

Falcon: Fair Active Learning using Multi-armed Bandits

Ki Hyun Tae¹, Hantian Zhang², **Jaeyoung Park**¹, Kexin Rong², Steven Euijong Whang¹

¹KAIST, ²Georgia Institute of Technology,



Fairness in Machine Learning: Do Not Discriminate

- ML models must be developed responsibly
- We focus on fairness

A.I. Bias Caused 80% Of Black Mortgage Applicants To Be Denied

Kori Hale Contributor

I'm the CultureBanx CEO, redefining business news for hip-hop culture



Updated Sep 3, 2021, 09:35am EDT

Forbes, 2021

Bias in the Mortgage Approval Process

Discrimination and inequalities persist

By DANIEL THOMAS MOLLENKAMP Updated August 18, 2022

Reviewed by EBONY HOWARD

Fact checked by SUZANNE KVILHAUG

Follow

Investopedia, 2022

Sep 2, 2021, 07:48am EDT

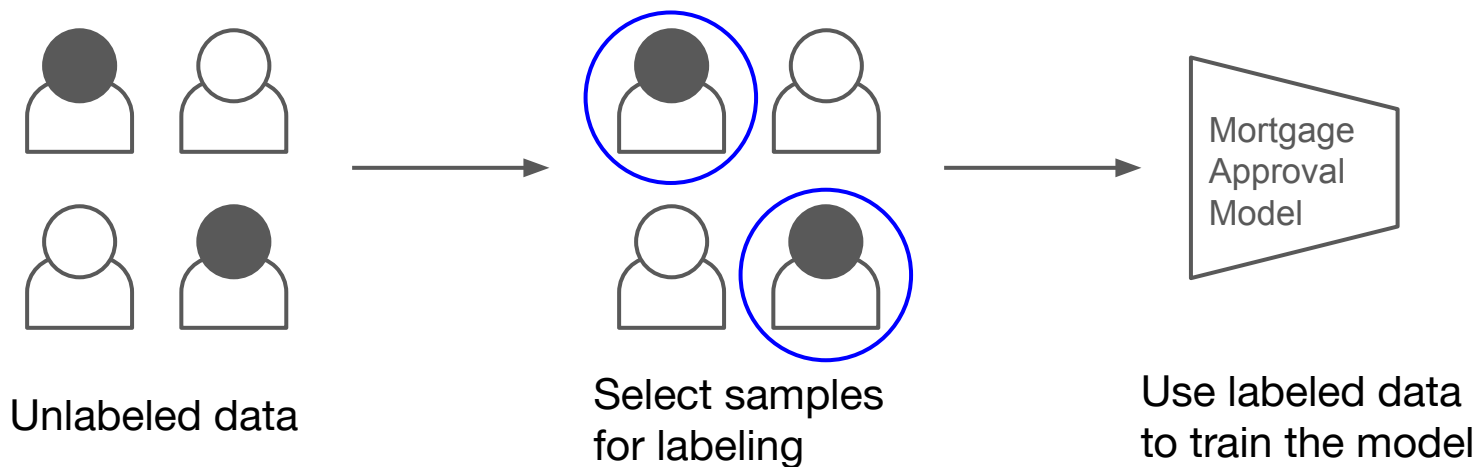
Traditional Active Learning

- Select samples from **unlabeled data** for labeling to maximize **accuracy**
- Minimize the data to label as labeling involves expensive human resources



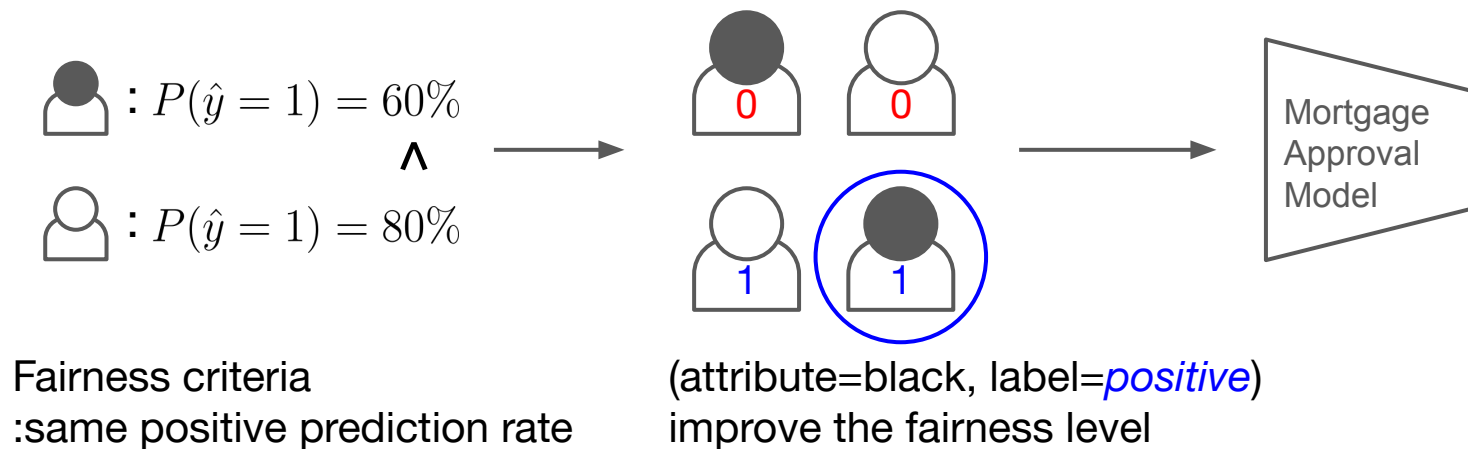
Active Learning for Fairness

- Key idea: we would like to label the samples that lead to better **fairness**
 - Targeted labeling can improve fairness



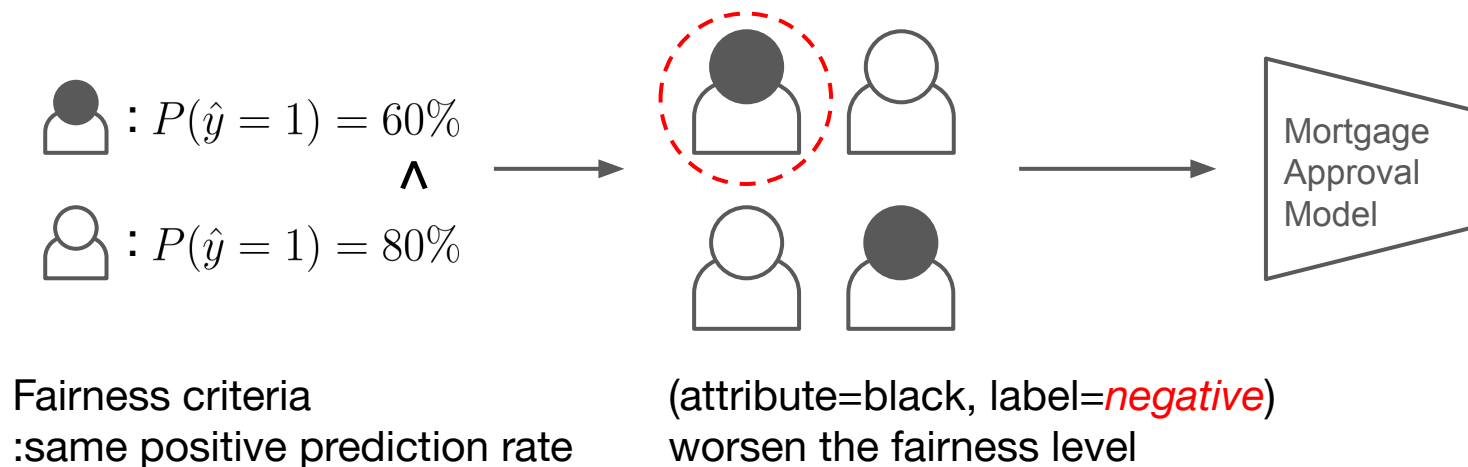
Main Challenge: Lack of Labels

- In the example below, we need to obtain samples with (attribute=*black*, label=*positive*).



Challenges in Fair Active Learning

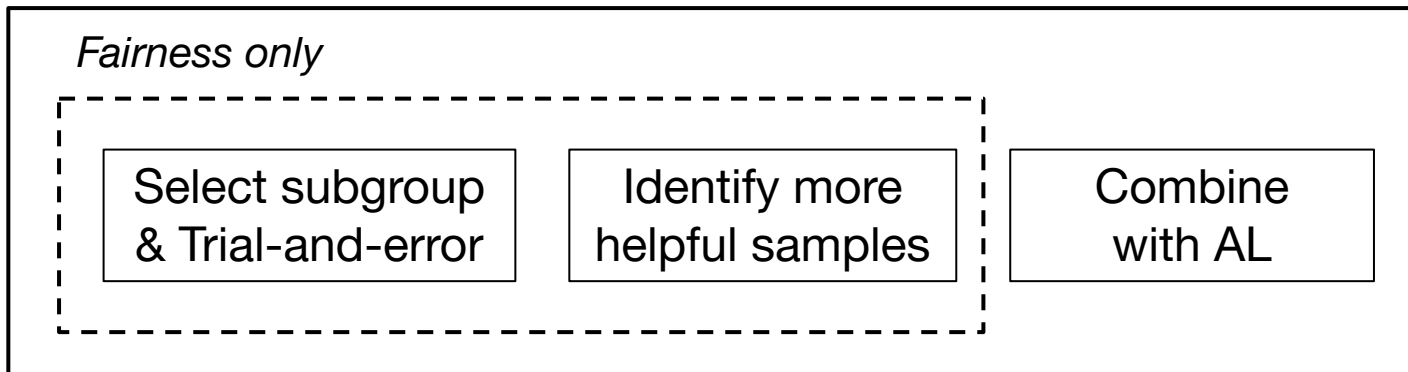
- Labeling the **wrong samples** with (attribute=*black*, **label=*negative***) decreases the positive prediction rate of black group and thus **worsens the fairness level**



Falcon

- Select subgroups to label (e.g., (attribute=*black*, label=*positive*)) & use a trial-and-error method to manage unknown ground-truth labels
- Identify more helpful samples from the subgroup using adversarial MABs
- Balances fairness and accuracy by alternating its selection with traditional AL

Falcon



Subgroup Labeling for Fairness

- Key strategy: increase the labeling of specific subgroups
 - Subgroup is defined using sensitive attributes and labels, e.g., (attribute=*black*, label=*positive*)

Step 1

Select subgroup
& Trial-and-error

Identify more
helpful samples

Combine
with AL

Subgroup Labeling for Fairness

- Key strategy: increase the labeling of specific subgroups
 - Subgroup is defined using sensitive attributes and labels, e.g., (attribute=*black*, label=*positive*)

Demographic Parity (DP): Similar positive prediction rate across sensitive groups

$$\text{●} : p(\hat{y} = 1) = 60\% < \text{○} : p(\hat{y} = 1) = 80\%$$

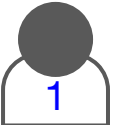
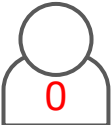
The goal is to close this gap

Subgroup Labeling for Fairness

- Key strategy: increase the labeling of specific subgroups
 - Subgroup is defined using sensitive attributes and labels, e.g., (attribute=*black*, label=*positive*)

Demographic Parity (DP): Similar positive prediction rate across sensitive groups

$$\text{Black icon} : p(\hat{y} = 1) = 60\% < \text{White icon} : p(\hat{y} = 1) = 80\%$$

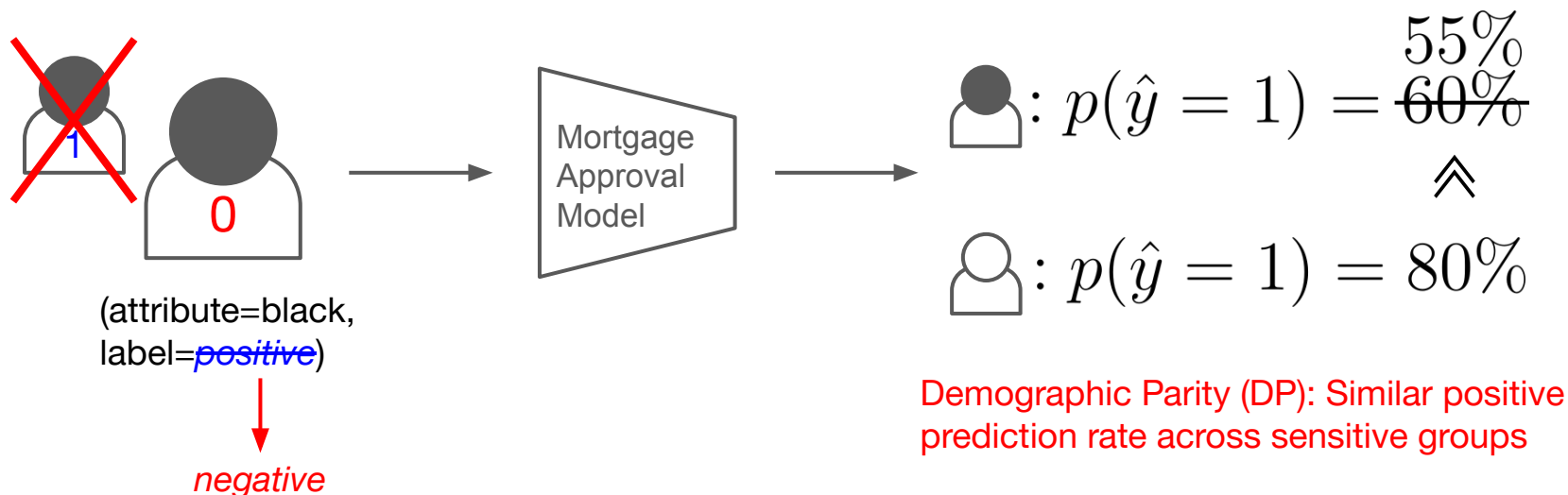
⇒ More  or  are needed

(attribute=*black*,
label=*positive*)

(attribute=*white*,
label=*negative*)

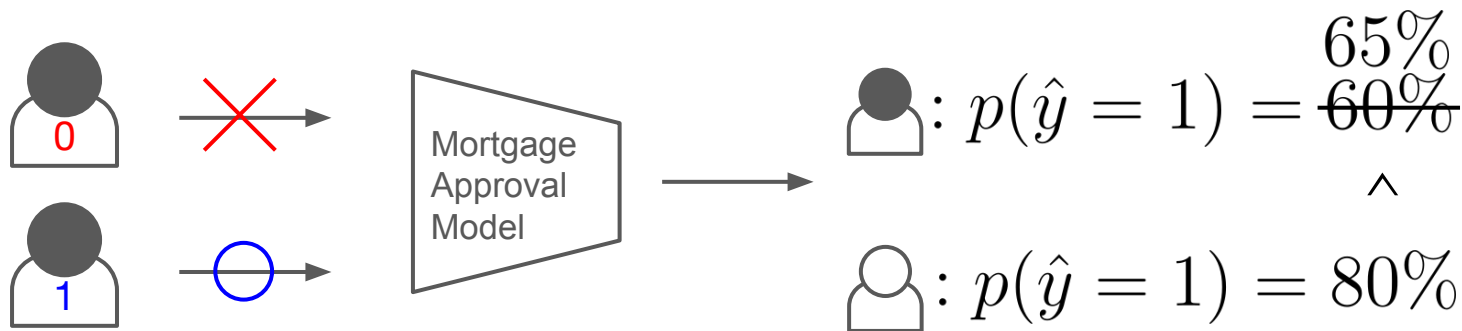
Unknown Ground Truth Labels

- However, ground truth labels are not available in an AL setting
- Adding samples with undesired labels can negatively affect fairness



Solution: Trial-and-error Strategy

- Select samples in the target sensitive group to label, but postpone using them in model training when they turn out to have undesirable labels
- Postponing undesired samples avoids worsening fairness



Demographic Parity (DP): Similar positive prediction rate across sensitive groups

Identify More Helpful Samples for Fairness

- Improve basic trial-and-error by choosing samples that are more likely to increase a target group's accuracy while also having the desired label

Step 2

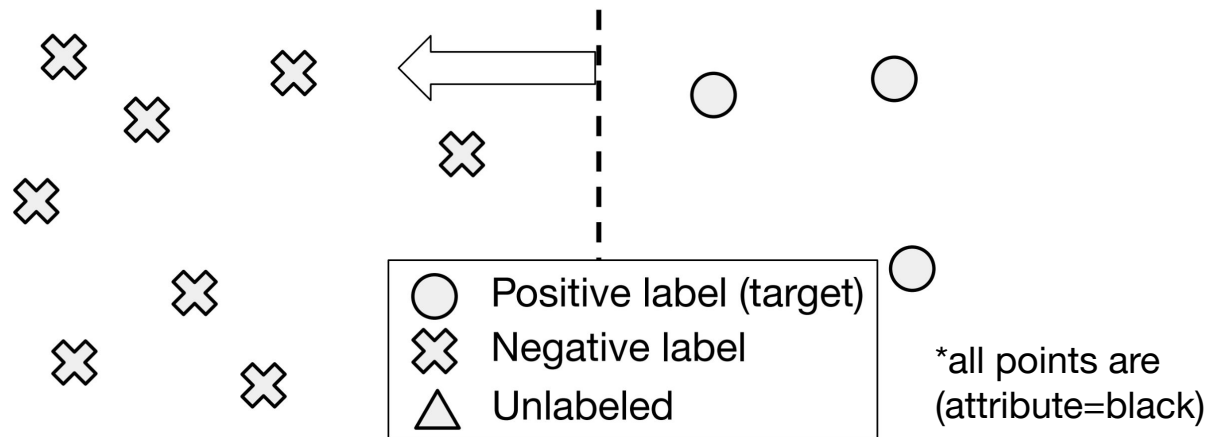
Select subgroup
& Trial-and-error

Identify more
helpful samples

Combine
with AL

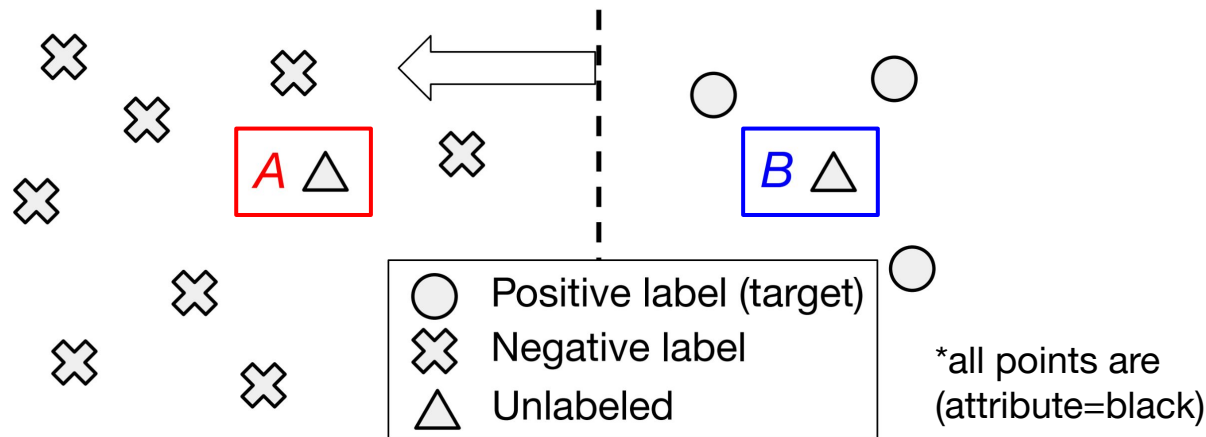
Trade-off b/w Informativeness and Postpone Rate

- Key observation: the more informative a sample is for improving the target group's positive prediction rate, the less likely it is to have the target label
 - Suppose the target subgroup is (attribute=black, label=*positive*)
 - To increase the positive prediction rate, the decision boundary must be shifted to the left



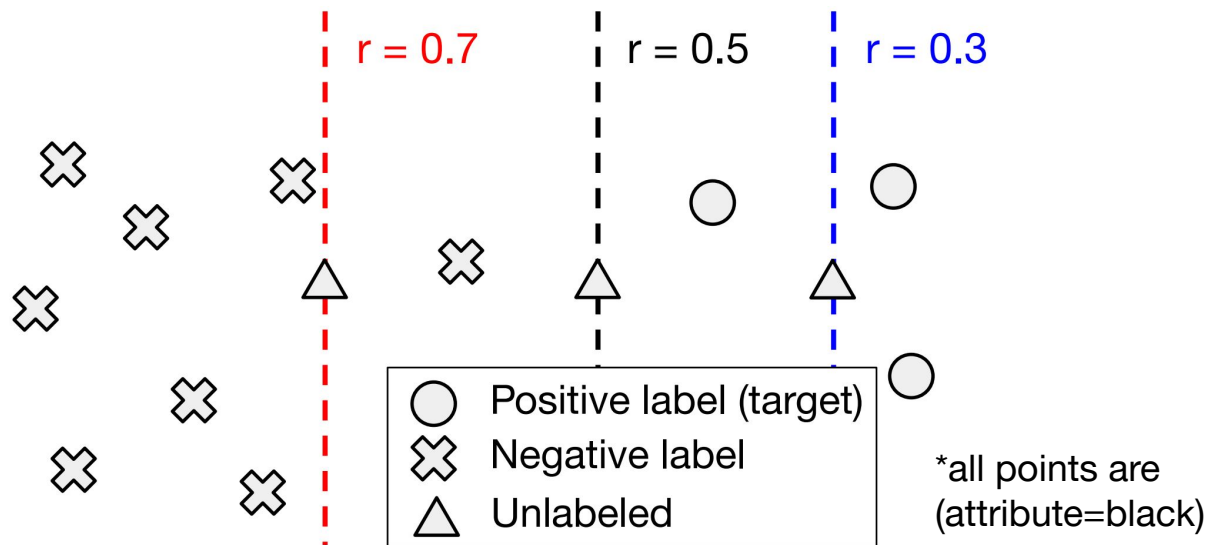
Trade-off b/w Informativeness and Postpone Rate

- Key observation: the more informative a sample is for improving the target group's positive prediction rate, the less likely it is to have the target label
 - Suppose the target subgroup is (attribute=black, label=*positive*)
 - To increase the positive prediction rate, the decision boundary must be shifted to the left
 - Sample **A** compared to **B**: better target group PPR, but lower chance of positive label



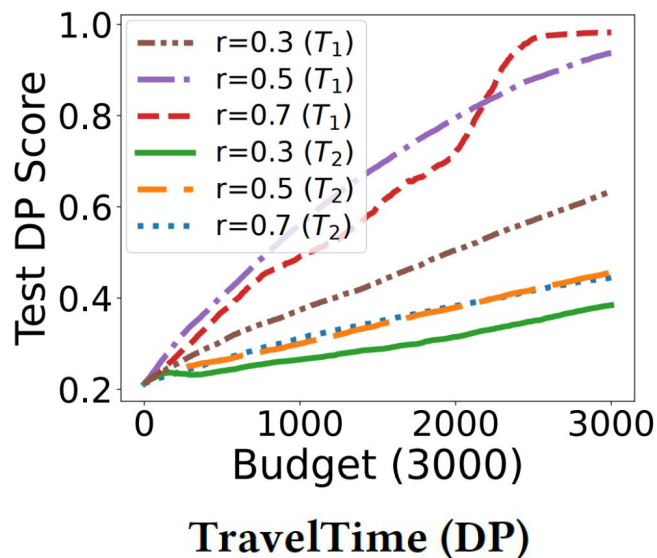
Policy: Amount of Risk Taking

- The more “risk” we are willing to take for finding an informative sample, the less likely it has the desired label
- We capture this risk taking as a policy “ r ” = c for each target group
 - Select a sample whose predicted probability for the target label closest to $(1 - c)$



Challenge: Optimal Policy Changes Over Time

- The optimal policy varies as we label more samples
- Need an adaptive algorithm to identify the optimal policy



Multi-armed Bandit (MAB) for Policy Search

- Arm: Policy
- Reward: Fairness improvement
- Unlike traditional MABs, the rewards do not follow time-invariant distribution
- Thus we use adversarial MABs*

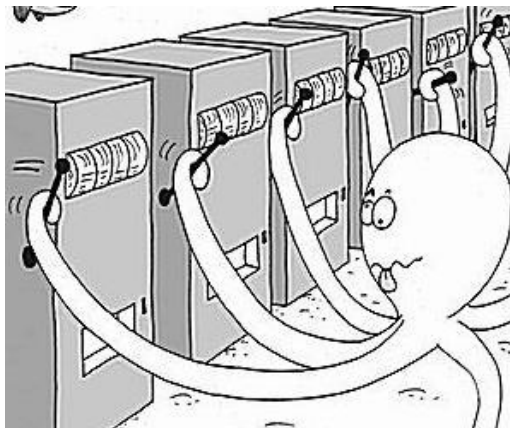
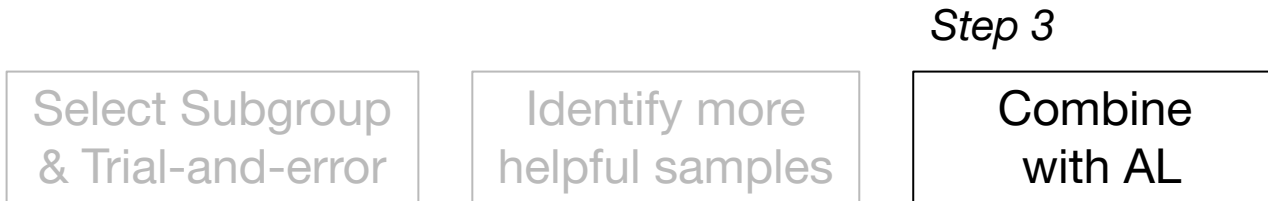


Image source: Microsoft Research

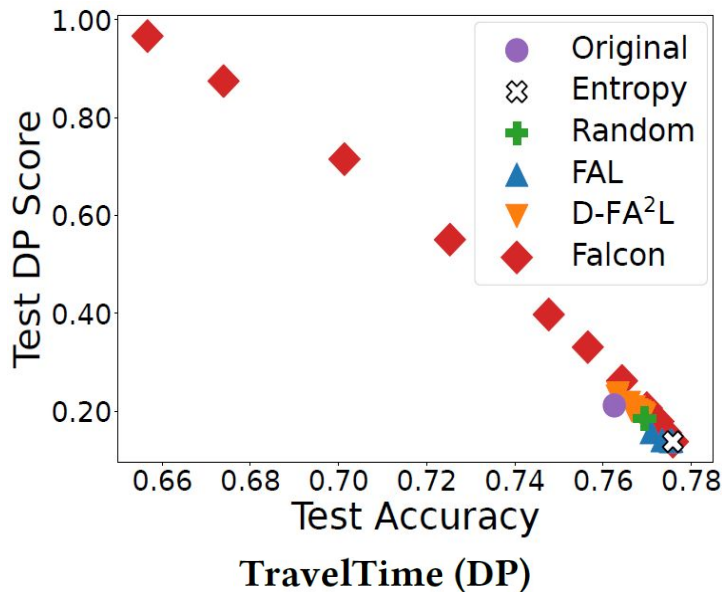
Combine with AL for Accuracy

- Alternates between fair and accurate labeling probabilistically
 - Improves fairness with λ probability and accuracy with $(1 - \lambda)$ probability
 - A higher λ indicates better fairness
- Does not require any modifications of the AL methods



Accuracy and Fairness Results

- Falcon shows the best accuracy and fairness trade-off
 - Also, similar results for other datasets, fairness measures, and ML models



Running Time Results

- Falcon is much faster than the fair AL baselines

Datasets	Avg. Running time (sec)				
	Entropy	Random	FAL	D-FA ² L	FALCON
TravelTime	139	91	1,420	179	126
Employ	114	76	1,411	140	98
Income	244	149	1,965	290	205
COMPAS	6.1	5.5	153	12	5.9

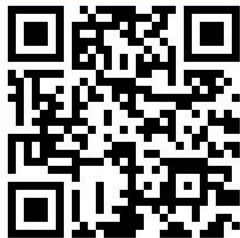
Summary

- Falcon selects samples to label for improving fairness and accuracy
 - Selects subgroups to label and handles unknown ground truth labels using trial-and-error
 - Automatically selects the best sampling policy using adversarial MABs
 - Balances fairness and accuracy by alternating its selection for fairness with traditional AL

Paper



Code



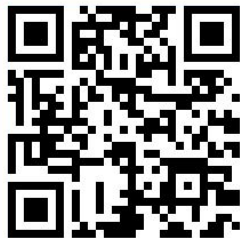
Summary

- Falcon selects samples to label for improving fairness and accuracy
 - Selects subgroups to label and handles unknown ground truth labels using trial-and-error
 - Automatically selects the best sampling policy using adversarial MABs
 - Balances fairness and accuracy by alternating its selection for fairness with traditional AL

Paper



Code



Thanks!