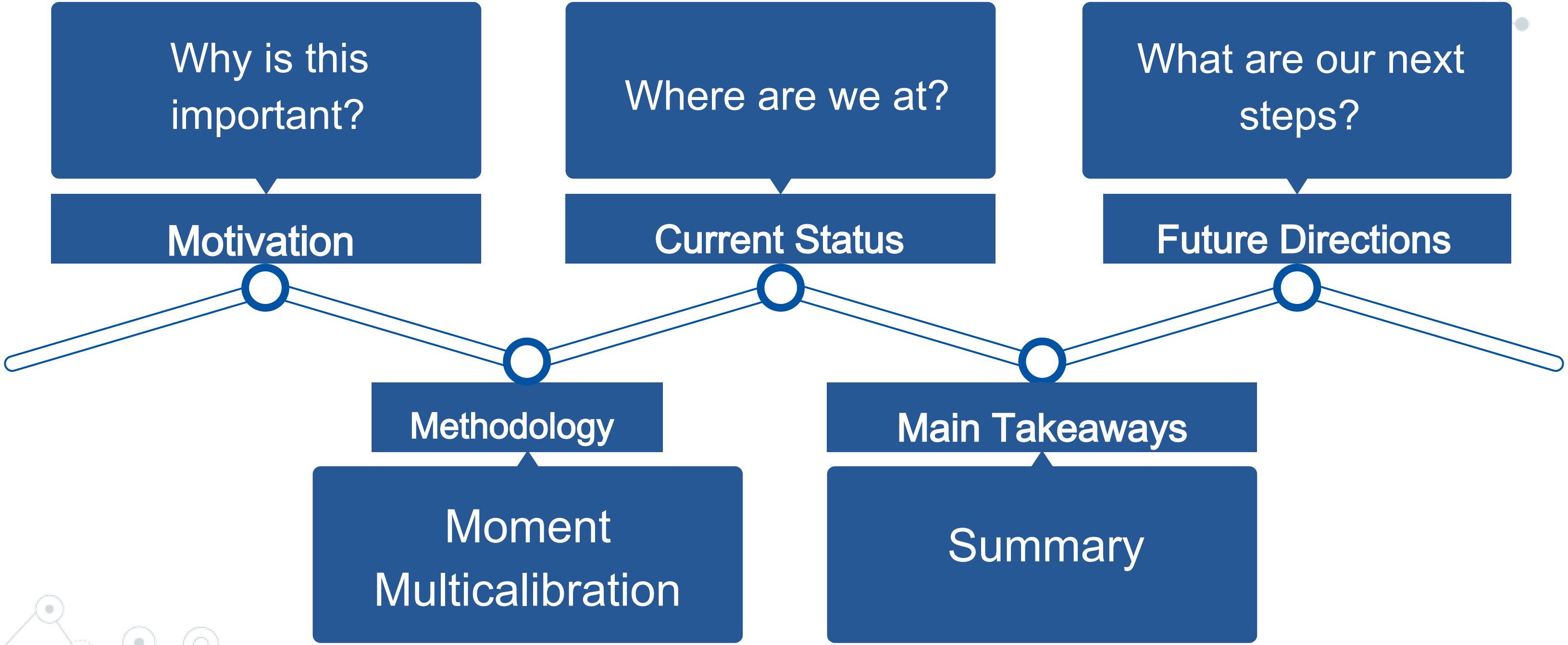
equalmodel

A Post-Processing Algorithm for Bias Reduction in Big Data Analytics

Roadmap



Why is this important?

Motivation

Methodology

Moment Multicalibration

Where are we at?

Current Status

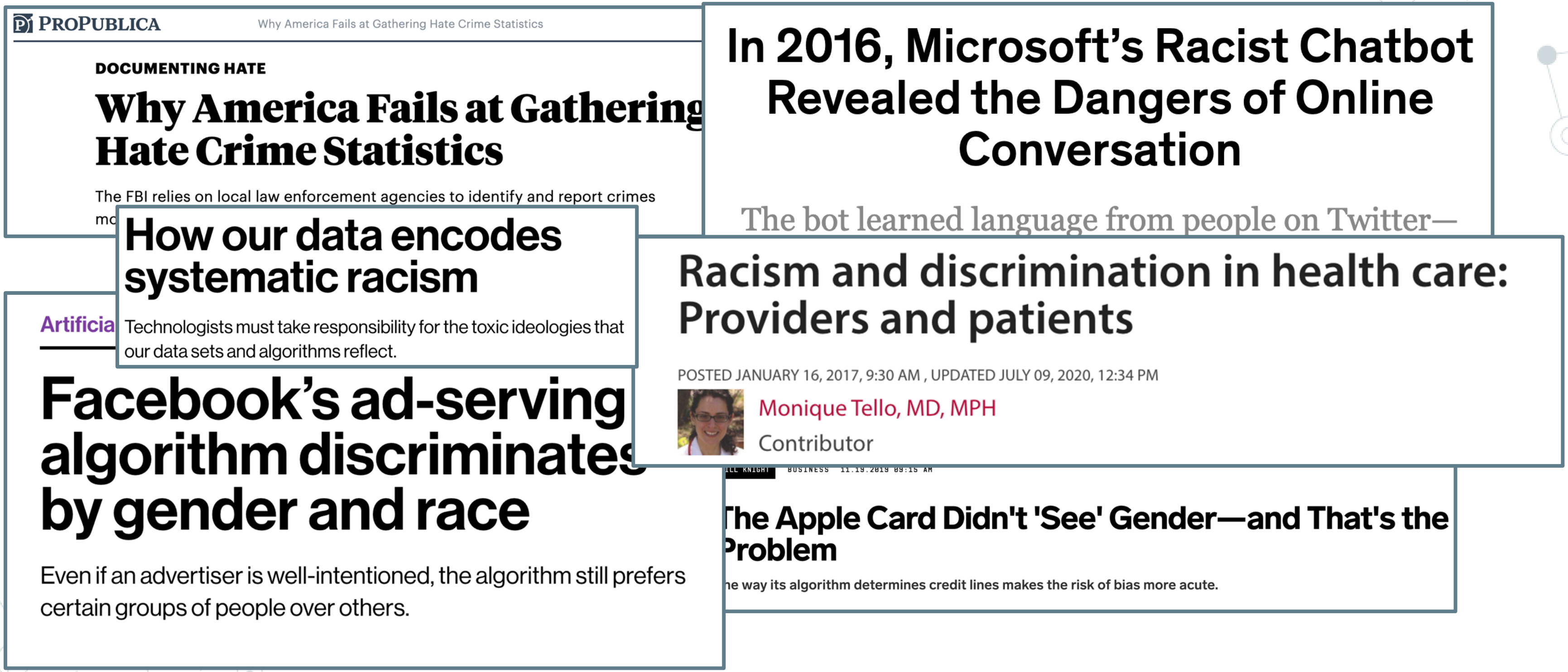
Main Takeaways

Summary

What are our next steps?

Future Directions

Machine Learning is becoming more prevalent but there are **consequences** …



PROPUBLICA Why America Fails at Gathering Hate Crime Statistics

DOCUMENTING HATE

Why America Fails at Gathering Hate Crime Statistics

The FBI relies on local law enforcement agencies to identify and report crimes mo

In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation

The bot learned language from people on Twitter—

How our data encodes systematic racism

Technologists must take responsibility for the toxic ideologies that our data sets and algorithms reflect.

Racism and discrimination in health care: Providers and patients

POSTED JANUARY 16, 2017, 9:30 AM , UPDATED JULY 09, 2020, 12:34 PM

Monique Tello, MD, MPH

Contributor

Artificia

Facebook's ad-serving algorithm discriminates by gender and race

Even if an advertiser is well-intentioned, the algorithm still prefers certain groups of people over others.

ILL KNIGHT BUSINESS 11.19.2019 09:15 AM

he Apple Card Didn't 'See' Gender—and That's the roblem

he way its algorithm determines credit lines makes the risk of bias more acute.

Let’s first ground our discussion...

Given features x, your dosage for a drug is f(x).

How sure are you?

The variance conditional on my estimate is g(x).

But I am part of a demographic representing less than 5% of the population

For Asian Americans under the age of 50, the confidence interval is [a,b]

For women with a family history of diabetes, the confidence interval is [c,d]

Key Observation

The dosage prediction is averaged over the population, *not* an individual, so the dose might not be accurate for an individual.

Problem

Implicit bias against underrepresented populations in the systems we rely on

Ethical standards of both fairness and privacy are breached

Goal

1) Make a tool that can supplement any existing algorithm, making it more fair
2) Reduce implicit bias

Solution: Multicalibration

- Calibration assures that our predictions are accurate overall
  - Fails to make the same guarantee for subpopulations
  - E.g. 90% accuracy for the total population does not guarantee 90% accuracy for a subpopulation
- Multicalibration offers the same assurance across all possible subgroups

Project Components & Resources

Software

PyTorch

Papers

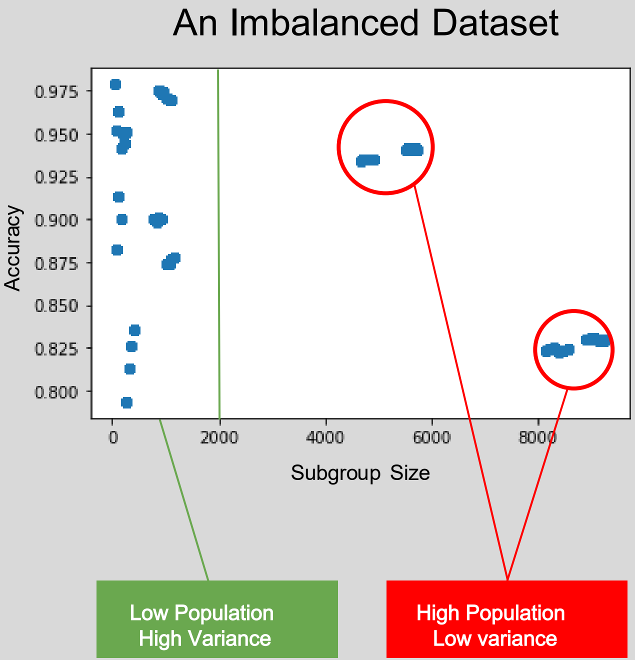
◎ Moment Multicalibration for Uncertainty Estimation (Jung, Lee, Pai, Roth, Vohra)
◎ Multiaccuracy: Black-Box Post-Processing for Fairness in Classification (Kim, Ghorbani, Zou)

Stakeholders

◎ Cary Coglianese, Edward B. Shils Professor of Law and Professor of Political Science
◎ The Defender Association of Philadelphia

The Problem with ML in Criminal Justice

- **Data Description**: Data combines socio-economic data, law enforcement data, and crime data
- **Goal**: predict violent crime number
- **Problem:** (1) High variance in accuracies for underrepresented people. (2) Models not calibrated to underrepresented people will only cause further harm

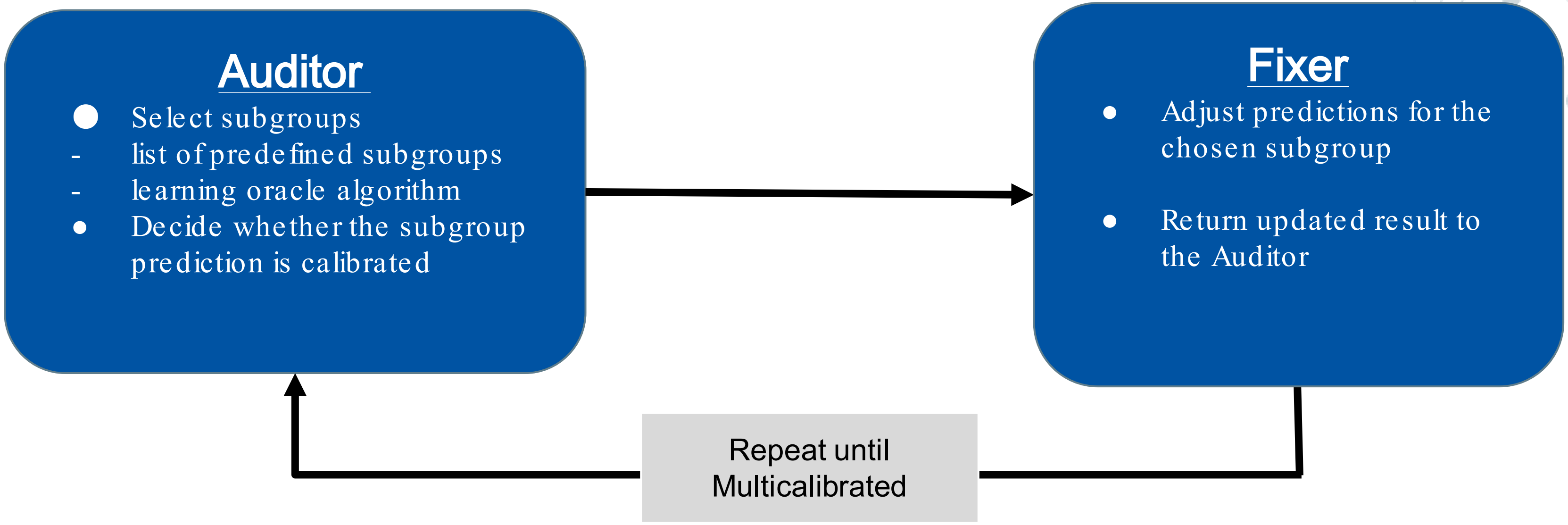


An Imbalanced Dataset

Low Population High Variance

High Population Low variance

Algorithm Overview



Auditor

- Select subgroups
  - list of predefined subgroups
  - learning oracle algorithm
- Decide whether the subgroup prediction is calibrated

Fixer

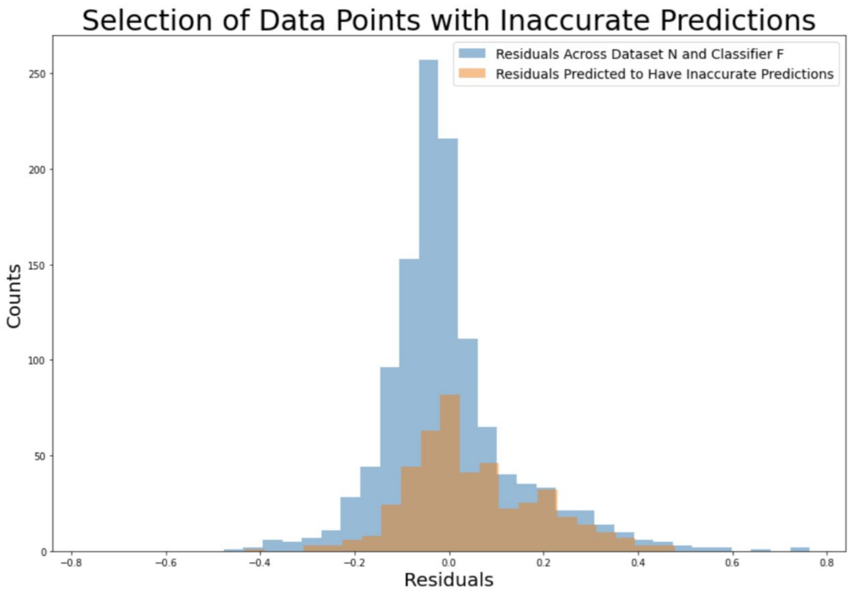
- Adjust predictions for the chosen subgroup
- Return updated result to the Auditor

Repeat until Multicalibrated

Auditor Visualization

Key Result

This algorithm crates a classifier to predict points in a dataset that will likely have inaccurate predictions

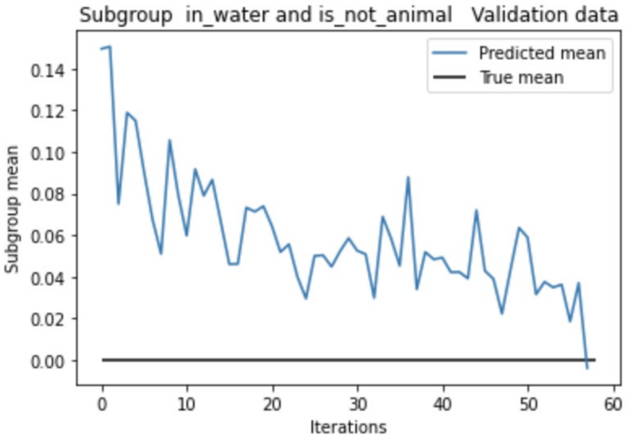


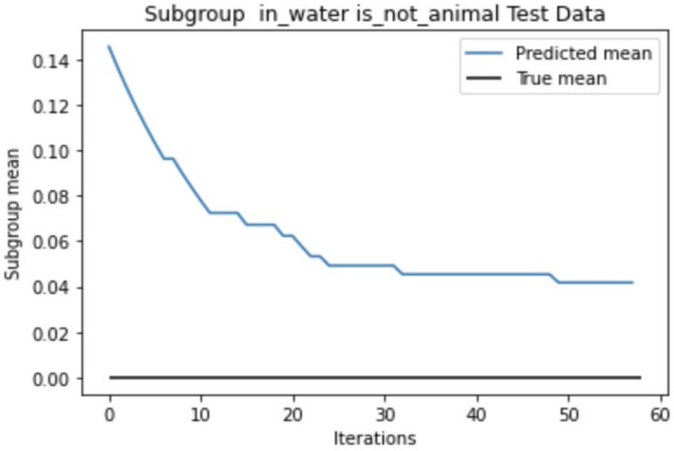
*residuals (n.)*The difference between the prediction and true label

Fixer Visualization

Key Result

After *T* iterations of post-processing, predicted mean is closer to the true mean





Mean predictions adjusted during post-processing for validation data

Mean predictions adjusted during post-processing for test data

Main Takeaways

1. **Implemented** mean multicalibration based on the algorithm in the paper by Jung, Lee, Pai, Roth, Vohra
2. **Tested** the auditor and fixer on two different datasets
3. **Demonstrated** the results showing that both components work

Future Steps

1. **Testing:** *evaluate* the algorithm on a variety of datasets
2. **Application:** *run* algorithm on specific use cases such as housing and medication
3. **Publishing** : *put* our code on a publicly accessible site like IBM AI Fairness 360 Package for ML developers to utilize

Contact Us @

eseseniordesign202021team06@gmail.com

Hyewon Lee

Trishla Pokharna

Brian Handen

Tashweena Heeramun

Margaret Ji