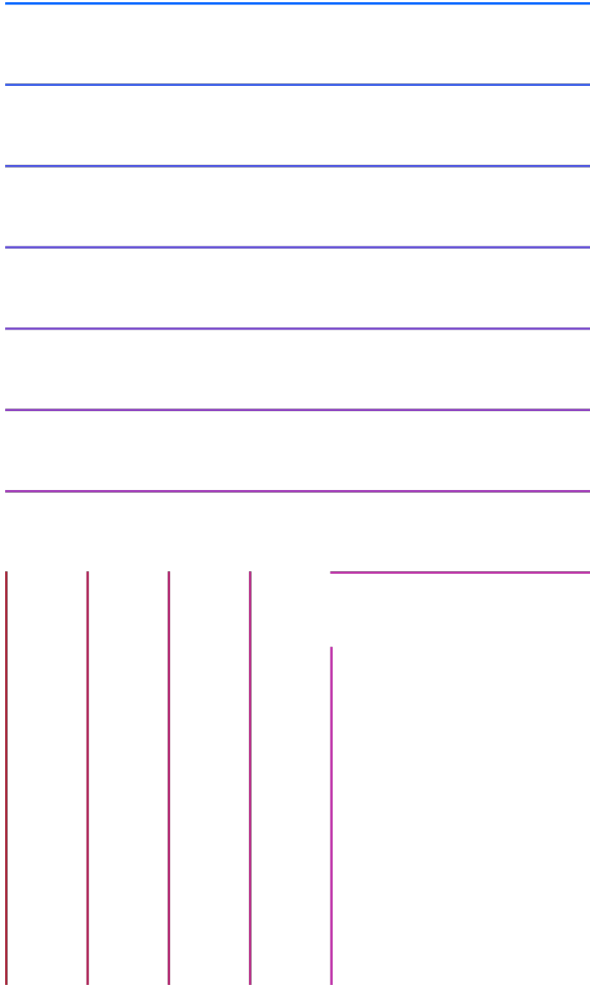
MIT-IBM
Watson AI Lab

IBM **Research**

# Fairness in Machine Learning

Mikhail Yurochkin



# Fairness: A case study

Example: Sentiment analysis – classify words as positive or negative

Positive: *admire, adorable, joy, lucky, talented, …*

Negative: *aggressive, distrust, nasty, radical, …*

Deep Learning + Word Embeddings -> **95%** test accuracy.

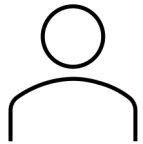
Success 

# Deployment Concerns

What is a sentiment of a name?

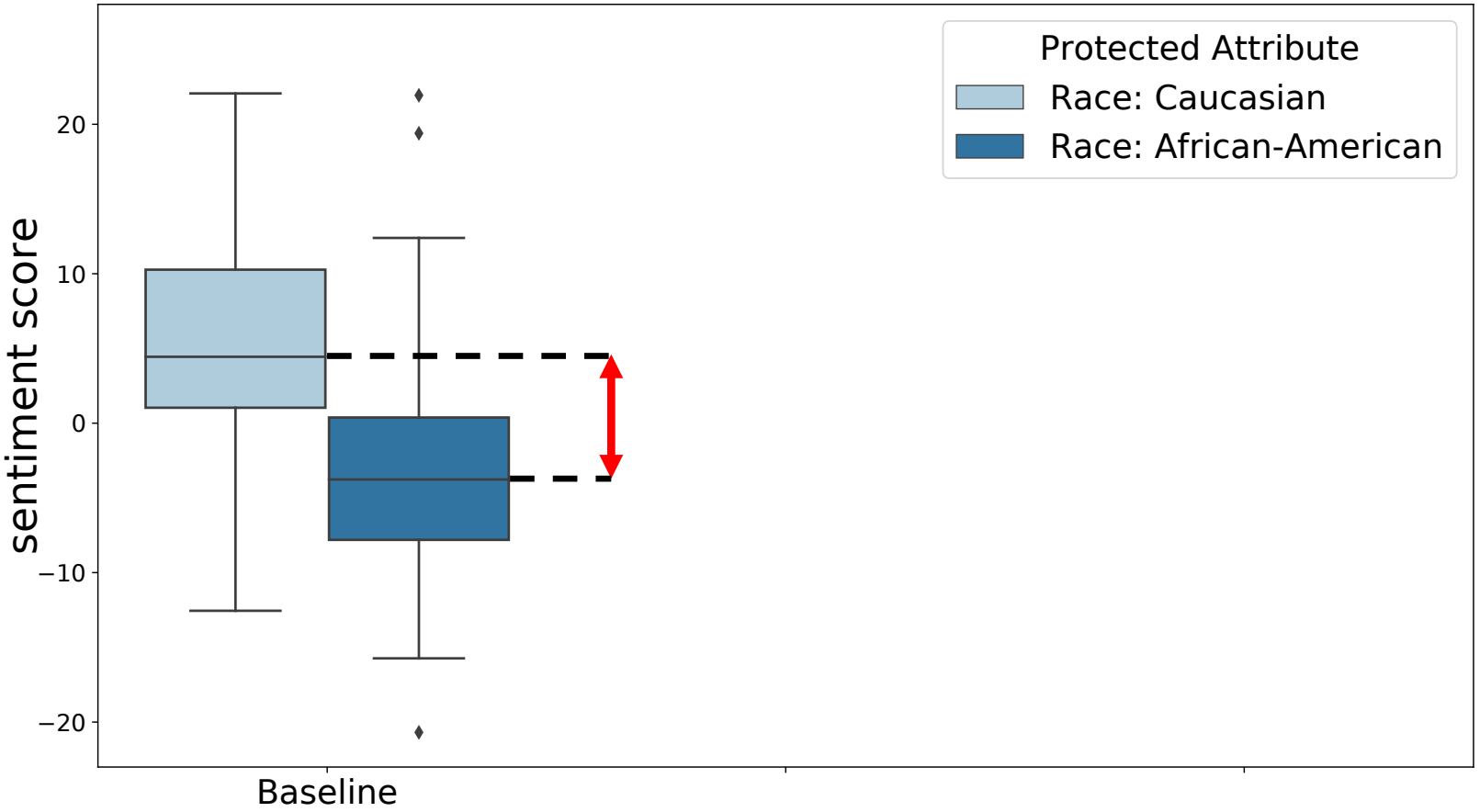
Common European-American names:
*Adam, Ryan, Paul, ... , Courtney, Meredith, Megan, ...*

Common African-American names:
*Alonzo, Leroy, Tyree, ... , Shereen, Sharise, Tawanda, ...*

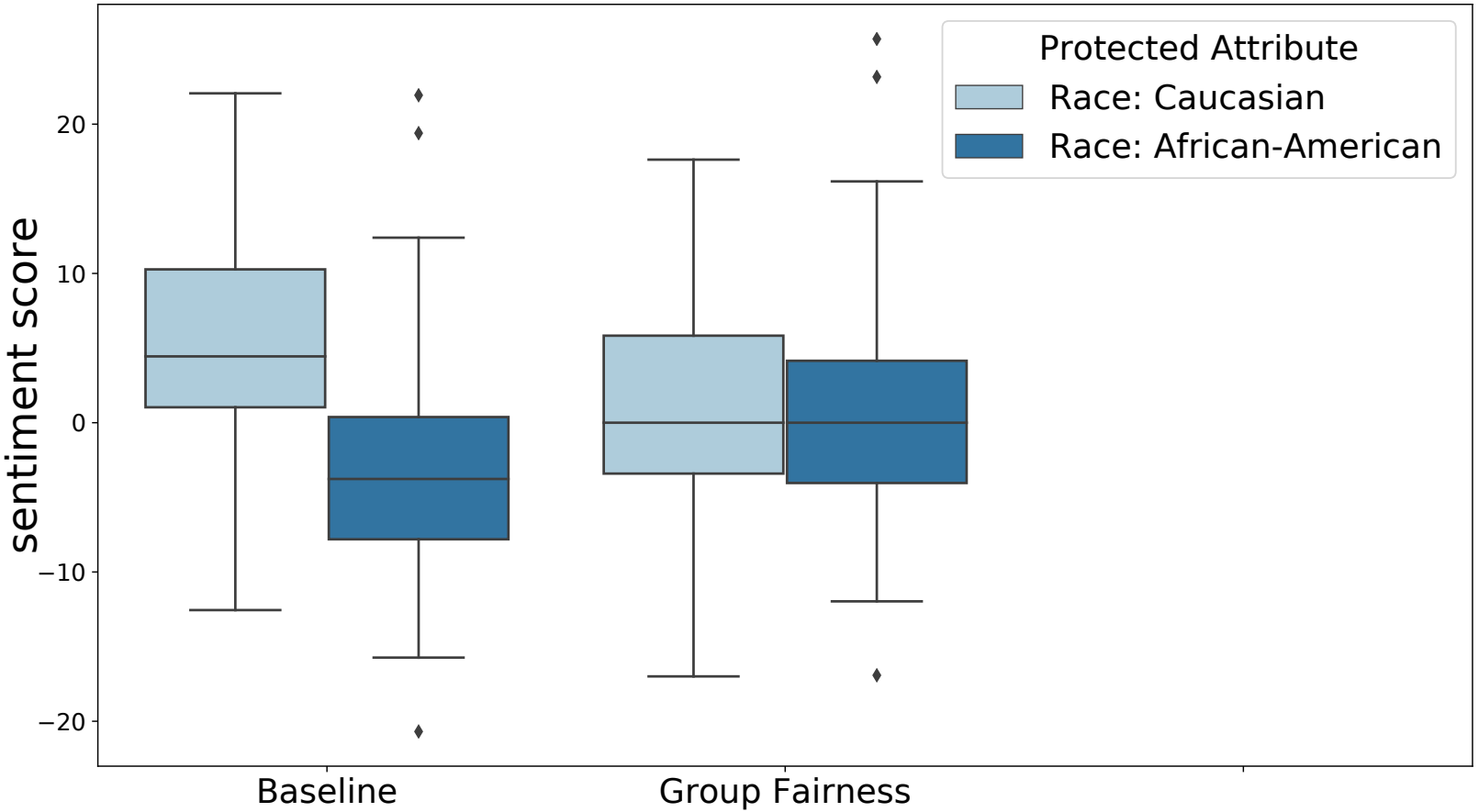


Names are from “Semantics derived automatically from language corpora contain human-like biases” (Caliskan et al., 2017)

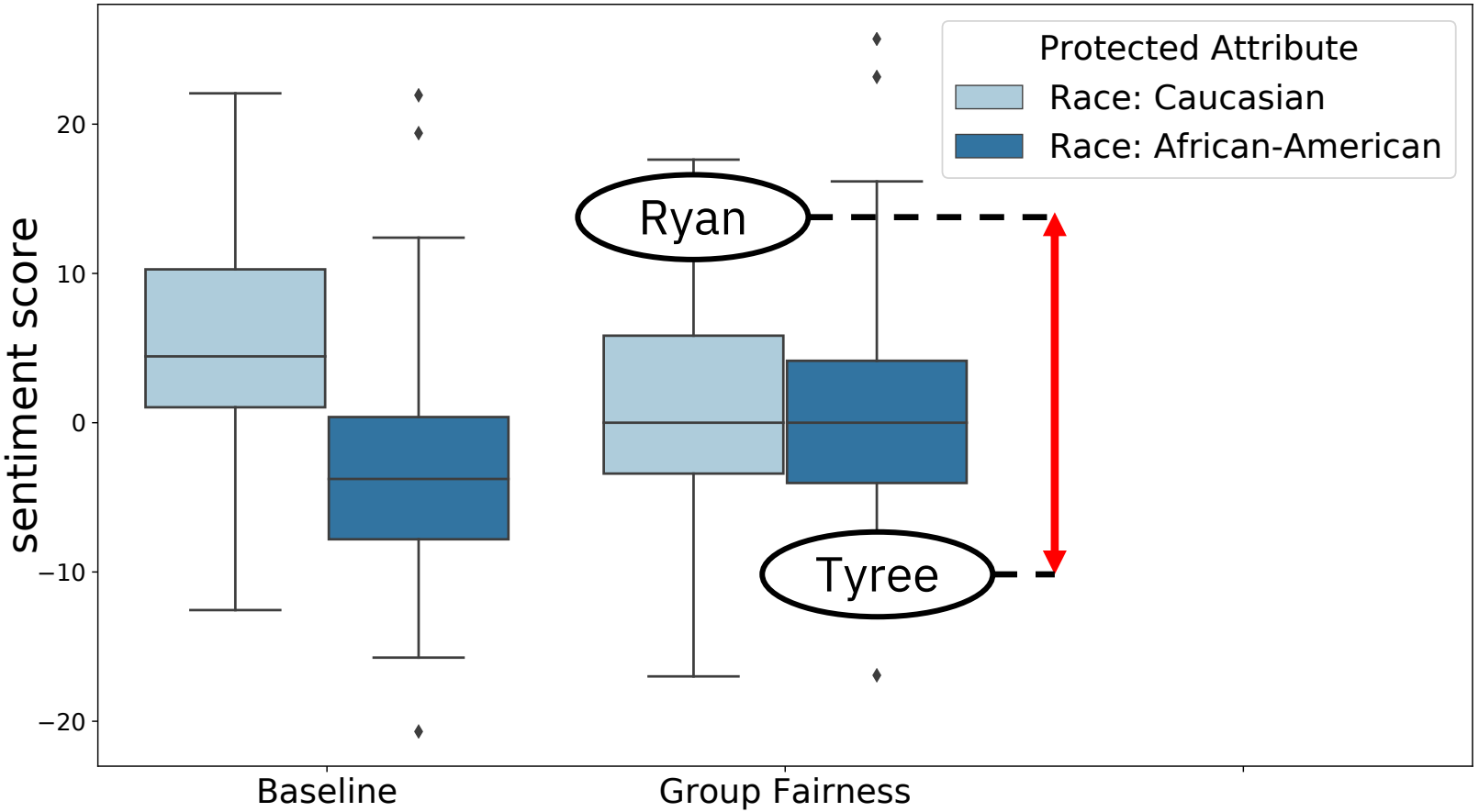
# Fairness Violations



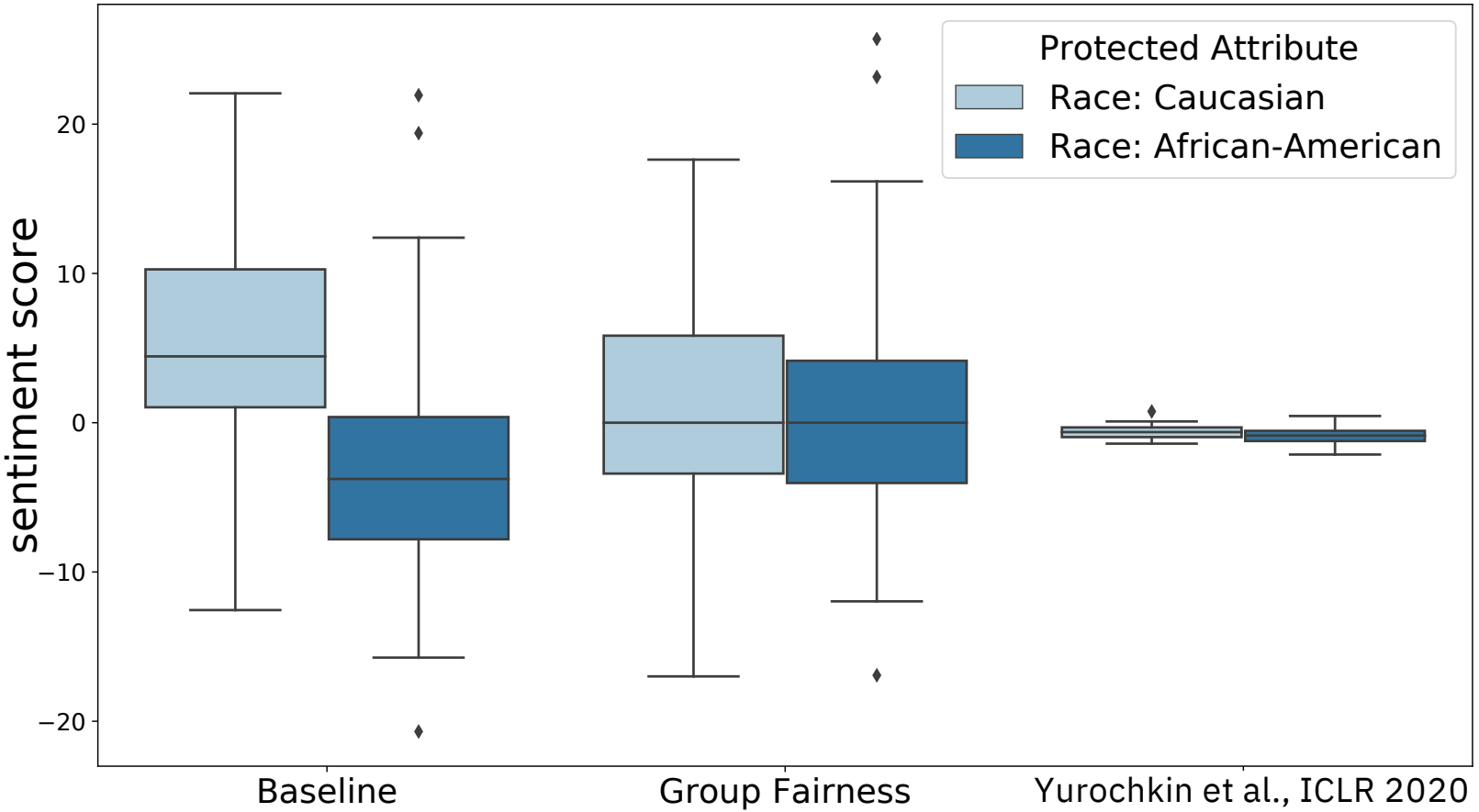
# Group Fairness



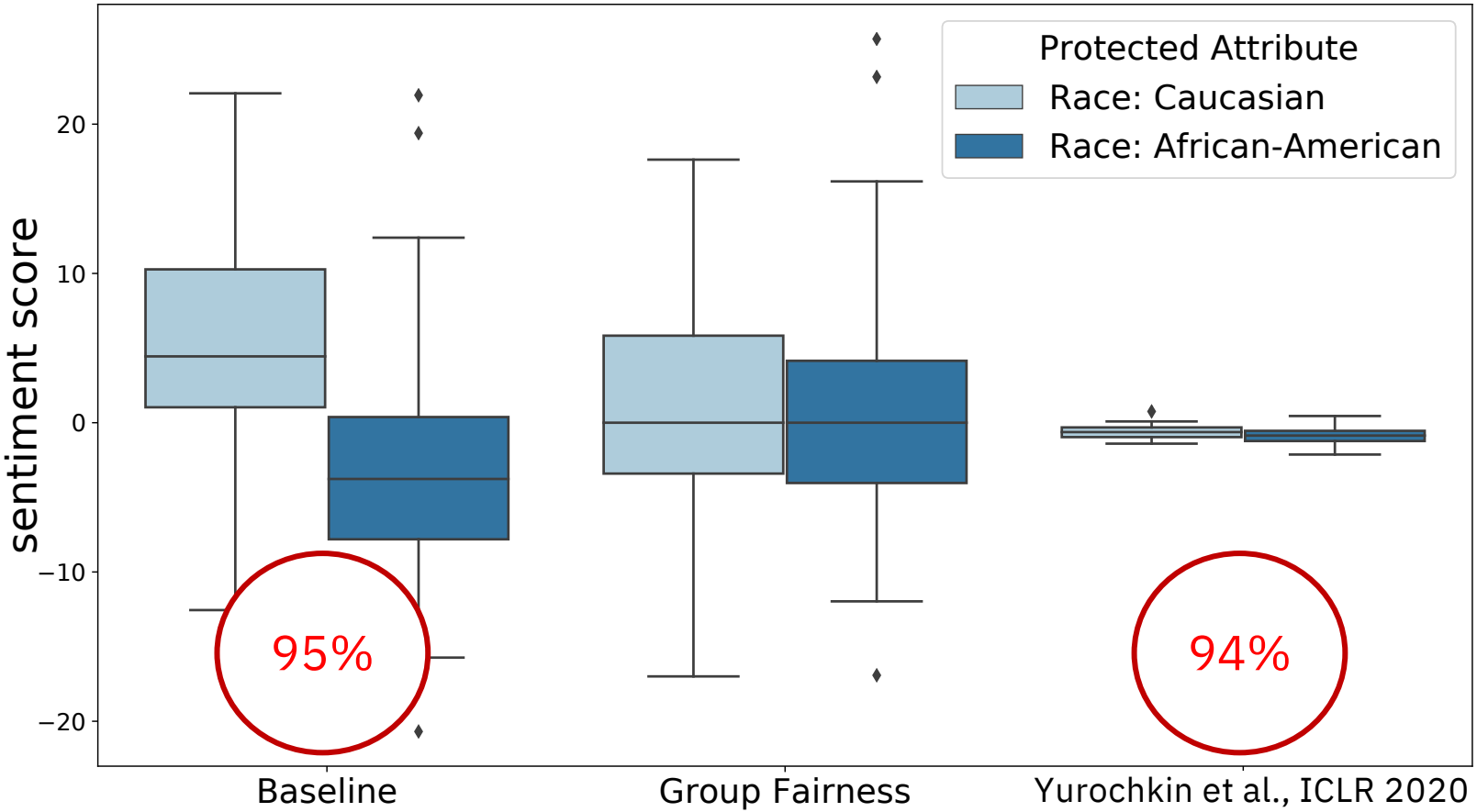
# Fairness is Violated for Individuals



# Individual Fairness



# Accuracy is Preserved



# Roadmap

AI is prone to biases

Definitions of algorithmic fairness

Practical fairness methods

- Identifying fairness violations
- Training fair models
- Post-processing for fairness

# Group Fairness

*Algorithm performs similarly on groups of individuals*

$$Y - \text{true label}$$

$$A - \text{protected attribute}$$

$$\hat{Y} - \text{prediction}$$

Demographic Parity: $\hat{Y}$ is independent of $A$

Equalized Odds: $\hat{Y}$ and $A$ are conditionally independent given $Y$

# Evaluating Group Fairness

Demographic Parity: $\hat{Y}$ is independent of $A$

*Compare average outcome for men and women*

Test data: $(x_1, a_1), \ldots, (x_N, a_N)$;
model to audit $h : \mathcal{X} \to \mathcal{Y}$

$$\text{Output DP} = \left| \frac{\sum_i \mathcal{I}(a_i{=}\text{male}, h(x_i){=}1)}{\sum_i \mathcal{I}(a_i{=}\text{male})} - \frac{\sum_i \mathcal{I}(a_i{=}\text{female}, h(x_i){=}1)}{\sum_i \mathcal{I}(a_i{=}\text{female})} \right|$$

# Evaluating Group Fairness

Equalized Odds: $\hat{Y}$ and $A$ are conditionally independent given $Y$

*Compare class accuracies for men and women*

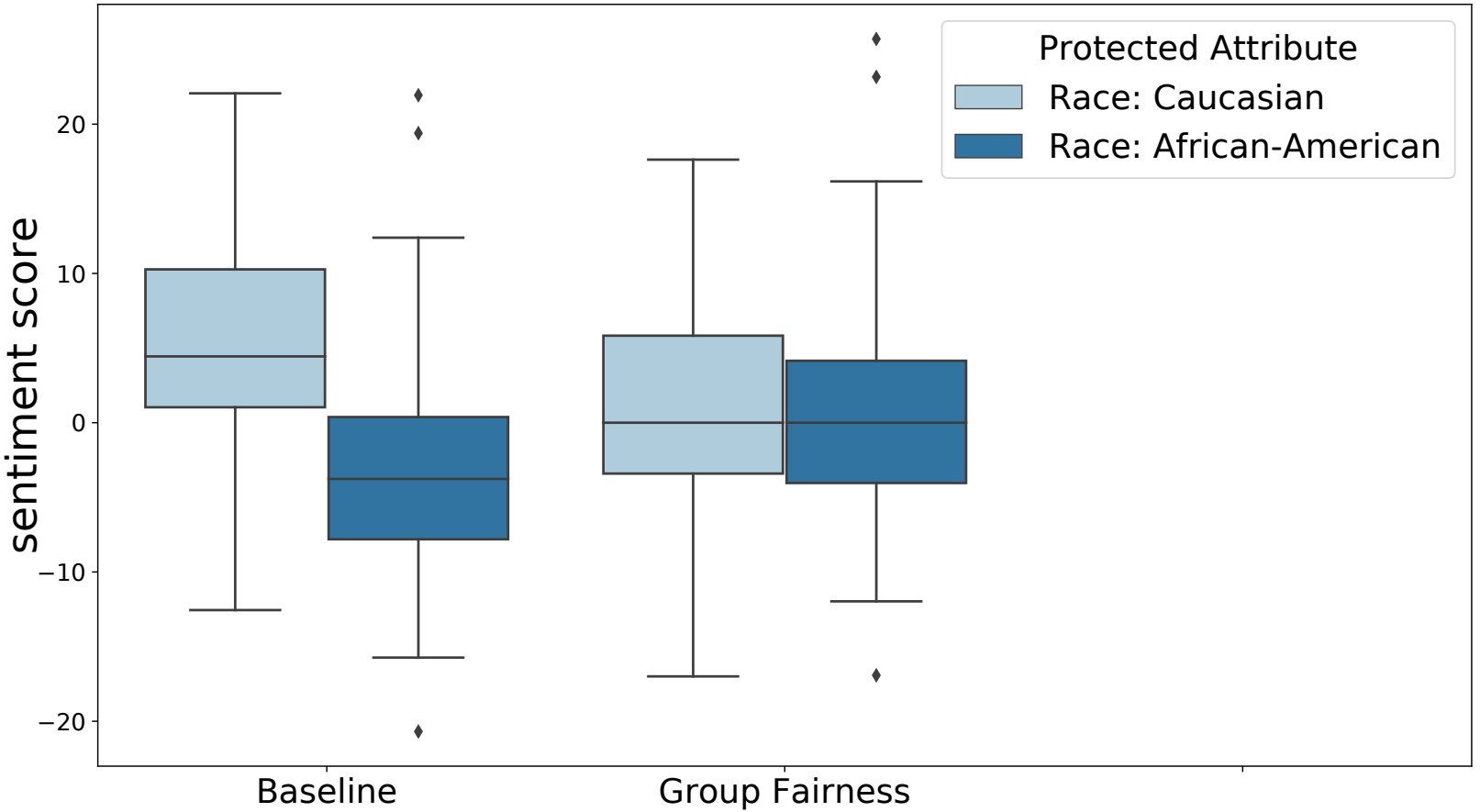
Test data: $(x_1, y_1, a_1), \ldots, (x_N, y_n, a_N)$;
model to audit $h : \mathcal{X} \to \mathcal{Y}$

$$\text{Measure EO}_0 = \left| \frac{\sum_i \mathcal{I}(a_i{=}\text{male}, y_i{=}0, h(x_i){=}0)}{\sum_i \mathcal{I}(a_i{=}\text{male}, y_i{=}0)} - \frac{\sum_i \mathcal{I}(a_i{=}\text{female}, y_i{=}0, h(x_i){=}0)}{\sum_i \mathcal{I}(a_i{=}\text{female}, y_i{=}0)} \right|$$

$$\text{Measure EO}_1 = \left| \frac{\sum_i \mathcal{I}(a_i{=}\text{male}, y_i{=}1, h(x_i){=}1)}{\sum_i \mathcal{I}(a_i{=}\text{male}, y_i{=}1)} - \frac{\sum_i \mathcal{I}(a_i{=}\text{female}, y_i{=}1, h(x_i){=}1)}{\sum_i \mathcal{I}(a_i{=}\text{female}, y_i{=}1)} \right|$$

$$\text{Output EO} = \tfrac{1}{2}(\text{EO}_0 + \text{EO}_1)$$

# What Group Fairness definition did we check?



# Individual Fairness

*Algorithm treats similar individuals similarly*

$$d_{\mathcal{Y}}(h(x_1), h(x_2)) \lesssim d_{\mathcal{X}}(x_1, x_2) \text{ for all } x_1, x_2 \in \mathcal{X}$$

- ML model is a map $h : \mathcal{X} \to \mathcal{Y}$
- $d_{\mathcal{Y}}$ measures similarity between outputs
- **Fair metric** $d_{\mathcal{X}}$ measures similarity between inputs

Dwork et al., ITCS 2012

# Evaluating Individual Fairness

## Prediction Consistency

*Compare predictions on similar inputs*

Toxic comment detection in online discussions:

*A sad day for American people everywhere* ➡ Non-Toxic

*A sad day for lesbian people everywhere* ➡ ???Toxic???

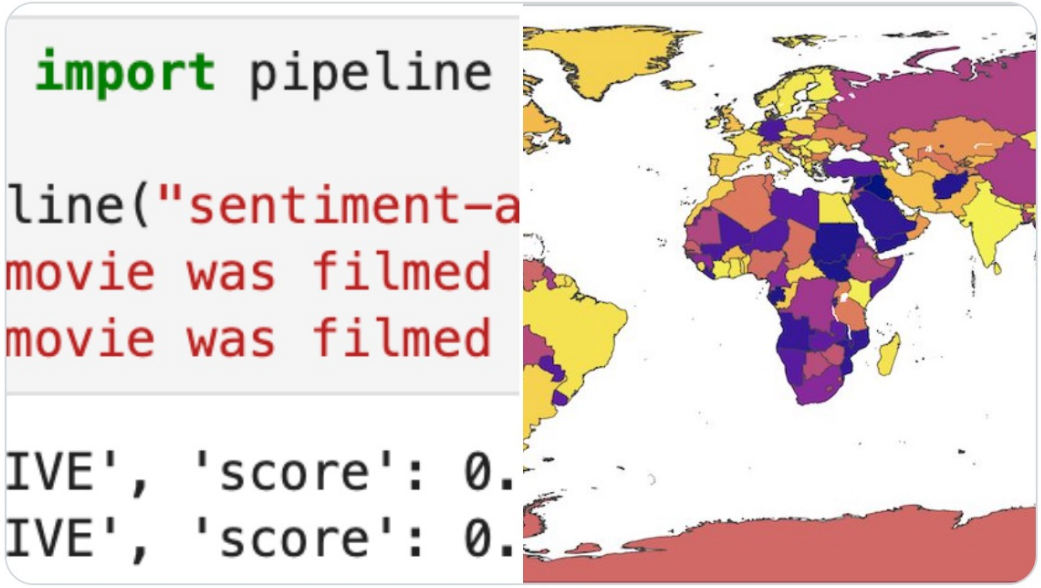
$$\text{Output PC} = \frac{\sum_i \mathcal{I}(h(x_i[american]) = h(x_i[lesbian]))}{N}$$

# Questions?

# Is “blindness” a solution?

**Aurélien Geron**
@aureliengeron

I noticed that DistilBERT loves movies filmed in India, but not in Iraq, so I plotted the result for each country: the resulting map is scary. #aibias




12:35 AM · Mar 20, 2022 · Twitter Web App

**374** Retweets **67** Quote Tweets **1,984** Likes

```
from transformers import pipeline

classifier = pipeline("sentiment-analysis")
classifier(["The movie was filmed in India.",
            "The movie was filmed in Iraq."])
```

```
[{'label': 'POSITIVE', 'score': 0.9783285856246948},
 {'label': 'NEGATIVE', 'score': 0.9872057437896729}]
```

# Roadmap

AI is prone to biases

Definitions of algorithmic fairness

Practical fairness methods

- Identifying fairness violations
- Training fair models
- Post-processing for fairness

# Identifying Fairness violations

## Group Fairness: measure DP (EO) on audit data

Test data: $(x_1, a_1), \ldots, (x_N, a_N)$;
model to audit $h : \mathcal{X} \to \mathcal{Y}$

$$\text{Output DP} = \left| \frac{\sum_i \mathcal{I}(a_i{=}\text{male}, h(x_i){=}1)}{\sum_i \mathcal{I}(a_i{=}\text{male})} - \frac{\sum_i \mathcal{I}(a_i{=}\text{female}, h(x_i){=}1)}{\sum_i \mathcal{I}(a_i{=}\text{female})} \right|$$

Four-Fifths Rule, US Equal Employment Opportunity Commission:
*"selection rate for any race, sex, or ethnic group [must be at least] four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate"*

# Identifying Fairness violations

## Individual Fairness: Prediction Consistency

Toxic comment detection in online discussions:

*A sad day for American people everywhere* ➡ Non-Toxic

*A sad day for lesbian people everywhere* ➡ ???Toxic???

$$\text{Output PC} = \frac{\sum_i \mathcal{I}(h(x_i[american]) = h(x_i[lesbian]))}{N}$$

# Individual Fairness in Social Science

## 2004 study of the racial bias in the US labor market

- The investigators responded to job ads in Boston and Chicago newspapers with fictitious resumes.
- They randomly assigned African-American or white sounding names to the resumes.
- The investigators concluded there is discrimination against African-Americans: the resumes assigned white names received 50% more callbacks for interviews.

Bertrand & Mullainathan, AER 2004

# Demonstration

# Find Individual Fairness Violations *Algorithmically*

$$d_{\mathcal{Y}}(h(x_1), h(x_2)) \lesssim d_{\mathcal{X}}(x_1, x_2) \text{ for all } x_1, x_2 \in \mathcal{X}$$

$x_1$: *A sad day for American people everywhere* $\xrightarrow{h(x_1)}$ Non-Toxic

$T(x_1) \to x_2$

Training an ML model to audit

- New sentence $x_2$ is similar in the fair metric, meaning $d_{\mathcal{X}}(x_1, x_2)$ is small
- New sentence $x_2$ results in a different prediction, meaning $d_{\mathcal{Y}}(h(x_1), h(x_2))$ is big

$x_2$: *A sad day for lesbian people everywhere* $\xrightarrow{h(x_2)}$ Toxic

Yurochkin & Sun, ICLR 2021

# Auditing for IF violations

Test data: $(x_1, y_1), \ldots, (x_N, y_N)$;
Auditor $T(x)$ for the model of interest $h$;
some loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$

Hypothesis ($h$ is individually fair)
$H_0$ : loss ratio on similar individuals is at most $\delta$

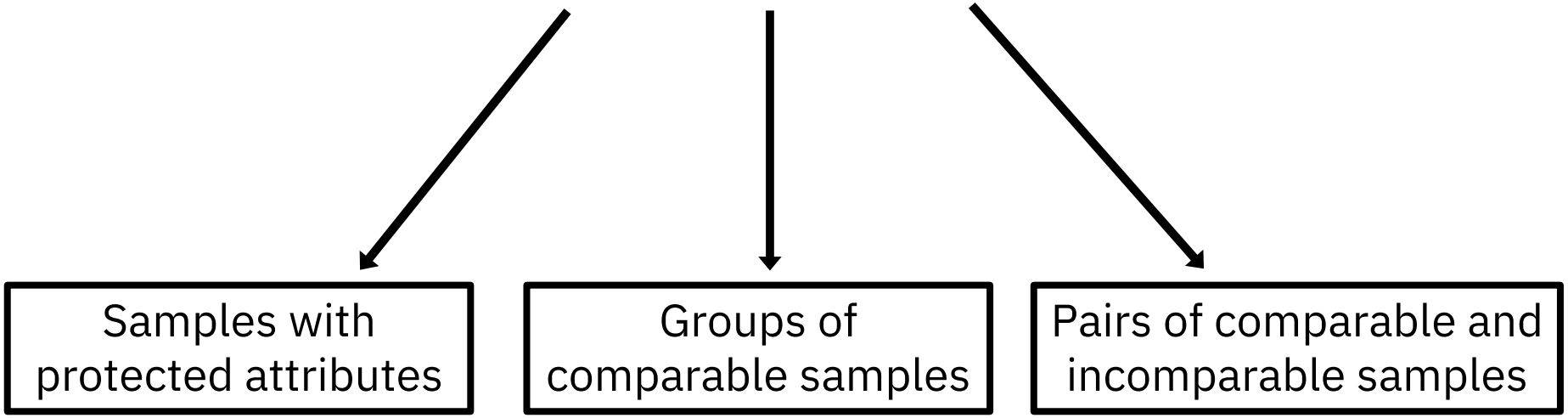
Compute loss ratios $R = \left\{ \frac{\ell(h(T(x_i), y_i)}{\ell(h(x_i), y_i)} \right\}_{i=1}^{N}$

Reject $H_0$ with confidence $(1 - \alpha)$ if $\text{Mean}(R) - \frac{z_{1-\alpha}}{\sqrt{N}} \text{Var}(R) > \delta$

Maity et al., ICLR 2021

# Demonstration

# Learning fair metrics from data




$$d_{\mathcal{X}}(x_1, x_2) = (x_1 - x_2)^\top \Sigma (x_1 - x_2)$$

Mukherjee et al., ICML 2020

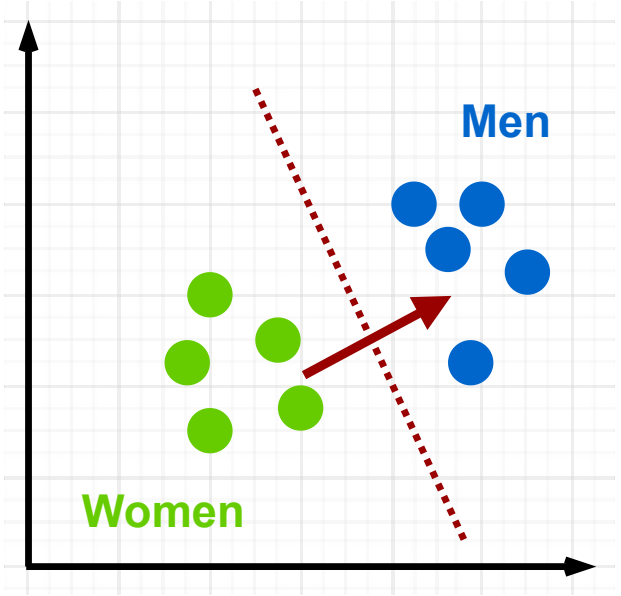
# Learning fair metrics from data

Samples with protected attributes:
gender/race information in the Adult dataset

Learn "sensitive" directions with Logistic Regression, i.e. $V = \{v_{\text{gender}}, v_{\text{race}}\}$.

Ignore them in the fair metric: $\Sigma = I - P_{\text{span}(V)}$.

$$d_{\mathcal{X}}(x_1, x_2) = (x_1 - x_2)^\top \Sigma (x_1 - x_2)$$




Yurochkin et al., ICLR 2020

# Learning fair metrics from data

Group of comparable samples:
word embeddings of popular baby names

Find directions of major variation with PCA, i.e. $V = \{v_1, \ldots, v_K\}$.

Ignore them in the fair metric: $\Sigma = I - P_{\mathrm{span}(V)}$.

$$d_{\mathcal{X}}(x_1, x_2) = (x_1 - x_2)^\top \Sigma (x_1 - x_2)$$

Mukherjee et al., ICML 2020

# Questions?

# Roadmap

AI is prone to biases

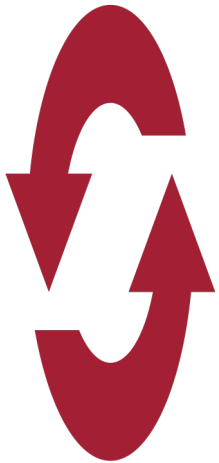
Definitions of algorithmic fairness

Practical fairness methods

- Identifying fairness violations
- Training fair models
- Post-processing for fairness

# Training Individually Fair models

*A variant of adversarial training*: Train model accurate on the available data **and** data similar in the fair metric



- Observe data
- Audit model: Find similar data where algorithm violates individual fairness
- Update model parameters to minimize prediction error **and** correct violations
- Repeat

Yurochkin & Sun, ICLR 2021

# Relation to Adversarial Robustness

*Adversarial training*: Train model accurate on the available data **and** visually similar data. Different "fair" metric.

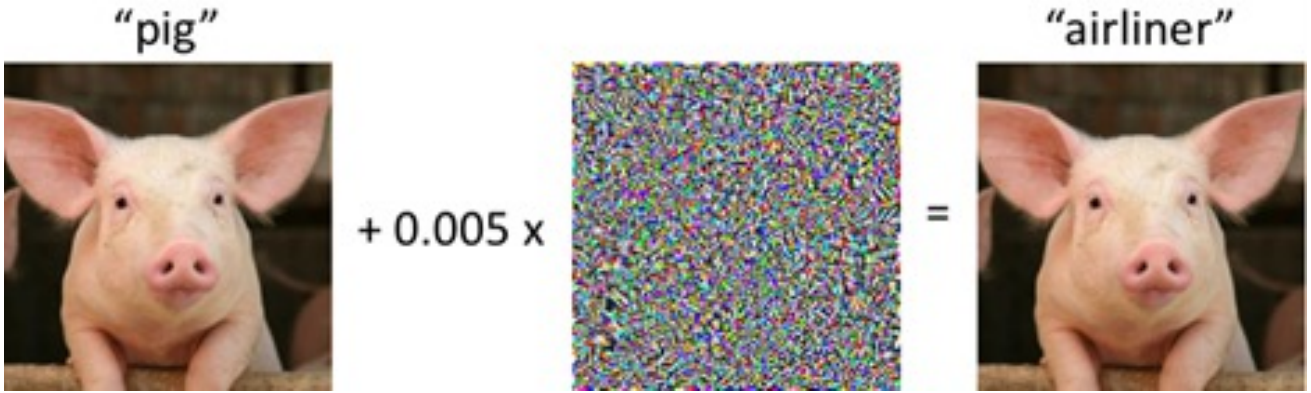



Image is from "A Brief Introduction to Adversarial Examples" (Mądry & Schmidt, 2018)

# Demonstration

# Training Group Fair models

*Optimization with (data-dependent) constraints*: Train model accurate on the available data **subject to** group fairness constraints

$$\min_{h \in \mathcal{H}} L(h)$$

$$\text{subject to DP} < \delta, \text{ where}$$

$$\text{DP} = \left| \frac{\sum_i \mathcal{I}(a_i = \text{male}, h(x_i) = 1)}{\sum_i \mathcal{I}(a_i = \text{male})} - \frac{\sum_i \mathcal{I}(a_i = \text{female}, h(x_i) = 1)}{\sum_i \mathcal{I}(a_i = \text{female})} \right|$$

# Demonstration

# Toxic Comment Detection

Individual Fairness (prediction consistency):

*A sad day for American people everywhere* ➡ Non-Toxic

*A sad day for lesbian people everywhere* ➡ ???Toxic???

Group Fairness (equalized odds):

Compare performance on sentences containing any of the group-words *"lesbian, gay, bisexual, …"* to the overall performance.

# Demonstration

fairbert.vizhub.ai

# What is Your type of Fairness?

*Group Fairness*:

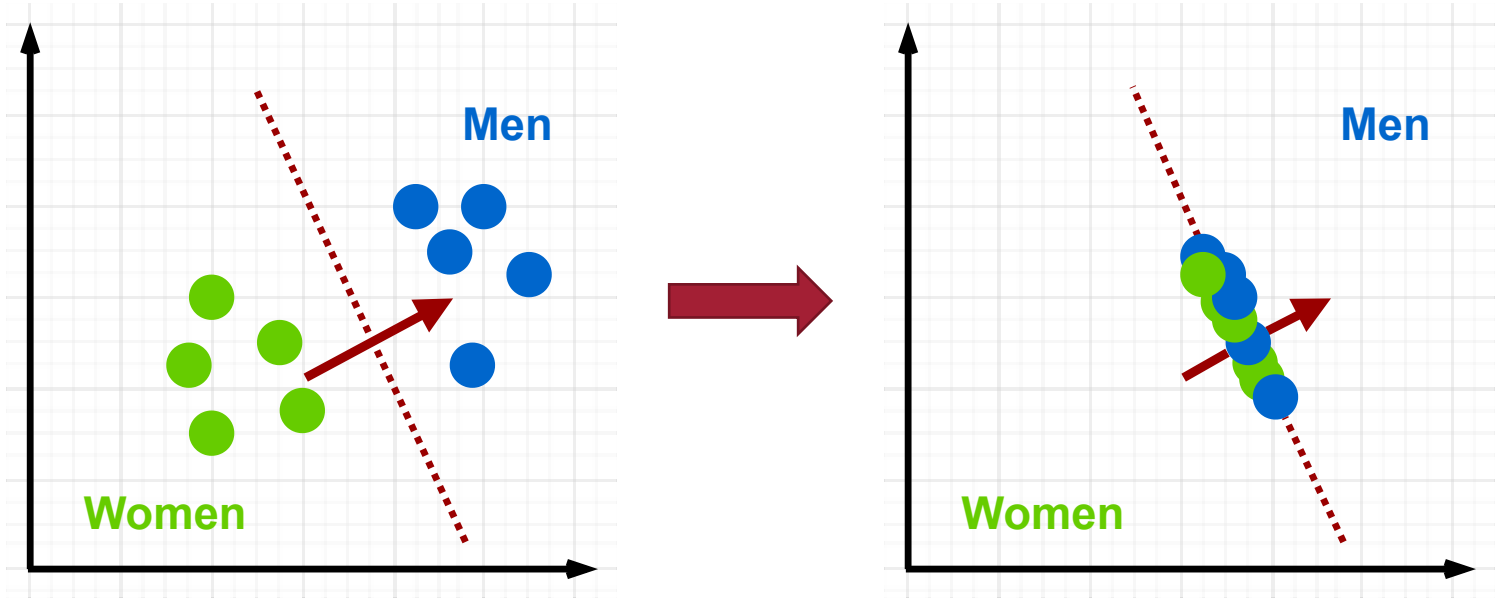
- Carefully choose GF notion appropriate for the application
- Several open-source solutions (AIF360, Fairlearn, TFCO)
- Check individual fairness!

*Individual Fairness*:

- Carefully choose data for learning the fair metric
- inFairness was open-sourced this summer
- Check group fairness!

# Pre-processing to Train Fair Models

*1. Modify data (features, sample weights, labels)*
*2. Train a regular model*




Bolukbasi et al., NeurIPS 2016; Yurochkin et al., ICLR 2020

# Questions?

# Roadmap

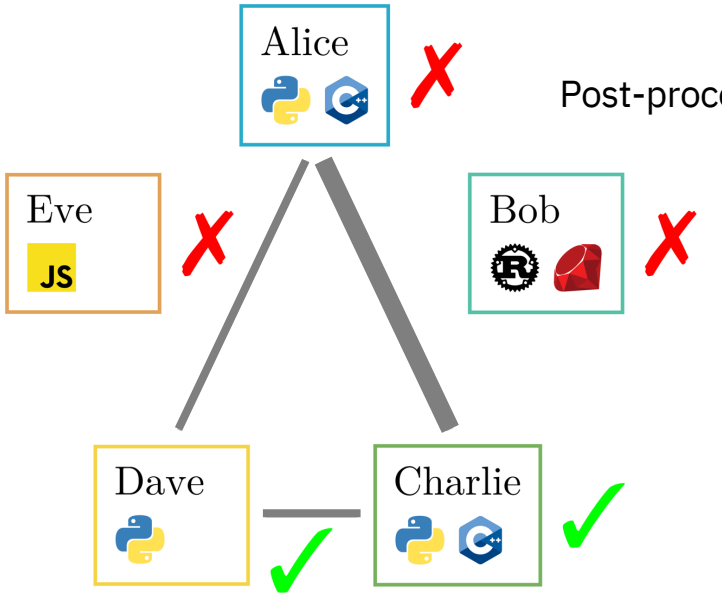
AI is prone to biases

Definitions of algorithmic fairness

Practical fairness methods

- Identifying fairness violations
- Training fair models
- Post-processing for fairness

# Post-processing for Individual Fairness




Original predictions

Graph Laplacian

Post-processed fair predictions

Minimize for $\mathbf{f}$: $\|\mathbf{f} - \hat{\mathbf{y}}\|_2^2 + \lambda\, \mathbf{f}^\top \mathbb{L} \mathbf{f}$

Stay close to original predictions

Penalize individual fairness violations

Measuring IF on a graph:

$$\sum_{i,j} W_{ij}(f_i - f_j)^2 = 2\mathbf{f}^\top \mathbb{L}\mathbf{f}$$

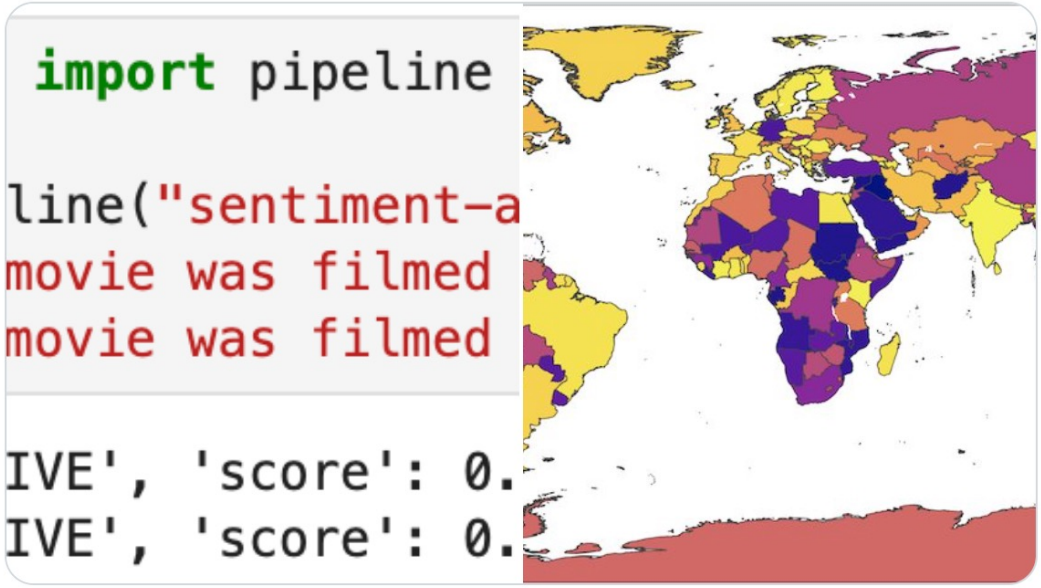
Closed-form solution!

$$\mathbf{f} = (I + \lambda \mathbb{L})^{-1} \hat{\mathbf{y}}$$

Petersen et al., NeurIPS 2021

**Aurélien Geron**
@aureliengeron

I noticed that DistilBERT loves movies filmed in India, but not in Iraq, so I plotted the result for each country: the resulting map is scary. #aibias



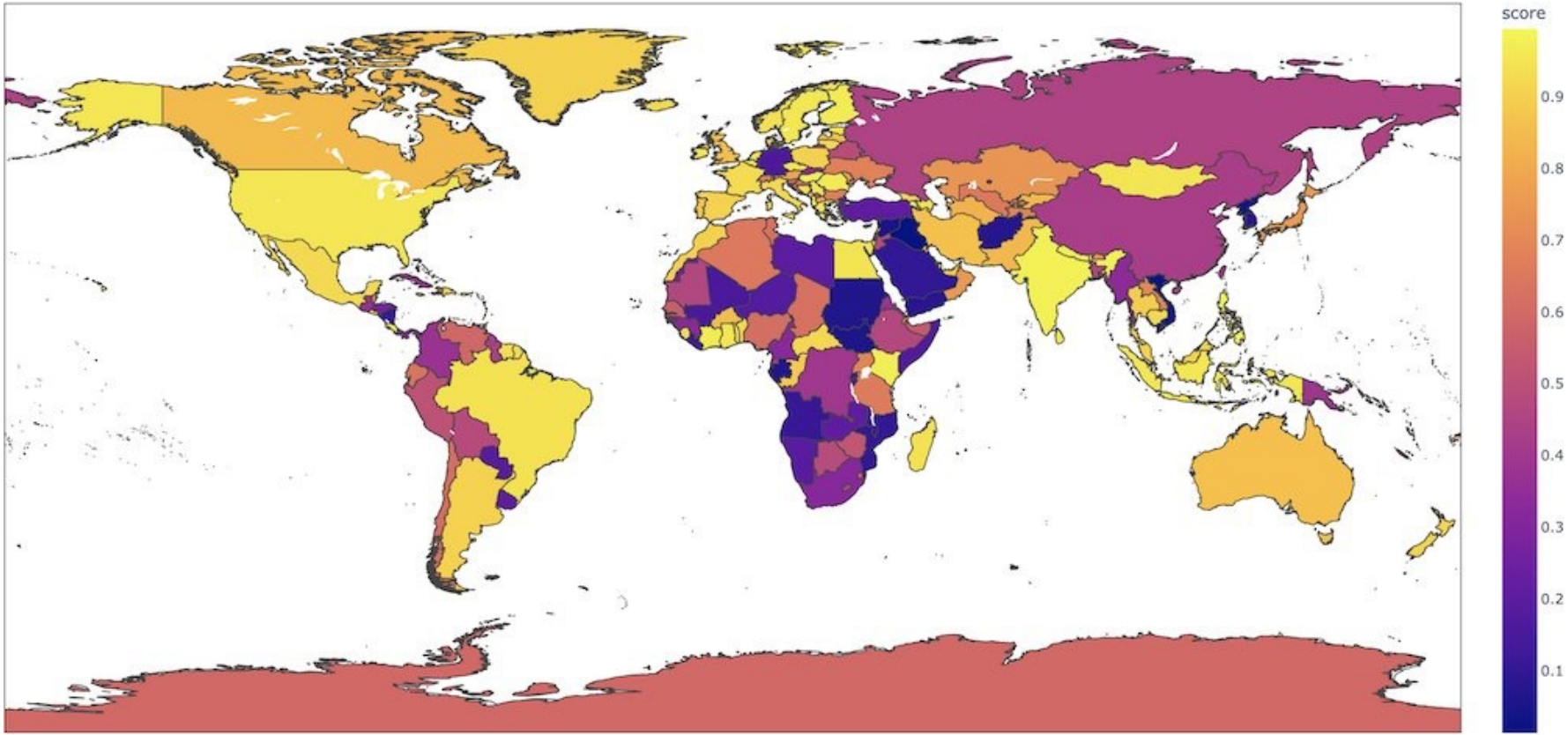

12:35 AM · Mar 20, 2022 · Twitter Web App

**374** Retweets **67** Quote Tweets **1,984** Likes

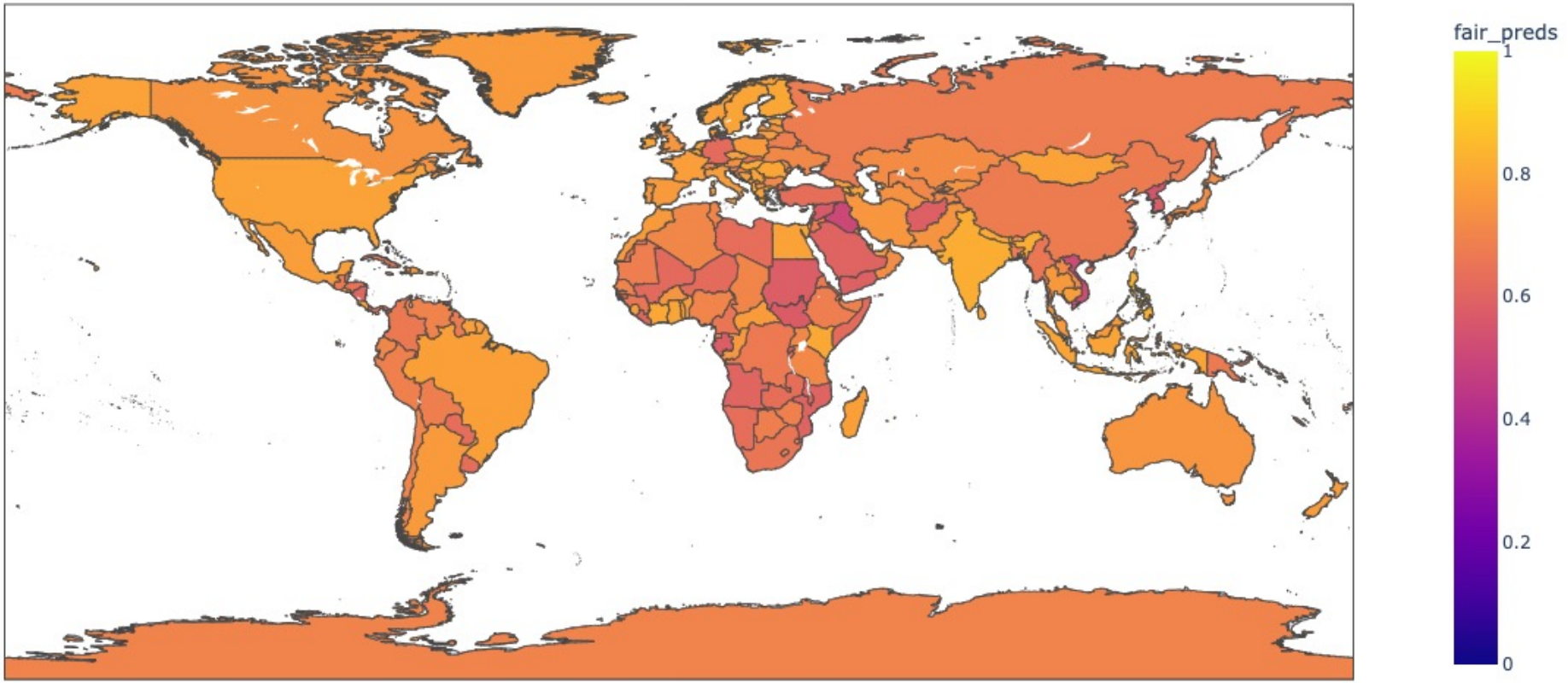
```
from transformers import pipeline

classifier = pipeline("sentiment-analysis")
classifier(["The movie was filmed in India.",
            "The movie was filmed in Iraq."])
```

```
[{'label': 'POSITIVE', 'score': 0.9783285856246948},
 {'label': 'NEGATIVE', 'score': 0.9872057437896729}]
```


score
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

# Post-processing for Individual Fairness



# Post-processing for Individual Fairness

| | Original model | Post-processed model |
|---|---|---|
| *This movie was filmed in Iraq* | Negative (0.01) | Positive (0.51) |
| *This movie was filmed in India* | Positive (0.98) | Positive (0.81) |
| *This movie was filmed in Germany* | Negative (0.16) | Positive (0.62) |

| | Original model | Post-processed model |
|---|---|---|
| *This movie is amazing* | Positive (0.99) | Positive (0.91) |
| *This movie is stunning* | Positive (0.99) | Positive (0.91) |
| *This movie is unimpressive* | Negative (0.01) | Negative (0.34) |
| *This movie is boring* | Negative (0.01) | Negative (0.33) |

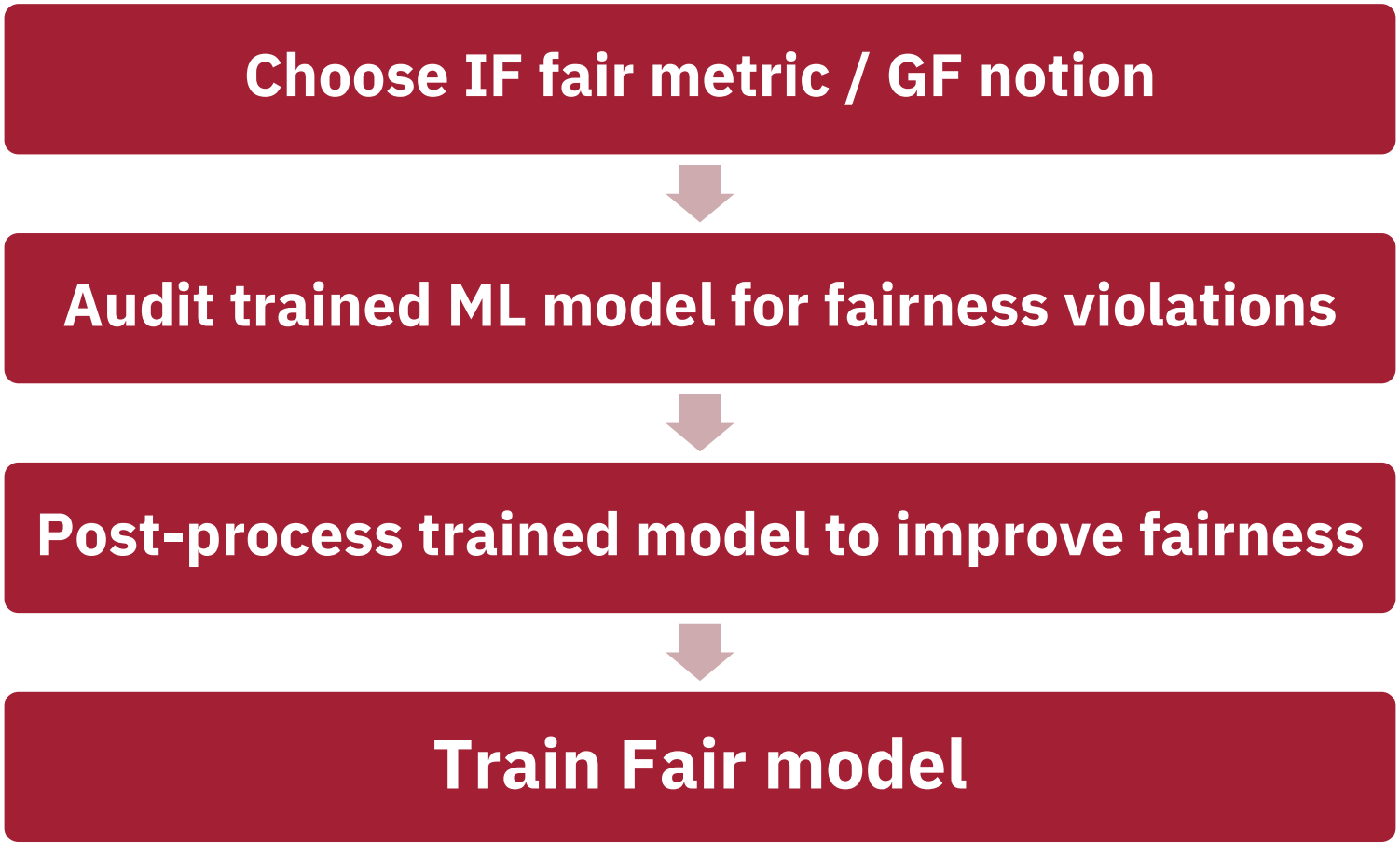
# Demonstration

# Post-processing for Group Fairness

Optimized Score Transformation for Consistent Fair Classification
*Wei et al., 2021*

FairScoreTransformer (FST): Available in AIF360

# Algorithmic Fairness pipeline



# Questions?

# Group Fairness References


- M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. NeurIPS 2016.
- B. Zhang, B, Lemoine, and M. Mitchell. Mitigating Unwanted Biases with Adversarial Learning. AAAI/ACM Conference on AI, Ethics, and Society 2018.
- A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. ICML 2018.
- D. Wei, K. Ramamurthy, F. Calmon. Optimized Score Transformation for Consistent Fair Classification. JMLR 2021.
- R. Bellamy et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias.
- A. Cotter, H. Jiang, K. Sridharan. Two-Player Games for Efficient Non-Convex Constrained Optimization. ALT 2019.
- AIF360: https://github.com/Trusted-AI/AIF360
- Fairlearn: https://github.com/fairlearn/fairlearn
- TFCO: https://github.com/google-research/tensorflow_constrained_optimization

# Individual Fairness References


- C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. ITCS 2012.
- M. Yurochkin, A. Bower, and Y. Sun. Training individually fair ML models with sensitive subspace robustness. ICLR 2020.
- S. Xue, M. Yurochkin, and Y. Sun. Auditing ML models for individual bias and unfairness. AISTATS 2020.
- D. Mukherjee, M. Yurochkin, M. Banerjee, and Y. Sun. Two Simple Ways to Learn Individual Fairness Metric from Data. ICML 2020.
- M. Weber, M. Yurochkin, S. Botros, V. Markov. Black Loans Matter: Distributionally Robust Fairness for Fighting Subgroup Discrimination. Fair AI in Finance Workshop, NeurIPS 2020.
- M. Yurochkin and Y. Sun. SenSeI: Sensitive Set Invariance for Enforcing Individual Fairness. ICLR 2021.
- A. Vargo, F. Zhang, M. Yurochkin, and Y. Sun. Individually Fair Gradient Boosting. ICLR 2021.
- A. Bower, H. Eftekhari, M. Yurochkin, and Y. Sun. Individually Fair Ranking. ICLR 2021.
- S. Maity, S. Xue, M. Yurochkin, and Y. Sun. Statistical inference for individual fairness. ICLR 2021.
- F. Petersen, D. Mukherjee, Y. Sun, and M. Yurochkin. Post-processing for Individual Fairness. NeurIPS 2021.
- inFairness: https://github.com/IBM/inFairness

# Blog-posts and Media

## AI fairness

In today's data-driven world, machine learning (ML) systems are increasingly used to make high-stakes decisions in domains like criminal justice, education, lending, and medicine. For example, a judge may use an algorithm to assess a defendant's chance of re-offending before deciding to detain or release the defendant. Although replacing humans with ML systems appear to eliminate human biases in the decision-making process, they can perpetuate or even exacerbate biases in the training data. Such biases are especially objectionable when it adversely affects underprivileged groups of users. The most obvious remedy is to remove the biases in the training data, but carefully curating the datasets that modern ML systems are trained on is impractical. This leads to the challenge of developing ML systems that remain "fair" despite biases in the training data.

### But what is fair?

There are two major families of definitions of fairness: **(1) group fairness; (2) individual fairness**. Group fairness requires certain constraints to be satisfied at the population level, e.g. proportion of hired job applicants should be similar across different demographic groups. Individual fairness (also known as *Lipschitz fairness*) states that hiring decisions for any pair of similar applicants (e.g. equally qualified applicants with different names) should be the same.

Research

5 minute read

## New research helps make AI fairer in decision-making

Our team developed the first practical procedures and tools for achieving Individual Fairness in machine learning (ML) and artificial intelligence (AI) systems.

# inFairness team

Mikhail Yurochkin

Onkar Bhardwaj

Mayank Agarwal

Aldo Pareja

# Collaborators

*University of Michigan*: Yuekai Sun, Amanda Bower, Songkai Xue, Debarghya Mukherjee, Moulinath Banerjee, Alexander Vargo, Fan Zhang, Subha Maity, Hamid Eftekhari

*IBM Research*: Mark Weber, Ben Hoover, Mayank Agarwal, Aldo Pareja, Onkar Bhardwaj, Uri Kartoun, Bum Chul Kwon, Kenney Ng, Zahra Ashktorab

*University of Konstanz*: Felix Petersen

*Wells Fargo*: Sherif Botros, Vanio Markov

# Thank You!

Paper links, videos, news, and code are on my website
moonfolk.github.io