}$, and all groups $k \in [K]$,

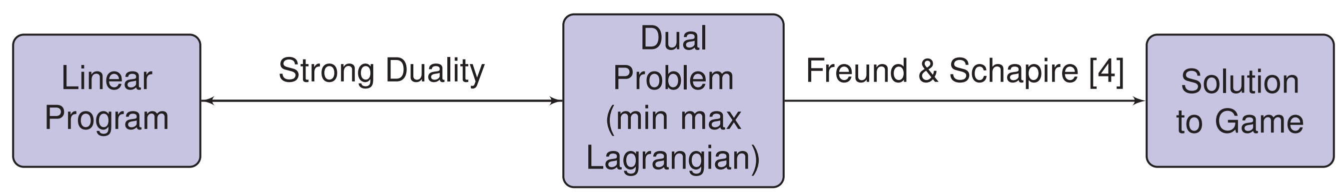
$$\left| \frac{\mathbb{E}_{(x,z)}[z_k 1[h(x) \neq y]]}{\mathbb{E}_{(x,z)}[z_k]} - \frac{\mathbb{E}_{(x,z)}[\hat{z}_k(x) 1[h(x) \neq y]]}{\mathbb{E}_{(x,z)}[\hat{z}_k(x)]} \right| \leq \alpha$$

Then to learn a proxy, we can solve the linear program:

$$\begin{aligned}\underset{\hat{z} \in \mathcal{G}}{\text{minimize}} \quad & \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}(x_i))^2 \\ \text{subject to} \quad & \frac{\sum_{i=1}^{n} z_i 1[h(x_i) \neq y_i]}{\sum_{i=1}^{n} z_i} = \frac{\sum_{i=1}^{n} \hat{z}(x_i) 1[h(x_i) \neq y_i]}{\sum_{i=1}^{n} \hat{z}(x_i)}, \quad \forall h \in \mathcal{H}\end{aligned} \qquad (1)$$

These constraints are multiaccuracy constraints [5, 6] – we want $\hat{z}$ to be an unbiased estimator for z on the set of points where h errs.

STRONG DUALITY AND LOW-REGRET DYNAMICS



EXPERIMENTS: OVERVIEW

Simulating a downstream learner, we train a model to be fair with respect to four representations of the sensitive feature and evaluate its performance:

- True Labels: Z
- Baseline Proxy: Logistic regression of Z on X
- **$\mathcal{H}$-Proxy: Solution to Program (1) without squared error objective**
- **MSE Proxy: Solution to Program (1) with squared error objective**

Conducted experiments on American Community Survey (ACS) datasets and tasks from [2].

Dataset	Sample Count	$\mathcal{X}$ Dim	Label
ACSEmployment	196104	12	Employment
ACSIncome	**101270**	**4**	**Income > $50K**
ACSIncomePovertyRatio	196104	15	Income-Poverty Ratio < 250%
ACSMobility	39828	17	Same address one year ago
ACSPublicCoverage	71379	15	Health Insurance
ACSTravelTime	89145	8	Commute > 20 minutes

EXPERIMENTS: ACS DATA

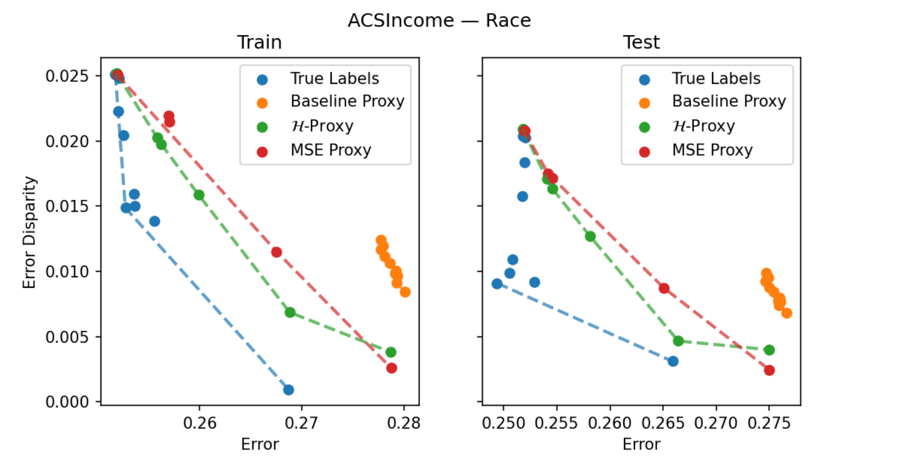


Figure: Proxy results on the ACSIncome dataset with race as sensitive feature

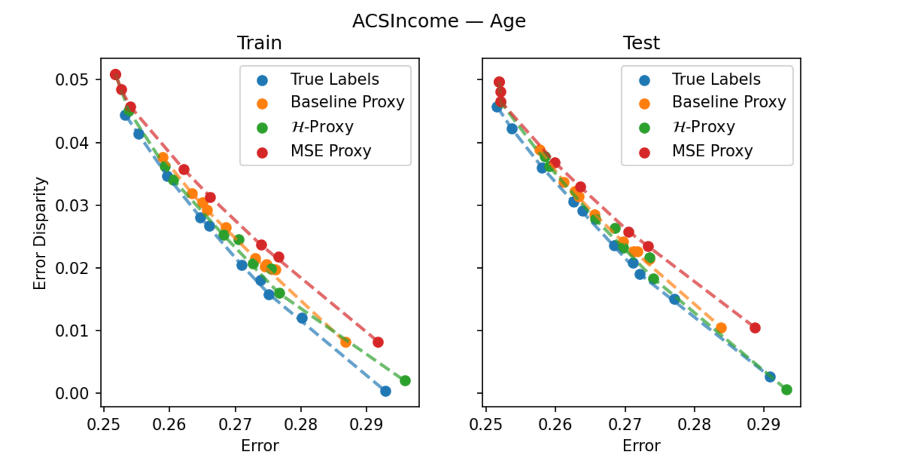


Figure: Proxy results on the ACSIncome dataset with age as sensitive feature

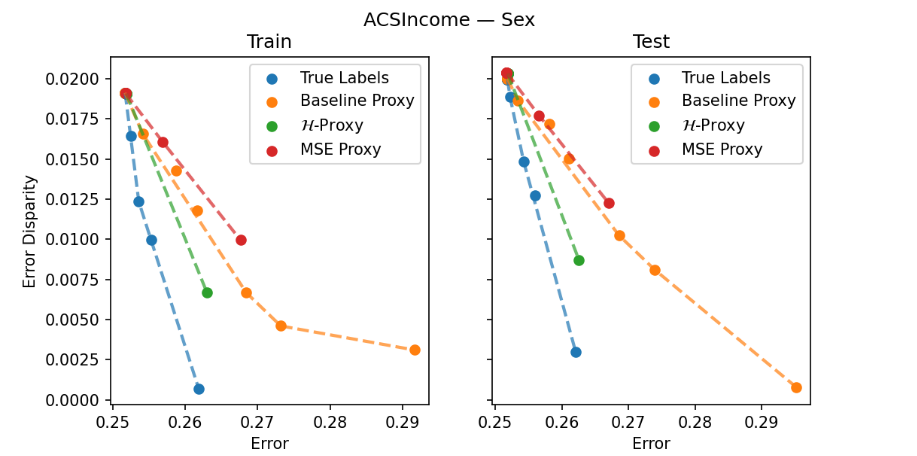


Figure: Proxy results on the ACSIncome dataset with sex as sensitive feature

CONCLUSION

- We have shown that it is possible to efficiently train proxies that can stand in for missing sensitive features to effectively train downstream classifiers subject to a variety of demographic fairness constraints.
- Our theoretical and empirical results demonstrate that proxies trained using our methods can stand in as near perfect substitutes for sensitive features in downstream training tasks.
- Results crucially depend on the assumption that the data that the Proxy Learner uses to train its proxy is distributed identically to the data that the Downstream Learner uses.
- In real applications, either of these assumptions can fail (or can become false due to distribution shift, even if they are true at the moment that the proxy is trained).

SELECTED REFERENCES

[1] Alexandra Chouldechova. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments". In: *Big Data* 5.2 (2017), pp. 153–163. DOI: 10.1089/big.2016.0047. URL: https://doi.org/10.1089/big.2016.0047.

[2] Frances Ding et al. "Retiring Adult: New Datasets for Fair Machine Learning". In: *CoRR* abs/2108.04884 (2021). arXiv: 2108.04884. URL: https://arxiv.org/abs/2108.04884.

[3] Cynthia Dwork et al. "Fairness through Awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. Cambridge, Massachusetts: Association for Computing Machinery, 2012, 214â226. ISBN: 9781450311151. DOI: 10.1145/2090236.2090255. URL: https://doi.org/10.1145/2090236.2090255.

[4] Yoav Freund and Robert E. Schapire. "Game Theory, On-line Prediction and Boosting". In: *Proceedings of the Ninth Annual Conference on Computational Learning Theory*. 1996.

[5] Ursula Hébert-Johnson et al. "Multicalibration: Calibration for the (computationally-identifiable) masses". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1939–1948.

[6] Christopher Jung et al. "Moment Multicalibration for Uncertainty Estimation". In: *Conference on Learning Theory*. PMLR. 2021.

[7] Jon Kleinberg. "Inherent Trade-Offs in Algorithmic Fairness". In: *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*. SIGMETRICS '18. Irvine, CA, USA: Association for Computing Machinery, 2018, p. 40. ISBN: 9781450358460. DOI: 10.1145/3219617.3219634. URL: https://doi.org/10.1145/3219617.3219634.

[8] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. "Discrimination-Aware Data Mining". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08. Las Vegas, Nevada, USA: Association for Computing Machinery, 2008, 560â568. ISBN: 9781605581934. DOI: 10.1145/1401890.1401959. URL: https://doi.org/10.1145/1401890.1401959.