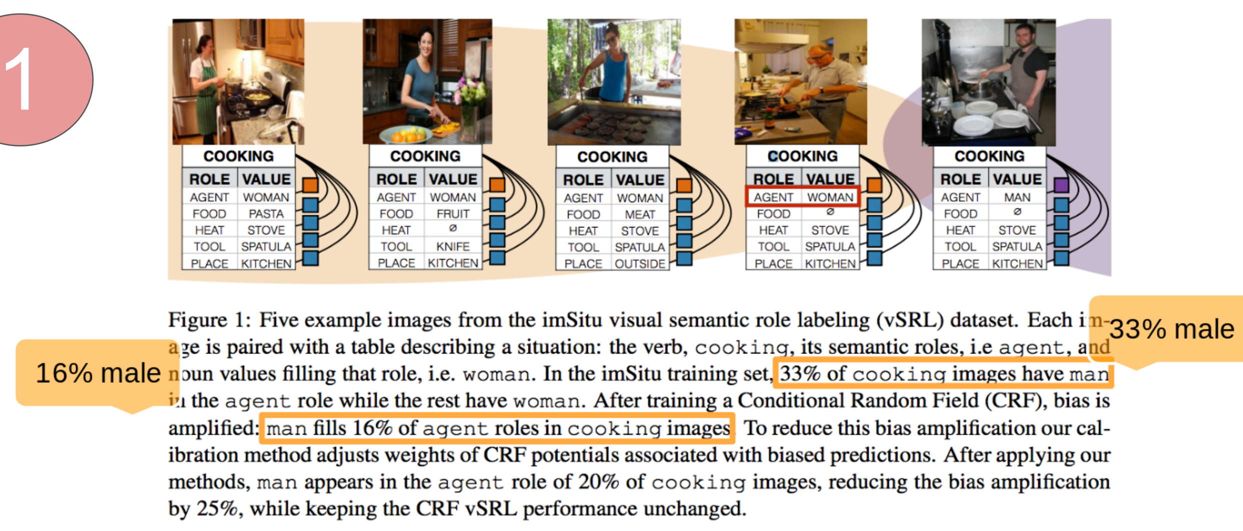
Pareto Efficient Fairness for Skewed Subgroup Data

Ananth Balashankar, Alyssa Lees, Chris Welty, Lakshminaryanan Subramanian
ananth@nyu.edu, {alyssalees, welty}@google.com, lakshmi@nyu.edu

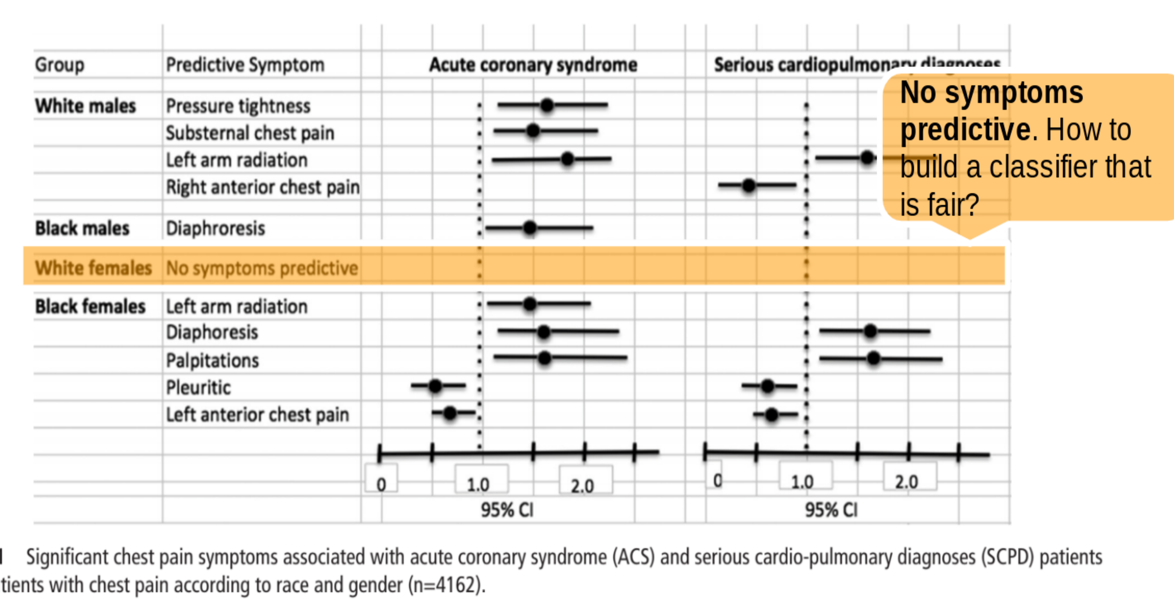
ML can amplify societal biases

1



"Men Also Like Shopping Reducing Gender Bias Amplification using Corpus-level Constraints" [Zhao et al., 2017]

2



"Gender, race and the presentation of acute coronary syndrome and serious cardiopulmonary diagnosis in ED patients with chest pain" [Allabban et al., 2017]

Perfect Fairness and Accuracy are at odds

- **Theorem**: If the class probabilities of the target (Y) and sensitive features (S) are *aligned,* there is an unavoidable degradation in accuracy owing to the fairness requirement [Menon and Williamson, 2018]

- Examples of Perfect Fairness Requirements (assuming Y is binary):
  - Demographic Parity: P(Ŷ=1| S=m) = P(Ŷ=1| S=n) [Calders and Verwer, 2010]
  - Equality of Opportunity: P(Ŷ=1|Y=1, S=m) = P(Ŷ=1|Y=1, S=n) [Hardt et al., 2016]
  - Equality of Odds: P(Ŷ=ŷ|Y=y, S=m) = P(Ŷ=ŷ|Y=y, S=n); ŷ, y ∊ {0,1}, ∀m, n ∊ S, m ≠ n [Hardt et al., 2016]

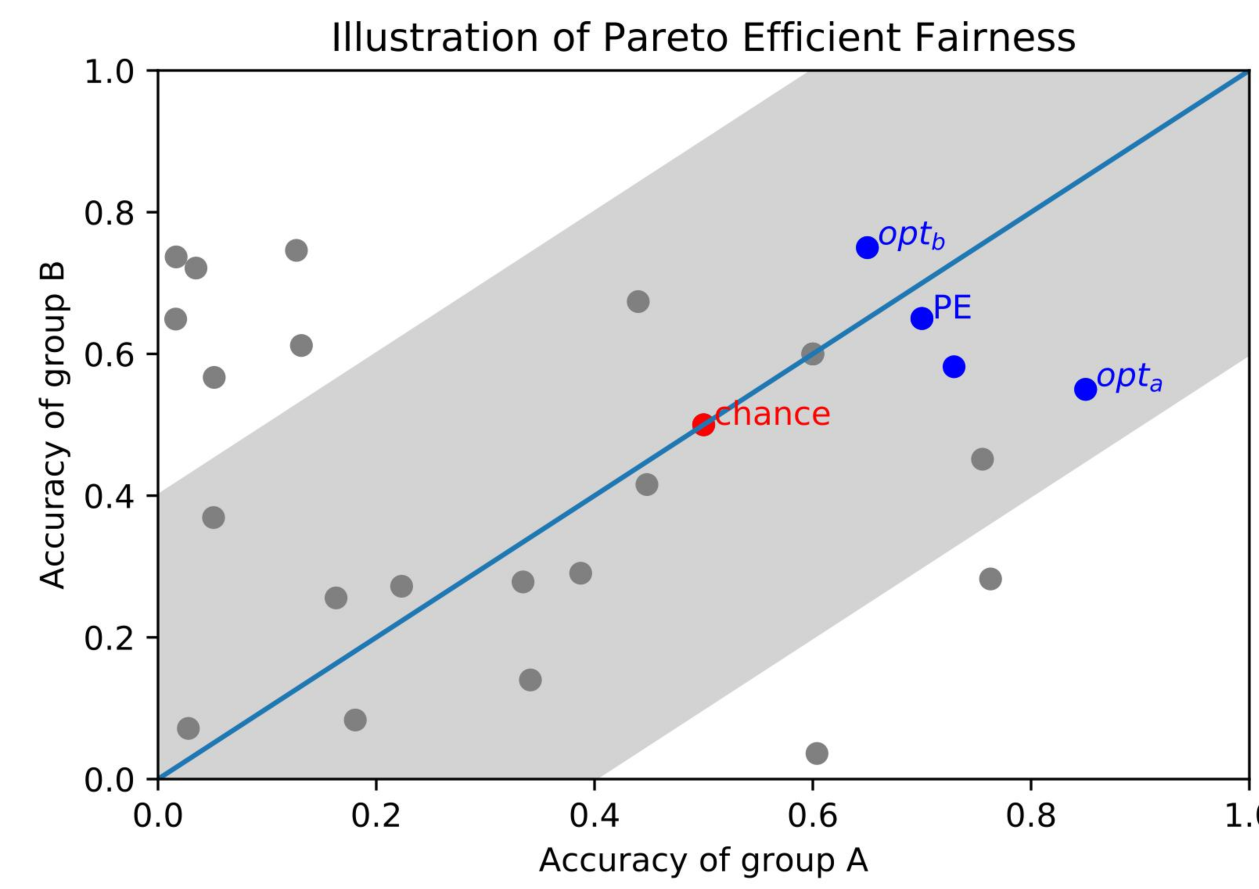
Existing Approaches

- Approximate Fairness with Lagrangian Constraints [Zhao et al., 2017]
  - Augments a fairness penalty term to the cross entropy loss
  - The penalty factor λ is hard to fine-tune by the domain expert
- Adversarial Multi-Task Learning [Beutel et al., 2017]
  - Learn the target and ensure that the sensitive feature is not learnt through negative gradients of multi-task learning
  - No control given to domain expert to control the trade-off meaningfully

What we want in many cases...

- Policies like affirmative action and UN Sustainable Development goals aim to improve performance of protected groups to meet the levels of the highest performing groups [Foster and Vohra, 1992]
- In skewed subgroup datasets, there might be an opportunity to choose performance for all groups that are better than the best perfectly fair one
- We want to give the domain expert the ability to search for Pareto-Efficient performance within a Fairness bound

Pareto-efficient fairness is better



- **Pareto-Efficient**: Set of points for which there does not exist another point, which is better performing across all sensitive groups
- **Fairness**: The deviation from each group's Pareto-optimal point is distributed equitably among groups

Pareto efficient bias mitigation

$$\mathcal{L}_p = \mathcal{L}_{ce} + \lambda(\alpha\|\mathcal{E}_G\|_1 + (1-\alpha)\sigma_G^2(\mathcal{E}_G))$$

G: set of sensitive groups, D: dataset, D_g: data of group $g \in G$
for $g \in G$ **do**
  $M_g = \arg\min \mathcal{L}_{ce}(D_g)$
  $f_{opt-g} = \text{eval}(M_g, D_g)$
  $f_g = \emptyset$
end for
while $\exists g \in G, f_g = \emptyset \vee f_g > f_{opt-g}$ **do**
  $f_{opt-g} = \max(f_g, f_{opt-g}), \forall g \in G$
  $M = \arg\min \mathcal{L}_p(D)$
  $f_g = \text{eval}(M, D_g), \forall g \in G$
end while

Evaluation on UCI Census data

- Data
  - 14 demographic features like age, education, occupation, etc from 1994 census
  - Predict if income is > 50K or not
  - Sensitive variables assumed are race and gender

- **Pareto Efficient loss is minimized, while achieving best overall accuracy**

Table 1. UCI Adult dataset with bias mitigation algorithms

Model	Accuracy	FPR	FNR	Discrepancy	Pareto Loss
Baseline (no bias loss)	0.630	0.253	0.747	0.199	0.016
Minimize Discrepancy	0.619	0.283	**0.712**	**0.167**	0.133
Adversarial Loss	0.648	0.224	0.769	0.226	0.077
Pareto-Efficient Loss	**0.678**	**0.165**	0.830	0.250	**0.000**

- **Pareto Efficient loss ensures best accuracy across all subgroups too!**

Table 2. Subgroup performance on UCI Adult dataset

Model	Subgroup 1	2	3	4	Pareto Loss
Baseline (no bias loss)	0.890	0.883	0.818	0.784	0.016
Minimize Discrepancy	0.853	0.856	0.806	0.778	0.133
Adversarial Loss	0.882	0.872	0.824	0.780	0.077
Pareto-Efficient Loss	**0.935**	**0.915**	**0.844**	**0.797**	**0.000**
Subgroup Pareto Frontier	0.934	0.894	0.815	0.783	N/A

NYU

Google AI