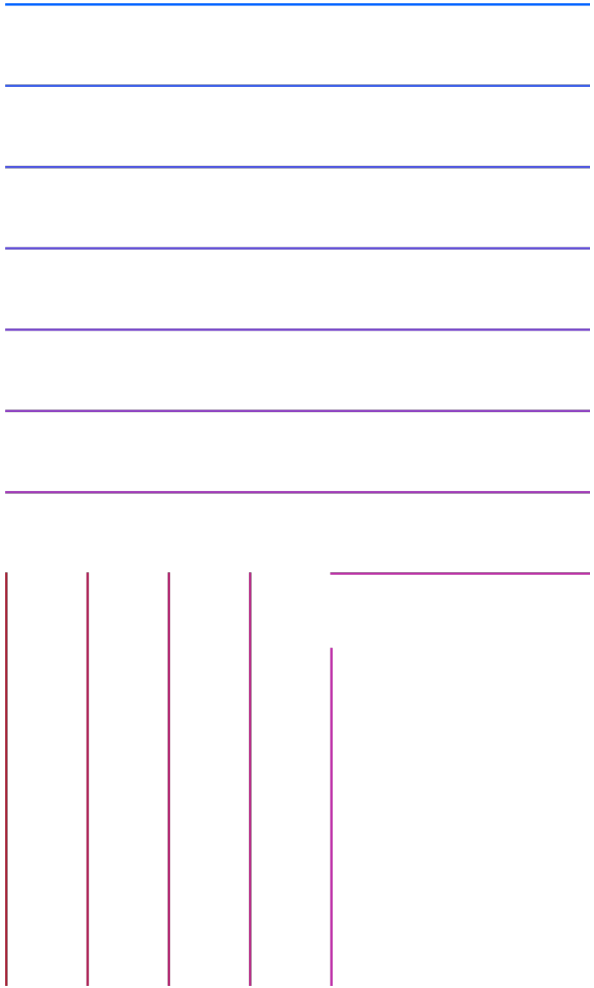
MIT-IBM
Watson AI Lab

IBM **Research**

Practical Individual Fairness

Mikhail Yurochkin



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

2019 study of 13.2 million mortgage and refinancing applications

BLACK & LATINX	EVERYONE ELSE
61% rejection rates	48%
+5.3 basis points (bps) on mortgage interest rates for fintech lending	
+7.9 bps on overall mortgage interest rates	→ $756M annual race premium

Bartlett, et al. “Consumer-Lending Discrimination in the FinTech Era.” NBER 2019.

The New York Times

ECONOMIC VIEW

Biased Algorithms Are Easier to Fix Than Biased People

Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.

Tim Cook

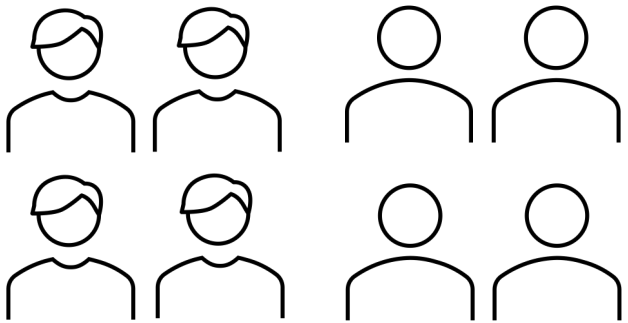
Roadmap

- AI is prone to biases
- What is a fair algorithm
- Distributional Individual Fairness (DIF)
- Enforcing DIF
- Subgroup Fairness
- Learning the Fair Metric

What is a fair algorithm?

Group Fairness:

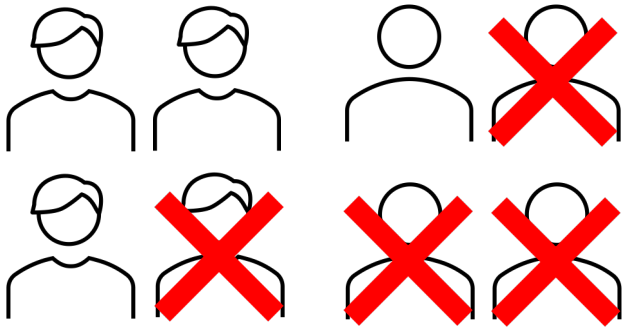
Algorithm is equally good on groups of individuals



What is a fair algorithm?

Group Fairness:

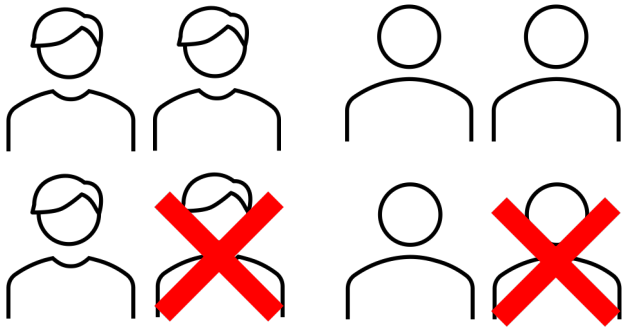
Algorithm is equally good on groups of individuals



What is a fair algorithm?

Group Fairness:

Algorithm is equally good on groups of individuals



What is a fair algorithm?

Group Fairness:

Algorithm is equally good on groups of individuals

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

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

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

Statistical Parity: $\hat{Y} \perp\!\!\!\perp A$

Equalized Odds: $\hat{Y} | Y \perp\!\!\!\perp A$

Most prior work is on group fairness

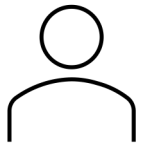
What is a fair algorithm?

Individual Fairness:

Algorithm treats similar individuals similarly

Ryan earns 50k and lives in a predominantly white zip-code

Tyree earns 50k and lives in a predominantly black zip-code

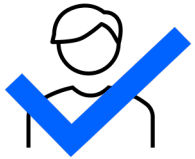


What is a fair algorithm?

Individual Fairness:

Algorithm treats similar individuals similarly

Ryan earns 50k and lives in a predominantly white zip-code



Tyree earns 50k and lives in a predominantly black zip-code

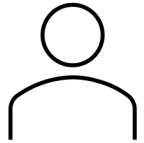
What is a fair algorithm?

Individual Fairness:

Algorithm treats similar individuals similarly

Ryan earns 50k and lives in a predominantly white zip-code

Tyree earns 50k and lives in a predominantly black zip-code



What is a fair algorithm?

Individual Fairness (Dwork et al. 2012):

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

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

Why Individual Fairness?

Example: Sentiment analysis – classify words as positive or negative

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

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

Why Individual Fairness?

Example: Sentiment analysis – classify words as positive or negative

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

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

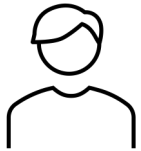
Deep Learning + Word Embeddings -> **95%** test accuracy.

Success 

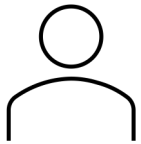
Why Individual Fairness?

What is a sentiment of a name?

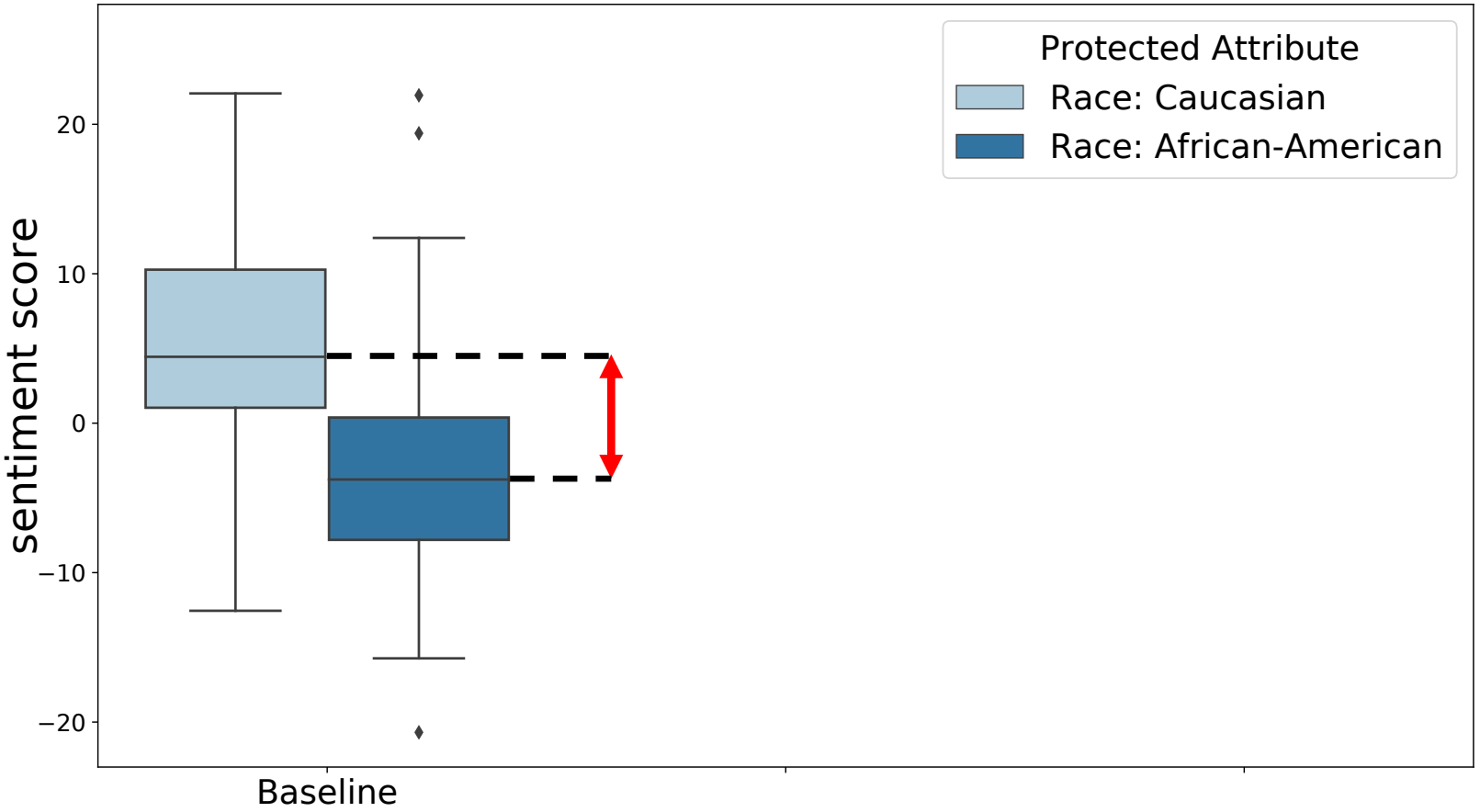
European-American names: *Adam, Ryan, Paul, ... , Courtney, Meredith, Megan, ...*



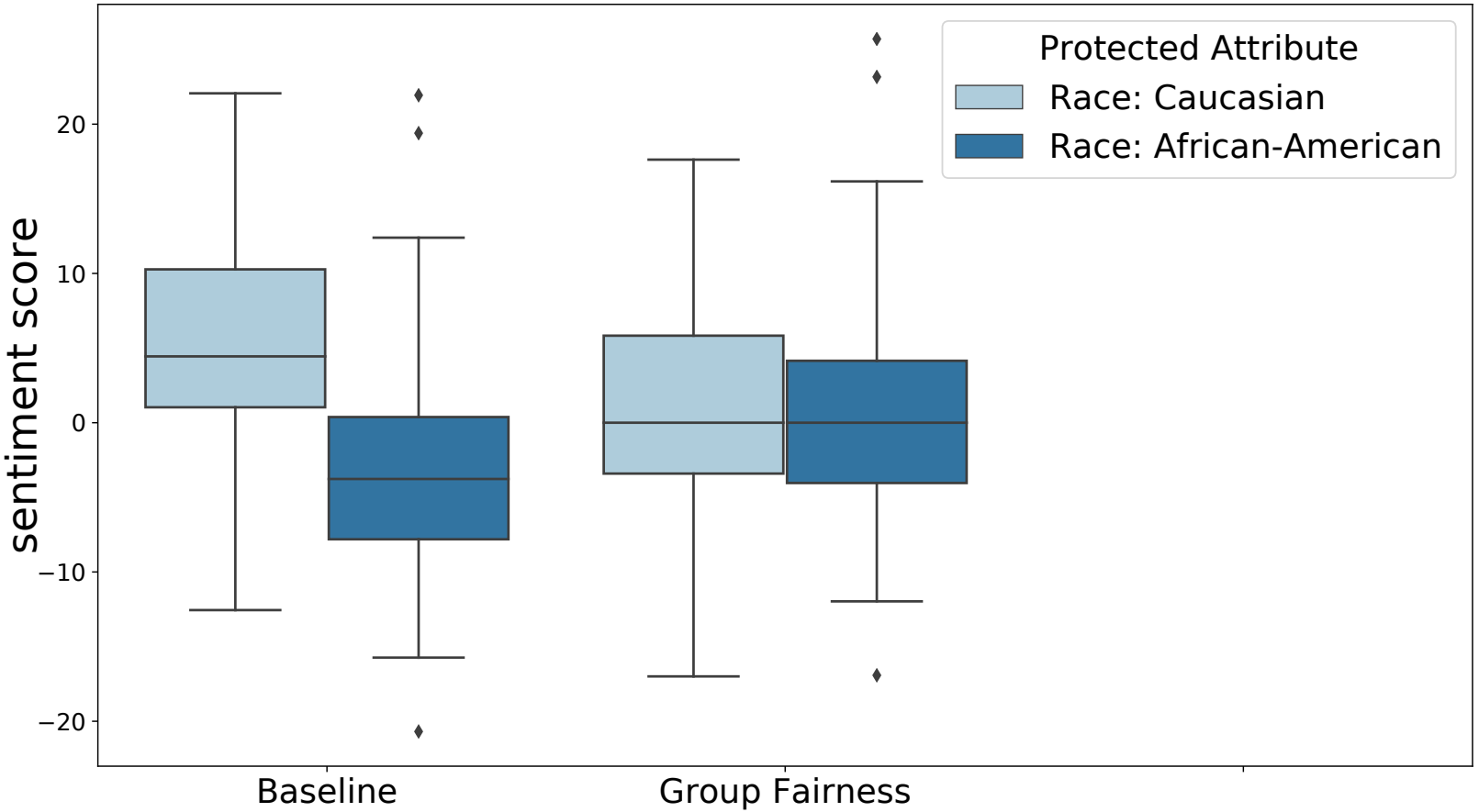
African-American names: *Alonzo, Leroy, Tyree, ... , Shereen, Sharise, Tawanda, ...*



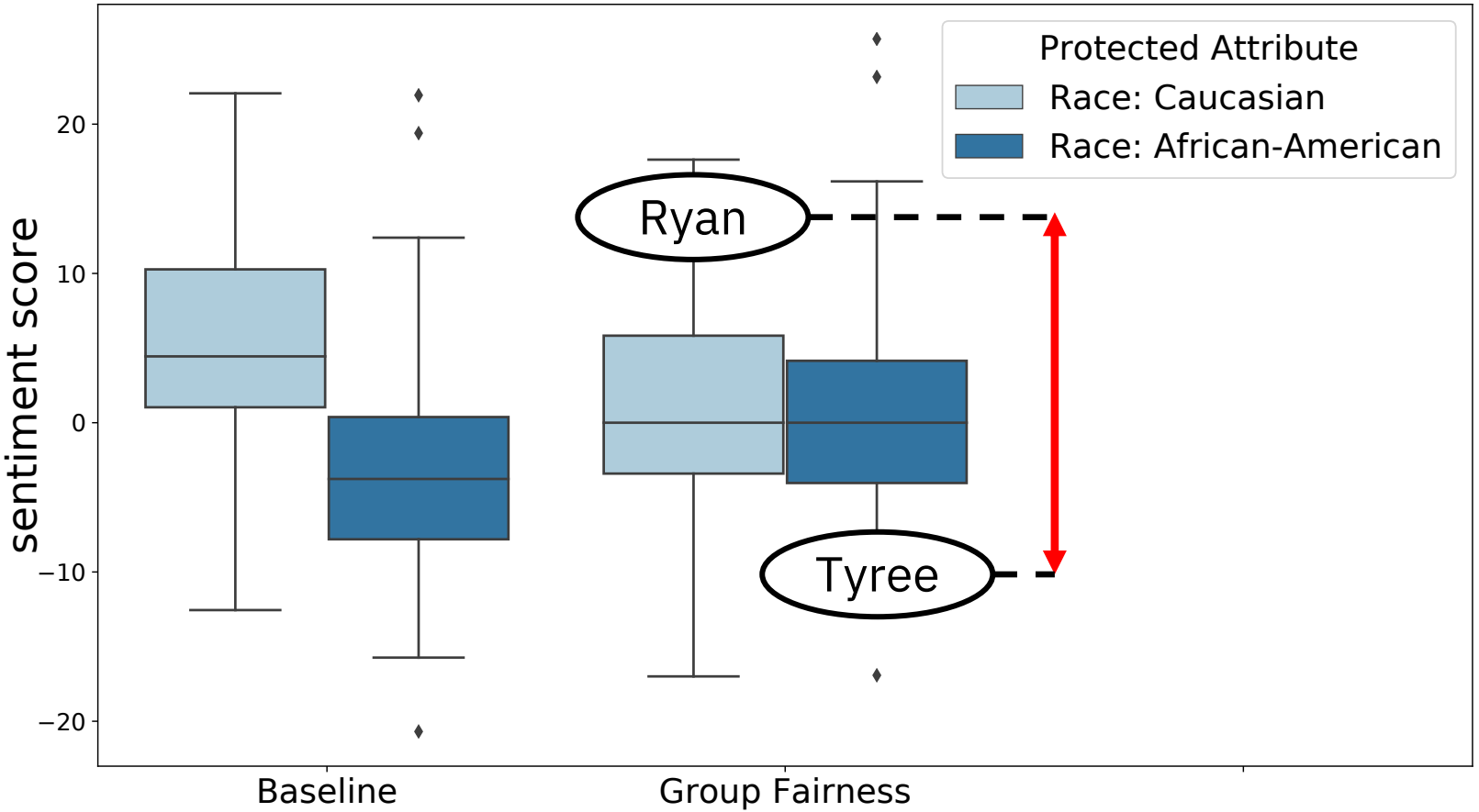
Why Individual Fairness?



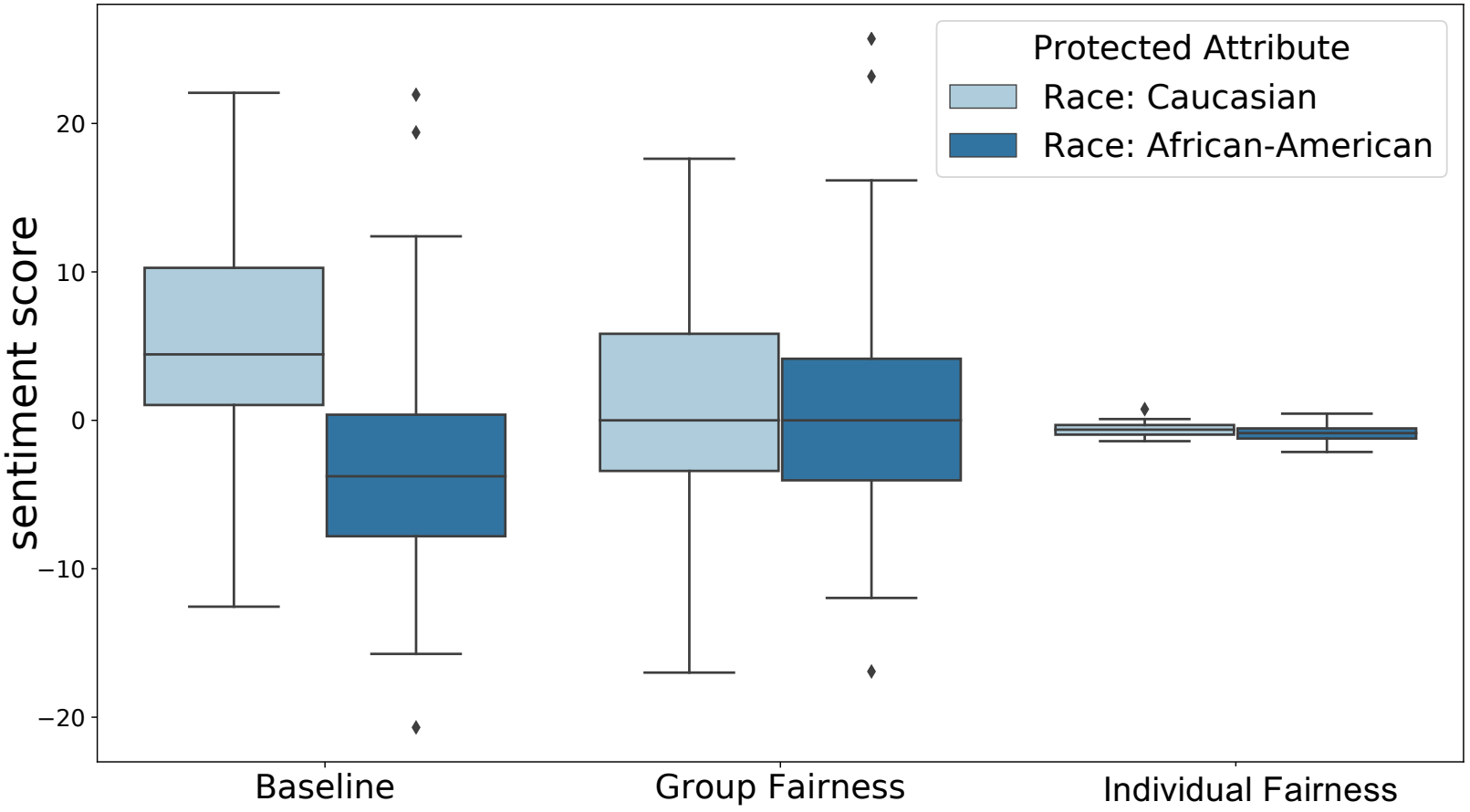
Why Individual Fairness?



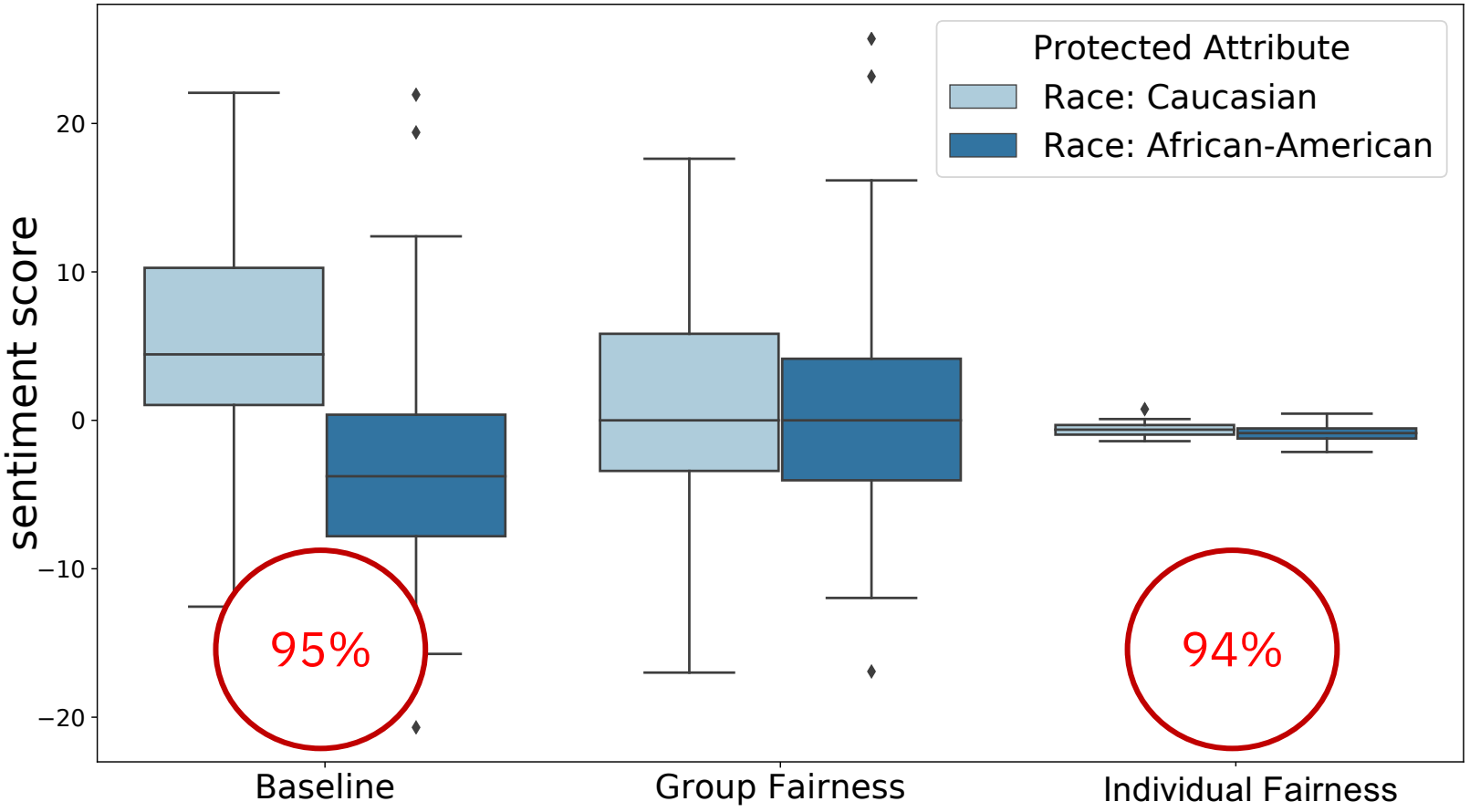
Why Individual Fairness?



Why Individual Fairness?



Why Individual Fairness?



Roadmap

- AI is prone to biases
- What is a fair algorithm
- Distributional Individual Fairness (DIF)
- Enforcing DIF
- Subgroup Fairness
- Learning the Fair Metric

Assessing Individual Fairness

$$\mathrm{IF}(h, x_i) \triangleq \left\{ \begin{array}{ll} \max\limits_{x_i' \in \mathcal{X}} & d_{\mathcal{Y}}(h(x_i), h(x_i')) \\ \text{subject to} & d_{\mathcal{X}}(x_i, x_i') \leq \epsilon \end{array} \right\}$$

- $d_{\mathcal{X}}$: fair metric
- $d_{\mathcal{Y}}$: metric on outputs
- ϵ: small tolerance parameter

Distributional Individual Fairness (DIF)

$$\mathrm{DIF}(h) \triangleq \begin{Bmatrix} \sup_{T:\mathcal{X}\to\mathcal{X}} & \mathbb{E}_{P_X}\left[d_{\mathcal{Y}}(h(x), h(T(x)))\right] \\ \text{subject to} & \mathbb{E}_{P_X}\left[d_{\mathcal{X}}(x, T(x))\right] \leq \epsilon. \end{Bmatrix}$$

- P_X (marginal) distribution of inputs
- Optimal T maps x_i to x_i'
- Constraint enforces $d_{\mathcal{X}}(x_i, x_i') \leq \epsilon$ **on average**

Comparing DIF and IF

Individual fairness map is feasible, but may not be optimal:

$$T_{\mathrm{IF}}(x_i) \triangleq \underset{d_{\mathcal{X}}(x_i, x_i') \leq \epsilon}{\arg\max} \; d_{\mathcal{Y}}(h(x_i), h(x_i')).$$

DIF may transport some points by more than ϵ.

IF does not imply DIF or vice a versa!

Comparing DIF and IF

Individual fairness map is feasible, but may not be optimal:

$$T_{\mathrm{IF}}(x_i) \triangleq \underset{d_{\mathcal{X}}(x_i, x_i') \leq \epsilon}{\arg\max}\ d_{\mathcal{Y}}(h(x_i), h(x_i')).$$

$$\text{Let } \mathrm{DIF}(h) < \delta, \text{ then}$$

$$P_X(d_{\mathcal{Y}}(h(x), h(T_{\mathrm{IF}}(x))) \geq \tau) \leq \tfrac{\delta}{\tau} \text{ for any } \tau > 0.$$

DIF implies IF with high probability

DIF in Social Science

$$\mathrm{DIF}(h) \triangleq \left\{ \begin{array}{ll} \sup_{T:\mathcal{X}\to\mathcal{X}} & \mathbb{E}_{P_X}\left[d_{\mathcal{Y}}(h(x), h(T(x)))\right] \\ \text{subject to} & \mathbb{E}_{P_X}\left[d_{\mathcal{X}}(x, T(x))\right] \leq \epsilon. \end{array} \right\}$$

- P_X (marginal) distribution of inputs
- Optimal T maps x_i to x_i'
- Large $\mathrm{DIF}(h)$ implies unfairness

DIF in Social Science

Bertrand & Mullainathan (2004) studied racial bias in the US labor market.

- P_X (marginal) distribution of inputs
- Optimal T maps x_i to x_i'
- Large DIF(h) implies unfairness

DIF in Social Science

Bertrand & Mullainathan (2004) studied racial bias in the US labor market.

- The investigators responded to job ads in Boston and Chicago newspapers with fictitious resumes.
- Optimal T maps x_i to x_i'
- Large DIF(h) implies unfairness

DIF in Social Science

Bertrand & Mullainathan (2004) studied 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.
- Large DIF(h) implies unfairness

DIF in Social Science

Bertrand & Mullainathan (2004) studied 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.

Roadmap

- AI is prone to biases
- What is a fair algorithm
- Distributional Individual Fairness (DIF)
- Enforcing DIF
- Subgroup Fairness
- Learning the Fair Metric

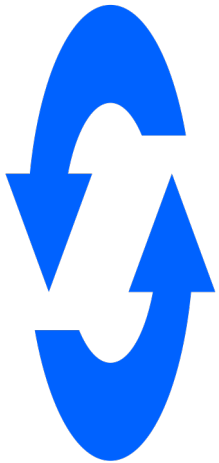
Training DIF models

Usual AI: train algorithm accurate on the available data

- Observe data
- Update parameters to minimize prediction error
- Repeat

Training DIF models

Individually Fair AI: train algorithm accurate on the available data **and** all possible similar data



- Observe data
- Generate similar data where algorithm performs differently to evaluate DIF
- Update parameters to minimize prediction error **and** DIF
- Repeat

Training DIF models

$$\min_{h \in \mathcal{H}} L(h) + \rho \mathrm{DIF}(h)$$

$$L(h) \triangleq \mathbb{E}\big[\ell(y, h(x))\big]$$

- $\mathcal{H}$: model class (e.g. neural nets with a certain architecture)
- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function
- ρ: regularization parameter

Training DIF models

$$\mathrm{DIF}(h) = \inf_{\lambda>0}\{\lambda\epsilon + \mathbb{E}_{P_X}\left[r_\lambda(h,x)\right]\}$$

$$r_\lambda(h,x) \triangleq \sup_{x'\in\mathcal{X}}\{d_\mathcal{Y}(h(x),h(x')) - \lambda d_\mathcal{X}(x,x')\}$$

DIF is infinite-dimensional, but its dual is univariate

Training DIF models

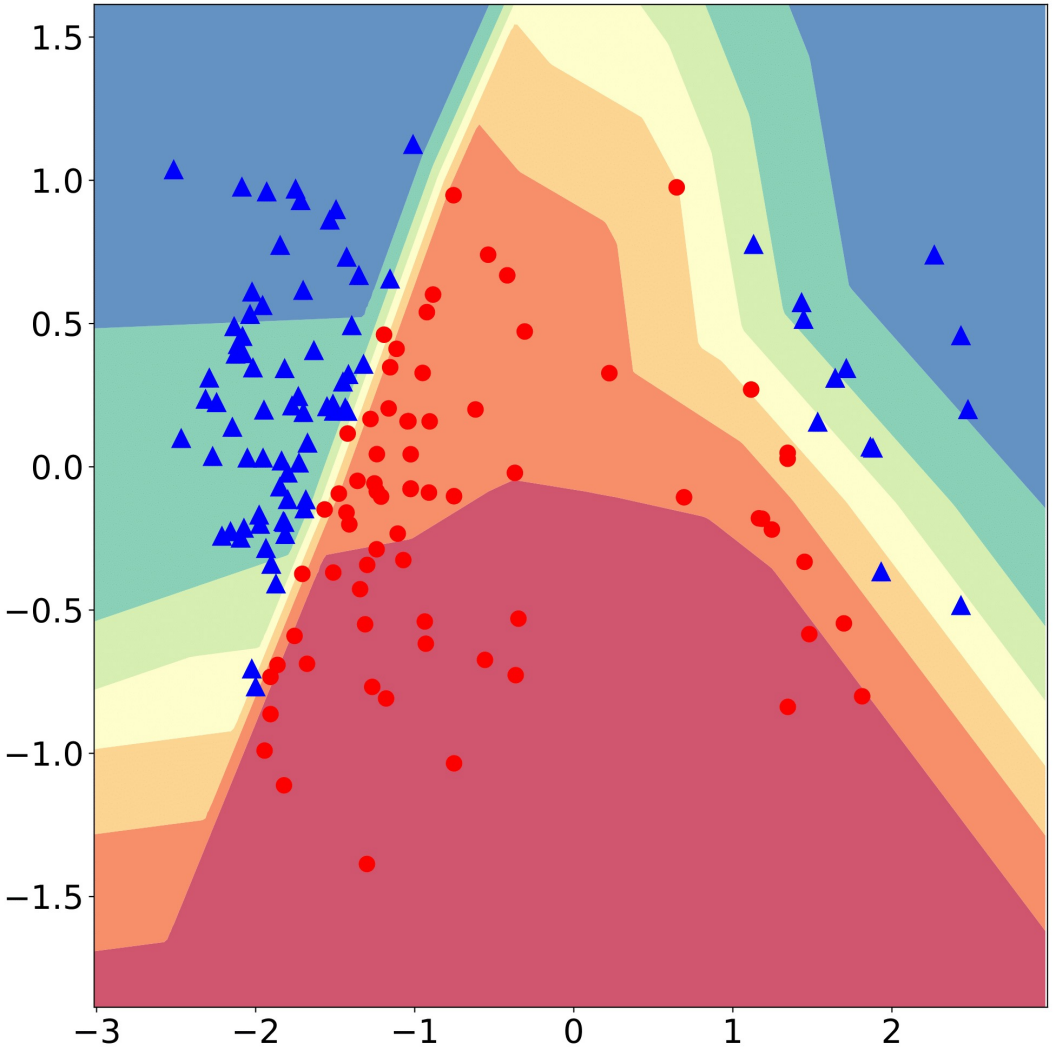
$$\min_{\theta \in \Theta, \lambda \in \mathbb{R}_+} \mathbb{E}\big[f(\theta, (x, y))\big]$$

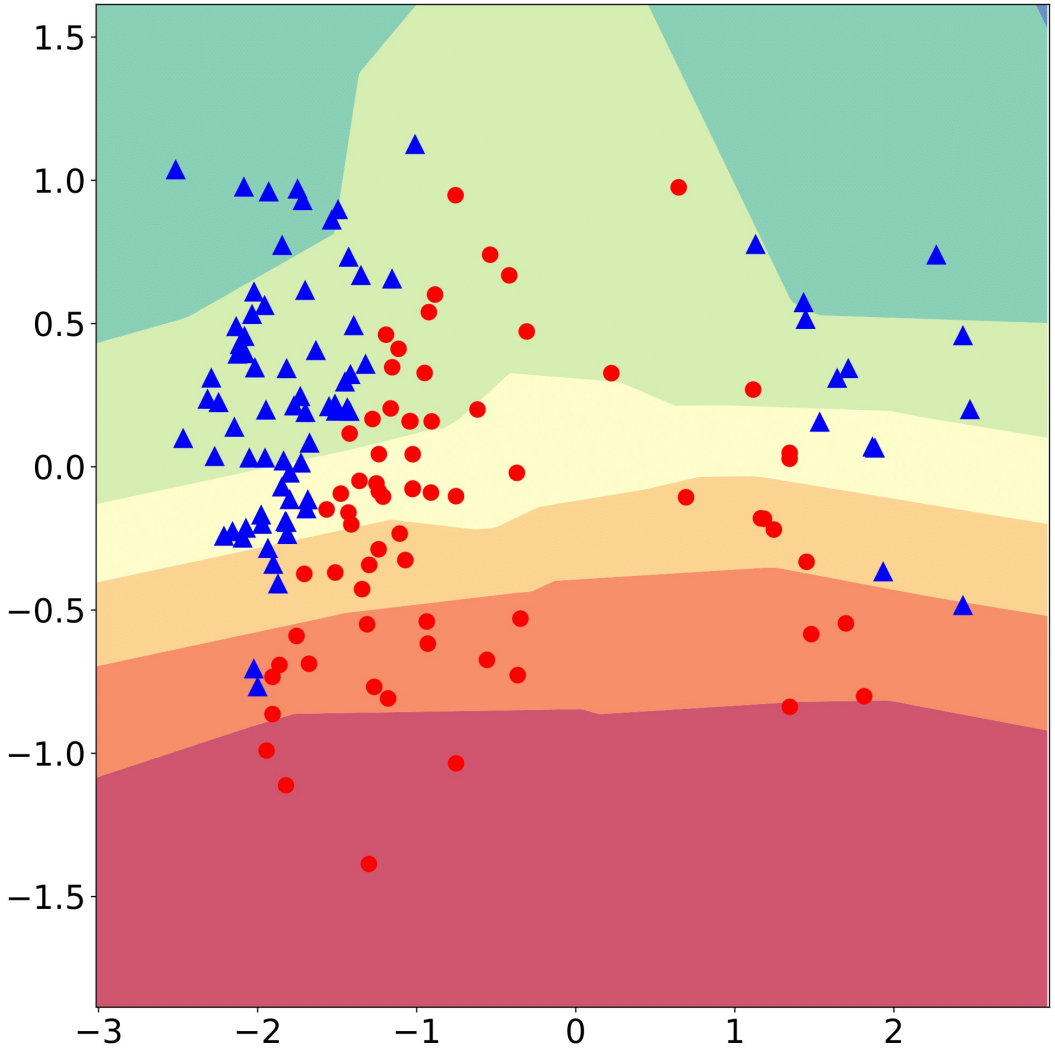
$$f(\theta, (x, y)) \triangleq \ell(y, h_\theta(x)) + \rho(\lambda\epsilon + r_\lambda(h_\theta, x))$$

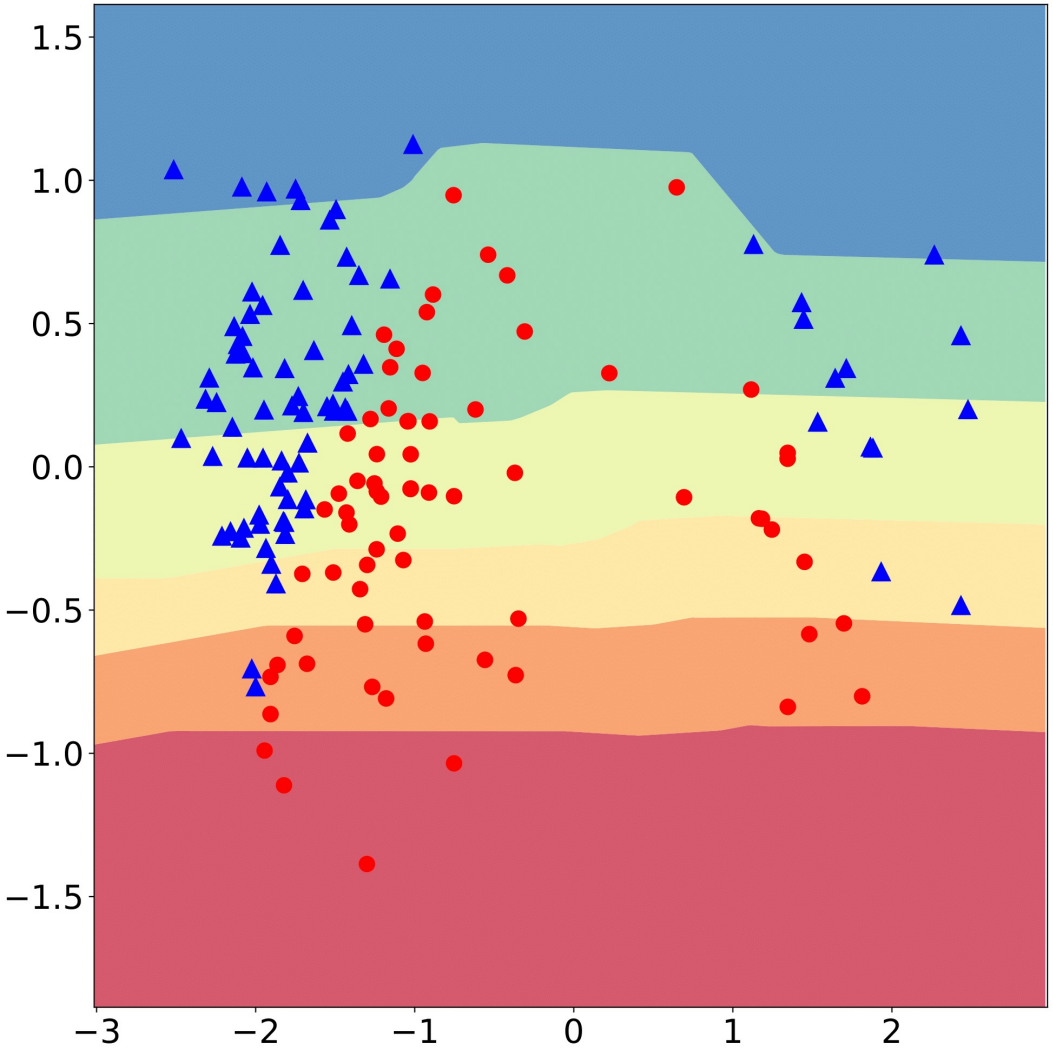
DIF is amendable to stochastic optimization

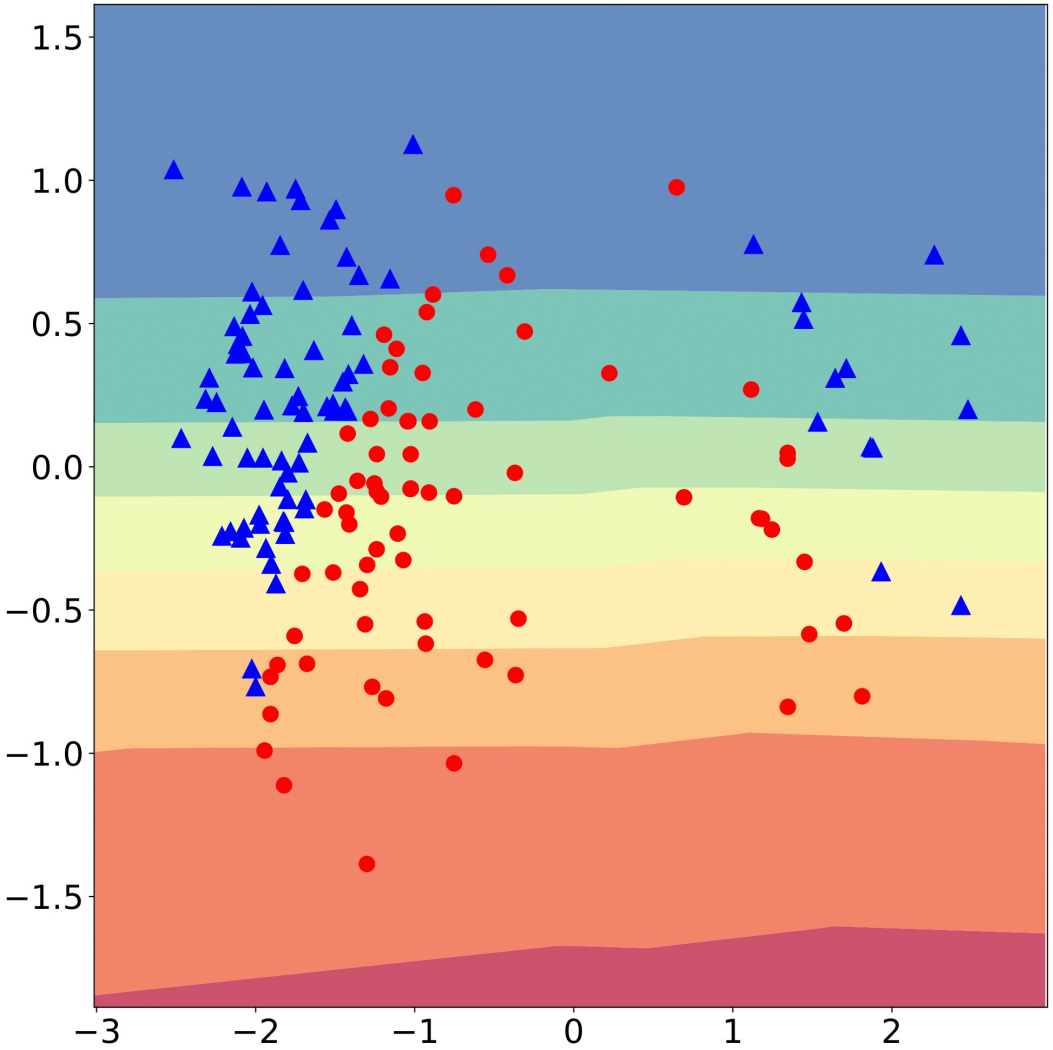
DIF Example

- Points on horizontal lines (identical y-values) are similar
- Training data is biased: $P_{Y|X}$ not constant on horizontal lines
- Use DIF regularizer to correct bias in training data


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2

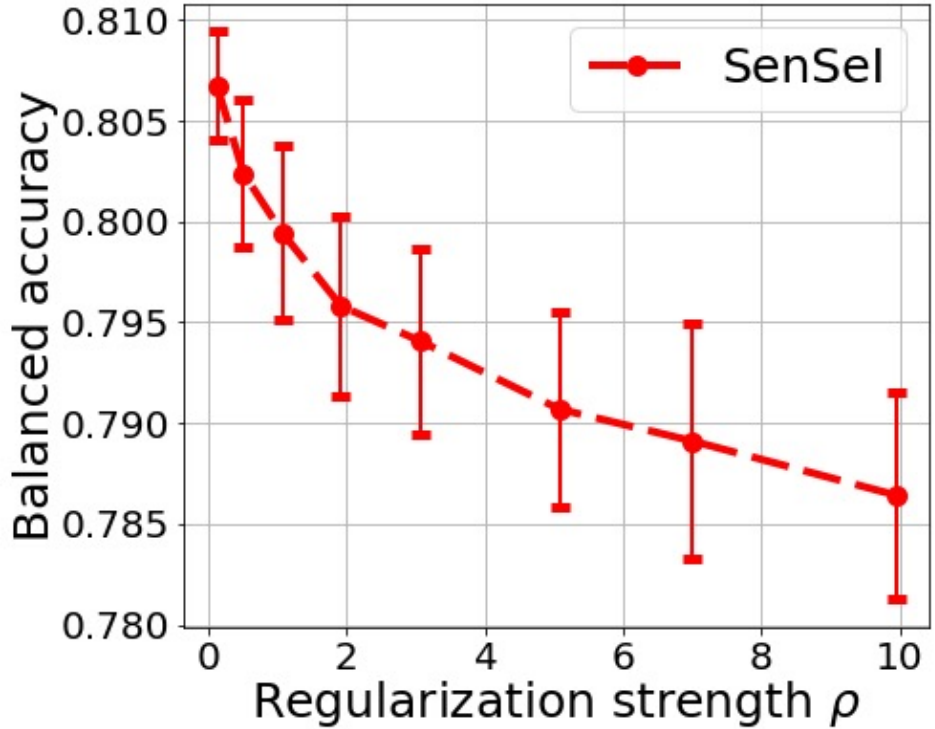

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2

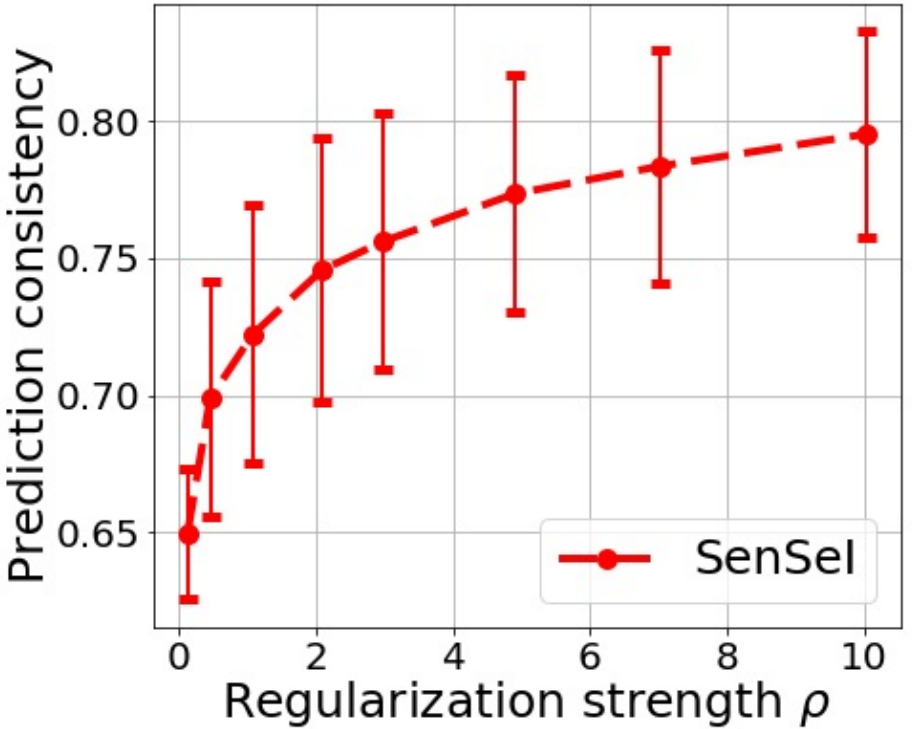

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2

Achieving IF on toxicity classification

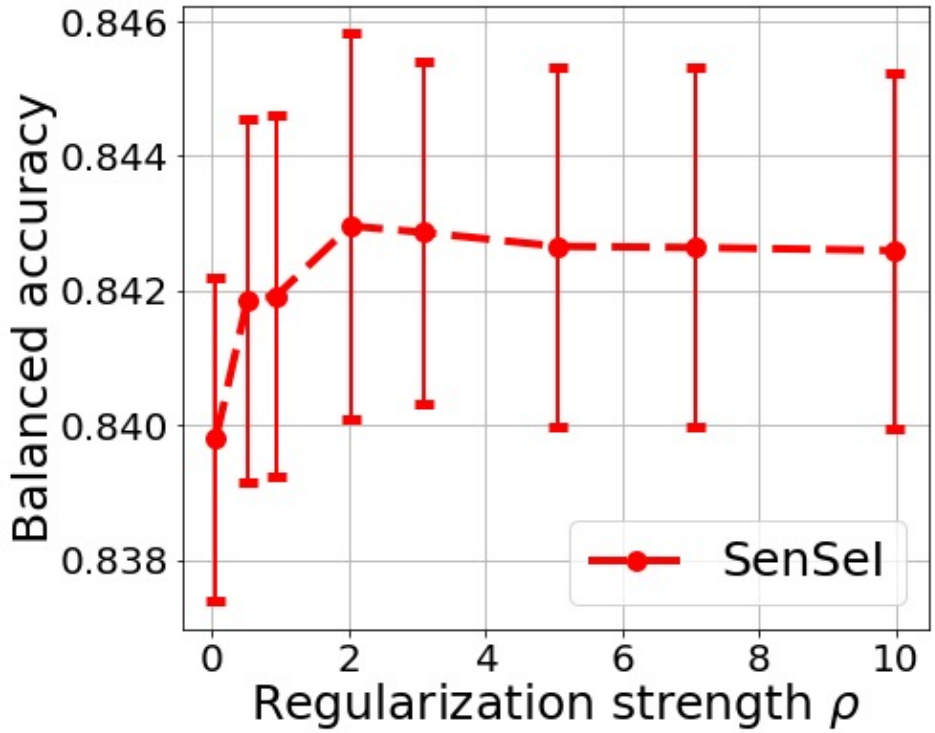
- Data: “Toxic Comment Classification Challenge”. 165k text comments
- Task: predict if a comment is toxic or not
- Measuring IF: does prediction change when changing identity tokens. Are predictions for “Some people are gay” and “Some people are straight” the same?
- 50 identity tokens: prediction on all variations should be the same to satisfy IF

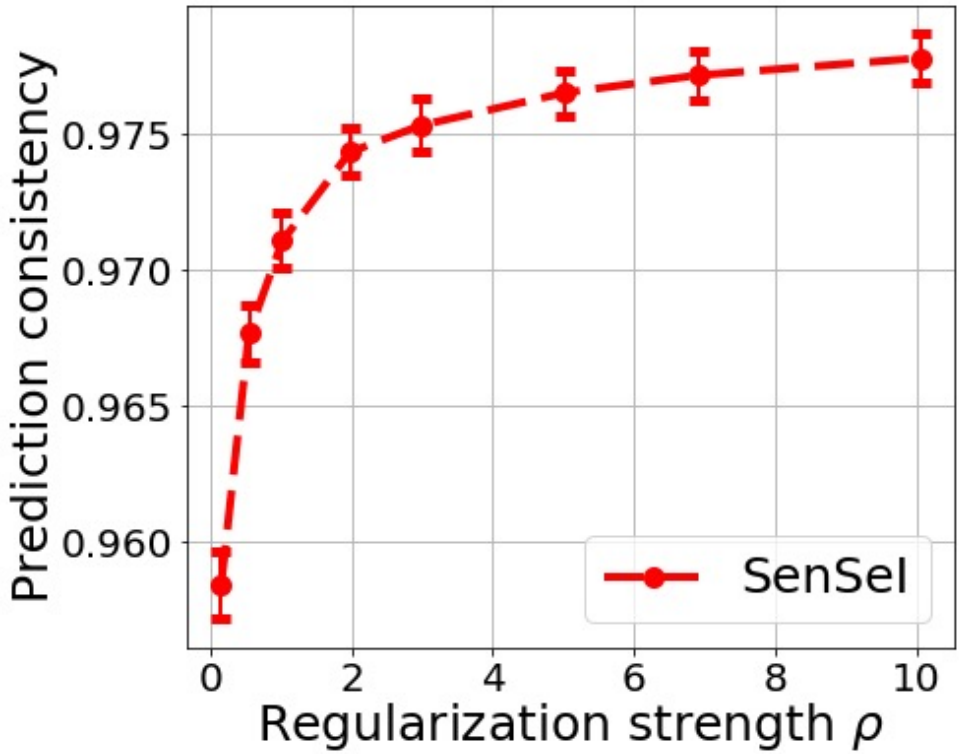

Balanced accuracy
0.810
0.805
0.800
0.795
0.790
0.785
0.780
0
2
4
6
8
10
Regularization strength ρ
SenSeI


Prediction consistency
0.80
0.75
0.70
0.65
0
2
4
6
8
10
Regularization strength ρ
SenSeI

Achieving IF on occupation prediction

- Data: Bias in Bios (De-Arteaga et al. 2019). 400k textual bio descriptions
- Task: predict one of the 28 occupations from a bio
- Measuring IF: does prediction change when changing names and gender pronouns in a bio?

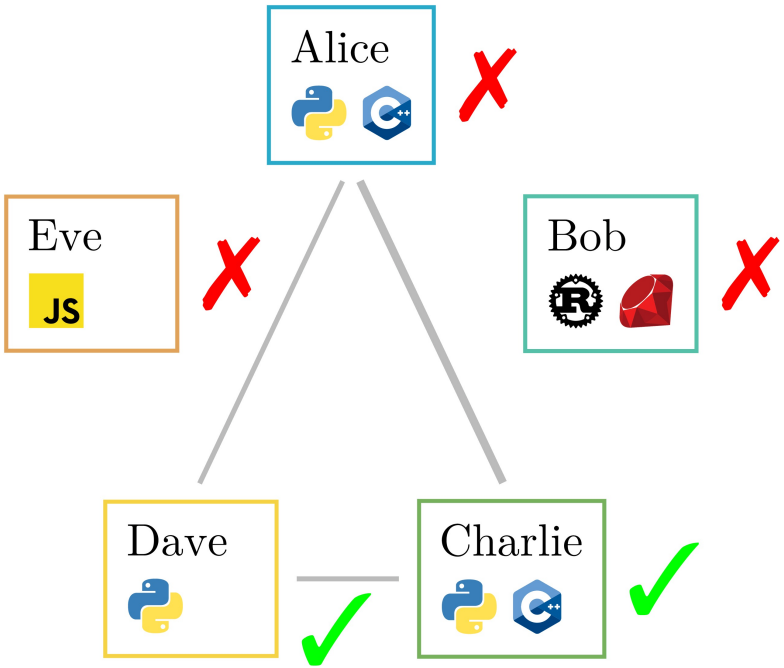

0.846
0.844
0.842
0.840
0.838
Balanced accuracy
0
2
4
6
8
10
Regularization strength ρ
SenSeI


0.975
0.970
0.965
0.960
Prediction consistency
0
2
4
6
8
10
Regularization strength ρ
SenSeI

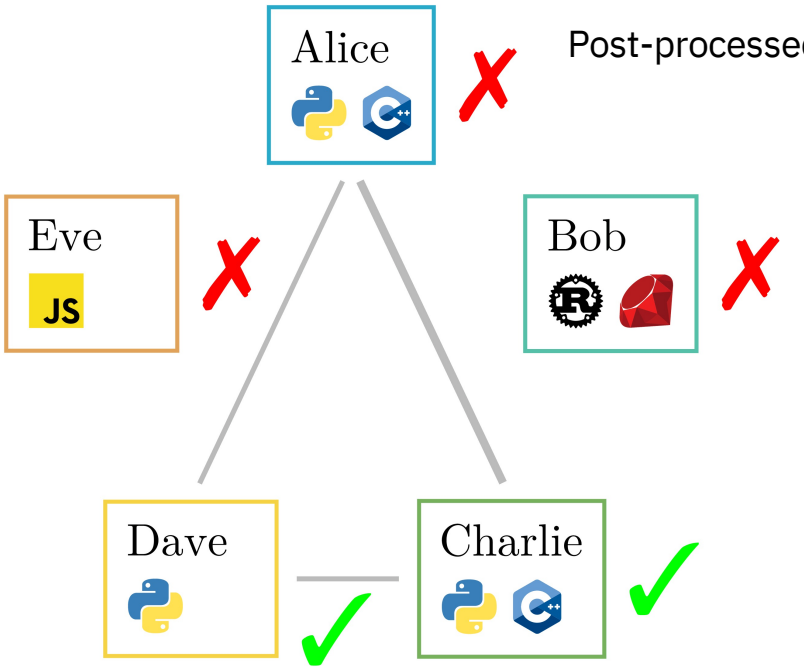
Post-processing for Individual Fairness

- Works with any trained model without re-training
- Fast and easy to implement
- Does not require knowledge of the model parameters
- Only needs access to the outputs
- Key idea: represent individuals as a graph and quantify fairness with the graph Laplacian quadratic form

Post-processing for Individual Fairness



Post-processing for Individual Fairness



Original predictions

Graph Laplacian

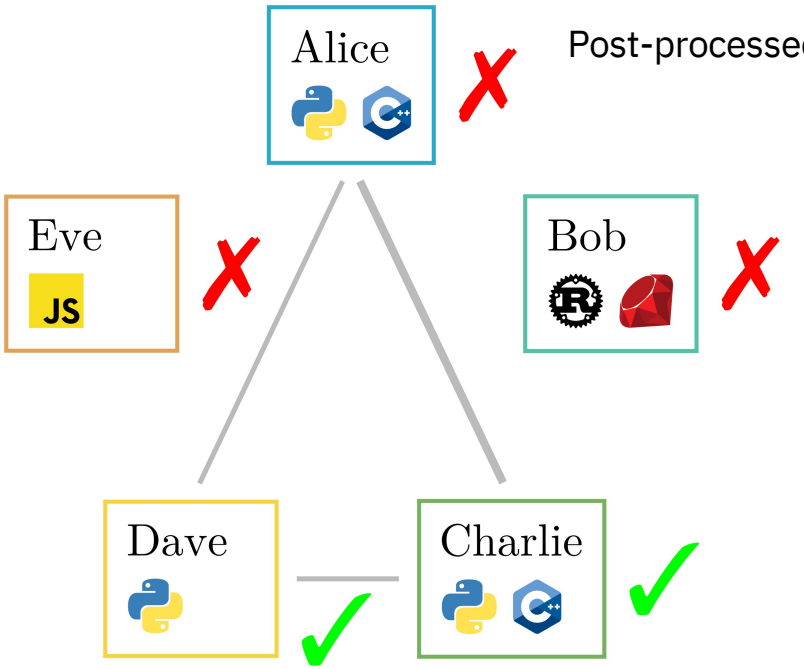
Post-processed fair predictions

$$\arg\min_{\mathbf{f}} \|\mathbf{f} - \hat{\mathbf{y}}\|_2^2 + \lambda\, \mathbf{f}^\top \mathbb{L} \mathbf{f}$$

Stay close to original predictions

Penalize individual fairness violations

Post-processing for Individual Fairness



Original predictions

Graph Laplacian

Post-processed fair predictions

$$\arg\min_{\mathbf{f}} \|\mathbf{f} - \hat{\mathbf{y}}\|_2^2 + \lambda\, \mathbf{f}^\top \mathbb{L} \mathbf{f}$$

Stay close to original predictions

Penalize individual fairness violations

Closed-form solution!

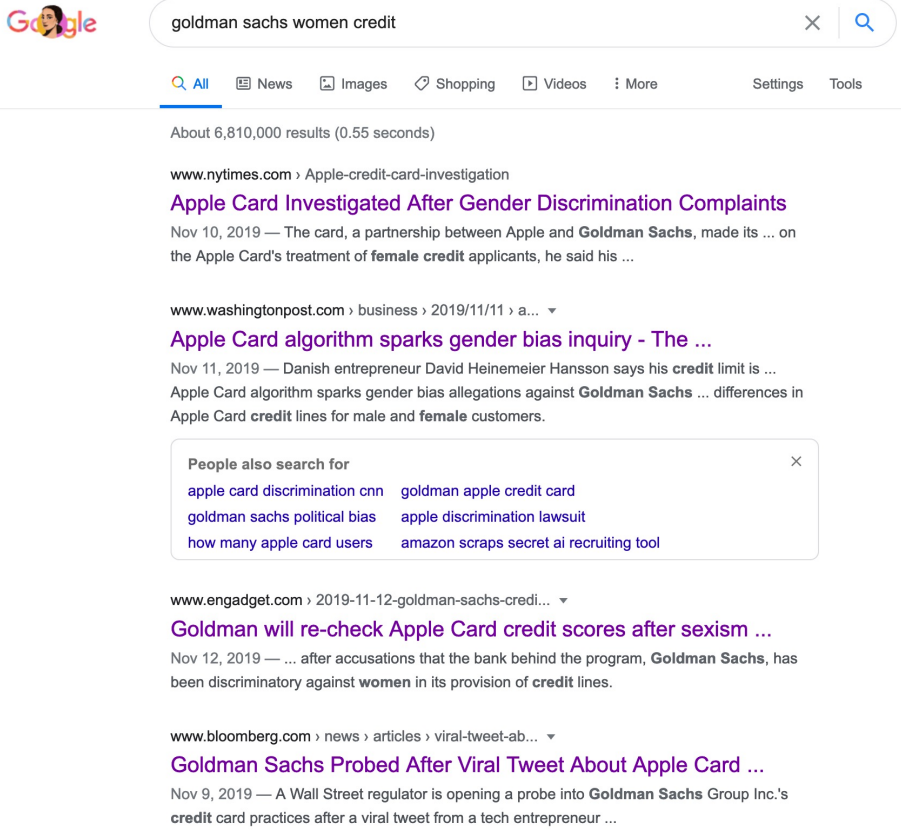
$$\mathbf{f} = (I + \lambda \mathbb{L})^{-1} \hat{\mathbf{y}}$$

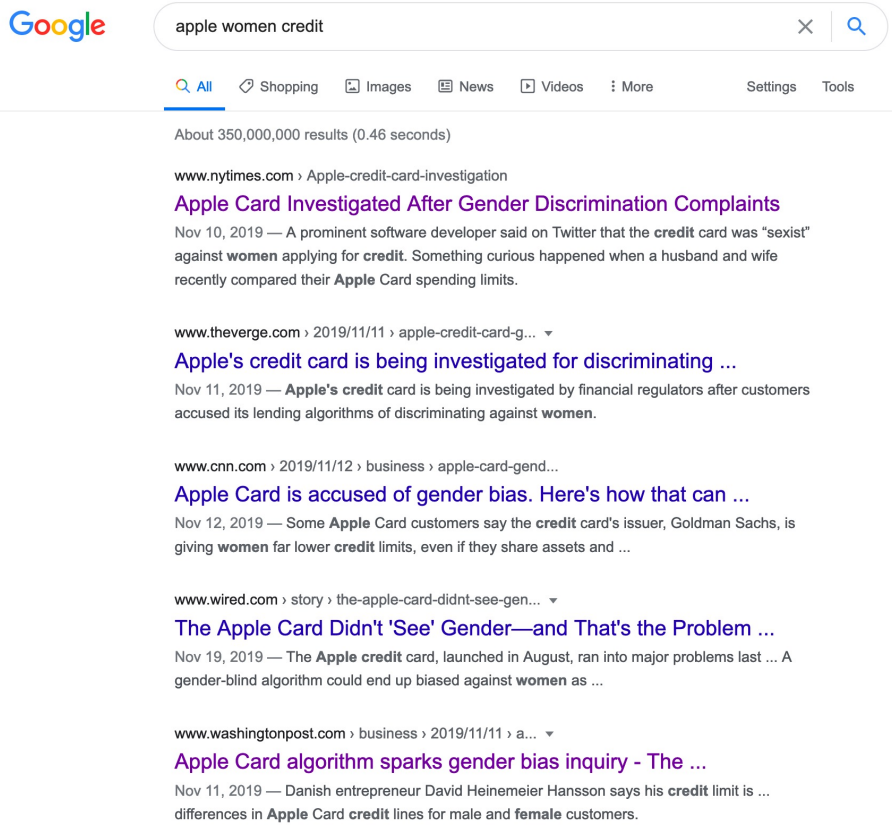
Roadmap

- AI is prone to biases
- What is a fair algorithm
- Distributional Individual Fairness (DIF)
- Enforcing DIF
- Subgroup Fairness
- Learning the Fair Metric

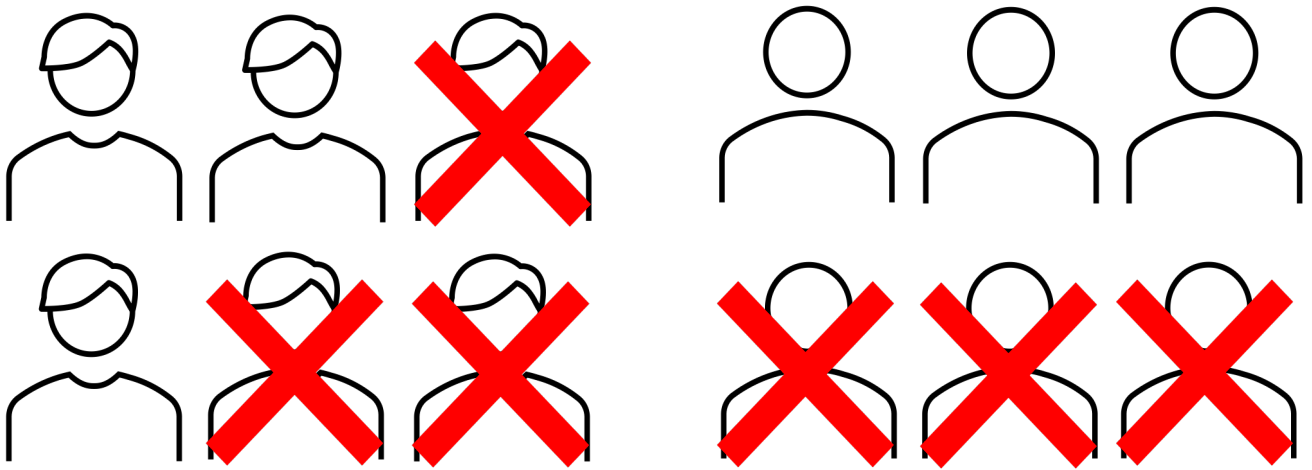
Fairness problems Today

Fairness problems Today

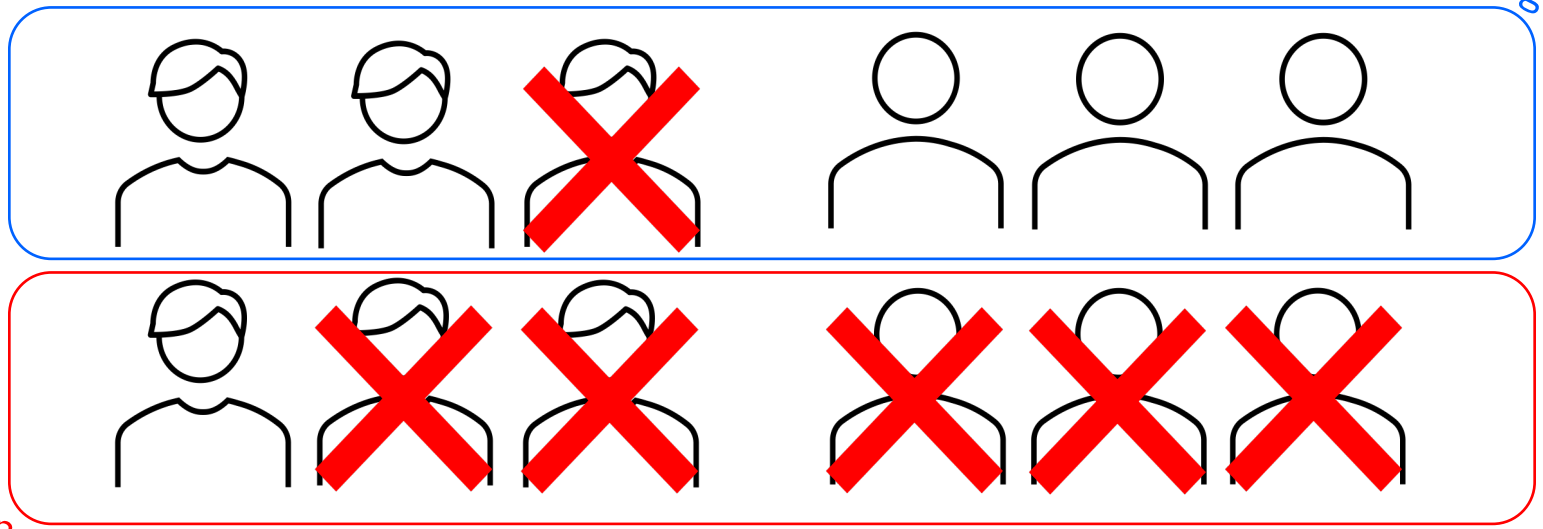




What is in-between?



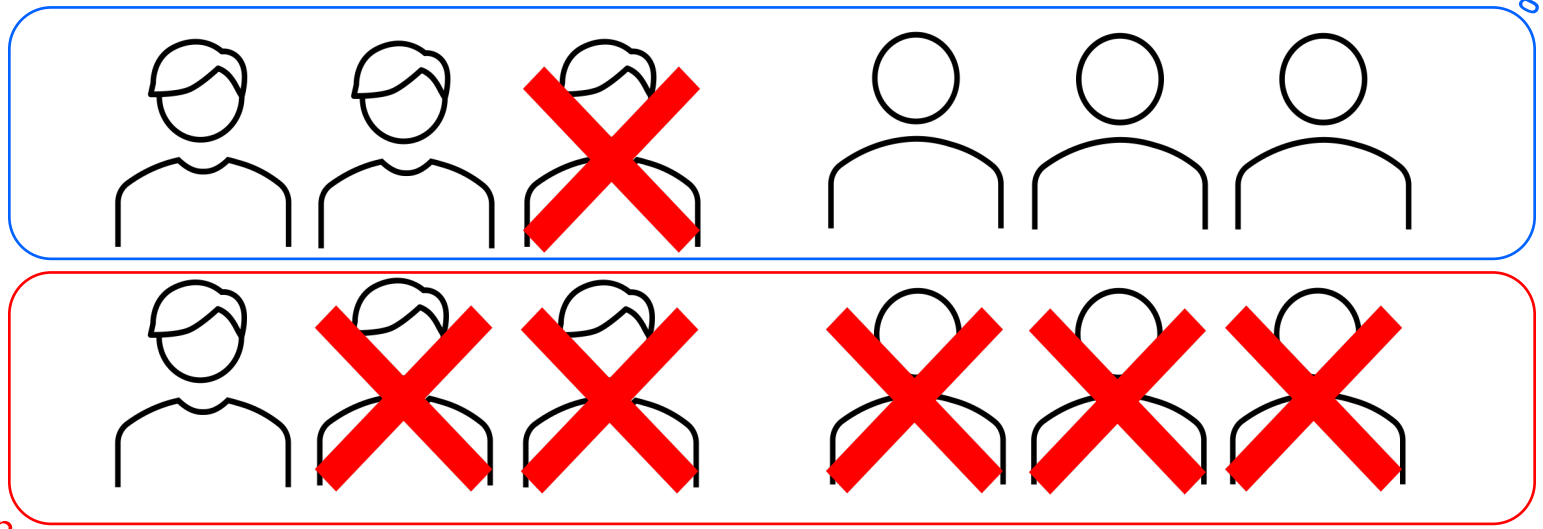
What is in-between?



Subgroup fairness violation

What is in-between?

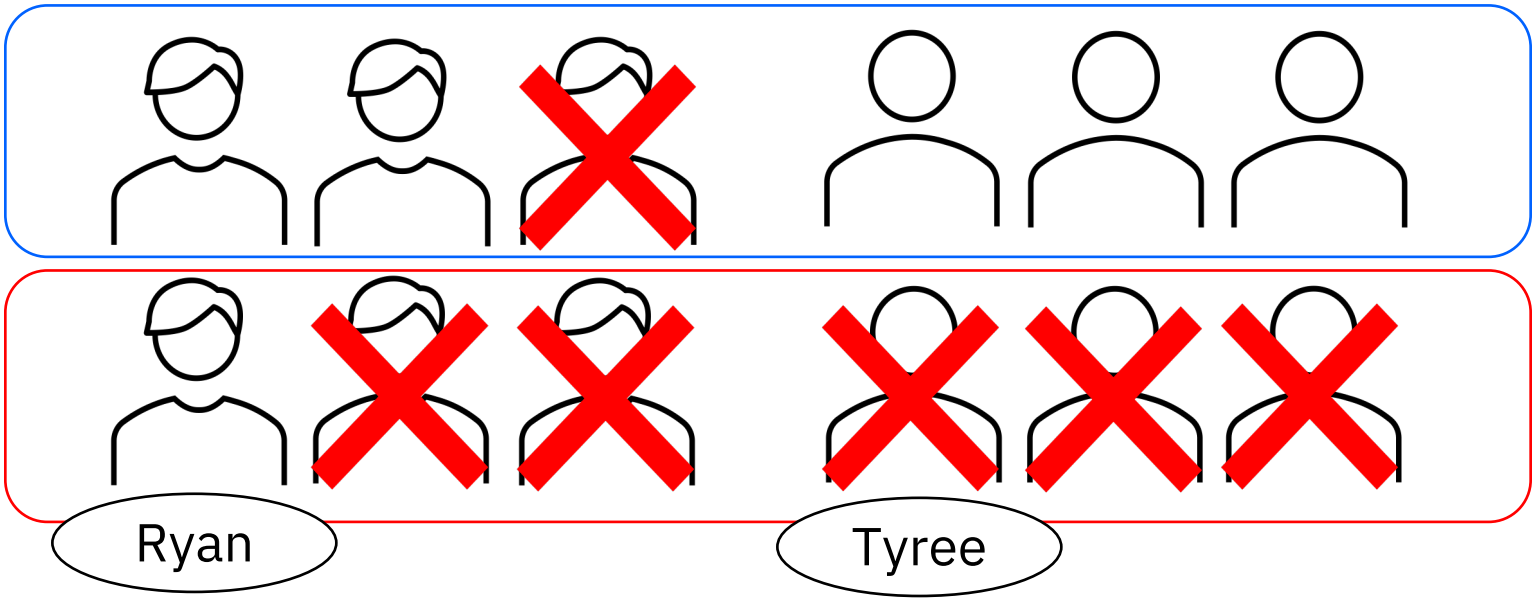
Wealthy neighborhood



Low-income neighborhood

Enforcing Subgroup fairness with Group fairness requires expert knowledge to define all subgroups

What is in-between?



Subgroup fairness can be enforced with Individual fairness *generalizing* to unforeseen subgroups

Example: Income prediction

Predict individuals' income based on Census data (ADULT dataset).

Evaluate SenSR against baseline and Adversarial Debiasing (group fairness).

	B-Acc,%	S-Con.	$\text{Gap}_G^{\text{RMS}}$	$\text{Gap}_R^{\text{RMS}}$	$\text{Gap}_G^{\text{max}}$	$\text{Gap}_R^{\text{max}}$
SenSR+EXPLORE	79.4	**.966**	**.065**	**.044**	**.084**	**.059**
SenSR	78.9	.934	.068	.055	.087	.067
Baseline	**82.9**	.848	.179	.089	.216	.105
Adv. debiasing	81.5	.807	.082	.070	.110	.078

Spousal Consistency

Spousal consistency measures counterfactual sensitivity to “relationship” status, i.e. how often the classification remains unchanged when relationship status is perturbed.

SenSR + EXPLORE is best for S-Con; the group fairness method is worse than baseline.

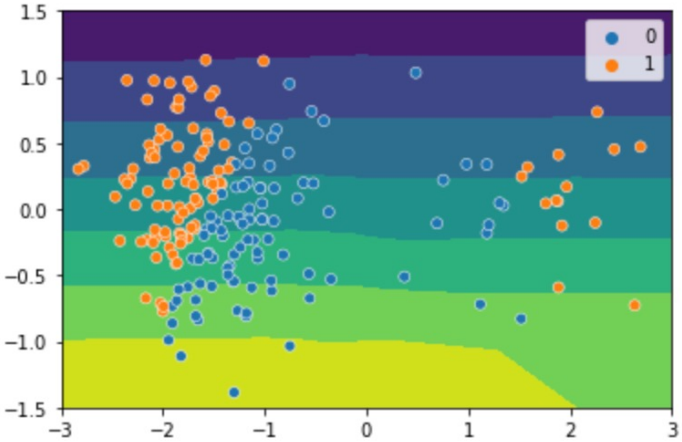
*SenSR is our earlier algorithm for enforcing individual fairness.

📣 Announcement 📣

inFairness Python package coming out next year

```
from inFairness.default_cfg import get_cfg_defaults
from inFairness.trainer import DefaultTrainer
```

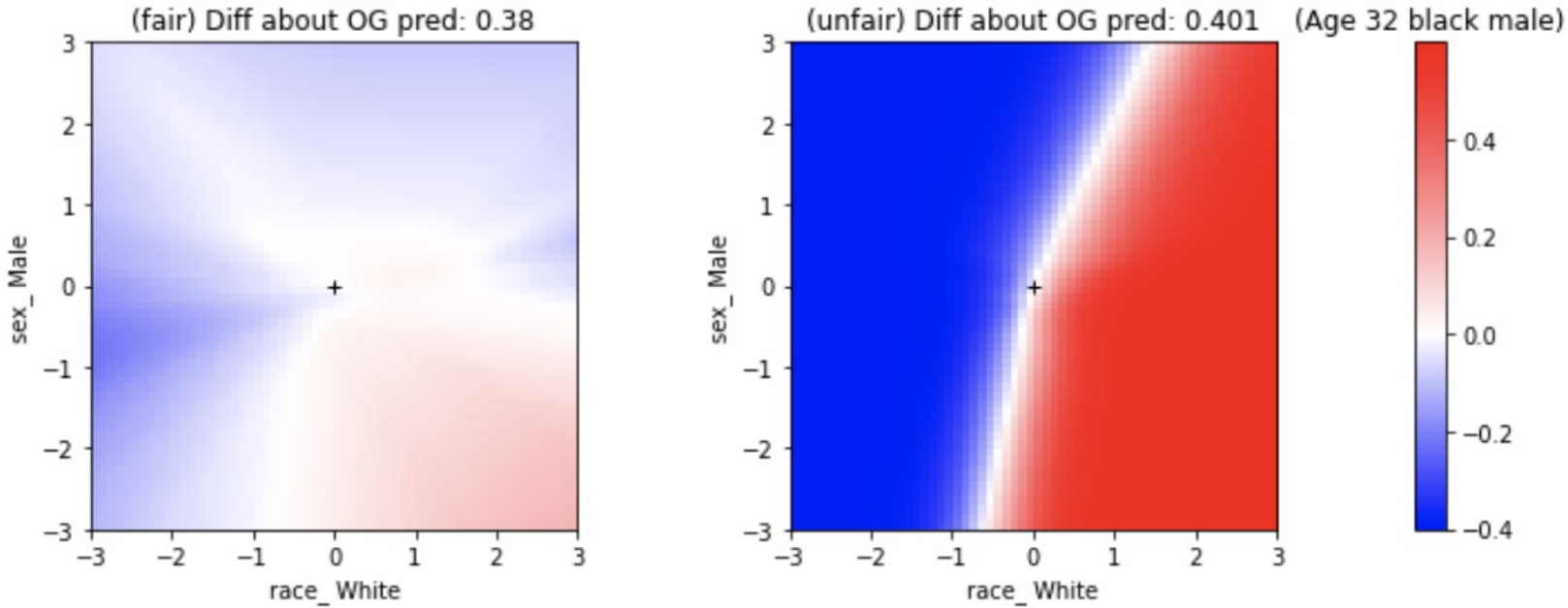
```
cfg = get_cfg_defaults()
cfg.merge_from_list(["MODEL.SENSEI_RHO", 5.0])
cfg.freeze()
print(cfg)

trainer50 = DefaultTrainer(cfg)
trainer50.train()
```



📣 Announcement 📣

Demonstrations and interactive visualization tools

▼ **Age 32 black male** working a full time job



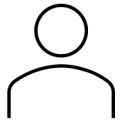
Roadmap

- AI is prone to biases
- What is a fair algorithm
- Distributional Individual Fairness (DIF)
- Enforcing DIF
- Subgroup Fairness
- Learning the Fair Metric

What are similar individuals?

Ryan earns 50k and lives in a predominantly white zip-code area

Tyree earns 50k and lives in a predominantly black zip-code area



Problem:

- Requires significant subject expertise to cover all possibilities
- Time consuming
- Very hard on some data types (images, customer records, ...)

Fair Metric Learning

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

Option 1 (EXPLORE)

Input: pairs of comparable $(y = 1)$ and incomparable $(y = 0)$ samples $\{x_{i1}, x_{i2}, y_i\}_{i=1}^n$

Fit a model to find Σ such that: $\mathbf{P}(y = 1|x_1, x_2) \propto \frac{1}{1+e^{d_x(x_1,x_2)}}$

Example: Loan applicants with similar income such as Ryan and Tyree

Fair Metric Learning

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

Option 2 (Sensitive subspace)

Input: group (or groups) of comparable samples

Find directions of major variation with PCA, i.e. $V = \{v_1, \ldots, v_K\}$.
Ignore them in the fair metric: $\Sigma = I - P_{\text{span}(V)}$.

Example: Word embeddings of popular baby names
in the sentiment classification experiment

References

- M. Yurochkin, A. Bower, and Y. Sun. Training individually fair ML models with sensitive subspace robustness. ICLR 2020 – *Spotlight Presentation.*
- 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 – *Spotlight Presentation*.
- M. Yurochkin and Y. Sun. SenSeI: Sensitive Set Invariance for Enforcing Individual Fairness. ICLR 2021 – *Oral Presentation*.
- A. Vargo, F. Zhang, M. Yurochkin, and Y. Sun. Individually Fair Gradient Boosting. ICLR 2021 – *Spotlight Presentation*.
- 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.

Blog-posts

- SenSR: the first practical algorithm for individual fairness.
  https://mitibmwatsonailab.mit.edu/research/blog/training-individually-fair-ml-models-with-sensitive-subspace-robustness

- Black Loans Matter: Fighting Bias for AI Fairness in Lending.
  https://mitibmwatsonailab.mit.edu/research/blog/black-loans-matter-fighting-bias-for-ai-fairness-in-lending

- New research helps make AI fairer in decision-making.
  https://www.research.ibm.com/blog/make-ai-fairer

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

Collaborators

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

University of Konstanz: Felix Petersen

IBM: Mark Weber, Ben Hoover, Hendrik Strobelt, Mayank Agarwal, Aldo Pareja, Onkar Bhardwaj, Ioana Baldini

Wells Fargo: Sherif Botros, Vanio Markov

Thank You!