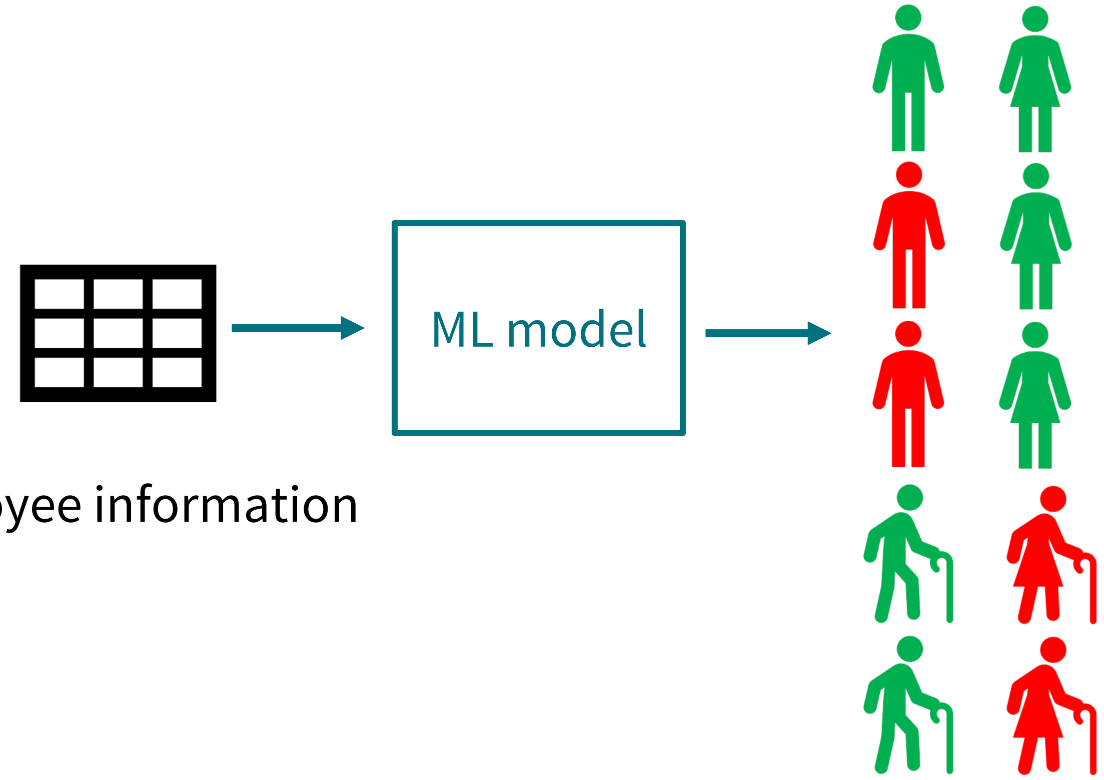
On the Fairness ROAD: Robust Optimization for Adversarial Debiasing

Vincent Grari*, Thibault Laugel*, Tatsunori Hashimoto, Sylvain Lamprier and Marcin Detyniecki

*Equal contribution

Context: algorithmic group fairness



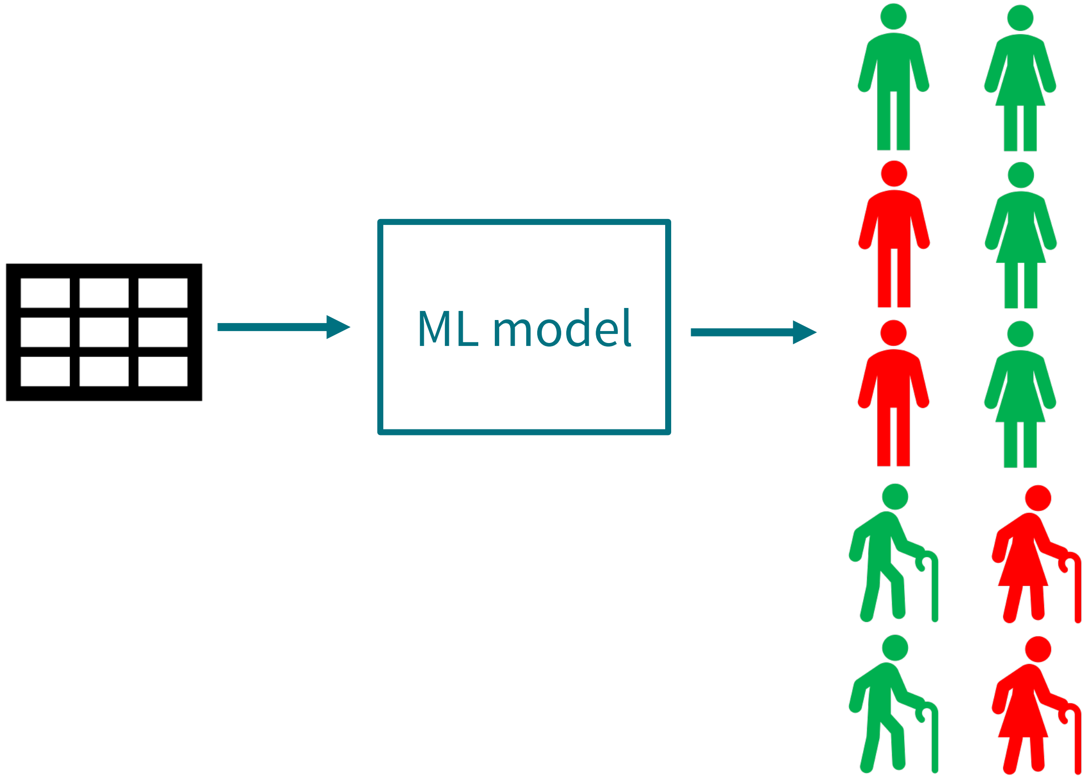
Employee information

Deserves a raise or not

Context: algorithmic group fairness

Traditional group fairness

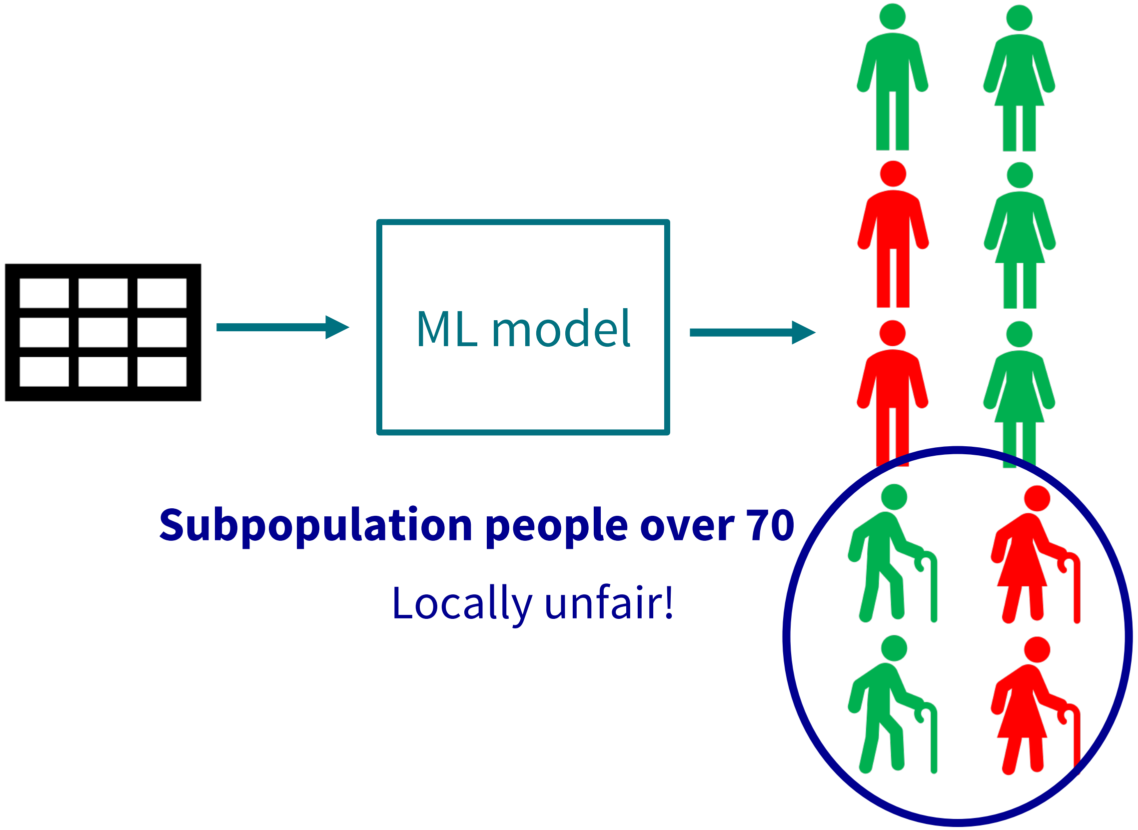
Globally fair model (DP): $\mathbb{P}(\hat{Y} = 1 | S = 1) = \mathbb{P}(\hat{Y} = 1 | S = 0)$



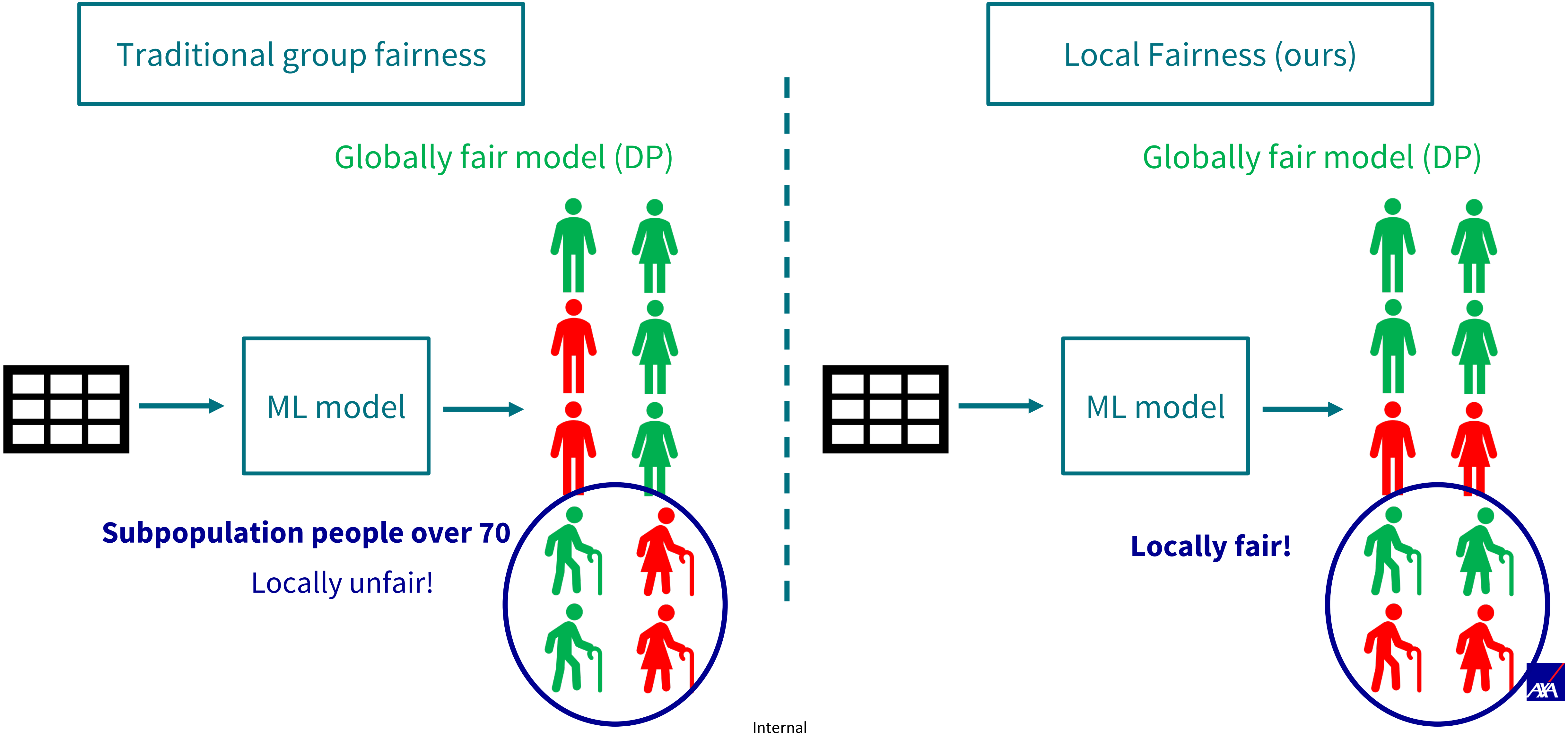
The Local (un)fairness problem

Traditional group fairness

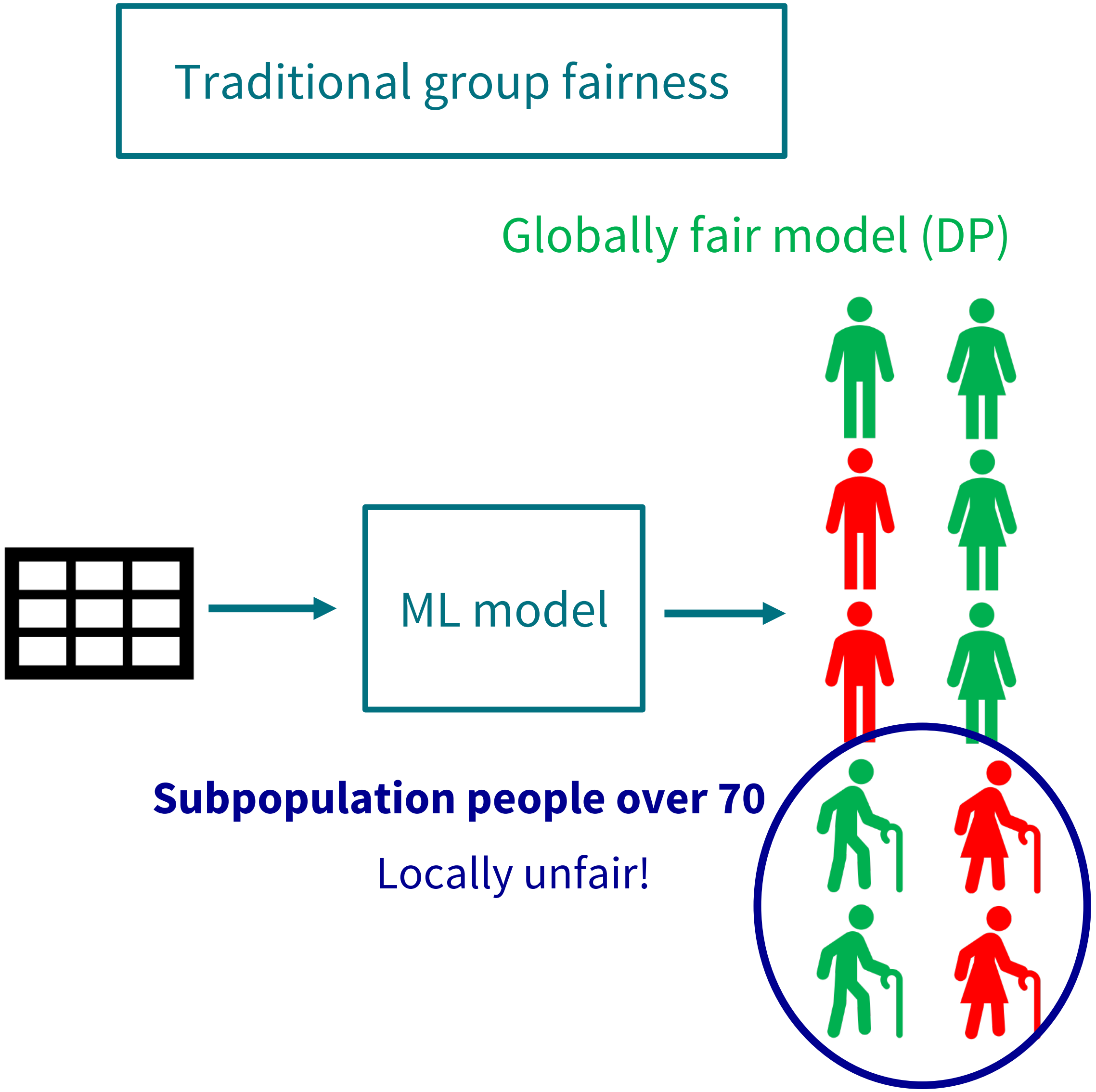
Globally fair model (DP)

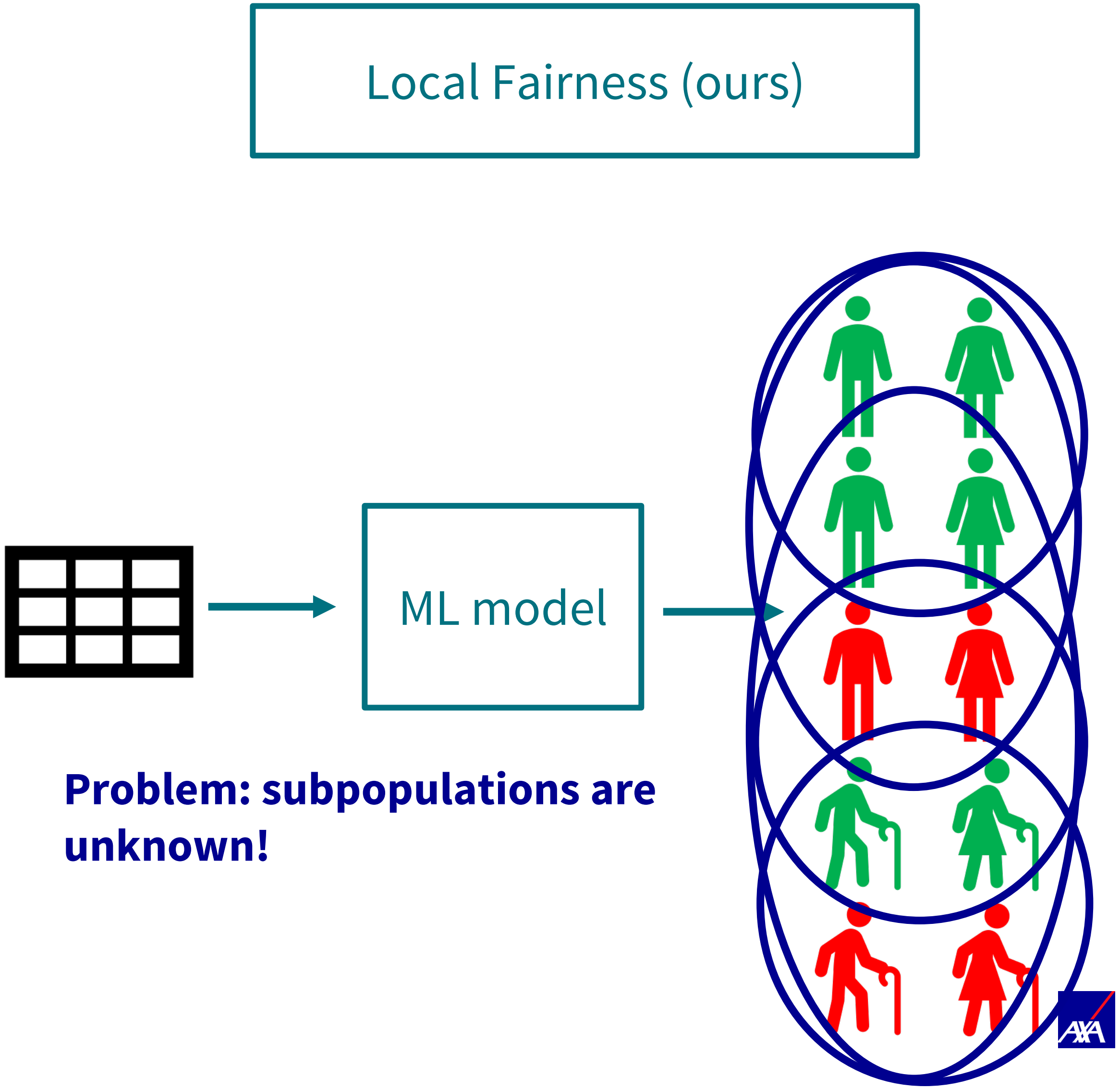


The Local (Un)fairness problem



The Local (Un)fairness problem





Distributionally Robust Optimization (DRO) for Fairness

Traditional group fairness

$$\min_{w_f} \mathbb{E}_p[L_Y\left(f_{w_f}(x), y\right)]$$

$$s.t. DI_{(x,s)\sim p}\left(f_{w_f}(x), s\right) < \epsilon$$

Local Fairness (ours)

$$\min_{w_f} \mathbb{E}_p[L_Y\left(f_{w_f}(x), y\right)]$$

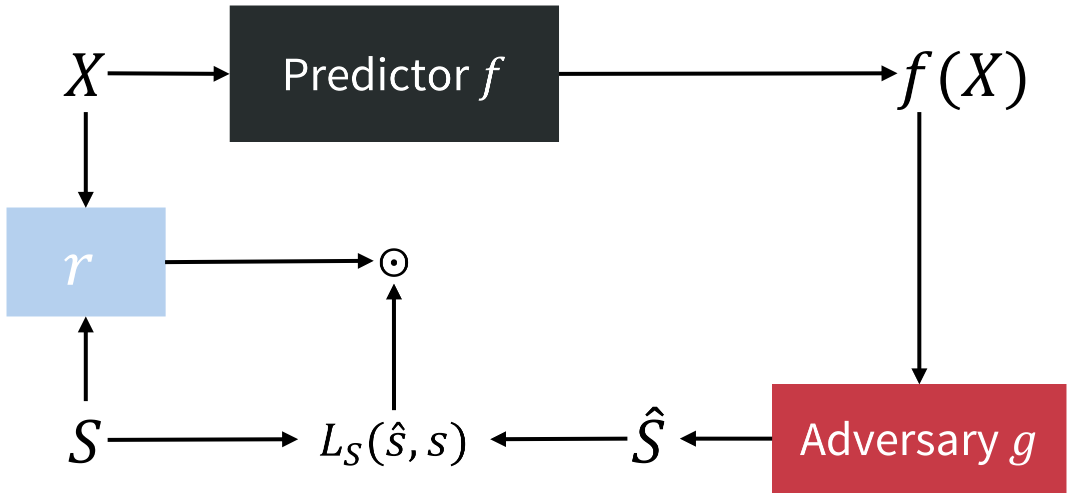
$$s.t. \max_{q\in Q} DI_{(x,s)\sim q}\left(f_{w_f}(x), s\right) < \epsilon$$

Q : set of "plausible" distributions
~set of subpopulations

In practice: KL divergence-ball around p

Adversarial model for Distributionally Robust Fairness

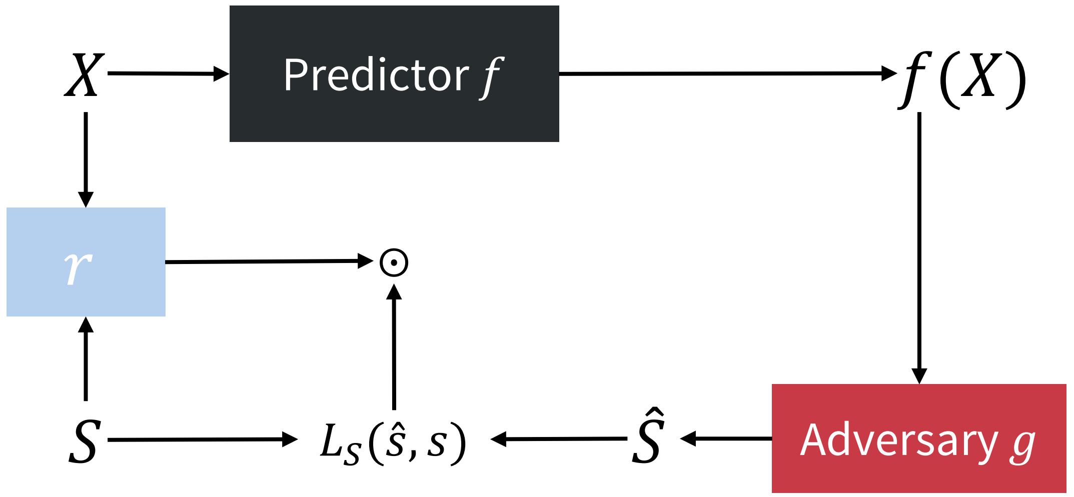
$$L_Y(f(x), y) - \lambda r(x, s) L_s(\hat{s}, s) + KL\ constraint$$



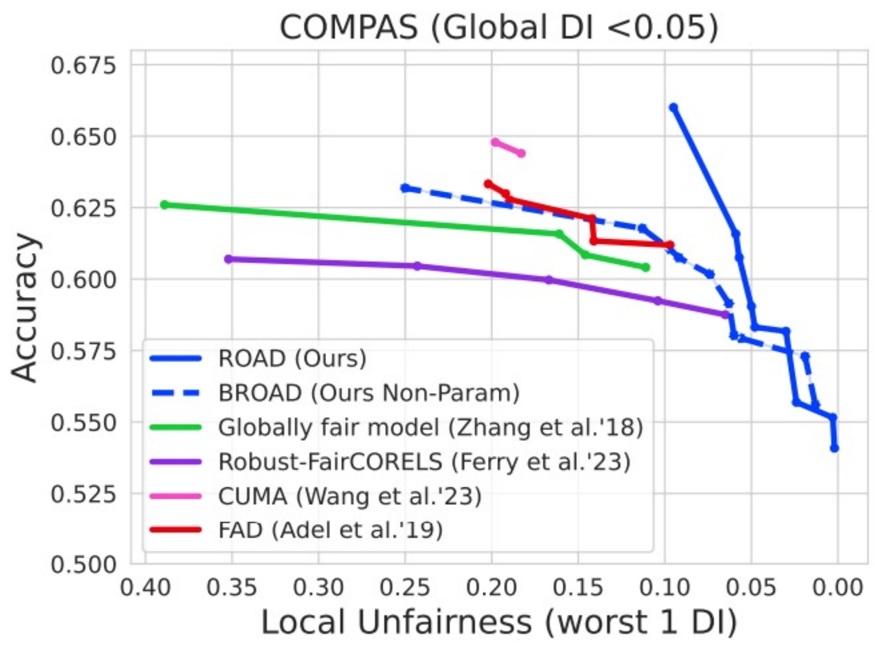
Classical adversarial approach for fairness

Adversarial model for Distributionally Robust Fairness

$$L_Y(f(x),y) - \lambda r(x,s)L_s(\hat{s},s) + KL\ constraint$$



Classical adversarial approach for fairness



Results: more fair locally for the same levels of group fairness and accuracy

Thanks for watching!

Paper : https://openreview.net/forum?id=xnhvVtZtLD
Code: https://github.com/axa-rev-research/ROAD-fairness/