HOW (NOT) TO USE FAIRNESS METRICS IN MACHINE LEARNING

Michėle Wieland
ETH Zürich

AGENDA

Why should we care about fairness?

How can we measure fairness?

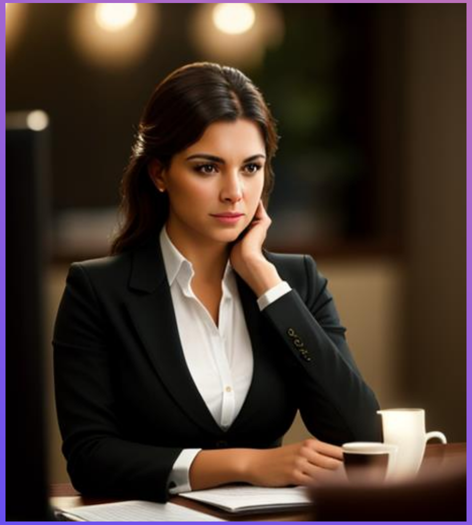
How can we improve fairness?

What could go wrong?

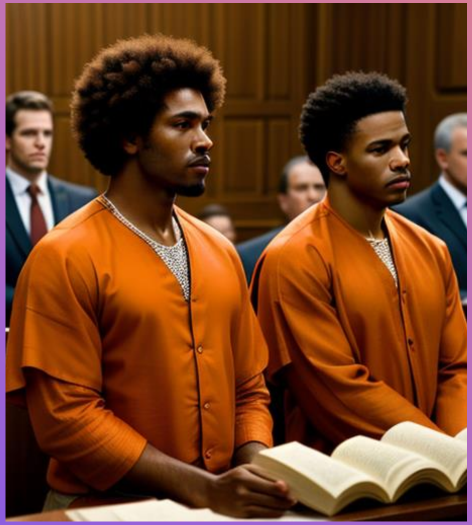
WHY SHOULD WE CARE ABOUT FAIRNESS?

WHY SHOULD WE CARE?

Images generated by: Stable Diffusion Web



Woman applying for a tech role



African-American men in court

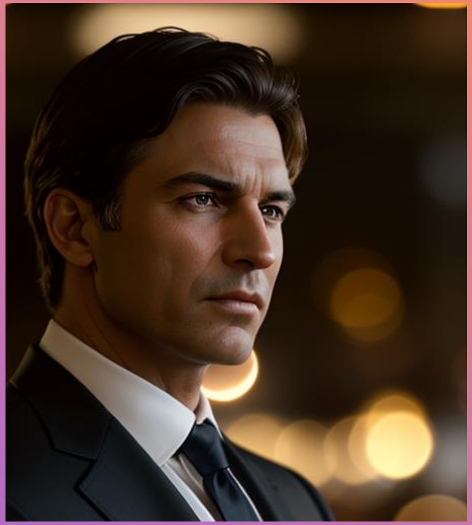
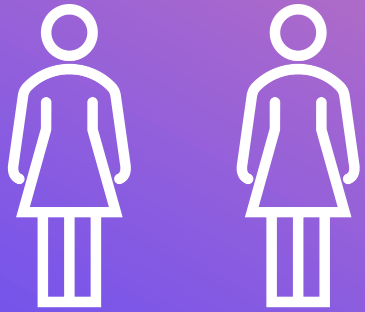


Image of a CEO

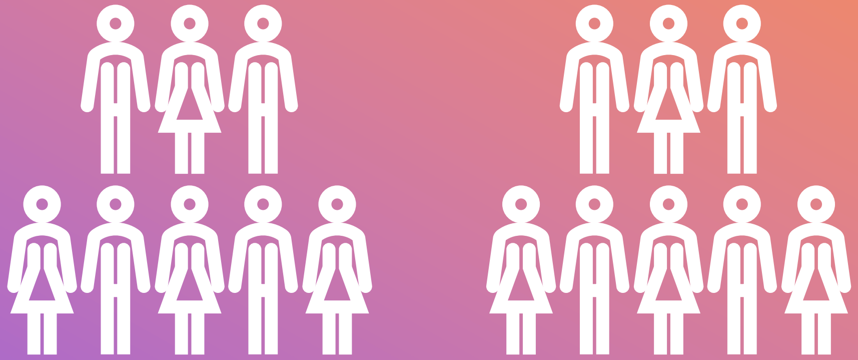
IMAGE OF A CEO

HOW CAN WE MEASURE FAIRNESS?

INDIVIDUAL VS GROUP FAIRNESS



Individual fairness



Group fairness

HOW TO MEASURE GROUP FAIRNESS?

- Demographic parity
- Disparate impact
- Equal opportunity
- Equalized odds
- Predictive parity
- Conditional demographic disparity
- Counterfactual fairness
- ..

COMMON GROUP FAIRNESS METRICS

- Demographic parity
- Equalized odds
- Equal opportunity

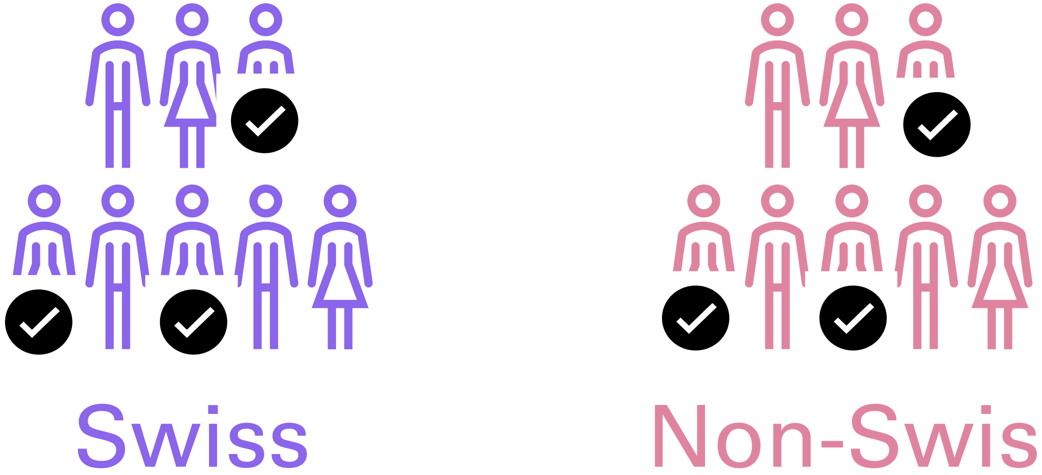
TOY EXAMPLE

- Applying for a loan at a bank
  - Y = 0: request denied
  - Y = 1: request accepted

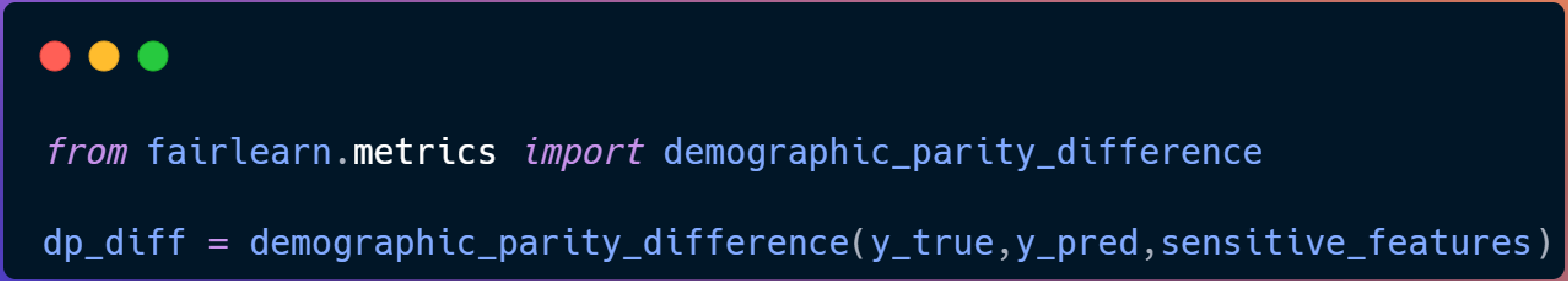
- Sensitive attribute
  - Z = 0: Swiss
  - Z = 1: Non-Swiss

Demographic parity

$$P(\widehat{Y} = 1 \mid Z = 0) = P(\widehat{Y} = 1 \mid Z = 1)$$



DEMOGRAPHIC PARITY IN PYTHON



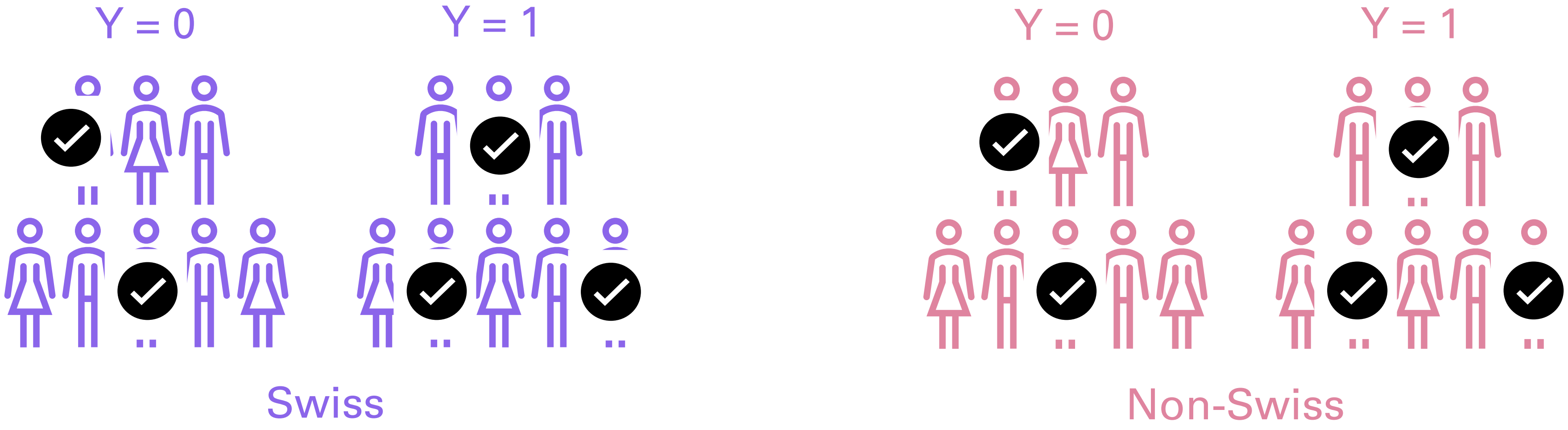
```python
from fairlearn.metrics import demographic_parity_difference

dp_diff = demographic_parity_difference(y_true,y_pred,sensitive_features)
```

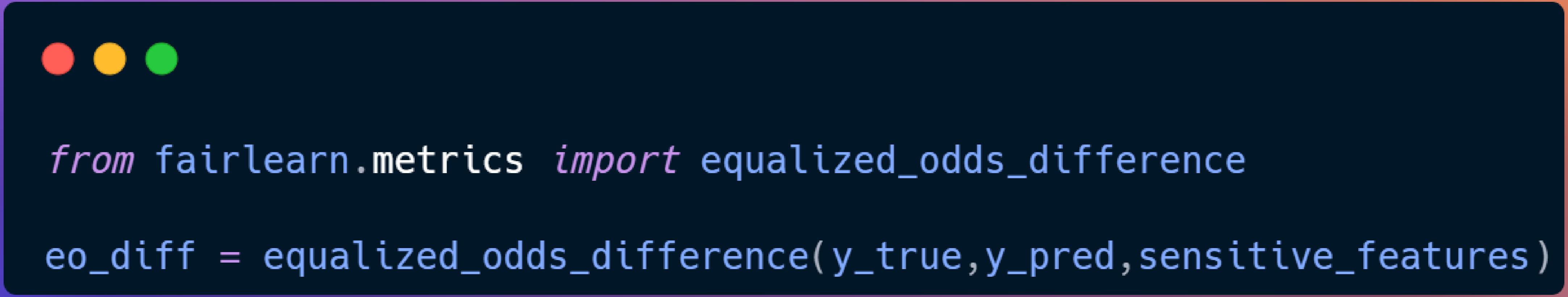
Equalized odds

$$P(\widehat{Y} = 1 \mid Y = 0, Z = 0) = P(\widehat{Y} = 1 \mid Y = 0, Z = 1)$$

$$P(\widehat{Y} = 1 \mid Y = 1, Z = 0) = P(\widehat{Y} = 1 \mid Y = 1, Z = 1)$$



EQUALIZED ODDS IN PYTHON

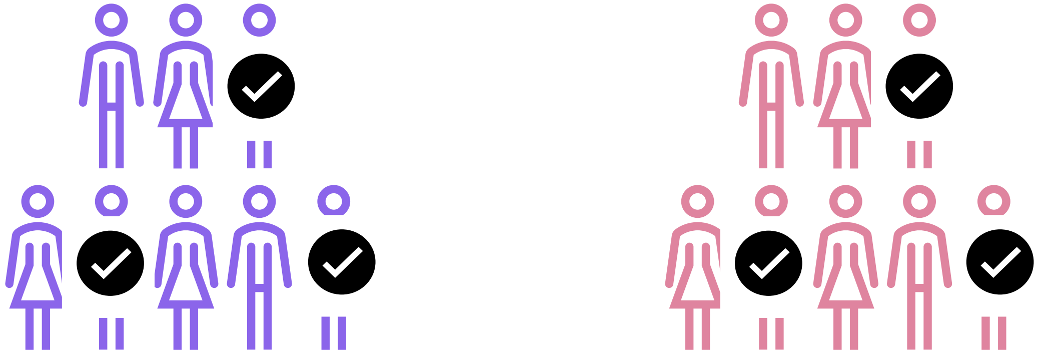


```python
from fairlearn.metrics  import equalized_odds_difference

eo_diff = equalized_odds_difference(y_true,y_pred,sensitive_features)
```

Equal opportunity

$$P(\widehat{Y} = 1 \mid Y = 1, Z = 0) = P(\widehat{Y} = 1 \mid Y = 1, Z = 1)$$



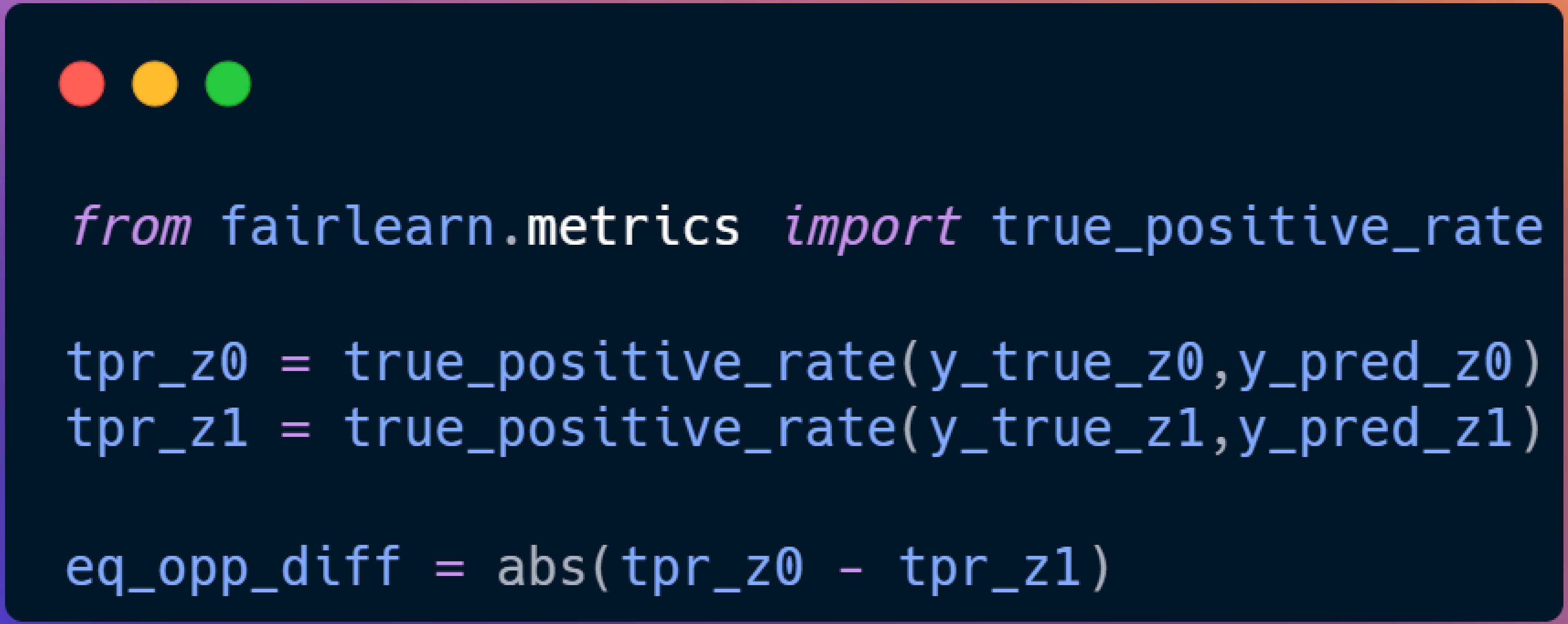
Y = 1

Swiss

Y = 1

Non-Swiss

EQUAL OPPORTUNITY IN PYTHON



```python
from fairlearn.metrics import true_positive_rate

tpr_z0 = true_positive_rate(y_true_z0,y_pred_z0)
tpr_z1 = true_positive_rate(y_true_z1,y_pred_z1)

eq_opp_diff = abs(tpr_z0 - tpr_z1)
```

HOW CAN WE IMPROVE FAIRNESS?

Deleting sensitive attributes

name	ZIP code	occupation	gender	age
Emilia*	8002	nurse	female	29
Roberto*	8155	firefighter	male	45
Dan*	8011	data scientist	diverse	22
Sarah*	8049	teacher	female	59

*data is fictional

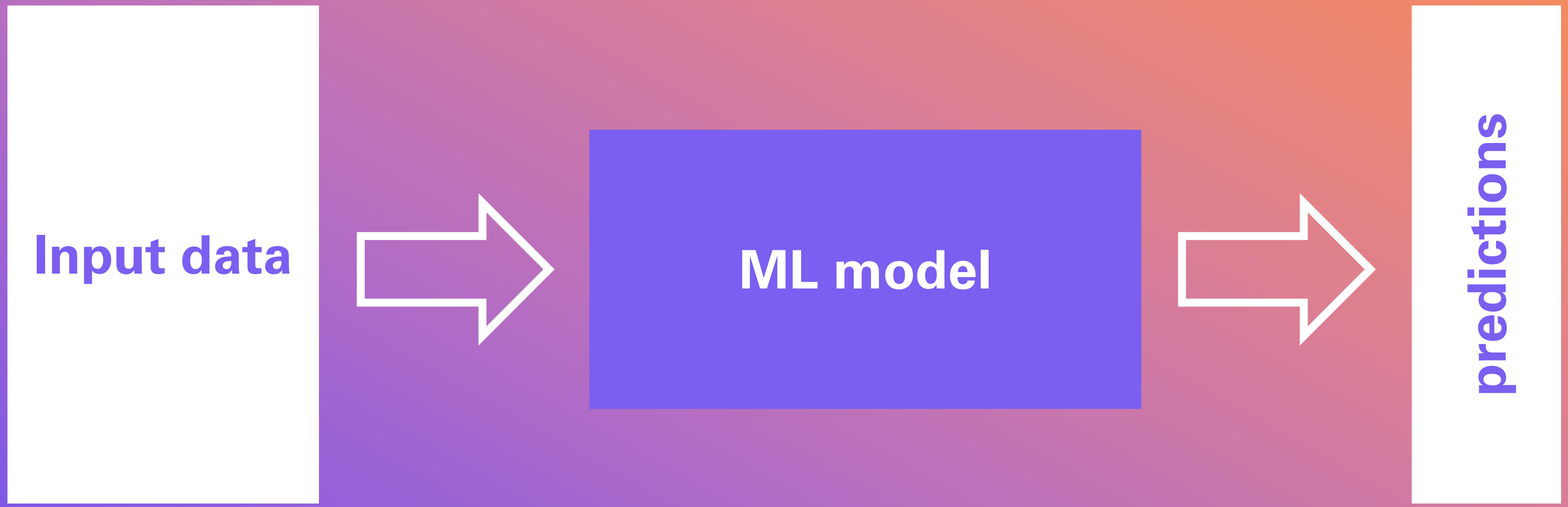
Deleting sensitive attributes

name	ZIP code	occupation	gender	age
Emilia*	8002	nurse		29
Roberto*	8155	firefighter		45
Dan*	8011	data scientist		22
Sarah*	8049	teacher		59

*data is fictional

FAIRNESS-PROMOTING ALGORITHMS

- Pre-processing
- In-processing
- Post-processing

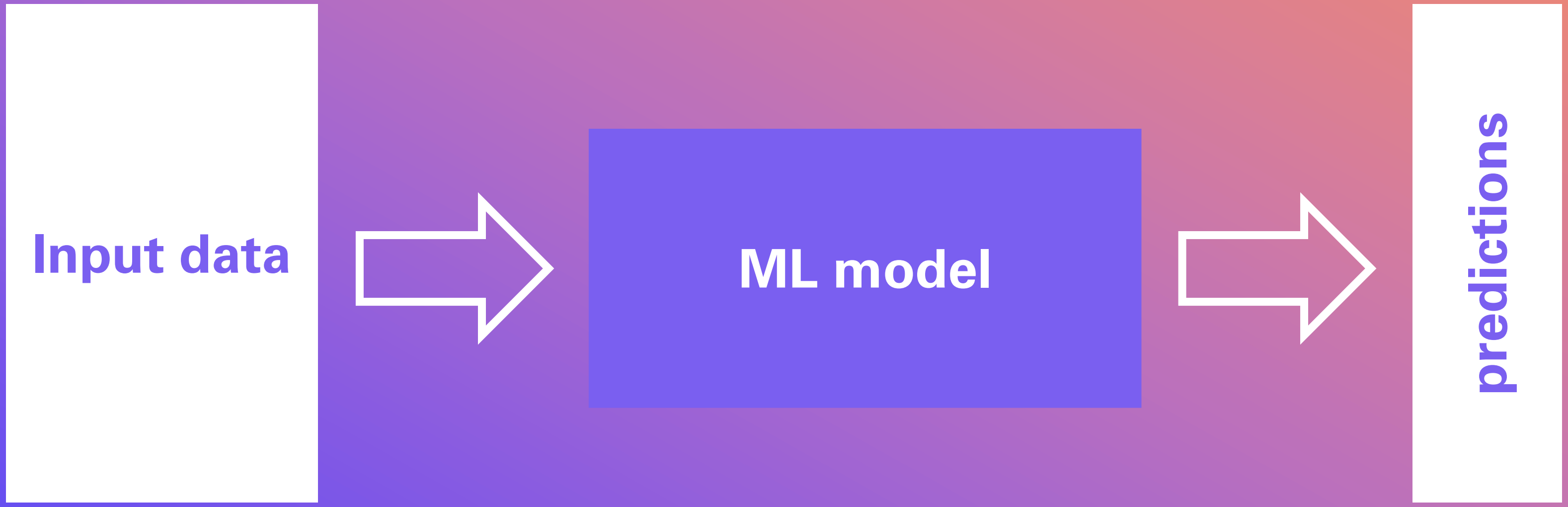


Input data

ML model

predictions

FAIRNESS-PROMOTING ALGORITHMS

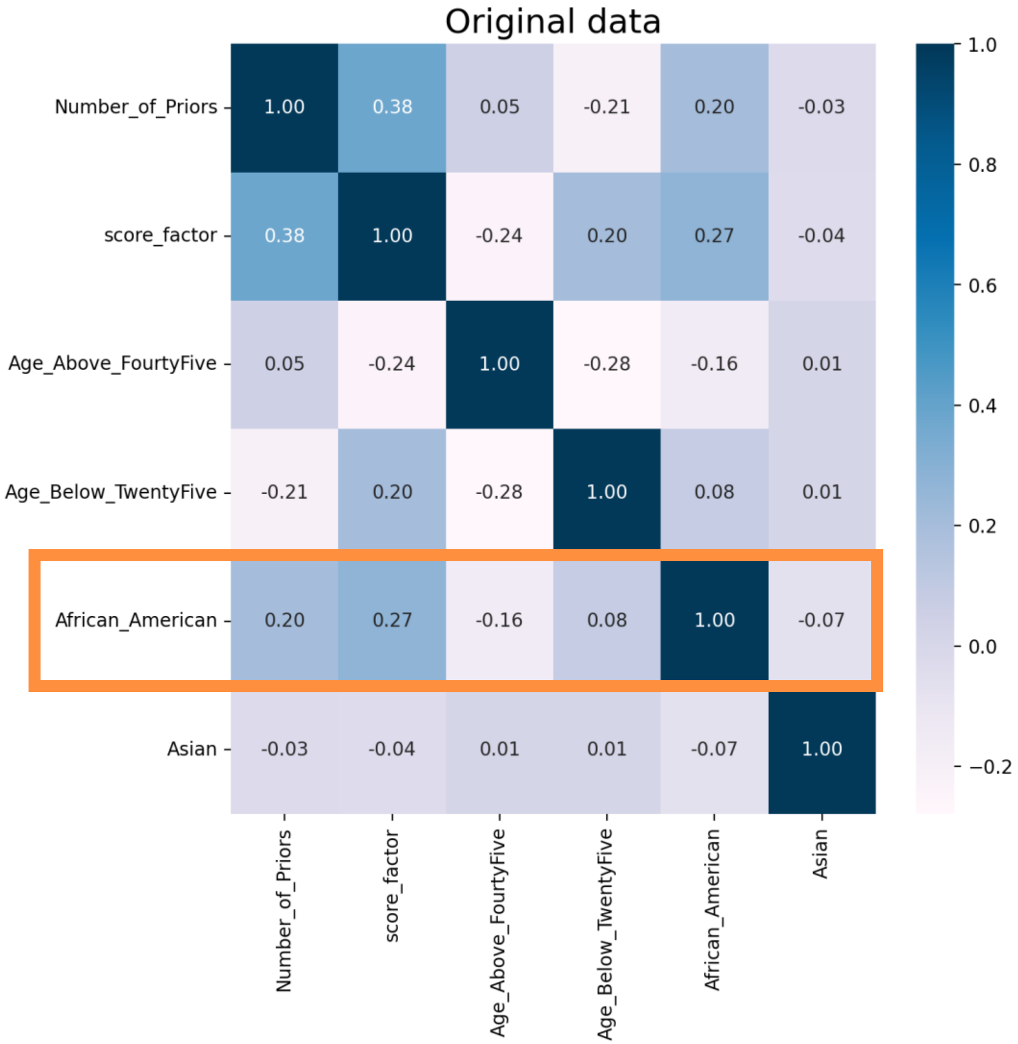


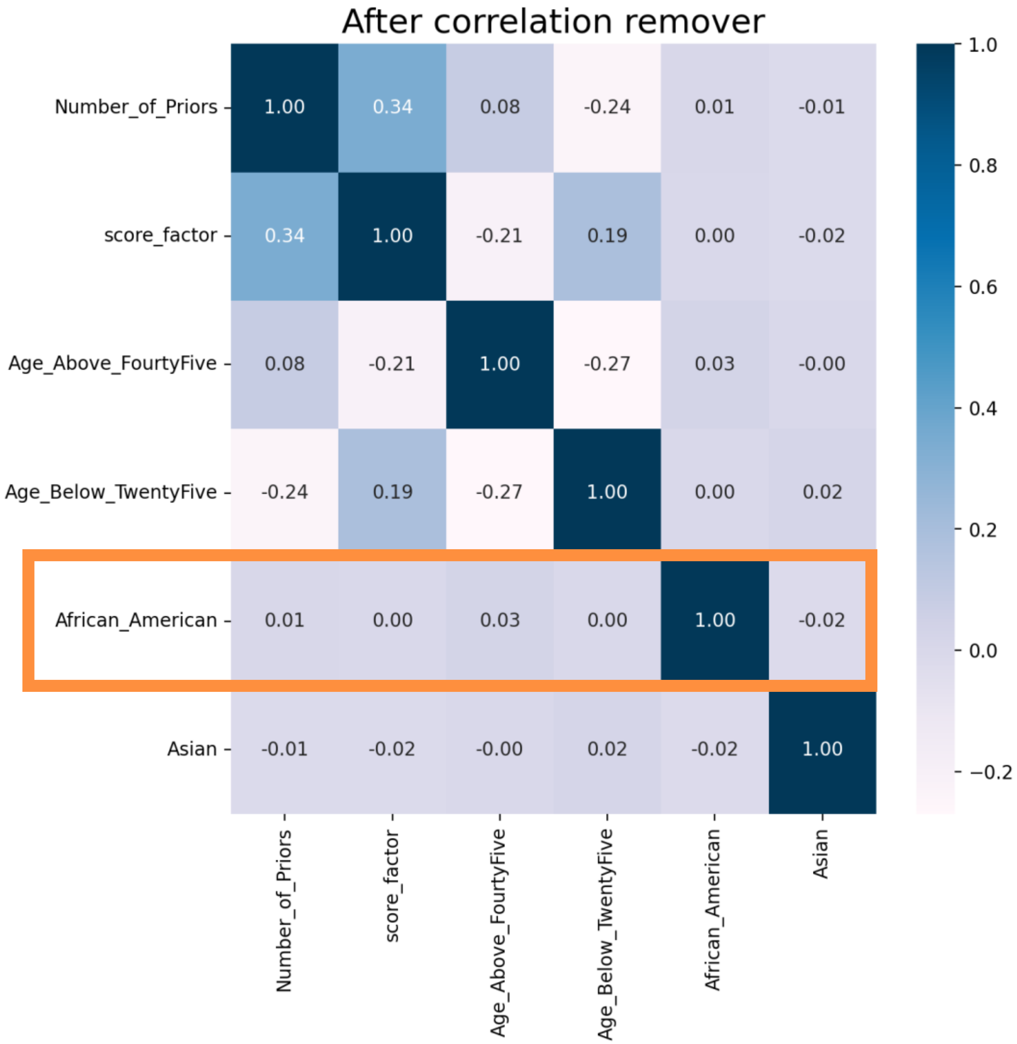
Pre-processing

In-processing

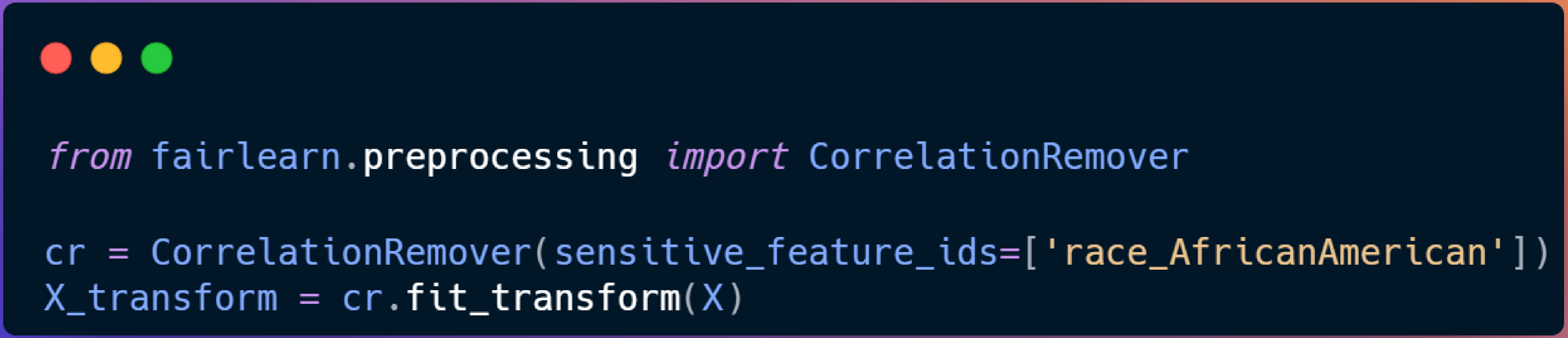
Post-processing

Pre-processing: correlation remover


Original data


After correlation remover

CORRELATION REMOVER IN PYTHON



```python
from fairlearn.preprocessing import CorrelationRemover

cr = CorrelationRemover(sensitive_feature_ids=['race_AfricanAmerican'])
X_transform = cr.fit_transform(X)
```

WHAT COULD GO WRONG?

SOLUTIONISM TRAP

*FAIRNESS AND ABSTRACTION IN SOCIOTECHNICAL SYSTEMS, SELBST ET AL.

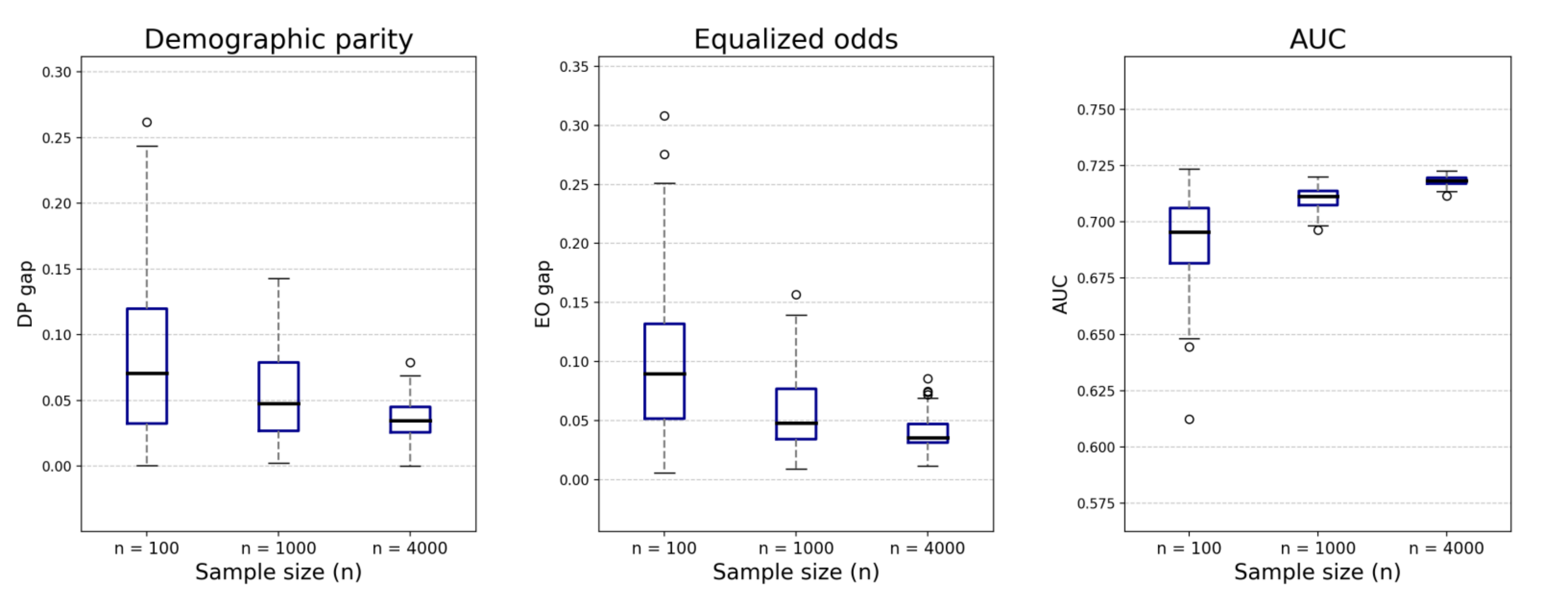
«Failure to recognize the possibility that the best solution to a problem may not involve technology»

CAN WE RELY ON POINT ESTIMATES?

- AUC: 0.71
- Demographic parity gap: 0.03
- Equalized odds gap: 0.04

UNCERTAINTY IN ESTIMATES

GITHUB - WIELANDMICHELE/UNCERTAINTY_FAIRNESS_ESTIMATES



PROTECTING ONE ATTRIBUTE

- Protecting a single attribute can increase unfairness for others
- Often recommended to protect multiple attributes simultaneously
- Not trivial to decide which attributes need protection

FAIRNESS-ACCURACY TRADEOFF

- Accuracy can drop with increased fairness
- Fairest model is random guessing
- Carefully decide which attributes to protect

THANK YOU

Michėle Wieland

wielandmichele@sunrise.ch

SOURCES

1. Common fairness metrics — Fairlearn 0.10.0.dev0 documentation
2. Preprocessing — Fairlearn 0.10.0.dev0 documentation
3. Credit Loan Decisions — Fairlearn 0.10.0.dev0 documentation
4. LECTURE12_GROUP_FAIRNESS (ethz.ch)
5. Fairness and Abstraction in Sociotechnical Systems (friedler.net)