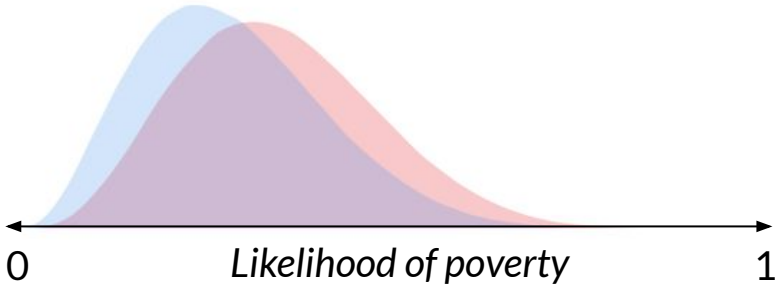
Fair Enough: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds

Michiel Bakker (MIT)
Humberto Riveron Valdes (MIT)
Patrick Tu (MIT)
Krishna Gummadi (MPI-SWS)
Kush Varshney (IBM Research)
Adrian Weller (University of Cambridge)
Alex Pentland (MIT)

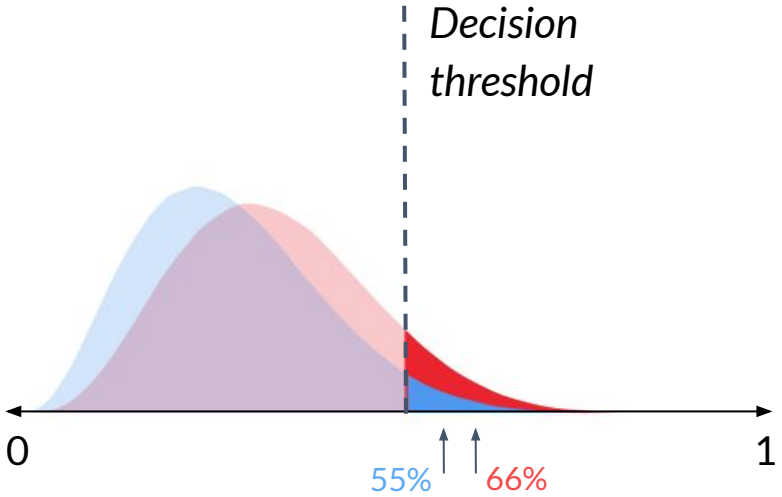
SafeAI 2020

A group-level measure of fairness



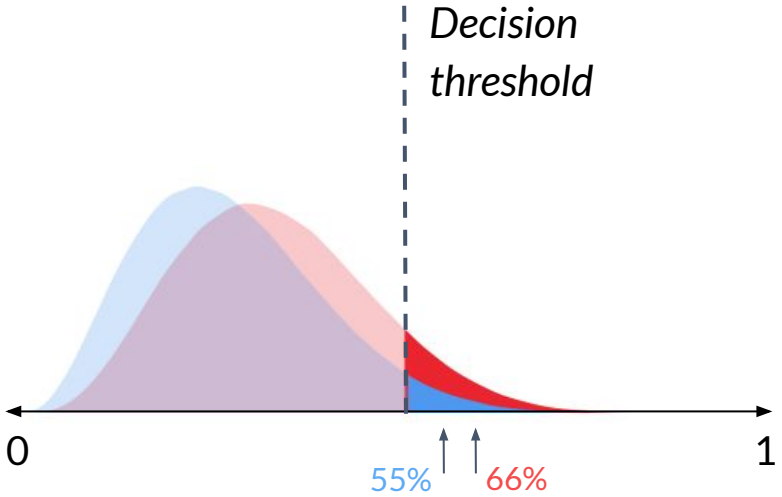
Two different subgroups are classified differently by the classifier

A group-level measure of fairness



Decisions for red group are more successful → false positive rate higher for blue group

A group-level measure of fairness



Group-level fairness measures

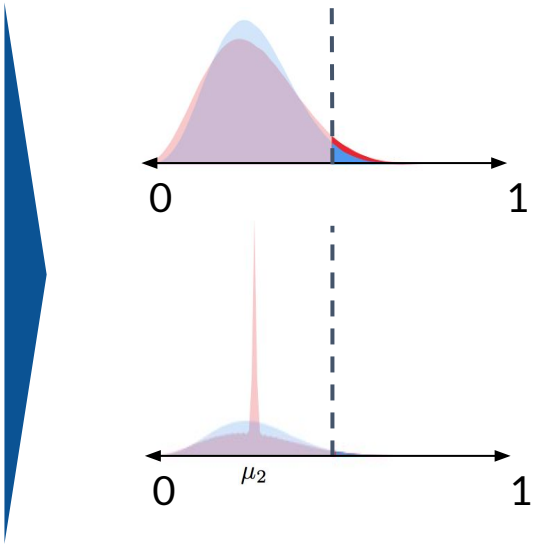
1. Equal false positive or false negative rate
   → "equal opportunity"
2. Equal error rates
   → "overall accuracy equality"
3. …

Decisions for red group are more successful → false positive rate higher for blue group

The problem with group-level fairness

Randomization

Prob =
- Prob = Prob IF 
- Prob = Mean IF 

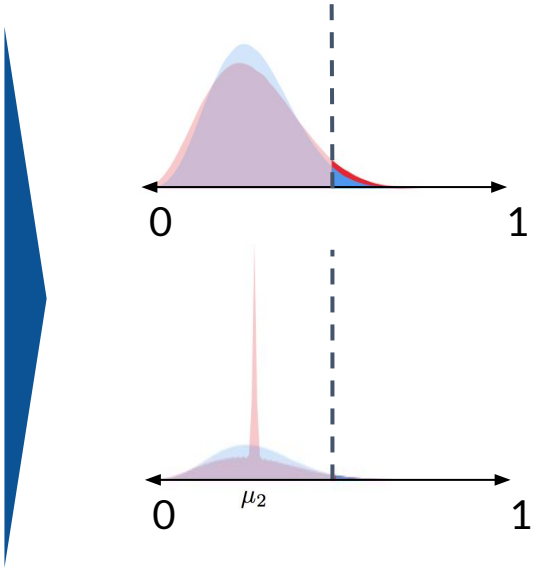


Pleiss et al 2017, Hardt et al 2016

The problem with group-level fairness

Randomization

Prob =
- Prob = Prob IF 
- Prob = Mean IF 



Issues

- Information inefficiency
- Pareto sub-optimality
- Intra-group unfairness

Pleiss et al 2017, Hardt et al 2016

Group versus individual notions of fairness

Group-level fairness

Some statistic of a classifier should be equalized across protected subgroups.

Easy to measure and understand but fails to provide guarantees to individual.

Individual-level fairness

Similar individuals should be treated similarly.

Or

Outcomes should be equally distributed.

Provides guarantees to individual, but difficult to enforce in practice and often at odds with group-level fairness.

Group versus individual notions of fairness

Group-level fairness

Some statistic of a classifier should be equalized across protected subgroups.

Easy to measure and understand but fails to provide guarantees to individual.

Individual-level fairness

Similar individuals should be treated similarly.
Or
Outcomes should be equally distributed.

Provides guarantees to individual, but difficult to enforce in practice and often at odds with group-level fairness.

How can we combine the best of both worlds?

Multicalibration (Herbert-Johnson et al 2018), Generalized Entropy Indices (Speicher et al 2018), Rich subgroup fairness (Kearns et al 2018), Average individual fairness (Kearns et al 2019)

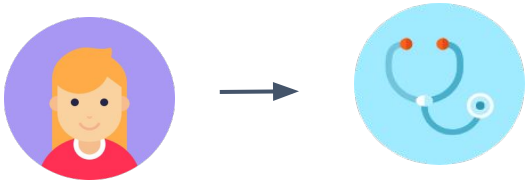
Prediction-time active feature-value acquisition (AFA)

Patient

t = 0

The goal of is to maximize accuracy while minimizing the cost spent on features.

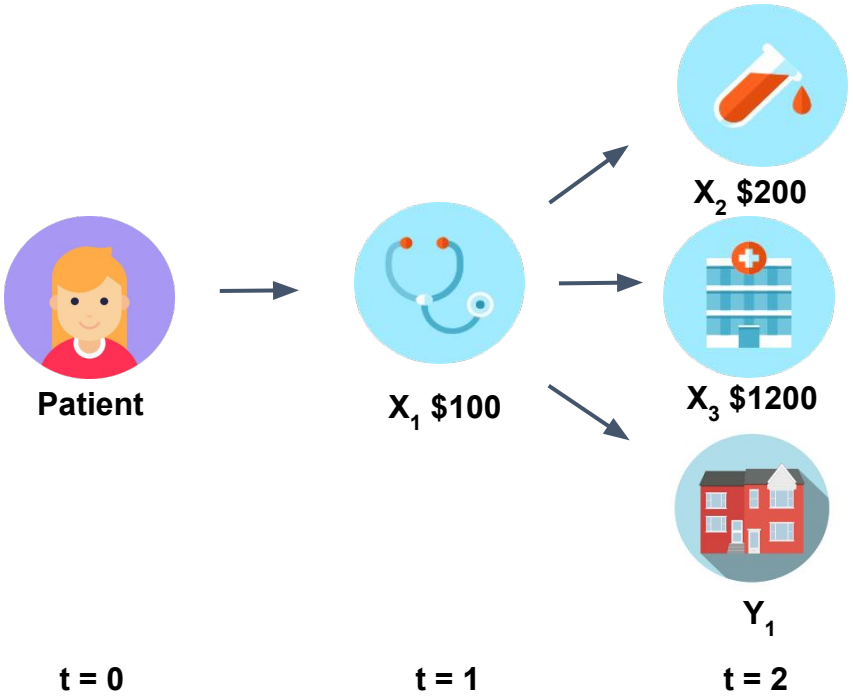
Prediction-time active feature-value acquisition (AFA)



t = 0 t = 1

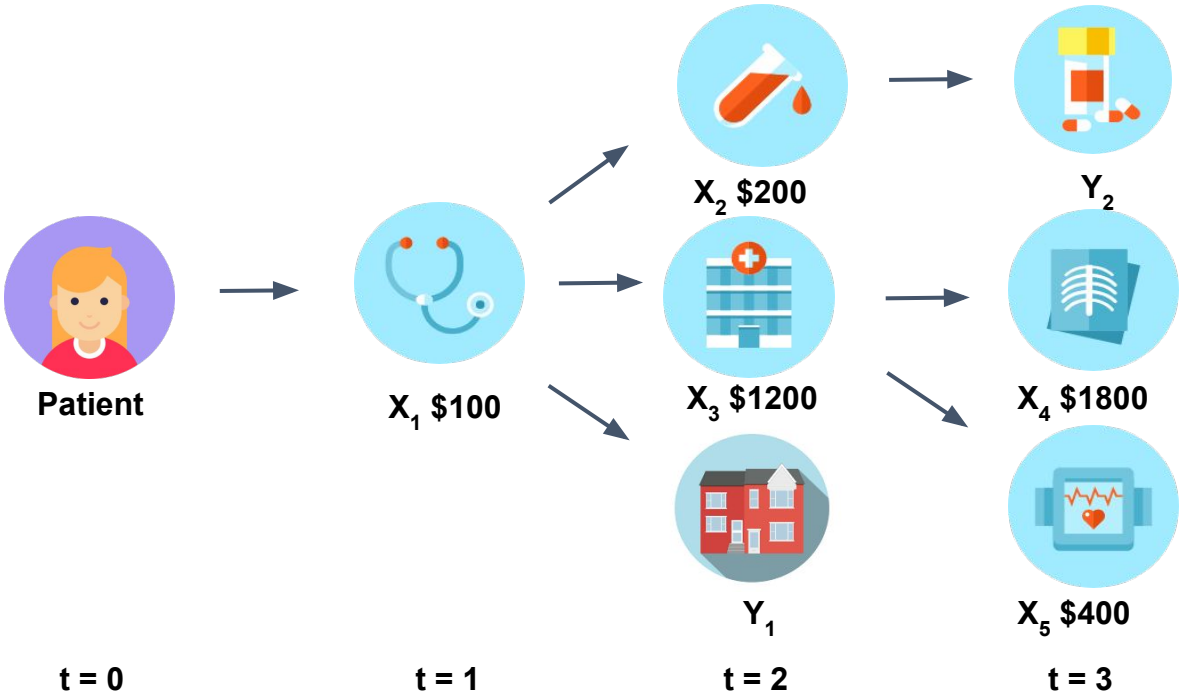
The goal of is to maximize accuracy while minimizing the cost spent on features.

Prediction-time active feature-value acquisition (AFA)



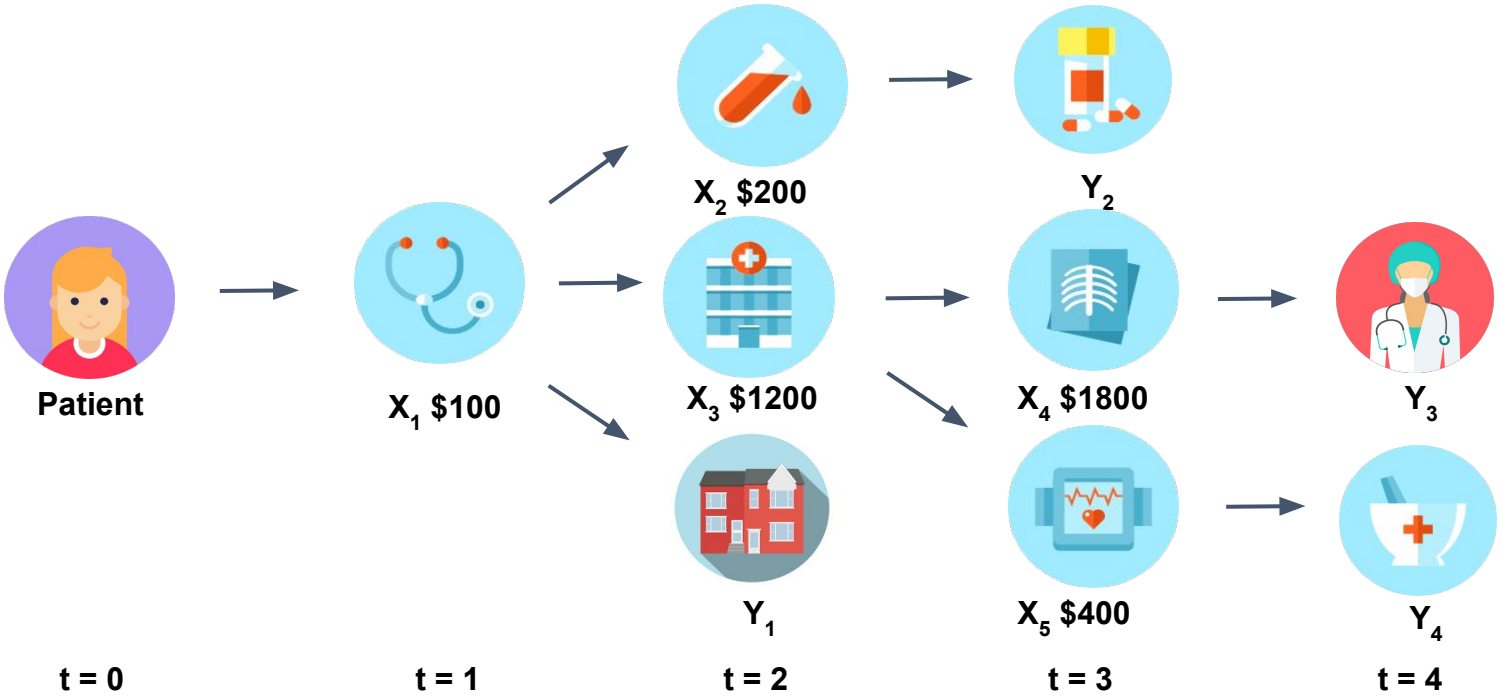
The goal of is to maximize accuracy while minimizing the cost spent on features.

Prediction-time active feature-value acquisition (AFA)



The goal of is to maximize accuracy while minimizing the cost spent on features.

Prediction-time active feature-value acquisition (AFA)

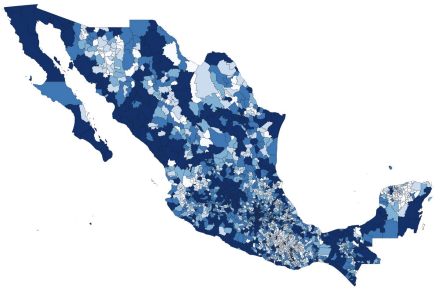


The goal of is to maximize accuracy while minimizing the cost spent on features.

AFA systems are everywhere

Recruiting

(Micro-) credit assessment



Poverty prediction

Not only in the medical domain but in many domains active feature acquisition is relevant

Three AFA components

1. **Classifier that can handle partial feature sets**

2. **Acquisition strategy– which unselected features to select?**

3. **Stopping criterion–when to stop selecting features and classify?**
   Confidence thresholds

We focus on achieving fairness by finding the right time to stop for each group

Three AFA components

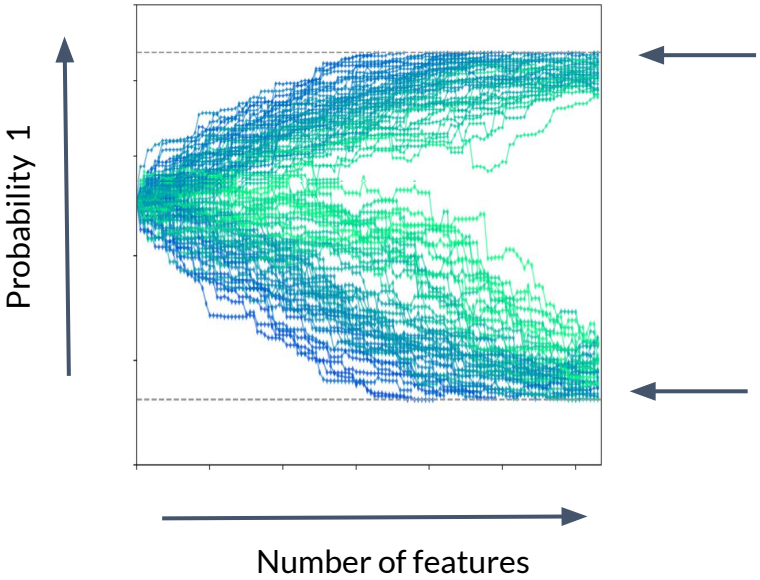
1. **Classifier that can handle partial feature sets**

2. **Acquisition strategy– which unselected features to select?**

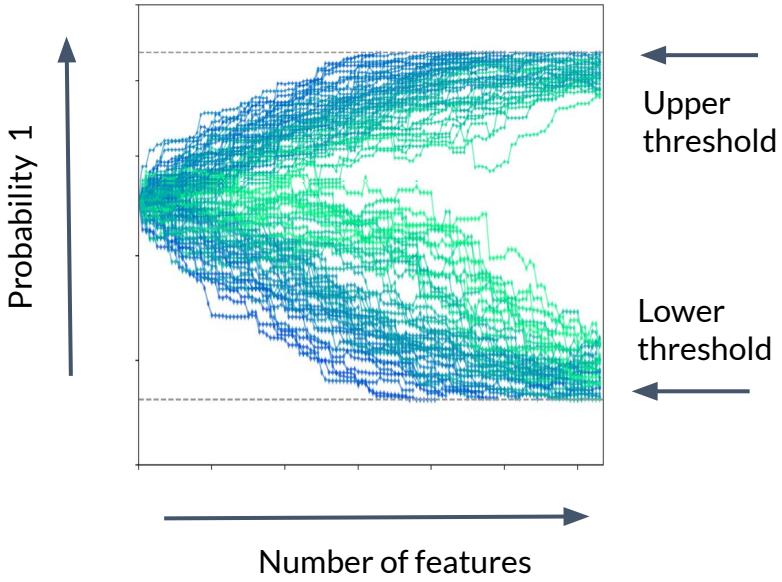
3. **Stopping criterion–when to stop selecting features and classify?**
   Confidence thresholds

We focus on achieving fairness by finding the right time to stop for each group

Confidence thresholds for fair decision making



Confidence thresholds for fair decision making



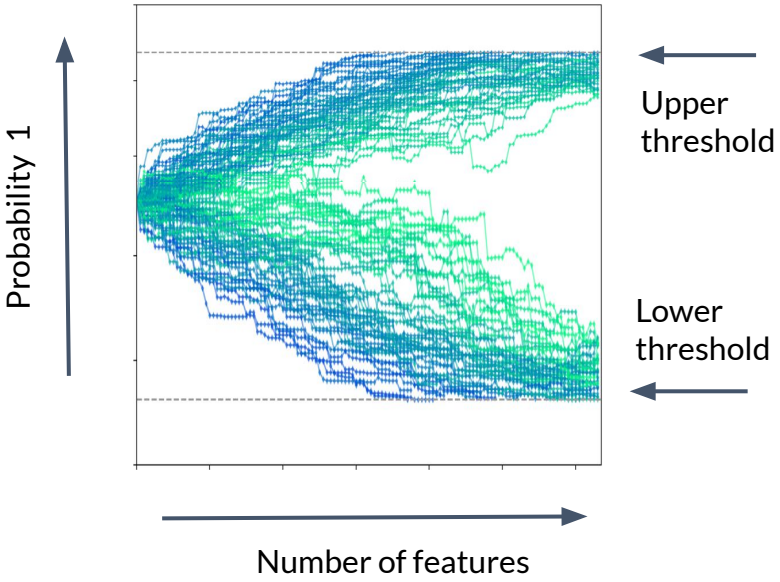
Intuitively, for fairness in this setting, we should have equal quality decision making for every individual.

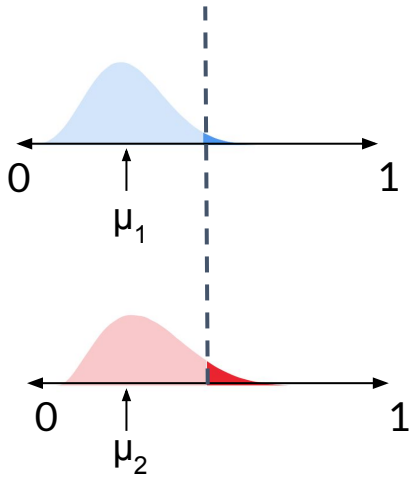
↓

Equal expected error rate for every individual.

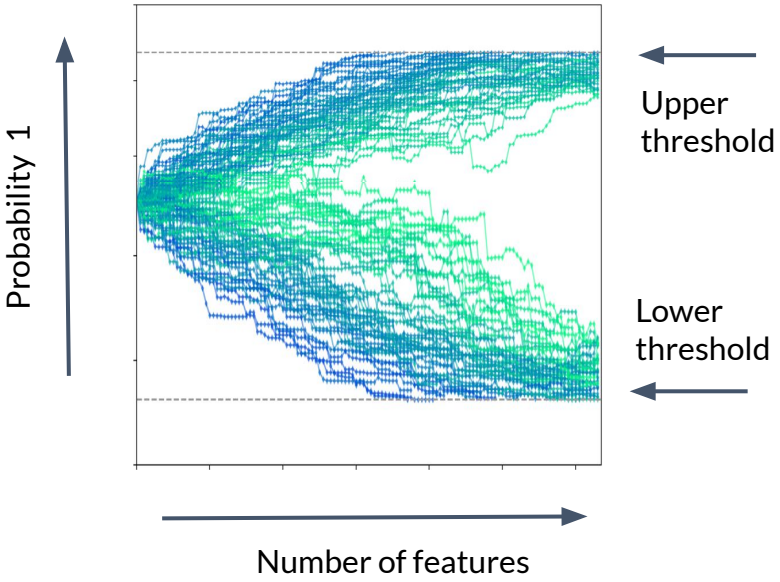
Intuitively, we require an equal expected error rate for every individual

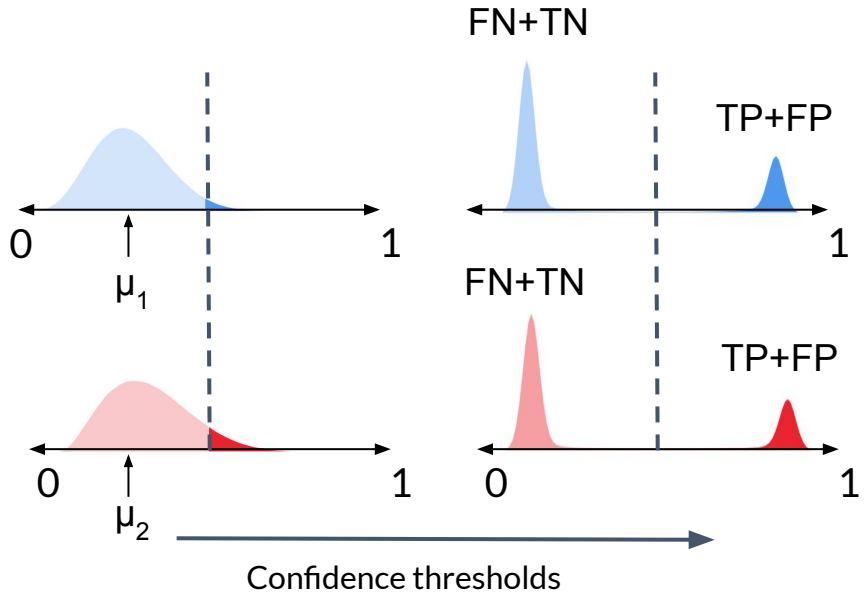
Confidence thresholds for fair decision making





Confidence thresholds for fair decision making

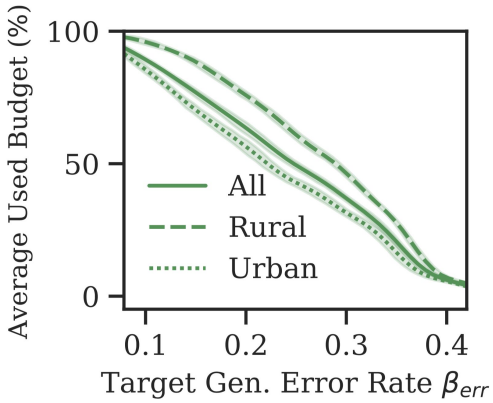


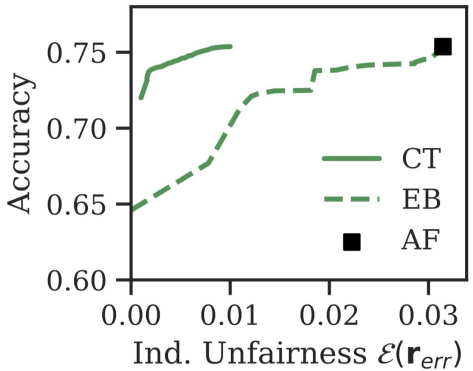


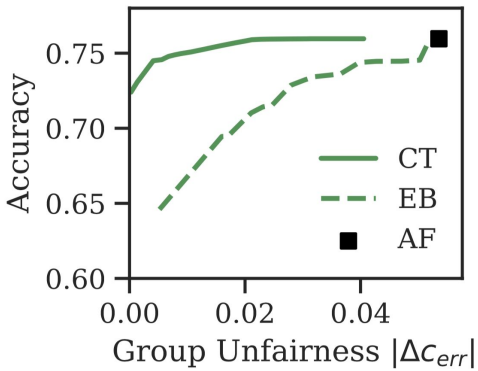
We can select the confidence thresholds to obtain **equal accuracy** or **equal opportunity** across groups and individuals.

Error disparity (Mexican Poverty)

We measure group unfairness using the **absolute difference in error rates across group**s and the individual unfairness by computing the inequality of expected error rates across individuals.





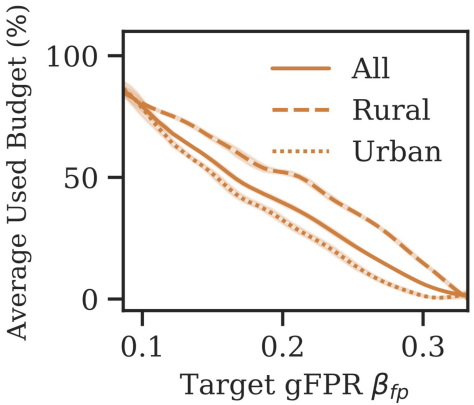


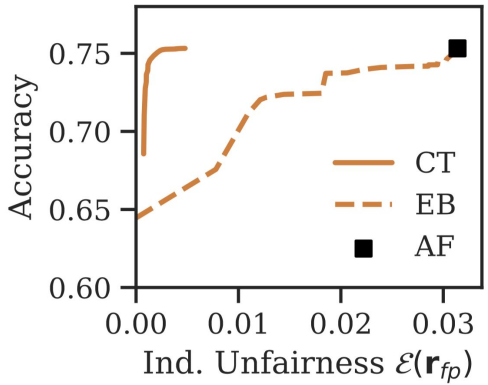
Confidence thresholds simultaneously **mitigate group and individual-level unfairness** by allocating more budget to individuals for which the classifier faces most uncertainty,

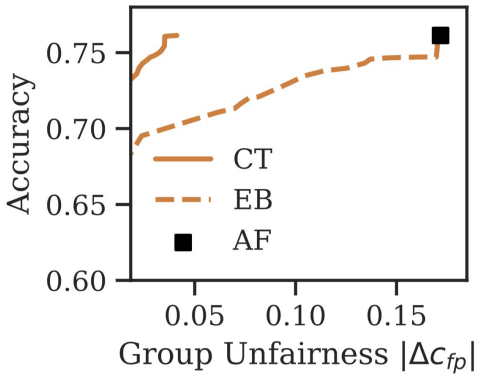
Hardt et al 2016, Speicher et al 2018

FPR disparity (Mexican Poverty)

We measure group unfairness using the **absolute difference in false positive rates** and individual unfairness by computing the inequality of expected false positive rates across individuals.







Confidence thresholds simultaneously **mitigate group and individual-level unfairness** by allocating more budget to individuals for which the classifier faces most uncertainty,

Hardt et al 2016, Speicher et al 2018

Conclusions And Future Work

Confidence thresholds **mitigate group disparities** by acquiring more information for those individuals for which the classifier faces most uncertainty. Effect for achieving **equal opportunity** and **equal accuracy**.

Conclusions And Future Work

Confidence thresholds **mitigate group disparities** by acquiring more information for those individuals for which the classifier faces most uncertainty. Effect for achieving **equal opportunity** and **equal accuracy**.

We achieve mitigate both **group and individual unfairness** with respect to a set of predefined subgroups and, in the case of equal accuracy, for any arbitrary subgroup.

Conclusions And Future Work

Confidence thresholds **mitigate group disparities** by acquiring more information for those individuals for which the classifier faces most uncertainty. Effect for achieving **equal opportunity** and **equal accuracy**.

We achieve mitigate both **group and individual unfairness** with respect to a set of predefined subgroups and, in the case of equal accuracy, for any arbitrary subgroup.

The framework has interesting implications on privacy as it automatically satisfies the GDPR **"data minimization"** principle: for each individual we only collects the minimum set of features needed for an accurate prediction.

Conclusions And Future Work

Confidence thresholds **mitigate group disparities** by acquiring more information for those individuals for which the classifier faces most uncertainty. Effect for achieving **equal opportunity** and **equal accuracy**.

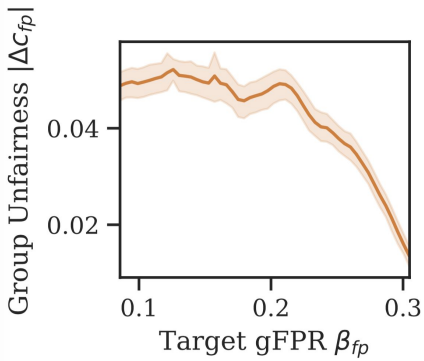
We achieve mitigate both **group and individual unfairness** with respect to a set of predefined subgroups and, in the case of equal accuracy, for any arbitrary subgroup.

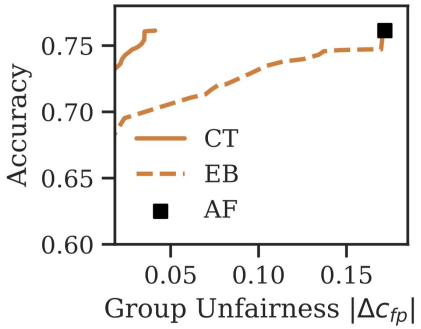
The framework has interesting implications on privacy as it automatically satisfies the GDPR **"data minimization"** principle: for each individual we only collects the minimum set of features needed for an accurate prediction.

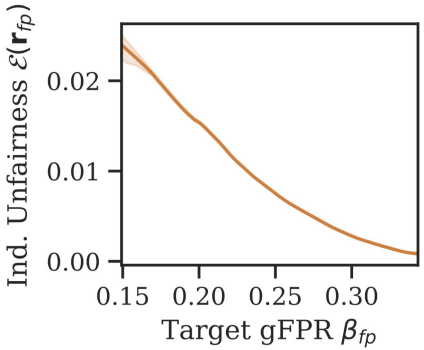
We encourage future work on **datasets** and **applications** that are more natural in an active feature-value setting.

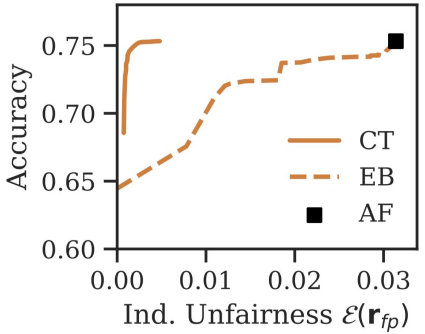
Backup

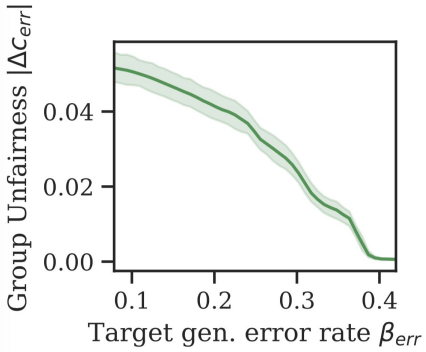
Implementation - probabilistic RF

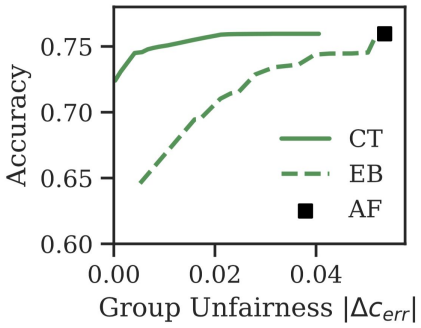


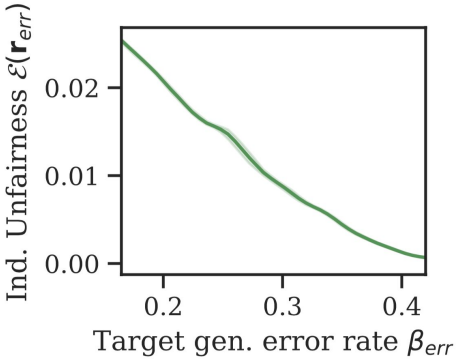


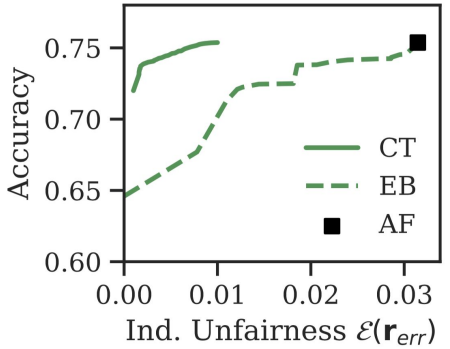




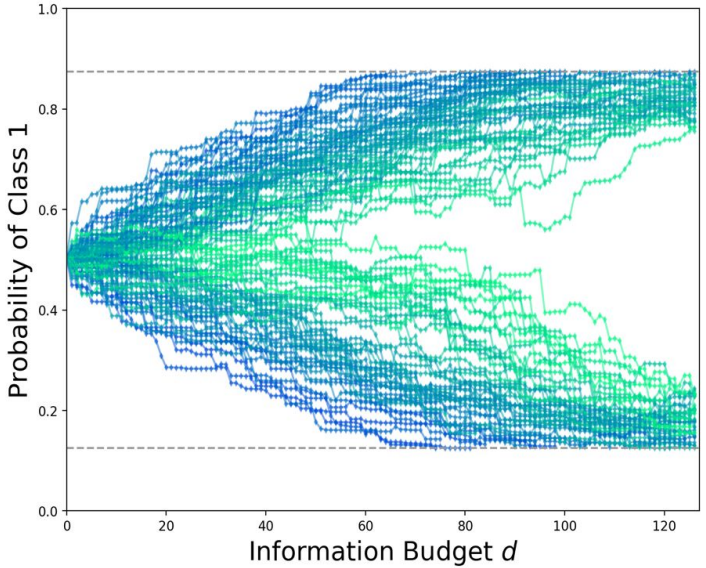




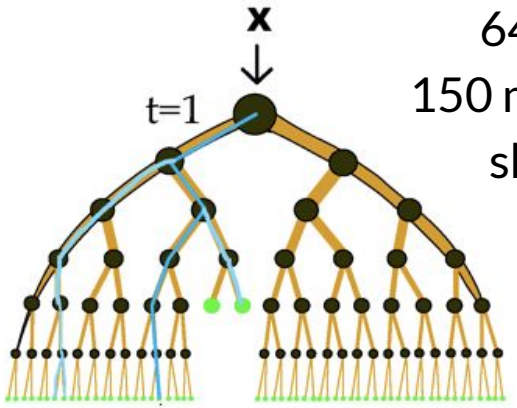




Implementation - probabilistic RF



Implementation



64 trees
150 max depth
sklearn

Probabilistic random forests

For each individual, features are added one-by-one in **until a confidence threshold is reached**.
This leads to an **individual-level budget** that changes according to the needs of each individual.

Implementation - feature acquisition strategy

Goal: selecting features that lead to high accuracy while keeping costs low. One way is through a heuristic that queries the features that maximizes an expected utility

$$EU(x_j) = \int_v P(x_j = v)\frac{U(x_j = v)}{c_j} \longrightarrow j' = \underset{j:j\notin Q_t, j\in[0,d]}{\arg\max} EU(x_j)$$

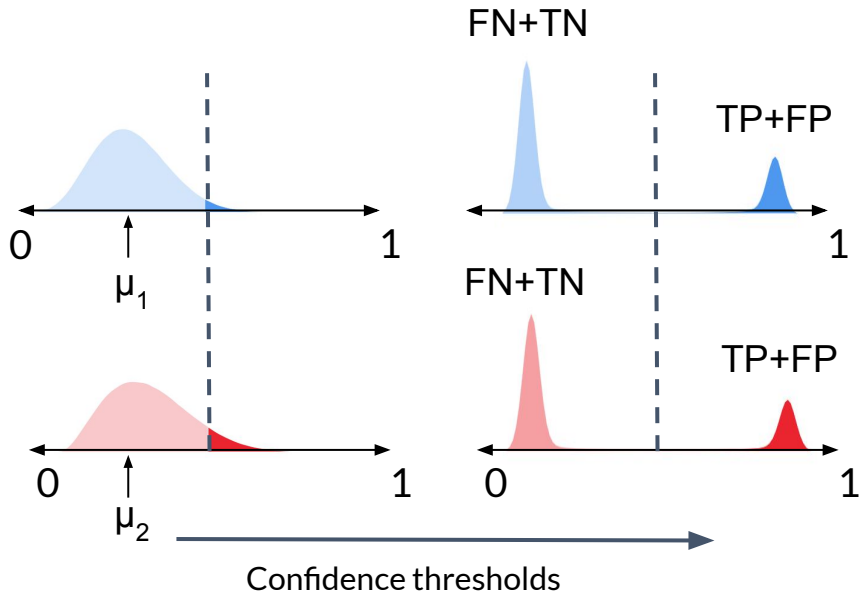
$$j' = \underset{j:j\notin \mathcal{O}_t}{\arg\max}\frac{1}{c_j}\sum_v P(x_j = v)\underbrace{|P(y = 1|\mathcal{O}_t \cup \{x_j = v\}) - P(y = 1|\mathcal{O}_t)|}_{\text{Utility } U(x_j=v)}$$

More info: [Kanani & Melville NIPS 2008]

Datasets

Dataset						Subgroup$_1$			Subgroup$_0$		
Name	$N_{samples}$	N_{feat}	Acc	AUC	μ	Label$_1$	n$_1$	μ_1	Label$_0$	n$_0$	μ_0
Mexican poverty [12, 22]	70,305	182	78.7%	0.856	35.5%	Urban	63.6%	34.9%	Rural	36.4%	36.6%
Adult income [18]	49,000	14	86.3%	0.911	23.9%	White	85.4%	25.4%	Non-white	14.6%	15.3%

Confidence thresholds with equal base rates



Equal base rates
$\mu_1 = \mu_2$

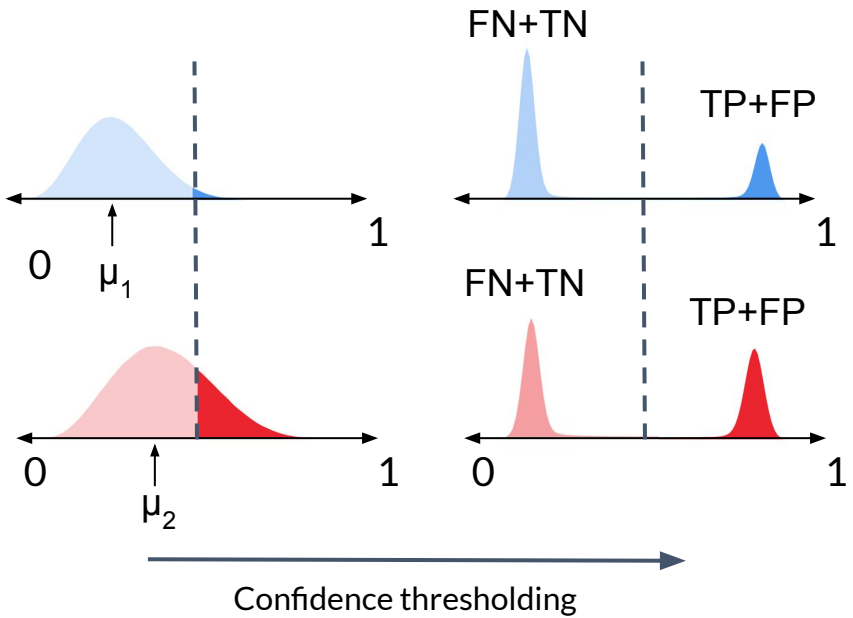
Equal Opportunity
$FNR_1 = FNR_2$

Equal Odds
$FPR_1 = FPR_2$ & $FNR_1 = FNR_2$

Equal Accuracy
$FPR_1 + FNR_1 = FPR_2 + FNR_2$

For equal base rates groups, confidence thresholds lead to **equal accuracy and equal odds**

Confidence thresholds with unequal base rates



Unequal base rates
$\mu_1 \neq \mu_2$

No Equal Opportunity
$FNR_1 \neq FNR_2$

No Equal Odds
$FPR_1 \neq FPR_2$ & $FNR_1 \neq FNR_2$

Equal Accuracy
$FPR_1+FNR_1 = FPR_2+FNR_2$

For different base rates, confidence thresholds ensure **Equal Accuracy**

Fair stopping criteria with different base rates

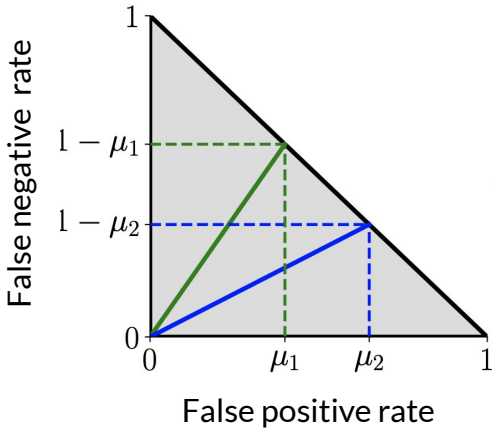
μ_1 - base rate for subgroup 1

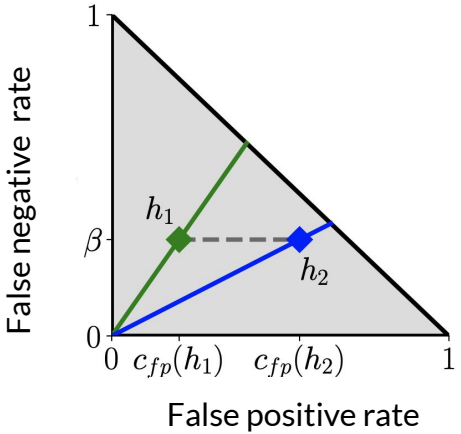
μ_2 - base rate for subgroup 2

h_1 - classifier for subgroup 1

h_2 - classifier for subgroup 2

β - the target false negative rate





$$\alpha_l = \frac{1}{2} - \frac{1}{2}\sqrt{1-4\beta\mu_t}$$

Fair lower threshold for group t

$$\alpha_u = \frac{1}{2} + \frac{1}{2}\sqrt{1-4\beta\mu_t}$$

Fair upper threshold for group t

For different base rates, one can only achieve calibration and equal opportunity