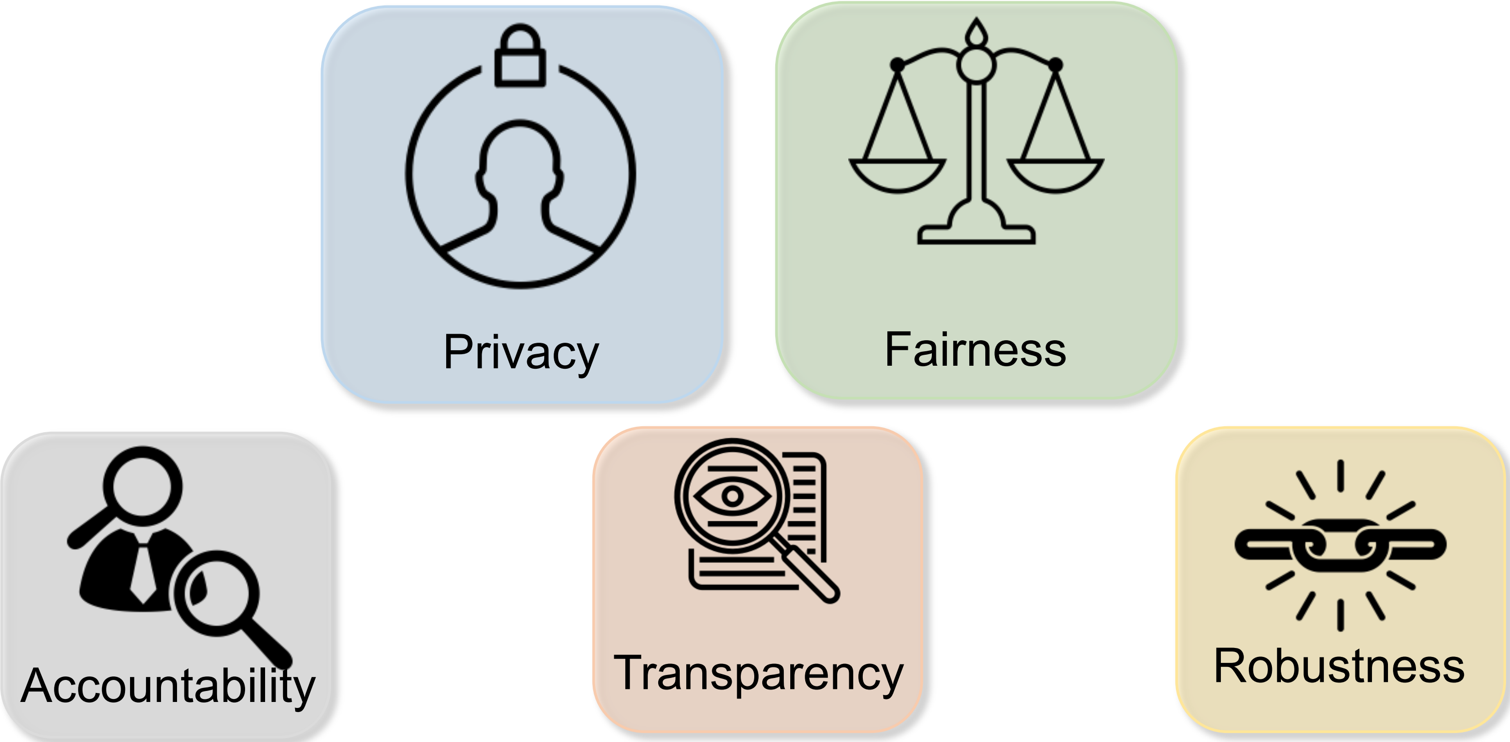
On the Privacy Risks of Algorithmic Fairness

Hongyan Chang, Reza Shokri
National University of Singapore

hongyan@comp.nus.edu.sg, reza@comp.nus.edu.sg

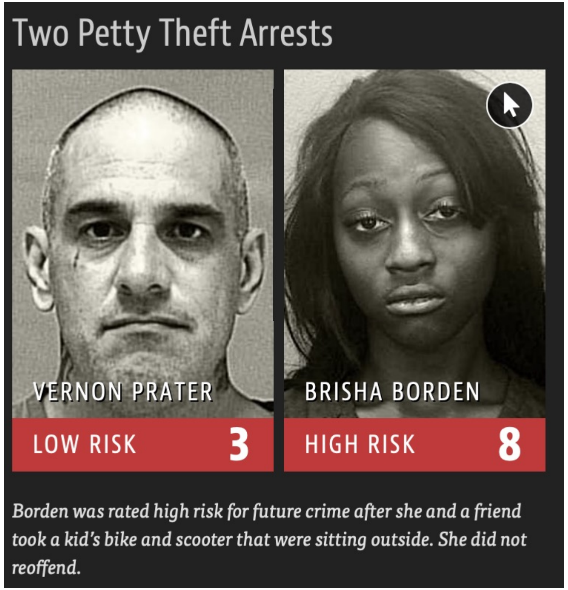
Ethical AI

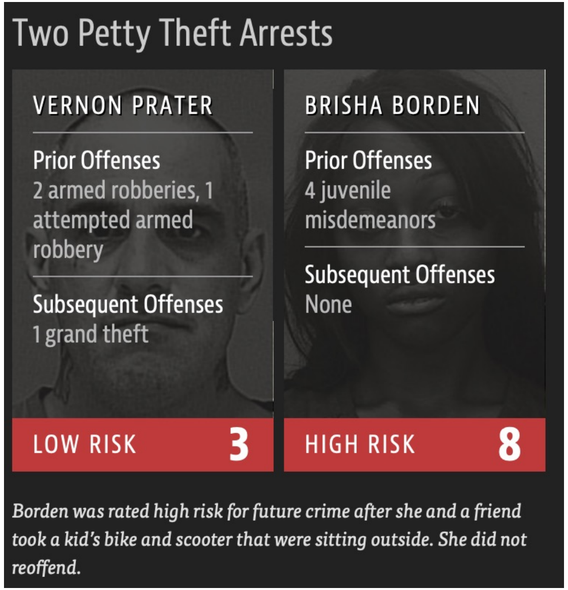


User Privacy by Yair Cohen, Scale by Douglas Machado, Search User by Francisco Garcia Gallegos, Transparency by Wichai Wi from the Noun Project

ML models are not neutral

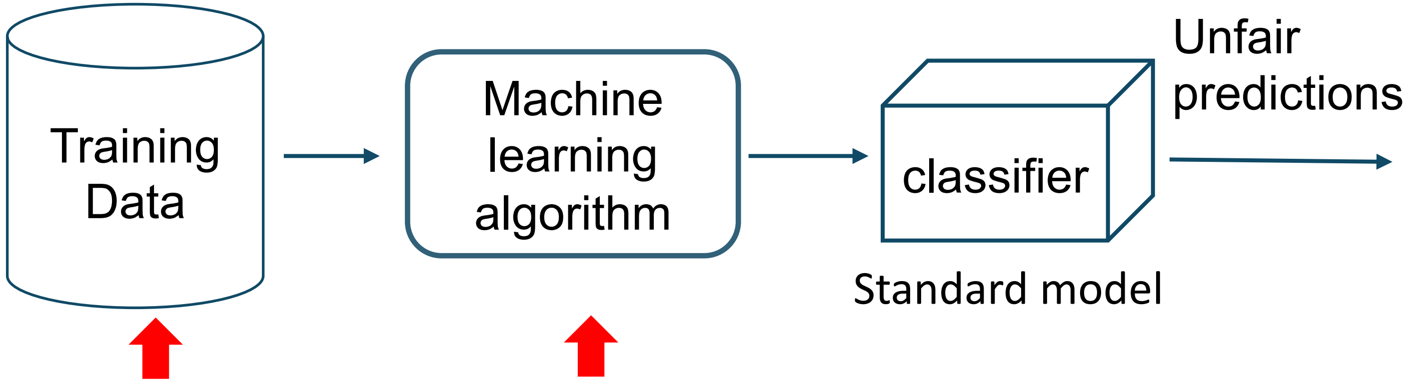
- Recidivism prediction

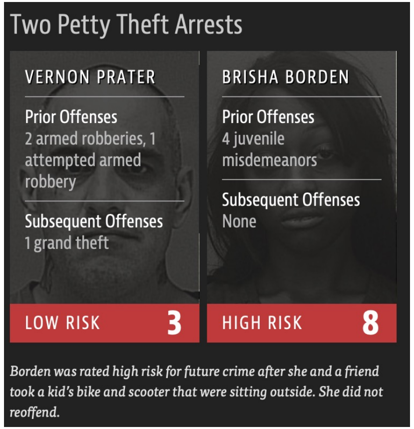




Source from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bias in ML

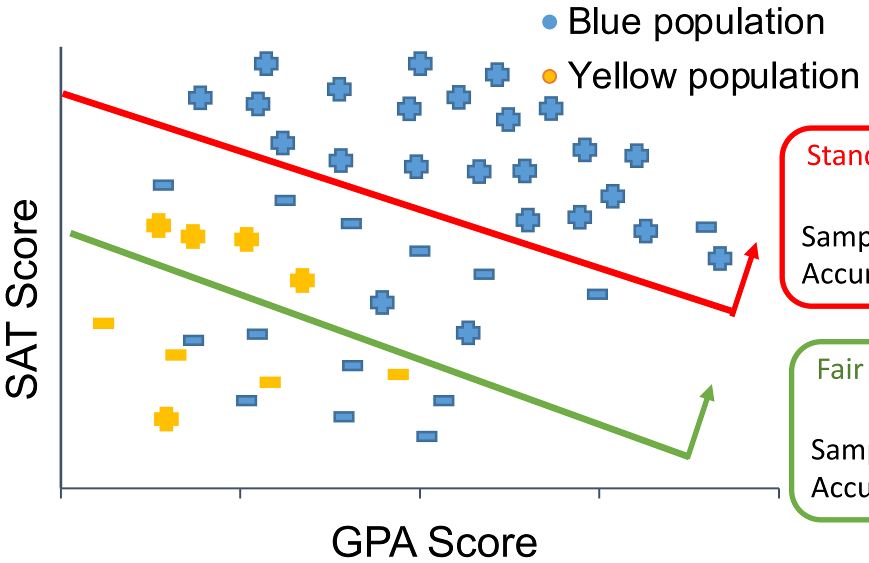




- **Data bias** : training data can reflect the human bias
- **Algorithmic bias:** the model is learned to minimize the overall loss

Fairness in ML

- **Equalized odds:** TPR and TNR should be similar across protected groups that are defined by a sensitive attribute (e.g., race, gender)



Standard model

	Blue | Yellow
Sample	80% | 20%
Accuracy	92% | 44%

Fair model

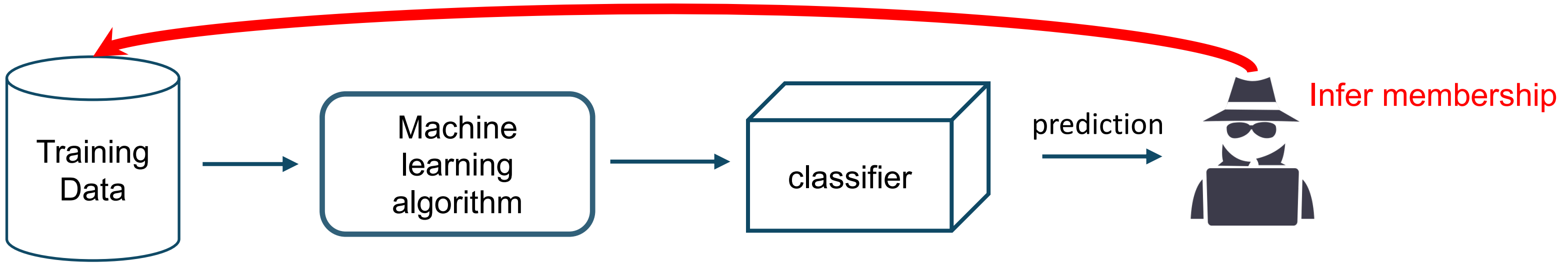
	Blue | Yellow
Sample	80% | 20%
Accuracy	81% | 88%

Side effect: **increase** the influence of the training data from underprivileged group on the learned model

Hardt, Moritz, Eric Price, and Nathan Srebro. “Equality of Opportunity in Supervised Learning.” NeurIPS 2016
Example is from Michael Kearns & Aaron Roth talk at Google

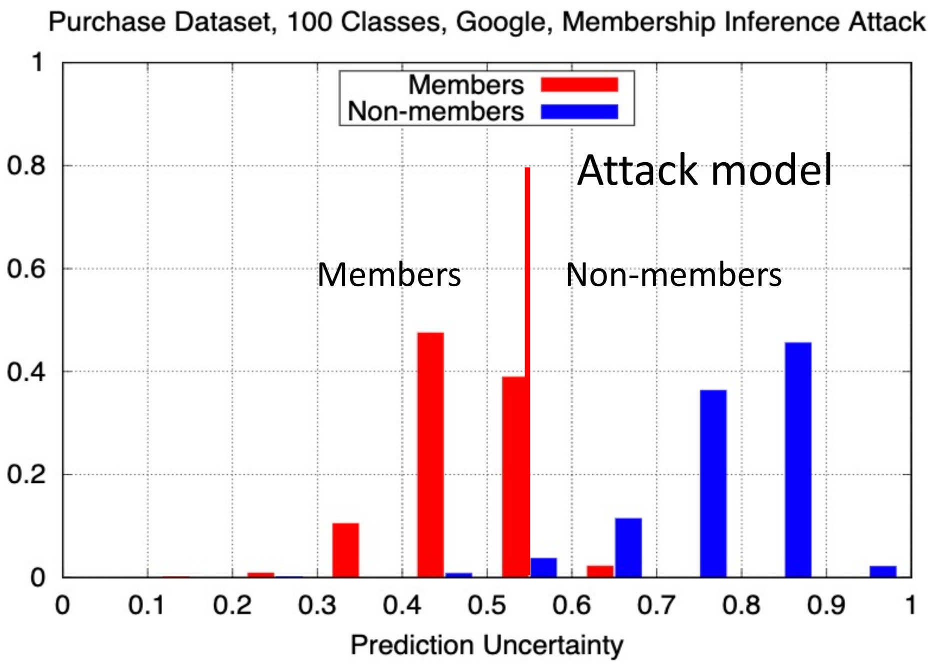
Fairness meets privacy

Membership inference attack: infer whether an individual's data is in the training dataset or not



Dwork, Cynthia, et al. "Calibrating noise to sensitivity in private data analysis." TCC, 2006.
Shokri, Reza, et al. "Membership inference attacks against machine learning models." *SP,* 2017.

Membership inference attack

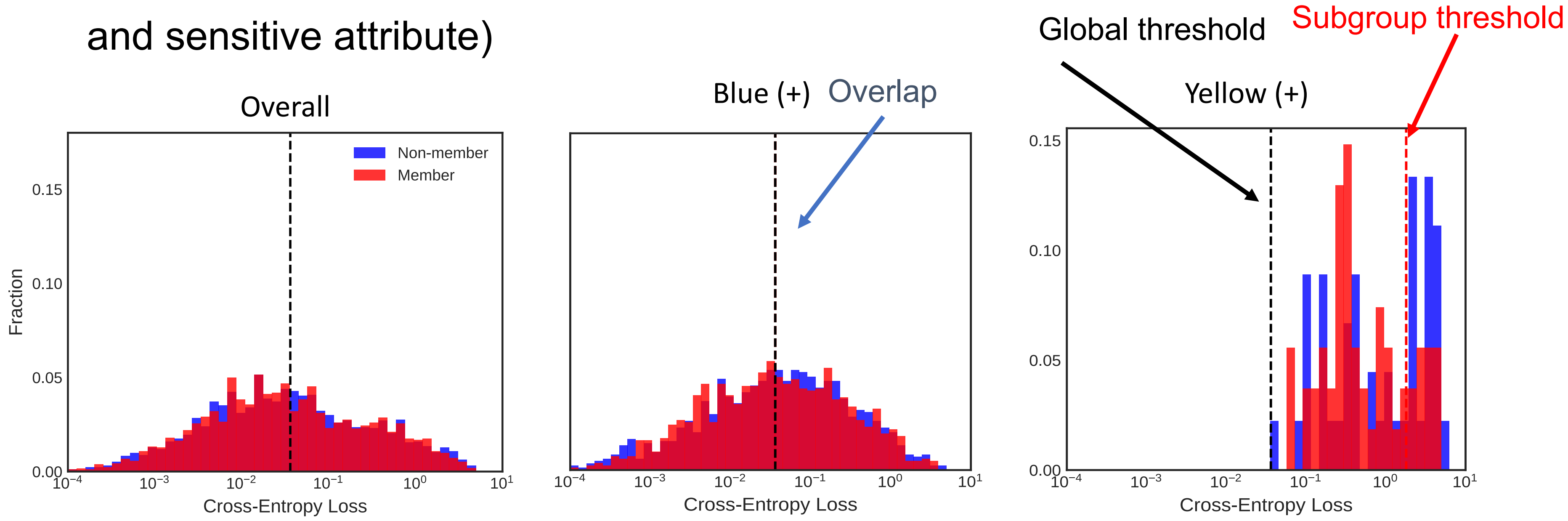


From Shokri et al.

Shokri, Reza, et al. "Membership inference attacks against machine learning models." *SP,* 2017.

Our attack strategy

- **Our proposal:** find an attack model for each subgroup (defined by the label and sensitive attribute)



Synthetic dataset

Attack accuracy

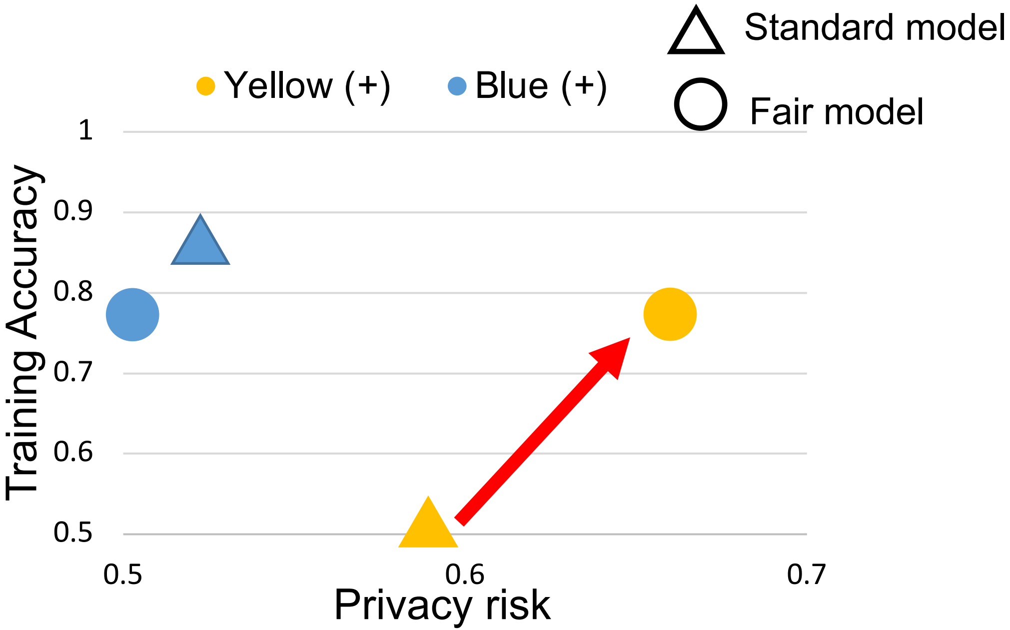
- Synthetic data with equalized odds (fairness gap is 0.001)

Attack Strategy	Target model	Yellow (+)	Blue (+)	Yellow (-)	Blue (-)
Single attack model	Standard	0.529	0.512	0.518	0.512
	Fair	0.608	0.528	0.524	0.522
Subgroup based attack	Standard	0.618	0.528	0.524	0.522
	Fair	0.692	0.534	0.525	0.515

Privacy cost = **Privacy risk on fair model** - **Privacy risk on unconstrained model**

Achieving group fairness increases the privacy risk of underprivileged subgroup

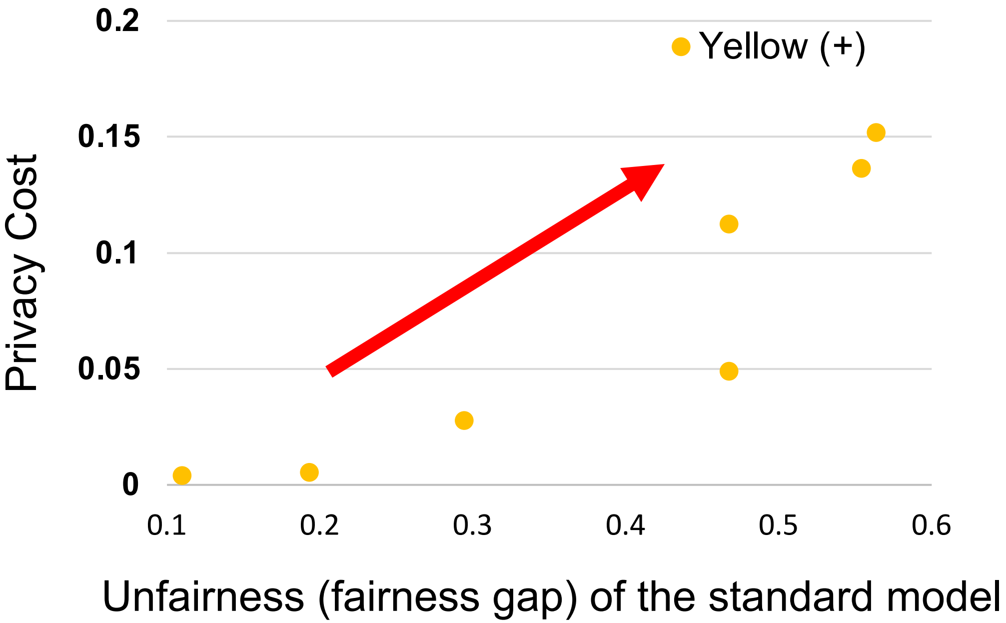
- Synthetic data with equalized odds (fairness gap is 0.001)



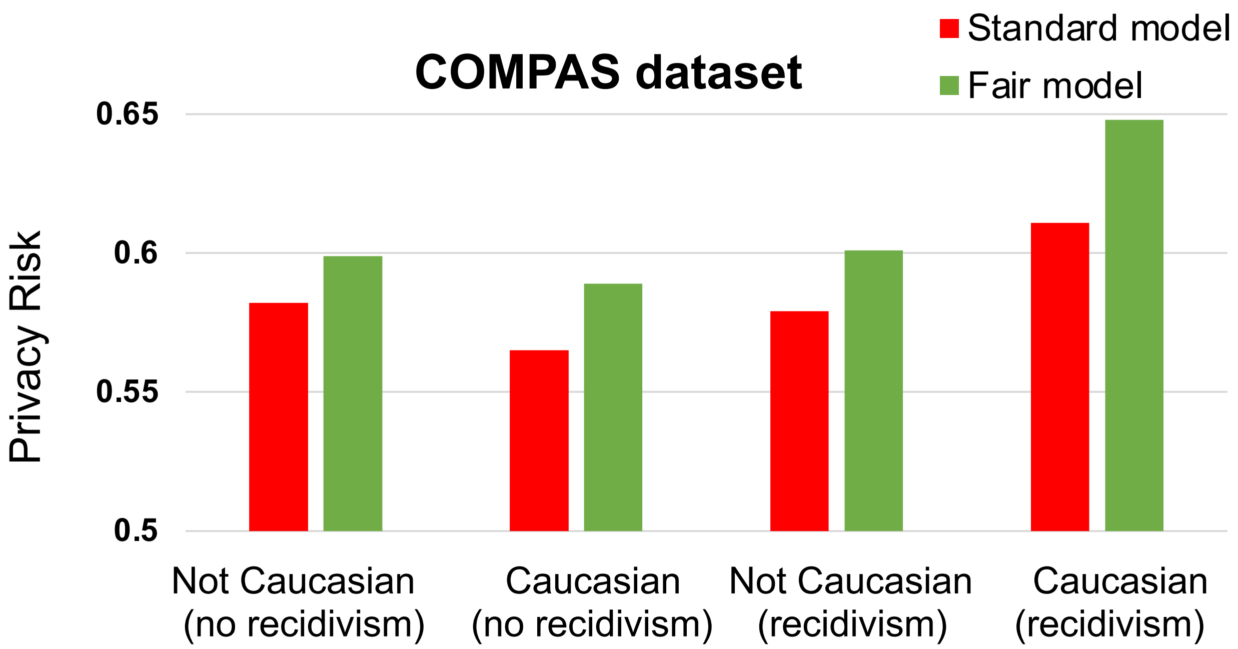
Fair model improves the accuracy but leaks more information on yellow population

Trade-off between group fairness and privacy

When there are more needs for fairness, privacy cost of achieving fairness is higher



Real world data



Takeaways

- Group fairness based on equalizing error across groups comes at the cost of privacy
- Privacy cost is not distributed equally across groups
- “Protecting” underprivileged groups using fair ML increases their privacy risks