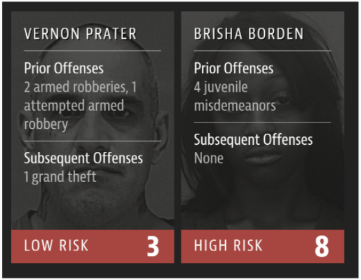
On The Impact of Machine Learning Randomness on Group Fairness

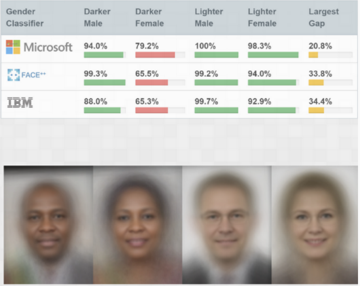
Prakhar Ganesh, Hongyan Chang, Martin Strobel, Reza Shokri

FAccT 2023

Machine Learning has a Fairness Problem



Bias in recidivism

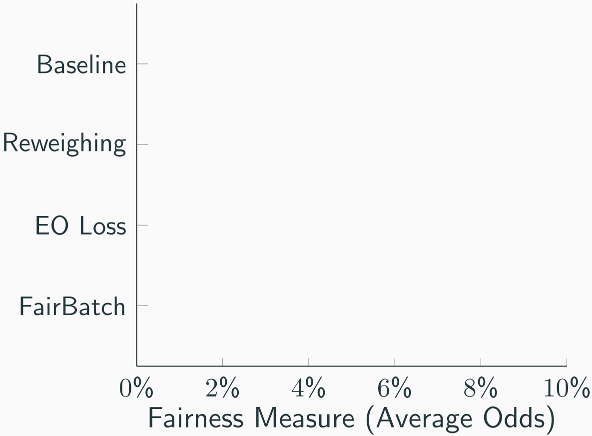


Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Bias in gender classification

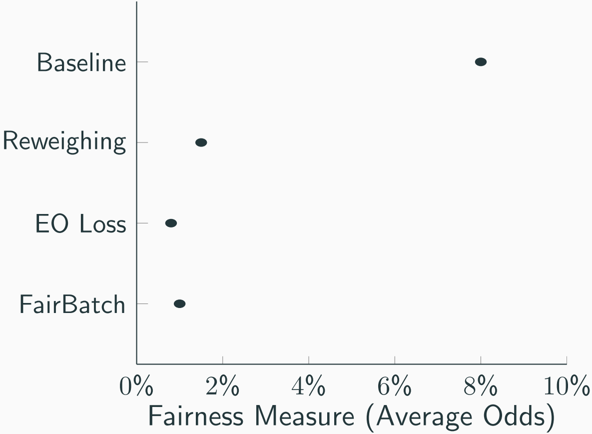
Source: (Angwin et al. 2016, Buolamwini, J., & Gebru, T. 2018)

But Fairness Measures Aren't Stable!



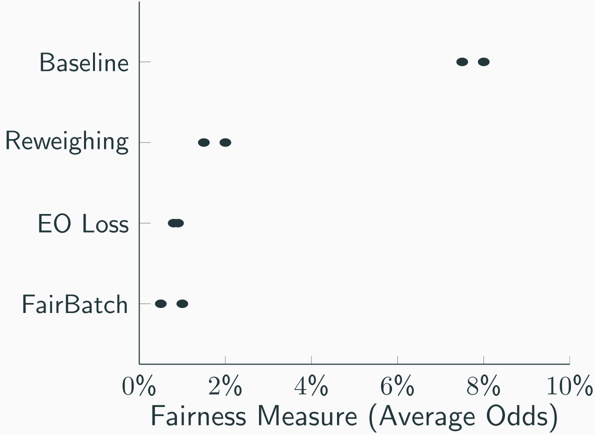
Source: (Amir et al. 2021, Sellam et al. 2022, Baldini et al. 2022)

But Fairness Measures Aren't Stable!



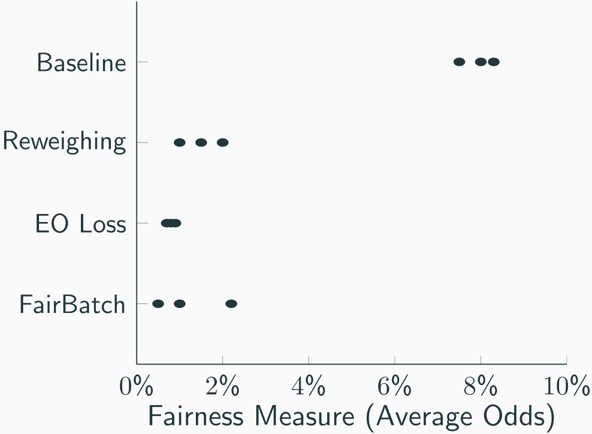
Source: (Amir et al. 2021, Sellam et al. 2022, Baldini et al. 2022)

But Fairness Measures Aren't Stable!



Source: (Amir et al. 2021, Sellam et al. 2022, Baldini et al. 2022)

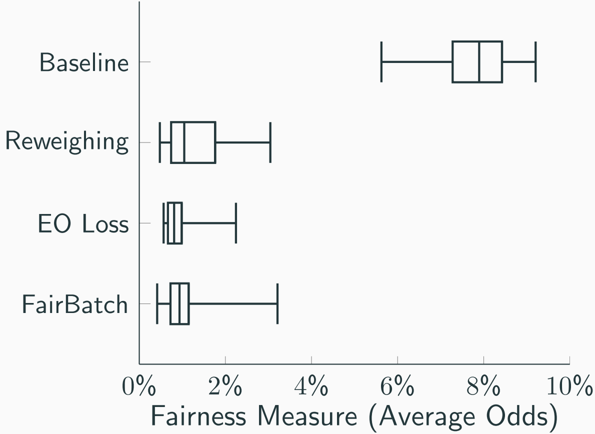
But Fairness Measures Aren't Stable!



Source: (Amir et al. 2021, Sellam et al. 2022, Baldini et al. 2022)

But Fairness Measures Aren't Stable!

Model fairness can vary significantly across random seeds.



Source: (Amir et al. 2021, Sellam et al. 2022, Baldini et al. 2022)

Existing Solutions

Executing multiple training runs with changing random seeds to capture overall fairness variance.

Existing Solutions

Executing multiple training runs with changing random seeds to capture overall fairness variance.

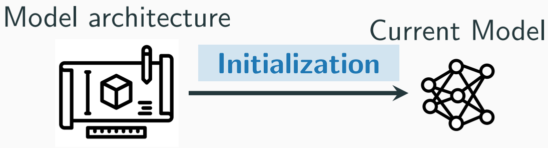
Blindly executing training runs

- is expensive,
- raises the bar to do fair ML research,
- **lacks the understanding of the underlying cause for high fairness variance.**

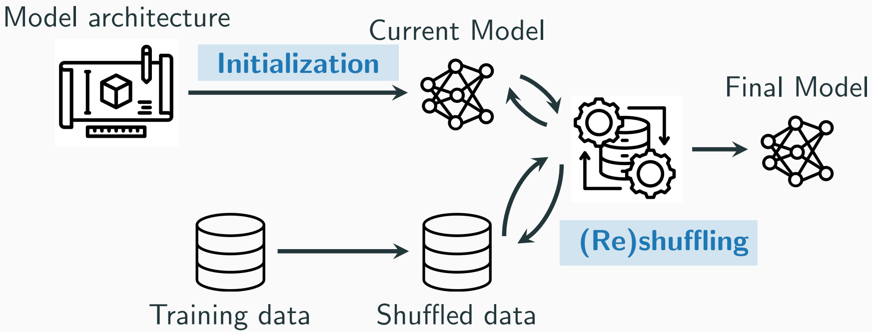
The Sources of Randomness

Weight Initialization and Data Reshuffling

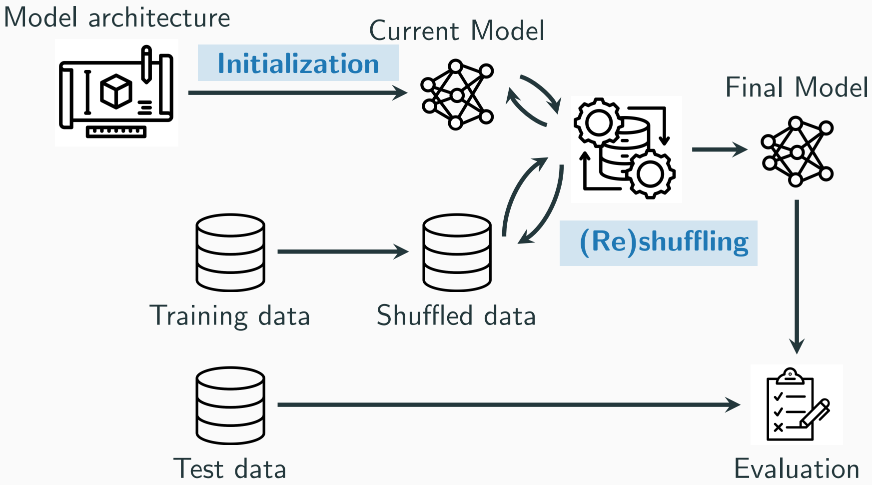
Weight Initialization and Data Reshuffling



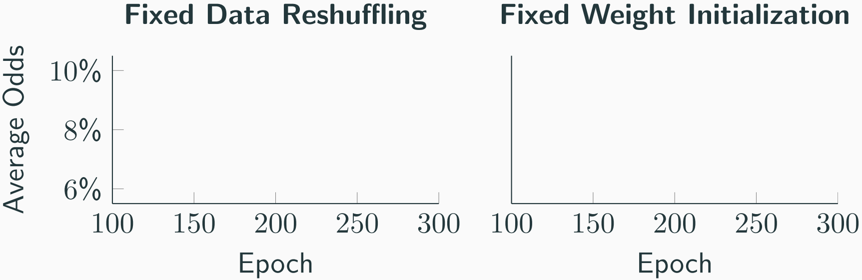
Weight Initialization and Data Reshuffling



Weight Initialization and Data Reshuffling

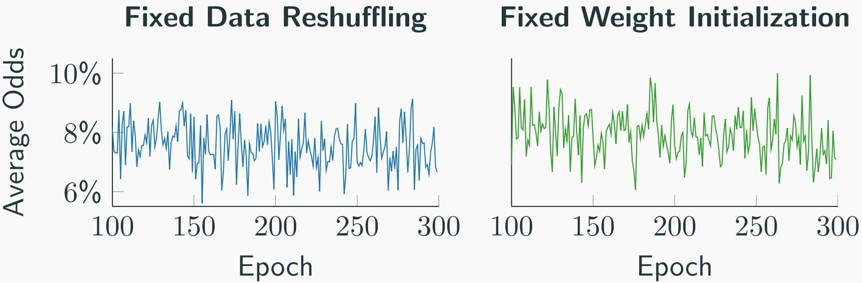


Variance Across Epochs



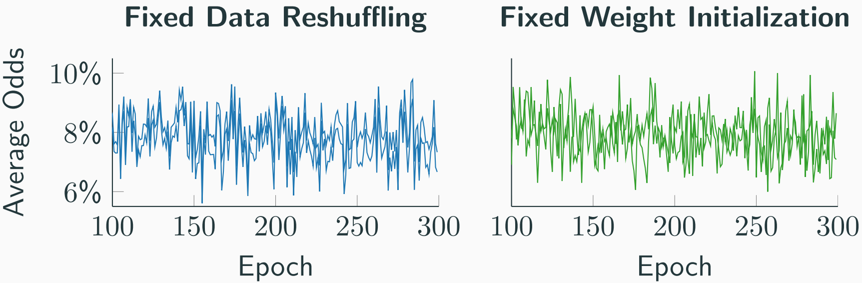
Variance Across Epochs

1 Run

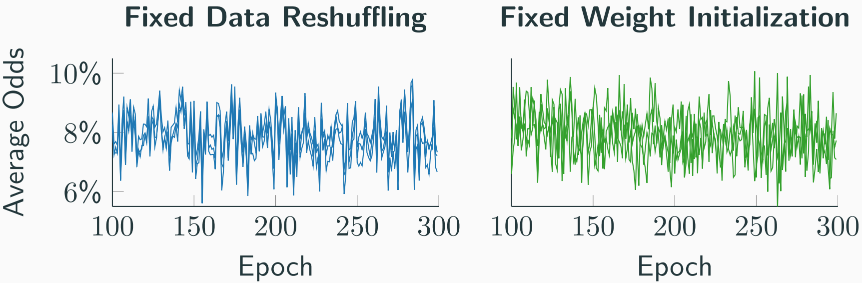


Variance Across Epochs

2 Runs

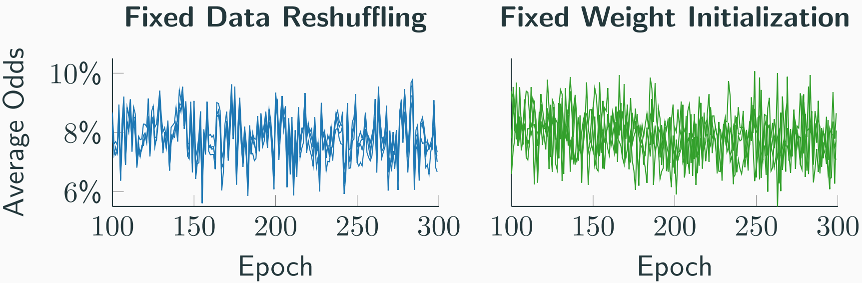


3 Runs


Fixed Data Reshuffling
Fixed Weight Initialization
Average Odds
10%
8%
6%
100
150
200
250
300
Epoch
100
150
200
250
300
Epoch

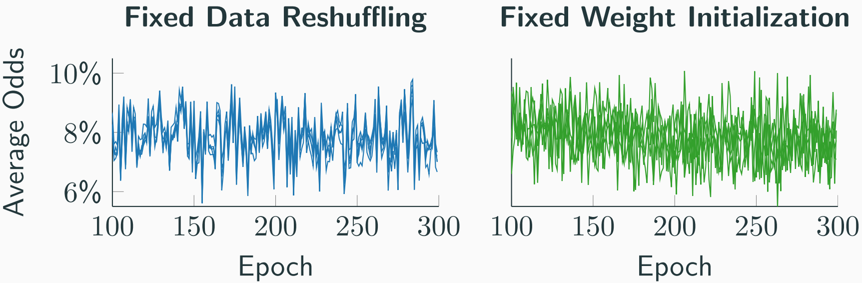
Variance Across Epochs

4 Runs

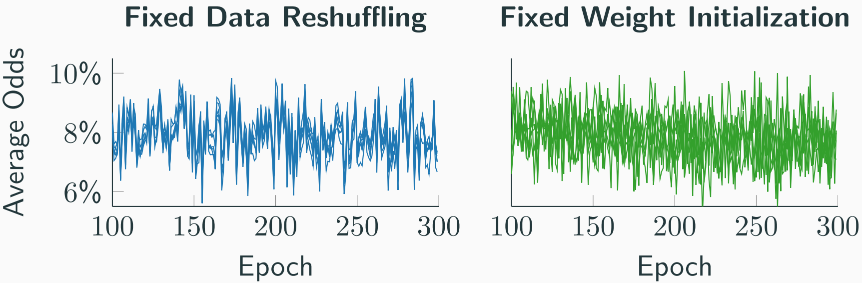


Variance Across Epochs

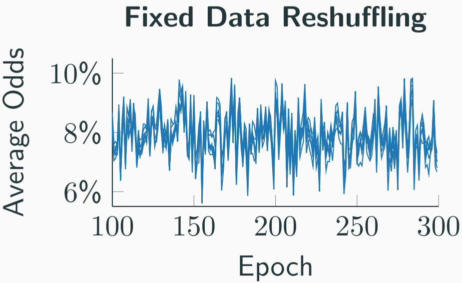
5 Runs

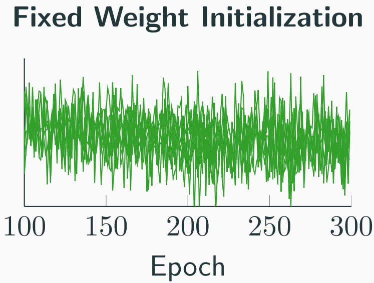


6 Runs

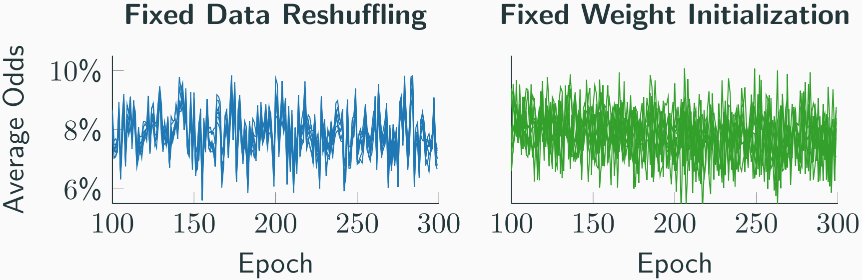

Fixed Data Reshuffling
Fixed Weight Initialization
Average Odds
10%
8%
6%
100
150
200
250
300
Epoch
100
150
200
250
300
Epoch

7 Runs

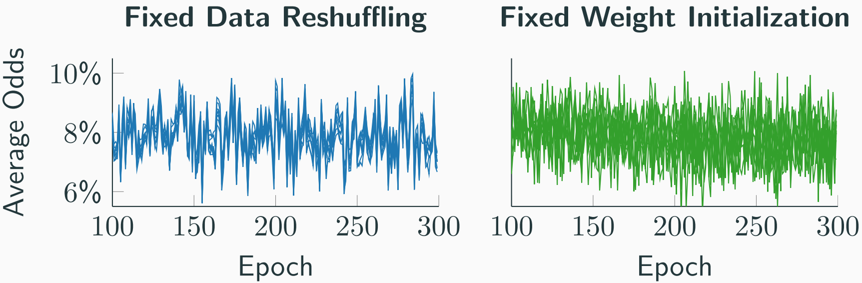

Fixed Data Reshuffling
Average Odds
10%
8%
6%
100
150
200
250
300
Epoch


Fixed Weight Initialization
100
150
200
250
300
Epoch

8 Runs

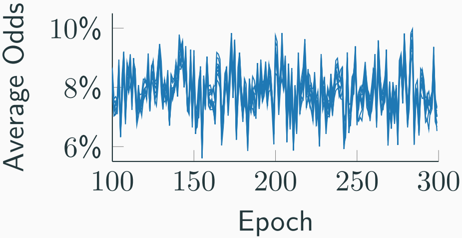

Fixed Data Reshuffling
Fixed Weight Initialization
Average Odds
10%
8%
6%
100
150
200
250
300
Epoch
100
150
200
250
300
Epoch

9 Runs

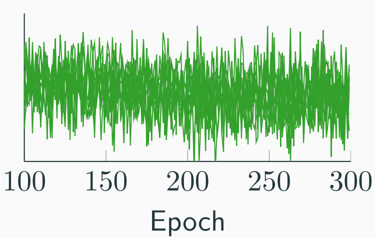

Fixed Data Reshuffling
Fixed Weight Initialization
Average Odds
10%
8%
6%
100
150
200
250
300
Epoch
100
150
200
250
300
Epoch

10 Runs

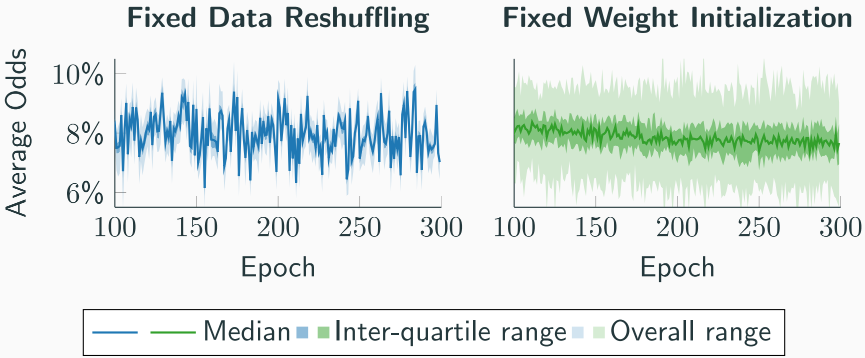
Fixed Data Reshuffling



Fixed Weight Initialization



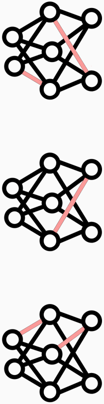
50 Runs


Fixed Data Reshuffling
Fixed Weight Initialization
Average Odds
10%
8%
6%
100
150
200
250
300
Epoch
Median
Inter-quartile range
Overall range

Fairness Variance Beyond Randomness

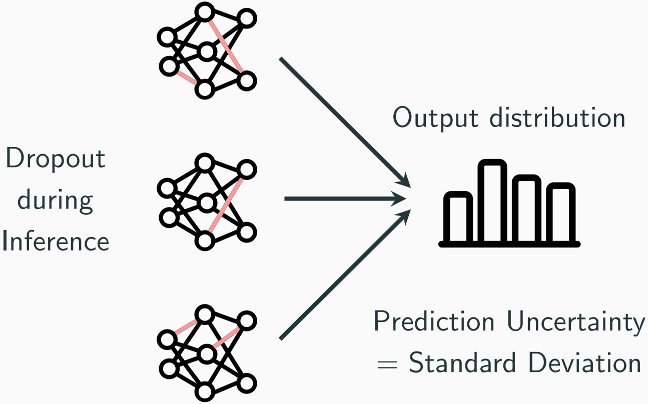
Measuring Uncertainty: Monte-Carlo Dropout

Dropout during Inference



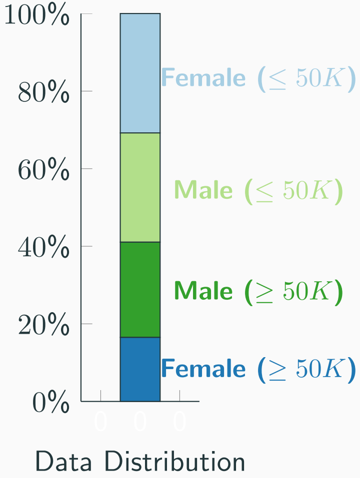
Source: (Gal, Y., & Ghahramani, Z. 2016)

Measuring Uncertainty: Monte-Carlo Dropout

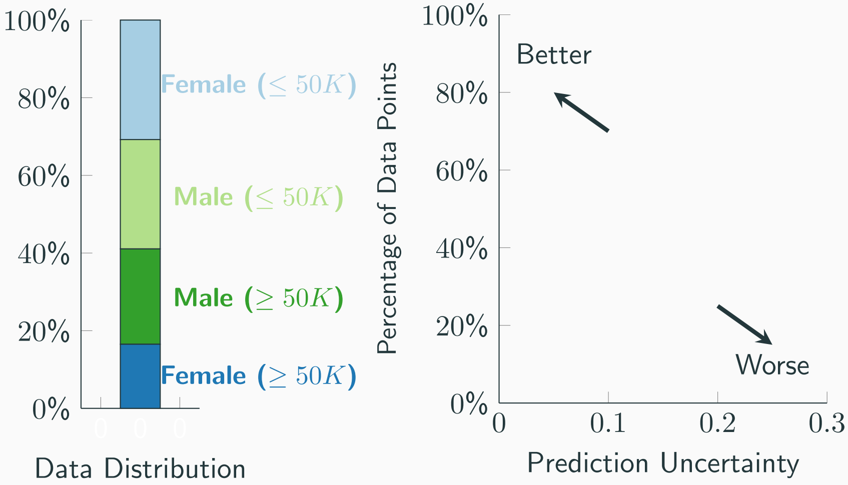


Source: (Gal, Y., & Ghahramani, Z. 2016)

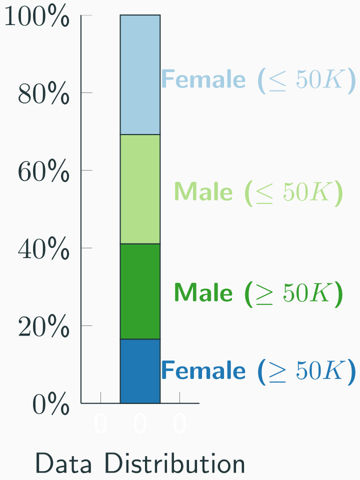
Prediction Uncertainty Across Groups

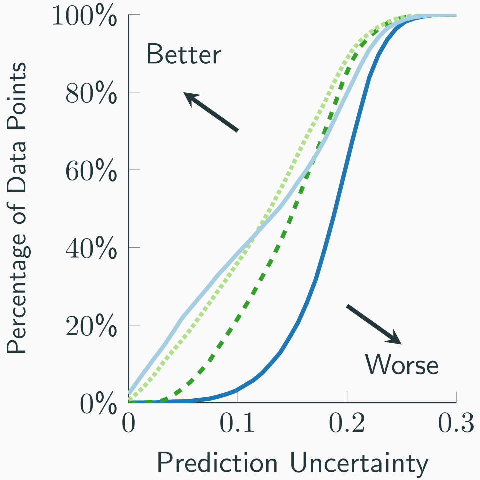


Prediction Uncertainty Across Groups



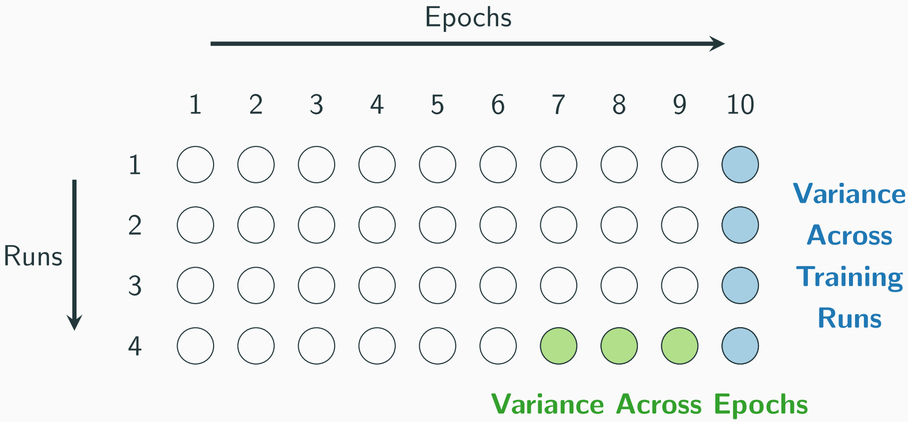
Prediction Uncertainty Across Groups





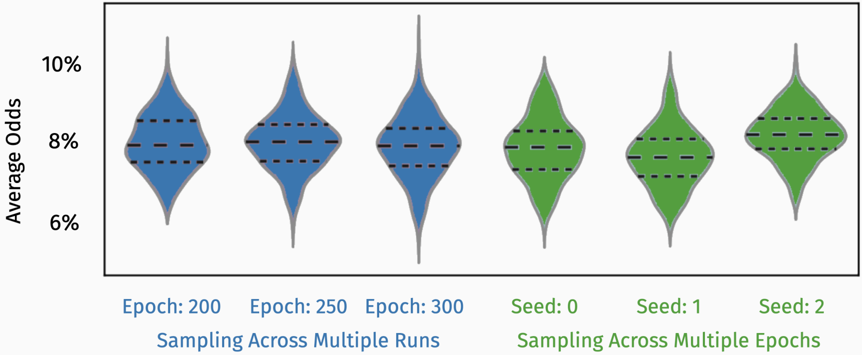
Applications

Variance Across Epochs vs Training Runs



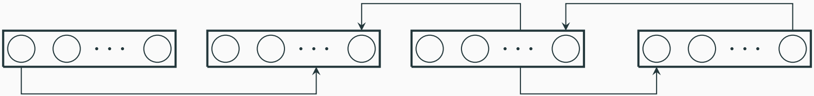
Variance Across Epochs vs Training Runs

The distribution of fairness scores **across multiple runs** is 'equal' to the distribution of fairness scores **across epochs in any single run.**

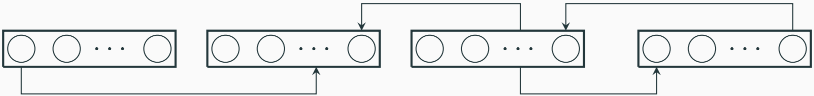


Manipulating Fairness with Data Order

Guiding Principle: The most recent gradient updates seen by the model have a significant influence on its fairness scores!



Guiding Principle: The most recent gradient updates seen by the model have a significant influence on its fairness scores!



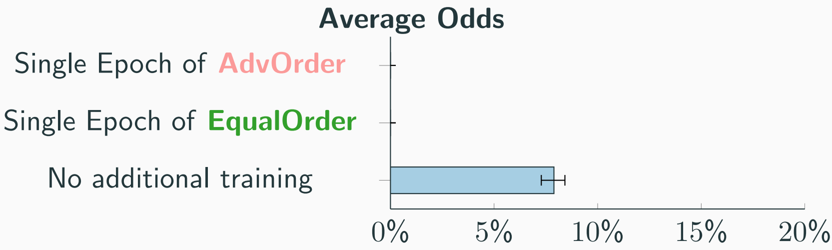
EqualOrder

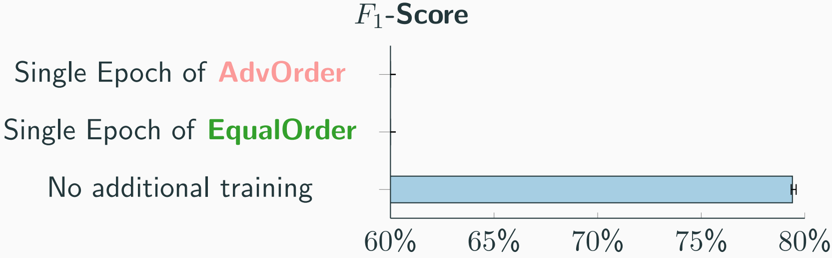
To improve fairness scores

AdvOrder

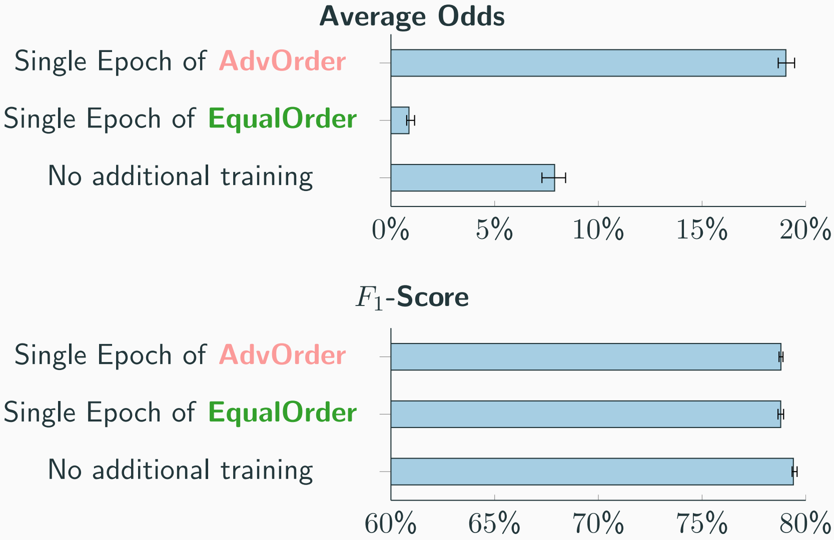
To adversarially introduce bias

Bias Mitigation with Data Order





Bias Mitigation with Data Order



Takeaways

Takeaways

- Data reshuffling is the dominant cause of fairness variance.

Takeaways

- Data reshuffling is the dominant cause of fairness variance.
- Under-represented groups are more sensitive to the randomness.

Takeaways

- Data reshuffling is the dominant cause of fairness variance.
- Under-represented groups are more sensitive to the randomness.
- Fairness distribution across multiple runs is equal to that across epochs within a single run, thus bypassing multiple training runs

Takeaways

- Data reshuffling is the dominant cause of fairness variance.
- Under-represented groups are more sensitive to the randomness.
- Fairness distribution across multiple runs is equal to that across epochs within a single run, thus bypassing multiple training runs
- Controlling data order alone can manipulate group-level accuracies.

Takeaways

- Data reshuffling is the dominant cause of fairness variance.
- Under-represented groups are more sensitive to the randomness.
- Fairness distribution across multiple runs is equal to that across epochs within a single run, thus bypassing multiple training runs
- Controlling data order alone can manipulate group-level accuracies.

Scan the QR Code for the paper

Feel free to contact us at: pganesh@u.nus.edu