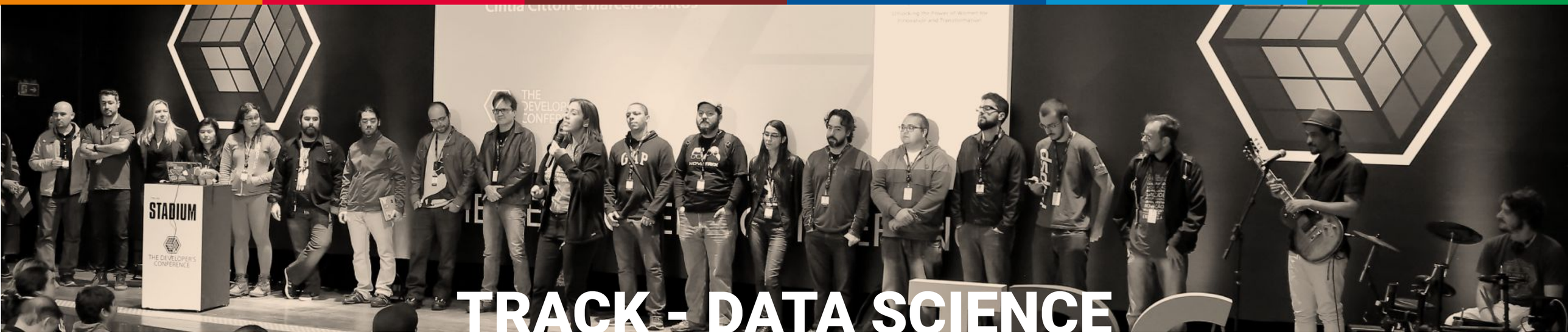


TRACK - DATA SCIENCE

How to Identify and Prevent AI Bias with ai360

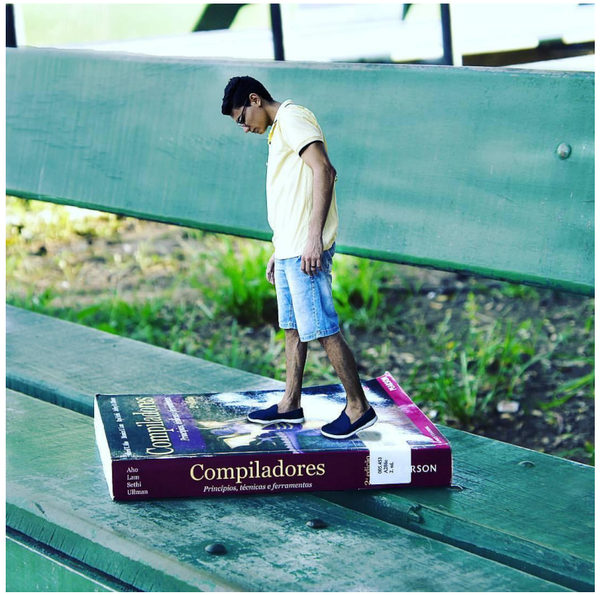
Crislânio Macêdo

THE DEVELOPER'S CONFERENCE

THE DEVELOPER'S CONFERENCE

Agenda

- Bias
- Fairness in ML
- High Visibility Fairness Examples
- AI Fairness 360
  - Metrics
  - Demo



whoami?

> Crislânio Macêdo
> Majored in Computer Science from Universidade Federal do Ceará and Mastering Degree from Universidade Estadual do Ceará.
> Kaggle Notebook Master

Bias

THE DEVELOPER'S CONFERENCE

Nonverbal Bias

We prefer to scrap own opinion in favour of the groups' opinion.

Beauty Bias

This is the view that we tend to think that the most handsome individual will be the most successful.

Conformity Bias

Occurs when a positive or negative evaluation is made of someone based on their body language, personal appearance or style of dress.

Similarity Bias

Naturally, we want to surround ourselves with people we feel are similar to us.

Fairness in ML

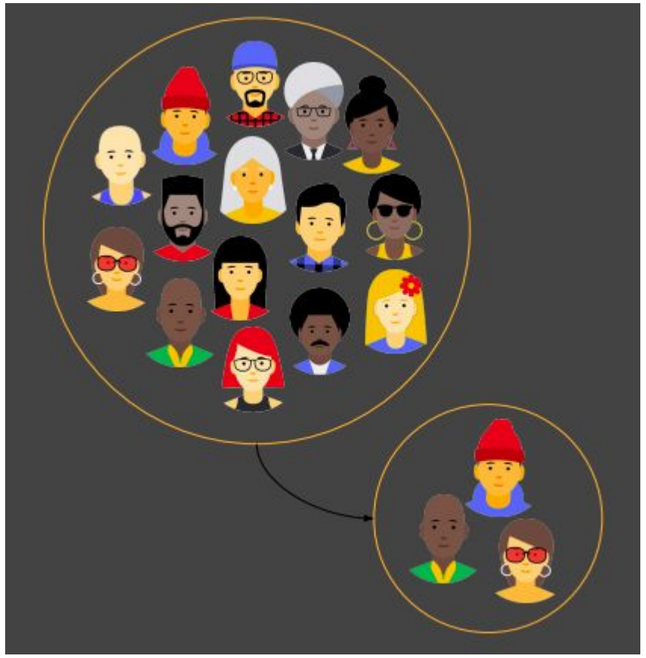
Biases in Data → Biased Labels

Wedding photographs (donated by Googlers), labeled by a classifier trained on the Open Images dataset. The classifier's label predictions are recorded below each image

source: https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

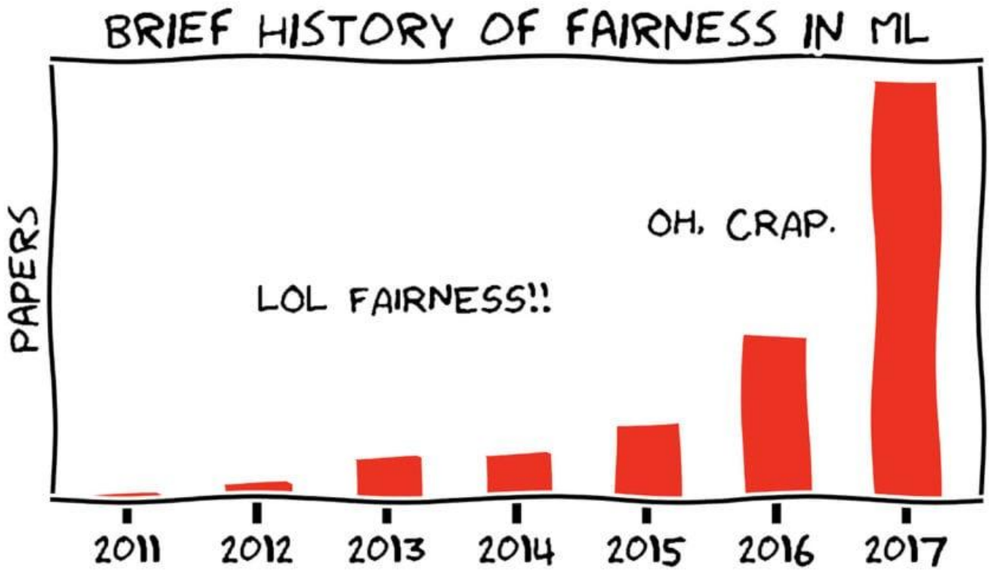
Fairness in ML

Biases in Data →
Biased Data
Representation



THE DEVELOPER'S CONFERENCE

Fairness in ML



The number of academic pubs on fairness, 2011-2017

Source: https://fairmlclass.github.io/1.html#/4

THE DEVELOPER'S CONFERENCE

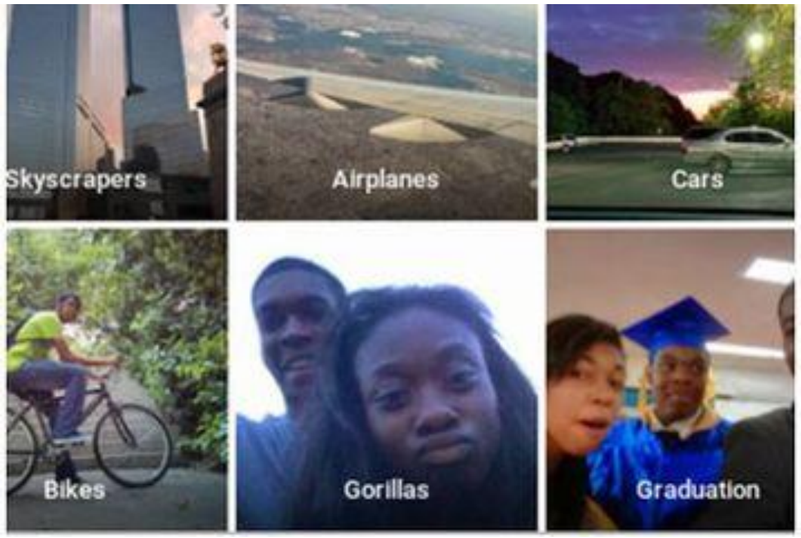
Algorithmic fairness is one of the hottest topics in the ML/AI research community

High Visibility Fairness Examples

Photo Classification Software -2016

Google has come under fire after the image-recognition feature in its Photos application mistakenly identified people with dark skin as "gorillas."

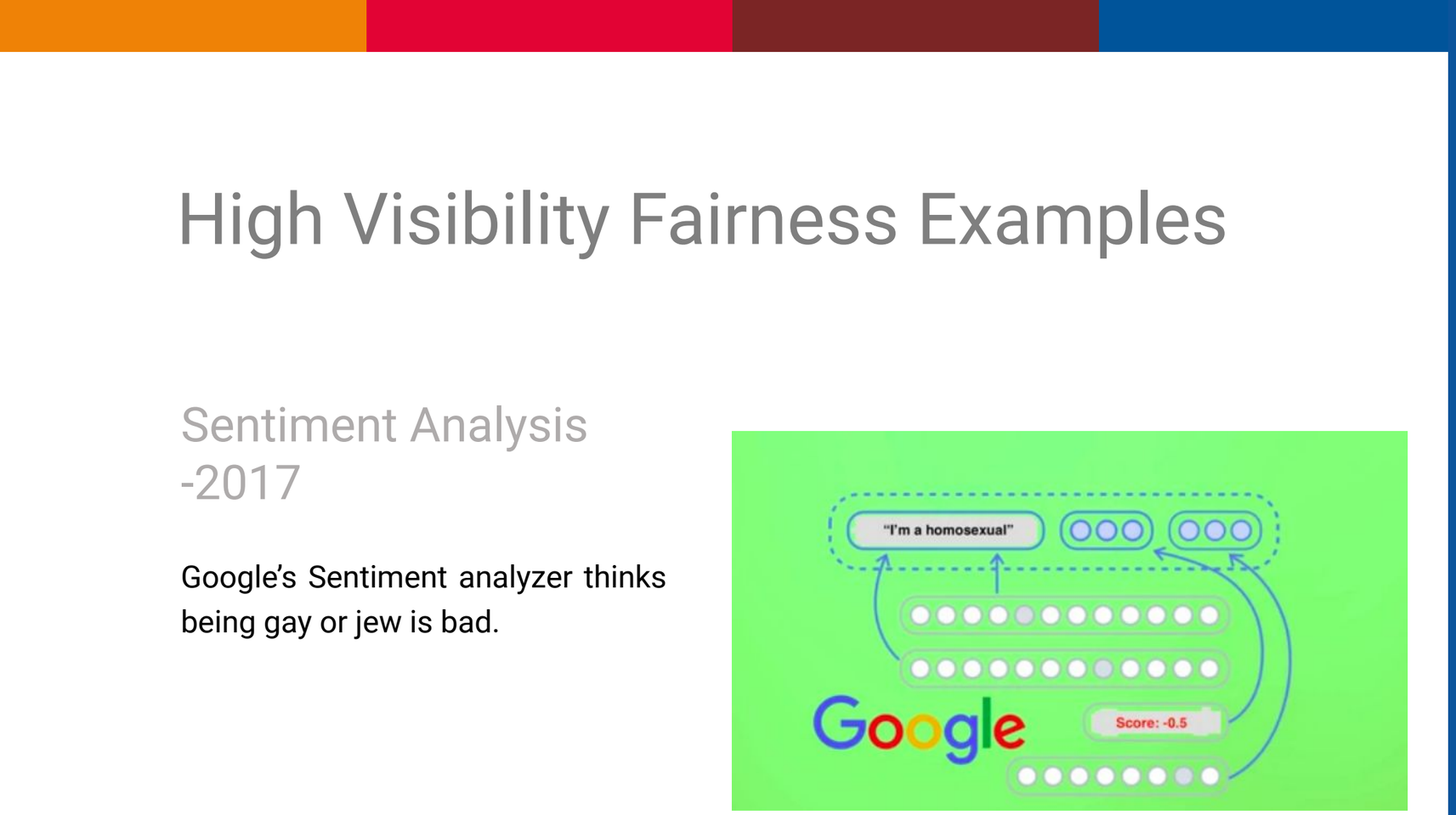
source: https://www.cbsnews.com/news/google-photos-labeled-pics-of-african-americans-as-gorillas/



THE DEVELOPER'S CONFERENCE

What does it take to trust a decision made by a machine?

High Visibility Fairness Examples



Sentiment Analysis -2017

Google's Sentiment analyzer thinks being gay or jew is bad.

source: https://www.vice.com/en/article/j5jmj8/google-artificial-intelligence-bias.

google api: https://cloud.google.com/natural-language/

This is an example of how bias creeps into artificial intelligence

High Visibility Fairness Examples

Amazon scraps secret AI recruiting tool that showed bias against women - 2018

The team had been building computer programs for 4 years ago to review job applicants' resumes with the aim of mechanizing the search for top talent"

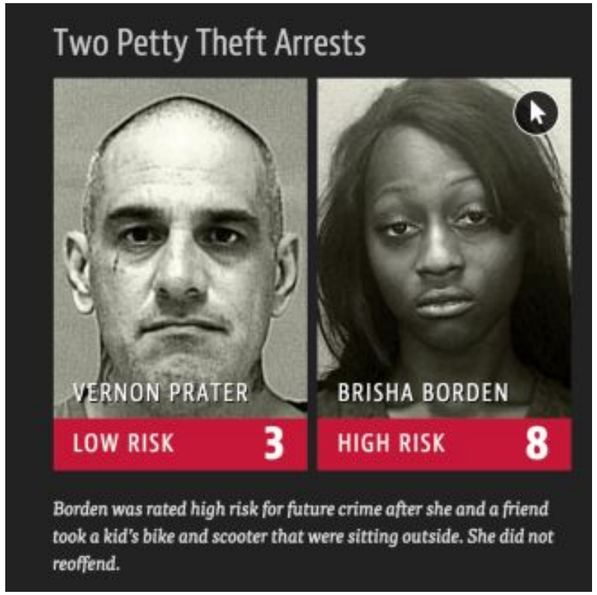
source:
https://www.hrkatha.com/recruitment/amazon-discreetly-abandoned-gender-biased-ai-based-recruiting-tool/

This is an example how an AI can perpetuate injustice in hiring

High Visibility Fairness Examples

Criminal Justice System - 2016

Since 2008, nearly every arrestee in Broward County, Florida has been assigned a risk score using Northpointe's COMPAS algorithm. Defendants with low risk scores are released on bail.



source:
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

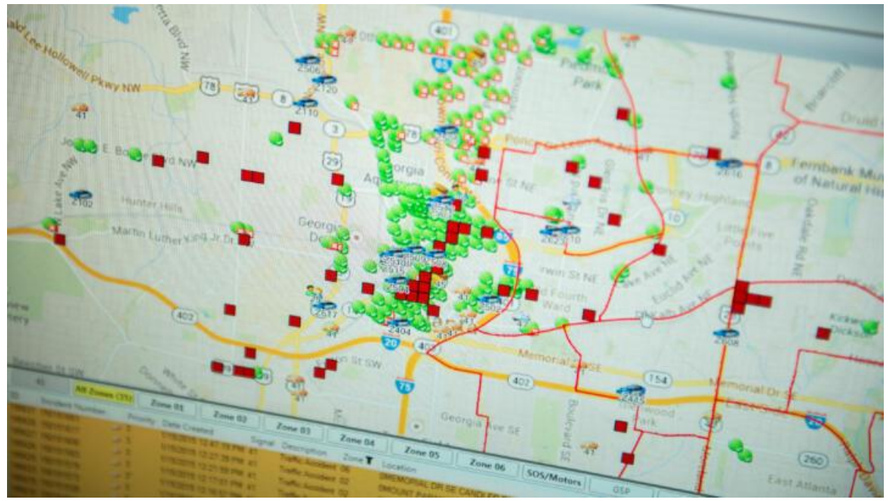
THE DEVELOPER'S CONFERENCE

This is an example of Bias in Recidivism Assessment

High Visibility Fairness Examples

Criminal Justice System - 2016

A map of Atlanta generated through PredPol that uses a predictive algorithm to map hotspots for potential crime.

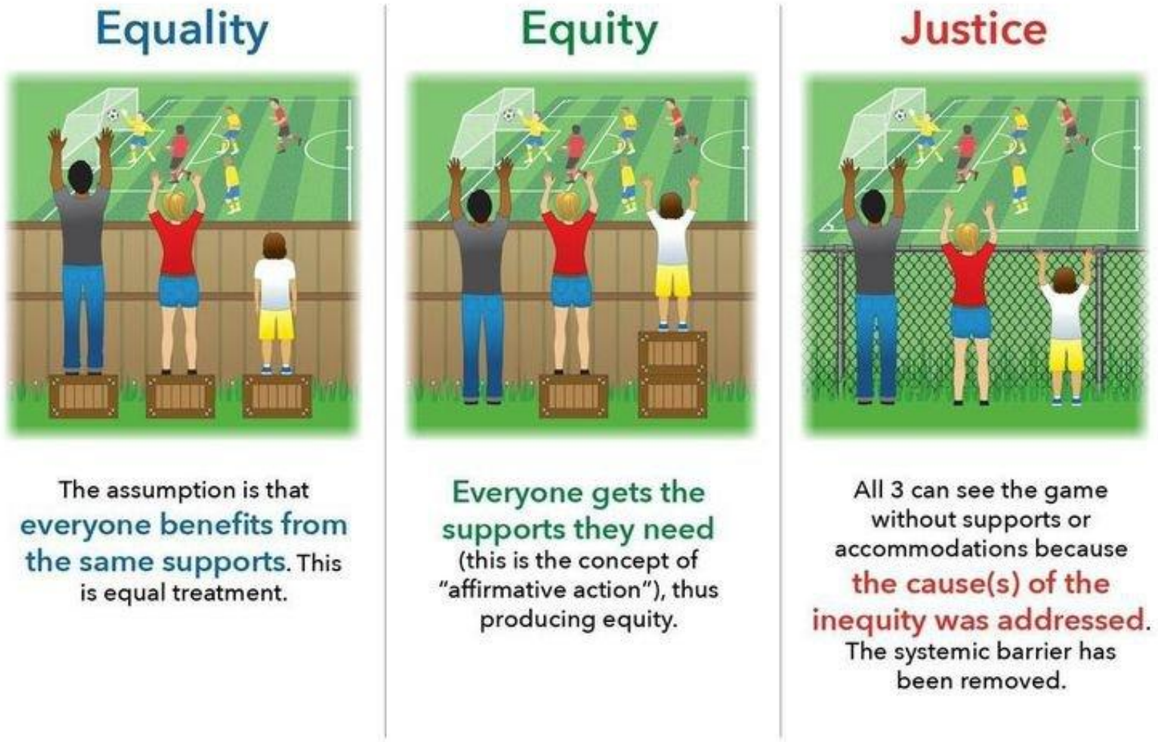


source:
https://themarkup.org/ask-the-markup/2020/08/20/does-predictive-police-technology-contribute-to-bias

THE DEVELOPER'S CONFERENCE

Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?

Fairness is Political

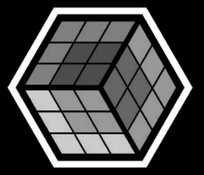


THE DEVELOPER'S CONFERENCE

Someone must decide

Decisions will depend on the product, company, laws, country, etc

There are at least 21 definitions of fairness



THE DEVELOPER'S CONFERENCE

There is no one definition of fairness applicable in all contexts

Some definitions even conflict

AI Fairness 360 (AI360)

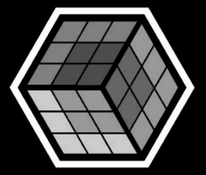
AI360 toolkit is an open-source library to help detect and remove bias in machine learning models.

Toolbox

- Fairness metrics (70+)
- Bias mitigation algorithms (10+)

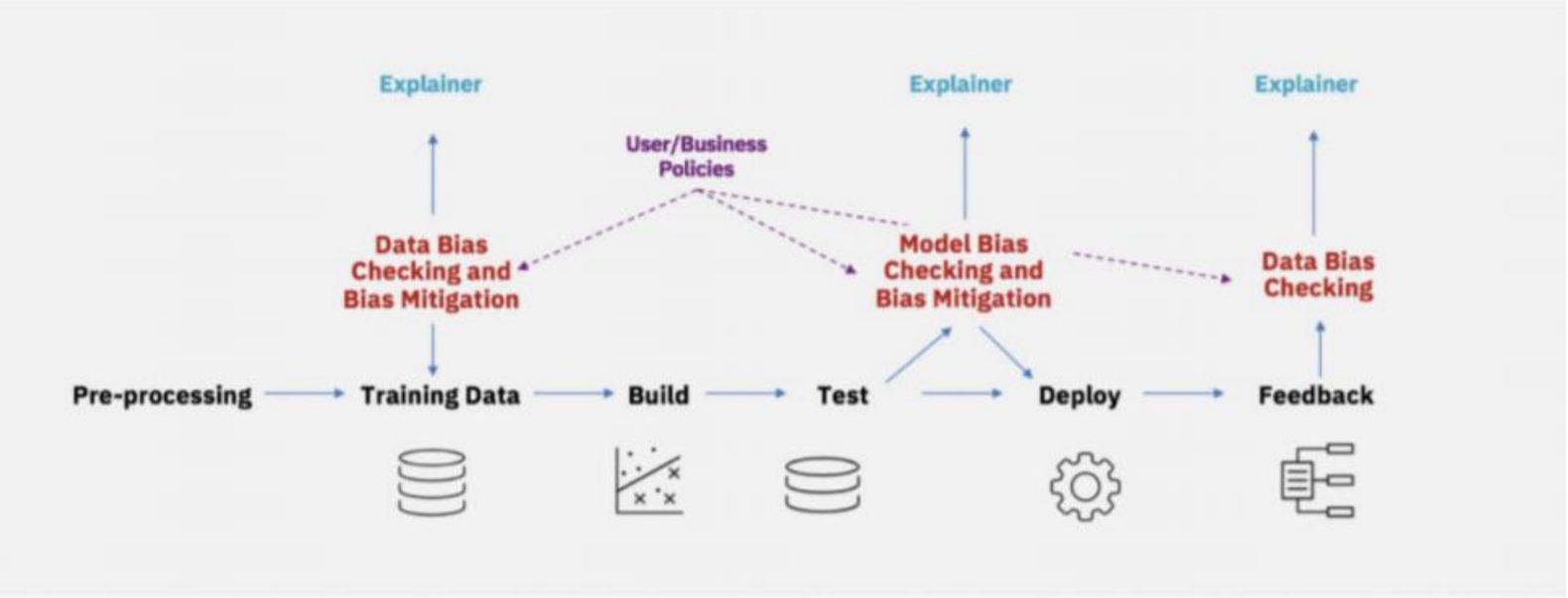
source: https://github.com/IBM/AIF360

AI Fairness 360 was created by IBM Research Additional research sites that advance other aspects of Trusted AI include: AI Explainability 360, AI Adversarial Robustness 360, AI FactSheets 360



THE DEVELOPER'S CONFERENCE

Checking and Mitigating Bias throughout the AI Lifecycle



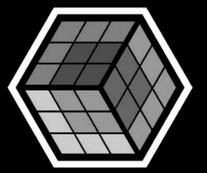
THE DEVELOPER'S CONFERENCE

Bias mitigation is not easy

Cannot simply drop protected attributes because features are correlated with them.

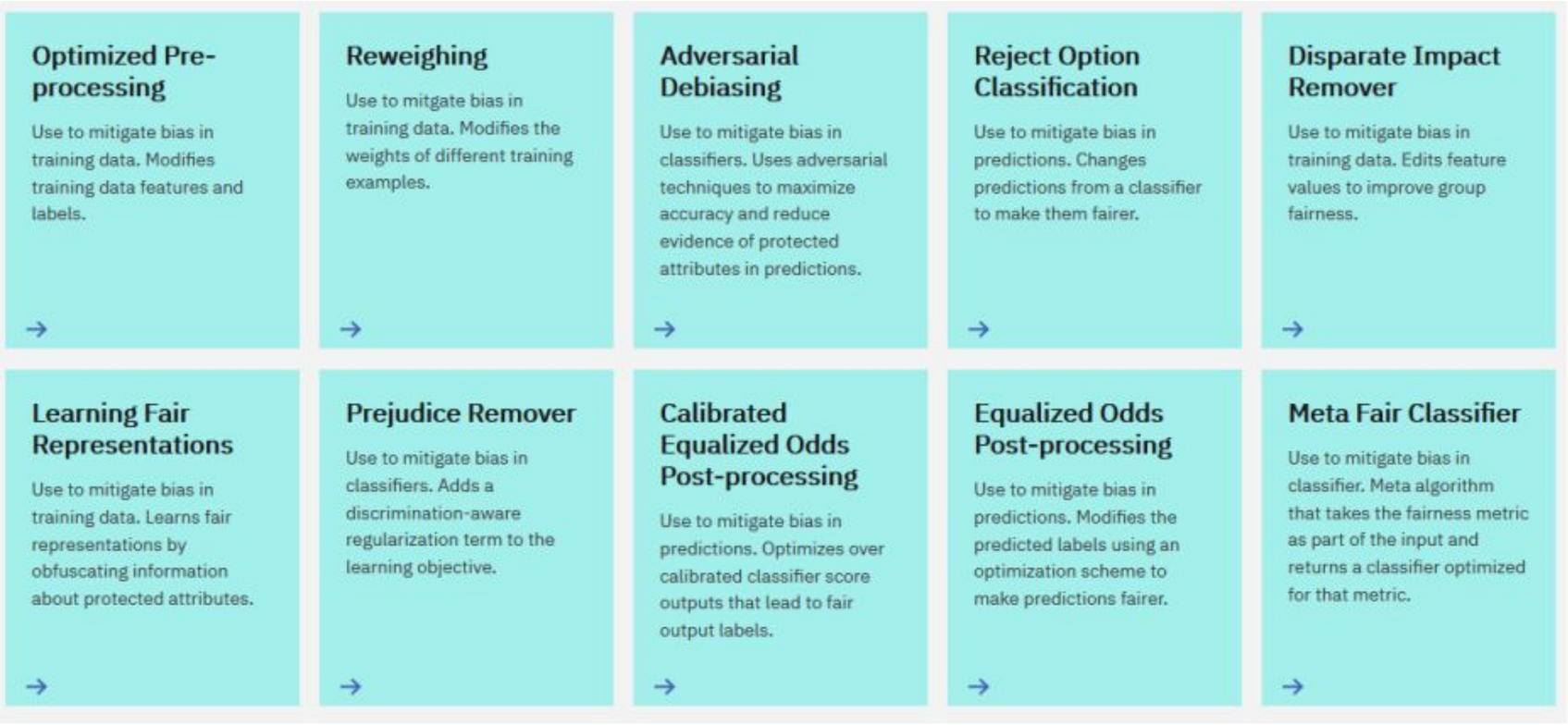
AI Fairness 360 - Resources

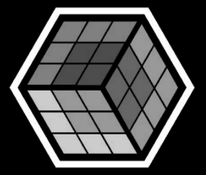
- Bias
- Group fairness
- Privileged protected attribute
- Protected attribute



THE DEVELOPER'S CONFERENCE

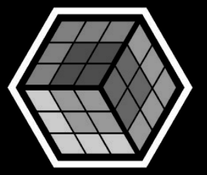
Some Algorithms





THE DEVELOPER'S CONFERENCE

Some Metrics



THE DEVELOPER'S CONFERENCE

Statistical Parity Difference

The difference of the rate of favorable outcomes received by the unprivileged group to the privileged group.

→

Equal Opportunity Difference

The difference of true positive rates between the unprivileged and the privileged groups.

→

Average Odds Difference

The average difference of false positive rate (false positives/negatives) and true positive rate (true positives/positives) between unprivileged and privileged groups.

→

Disparate Impact

The ratio of rate of favorable outcome for the unprivileged group to that of the privileged group.

→

Theil Index

Measures the inequality in benefit allocation for individuals.

→

Euclidean Distance

The average Euclidean distance between the samples from the two datasets.

→

Mahalanobis Distance

The average Mahalanobis distance between the samples from the two datasets.

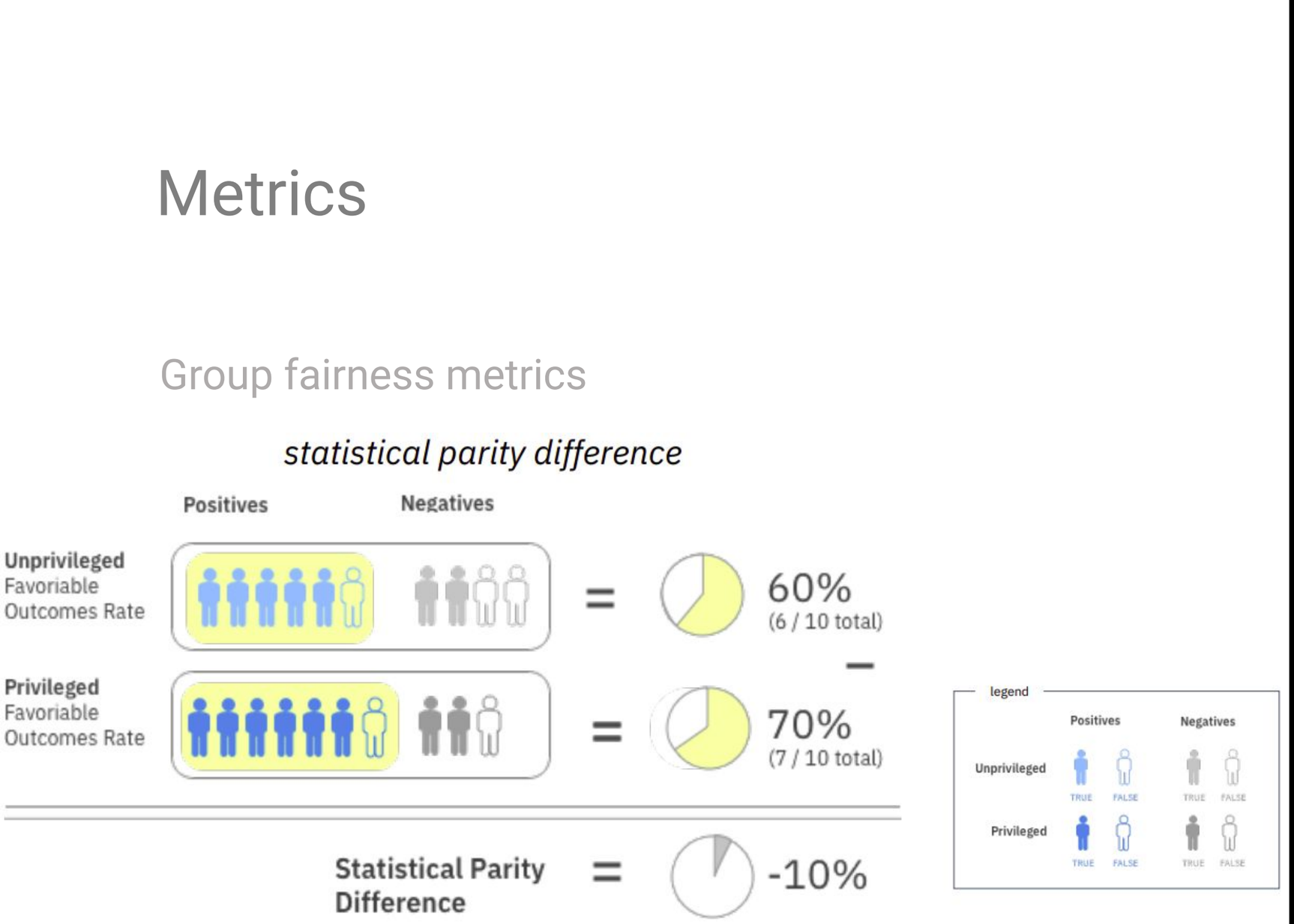
→

Manhattan Distance

The average Manhattan distance between the samples from the two datasets.

→

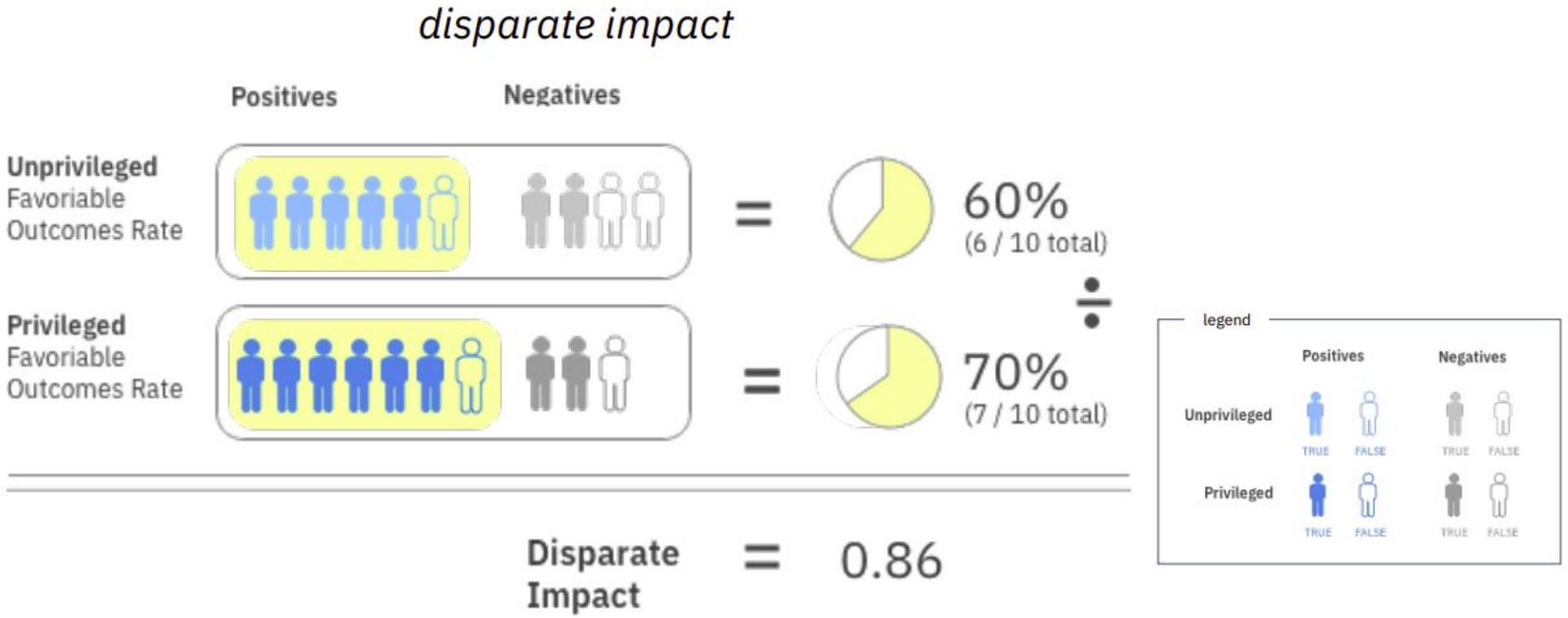
Metrics



Group fairness metrics

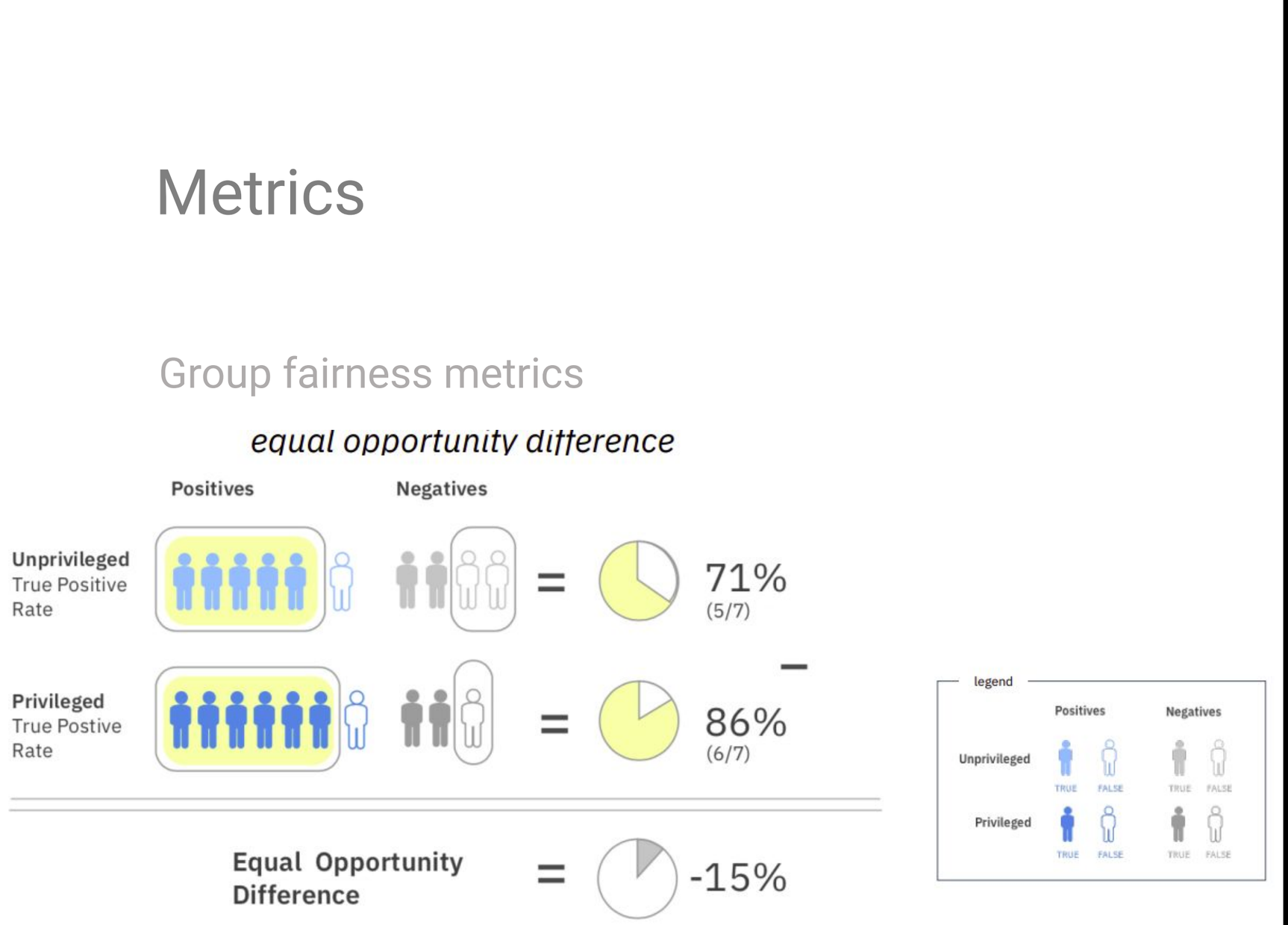
statistical parity difference

Metrics



THE DEVELOPER'S CONFERENCE

Metrics



THE DEVELOPER'S CONFERENCE

THE DEVELOPER'S CONFERENCE

Demo: AI Fairness 360 Web Application

Dataset: Adult census income

Predict whether income exceeds $50K/yr based on census data.

link:
https://archive.ics.uci.edu/ml/datasets/adult

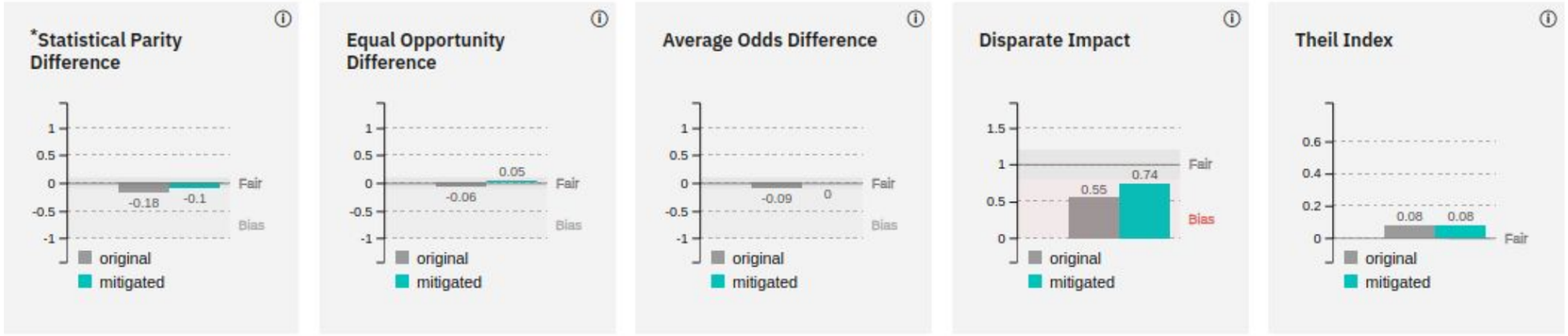
http://aif360.mybluemix.net/

Compare original vs. mitigated results

Dataset: Adult census income
Mitigation: Reweighing algorithm applied

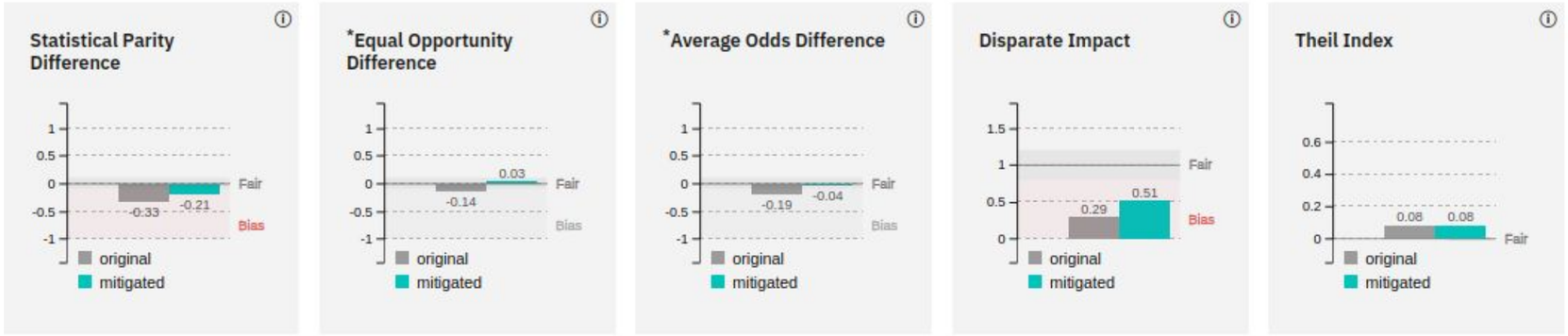
Protected Attribute: Race

Privileged Group: ***White***, Unprivileged Group: ***Non-white***
Accuracy after mitigation changed from 83% to 82%
Bias against unprivileged group was reduced to acceptable levels* for 1 of 2 previously biased metrics (1 of 5 metrics still indicate bias for unprivileged group)



Protected Attribute: Sex

Privileged Group: ***Male***, Unprivileged Group: ***Female***
Accuracy after mitigation changed from 83% to 81%
Bias against unprivileged group was reduced to acceptable levels* for 2 of 4 previously biased metrics (2 of 5 metrics still indicate bias for unprivileged group)



THE DEVELOPER'S CONFERENCE

References

Audit-AI Python library built on top of scikit-learn with various statistical tests for classification and regression tasks
https://github.com/pymetrics/audit-ai

Aequitas Web audit tool as well as python lib. Generates bias report for given model and dataset **https://github.com/dssg/aequitas**

Fairtest Tests for associations between algorithm outputs and protected populations **https://github.com/columbia/fairtest**

Themis Takes a black-box decision-making procedure and designs test cases automatically to explore where the procedure might be exhibiting group-based or causal discrimination **https://github.com/LASER-UMASS/Themis**

Fairness Measures Framework to test given algorithm on variety of datasets and fairness metrics https://github.com/megantosh/fairness_measures_code
Fairness Comparison Extensible test-bed to facilitate direct comparisons of algorithms with respect to fairness measures. Includes raw & preprocessed datasets https://github.com/algofairness/fairness-comparison

FairML Looks at significance of model inputs to quantify prediction dependence on inputs **https://github.com/adebayoj/fairml**

Themis-ML Python library built on scikit-learn that implements fairness-aware machine learning algorithms https://github.com/cosmicBboy/themis-ml

About Generative and Discriminative models https://medium.com/@jordi299/about-generative-and-discriminative-models-d8958b67ad32
Programming Fairness in Algorithms https://www.topbots.com/programming-fairness-in-algorithms/
Fair AI in Practice https://www.umsec.umn.edu/sites/www.umsec.umn.edu/files/presentations/Bellamy.FairAIinPractice.sm_.pdf
IBM AI workflow https://www.coursera.org/specializations/ibm-ai-workflow

Tutorial Credit Scoring https://nbviewer.jupyter.org/github/IBM/AIF360/blob/master/examples/tutorial_credit_scoring.ipynb

https://www.kaggle.com/nathanlauga/ethics-and-ai-how-to-prevent-bias-on-ml

https://jeremykun.com/2015/10/19/one-definition-of-algorithmic-fairness-statistical-parity
https://aif360.readthedocs.io/en/latest/modules/metrics.html#aif360.metrics.ClassificationMetric.theil_index
/

THE DEVELOPER'S CONFERENCE

Thanks

How to reach me:
LinkedIn:/crislanio
Kaggle:/caesarlupum
GitHub: @crislanio
Medium: @crislanio.ufc
Twitter: @crs_macedo

THE DEVELOPER'S CONFERENCE

Q&A section