Removing Unfair Bias in Machine Learning

AI Fairness 360 Open Source Toolkit

Today's Agenda

1. Intro to Fairness & Bias
2. Fairness Metrics & Algorithms
3. Fairness Guidelines
4. Metrics Interactive Demo
5. Medical Use Case - Python Tutorial

October 15, 2019 / © 2019 IBM Corporation

What is Fairness?

- There are 21 definitions of fairness (Narayanan, 2018)
- Many of the definitions conflict
- The way you define fairness impacts bias

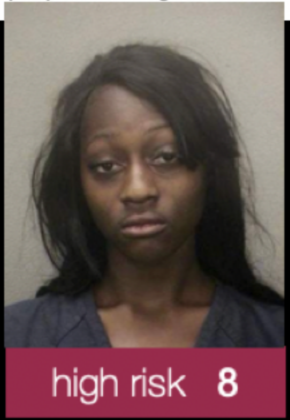
Fairness in Machine Learning Algorithms

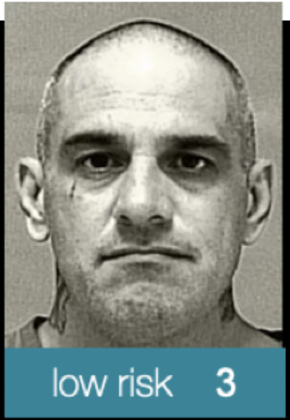
Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

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





Amazon’s AI Recruiting Tool – Taught Gender Bias to Itself

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc’s (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

“How to ensure that the algorithm is fair, how to make sure the algorithm is really interpretable and explainable - that’s still quite far off.”

Apple's 'sexist' credit card investigated by US regulator

DHH @dhh

The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

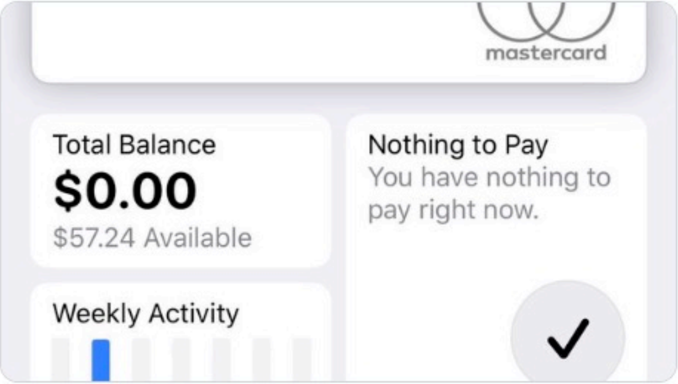
12:34 PM · Nov 7, 2019 · Twitter for iPhone

9.6K Retweets **29.1K** Likes

DHH @dhh · Nov 7, 2019
Replying to @dhh
I'm surprised that they even let her apply for a card without the signed approval of her spouse? I mean, can you really trust women with a credit card these days??!

89 246 4.5K

DHH @dhh · Nov 7, 2019
It gets even worse. Even when she pays off her ridiculously low limit in full, the card won't approve any spending until the next billing period. Women apparently aren't good credit risks even when they pay off the fucking balance in advance and in full.



77 388 3.8K

Steve Wozniak @stevewoz
Replying to @dhh

The same thing happened to us. We have no separate bank accounts or credit cards or assets of any kind. We both have the same high limits on our cards, including our AmEx Centurion card. But 10x on the Apple Card.

10:58 PM · Nov 9, 2019 · Twitter Web App

AI Fairness 360

[Open Source Toolbox to Mitigate Bias]()

– Demos & Tutorials on Industry Use Cases

– Fairness Guidance

– Comprehensive Toolbox

- 75+ Fairness metrics
- 10+ Bias Mitigation Algorithms
- Fairness Metric Explanations

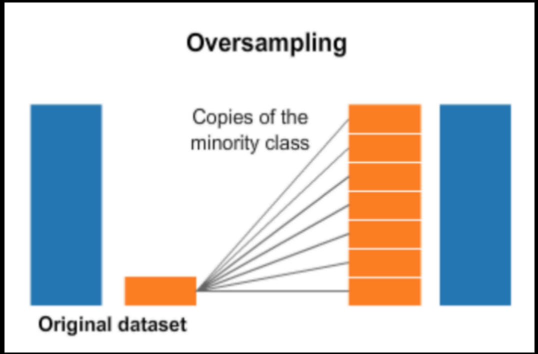
Extensible Toolkit for Detecting, Understanding, & Mitigating Unwanted **Algorithmic Bias**

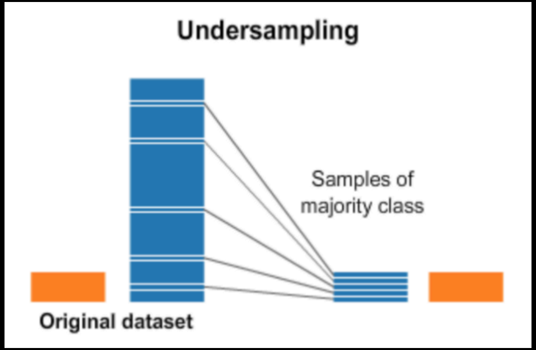
Leading Fairness Metrics and Algorithms from **Industry** & **Academia**

Designed to **translate new research** from the **lab to industry practitioners** (using Scikit Learn's fit/predict paradigm)

Next Section:
How Do You Measure Bias & Where Does it Come From?

Most Bias Come From Your Data –
Over /Under Sampling, Label & User Generated Bias





MIT Study of Top Face Recognition Services

99% accurate for lighter-skinned males

65% accurate for darker-skinned females

Most Bias Come From Your Data –
Over /Under Sampling, Label & User Generated Bias

IBM hopes 1 million faces will help fight bias in facial recognition

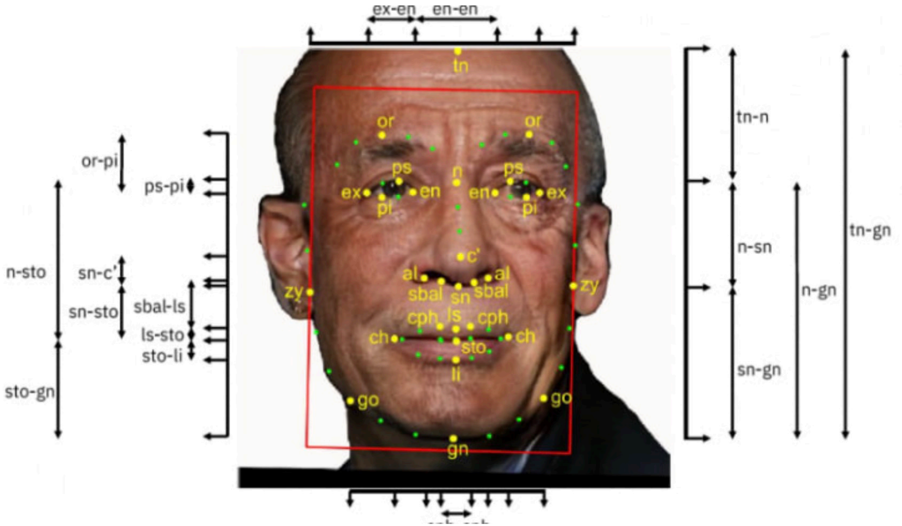
PUBLISHED TUE, JAN 29 2019•12:35 PM EST | UPDATED THU, JAN 31 2019•9:59 AM EST

Ryan Browne
@RYAN_BROWNE_

SHARE

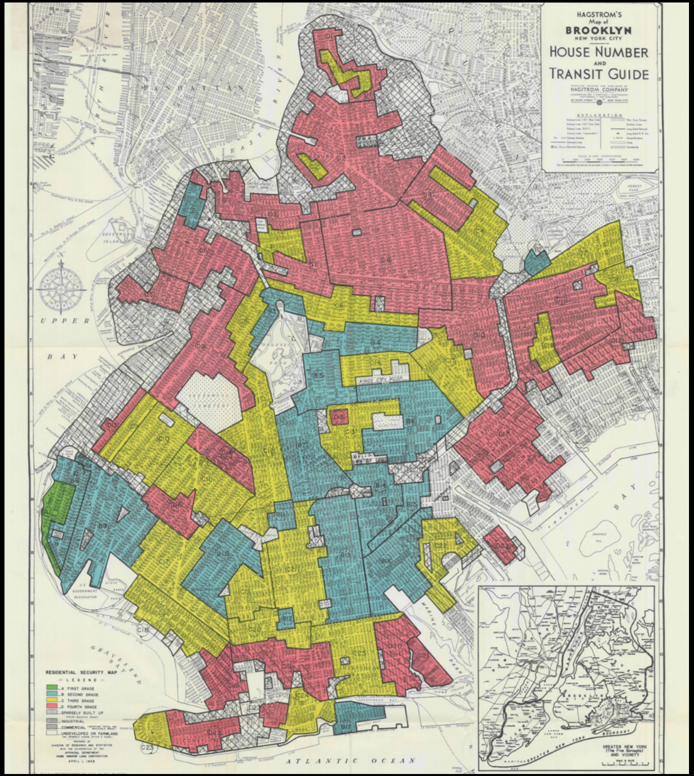
KEY POINTS

- IBM released a trove of data containing 1 million images of faces taken from a Flickr dataset with 100 million photos and videos.
- The images are annotated with tags related to features including craniofacial measurements, facial symmetry, age and gender.
- Researchers at IBM hope this will help developers train their AI-powered facial recognition systems to identify faces more fairly and accurately.



Why Not Just Remove Protected Attributes?

- You can’t just drop protected attributes (gender, race); other features correlated with them
- Example: Buy using zip codes you can deconstruct individual’s race or income



Fairness Terms You Need To Know

Protected Attribute – an attribute that partitions a population into groups whose outcomes should have parity (ex. race, gender, caste, and religion)

Privileged Protected Attribute – a protected attribute value indicating a group that has historically been at systemic advantage

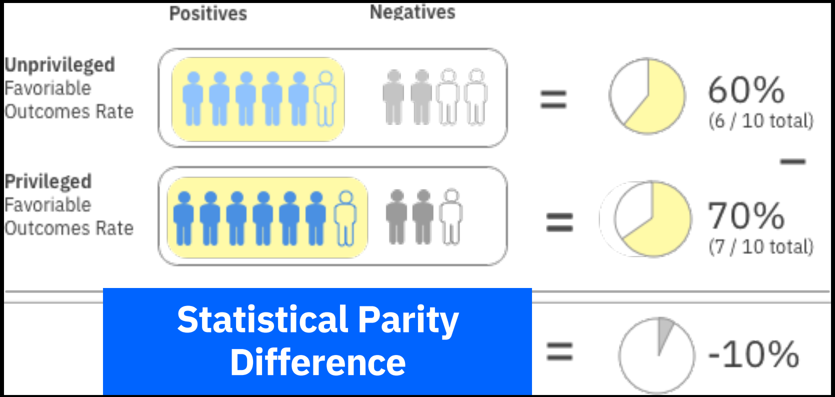
Group Fairness – Groups defined by protected attributes receiving similar treatments or outcomes

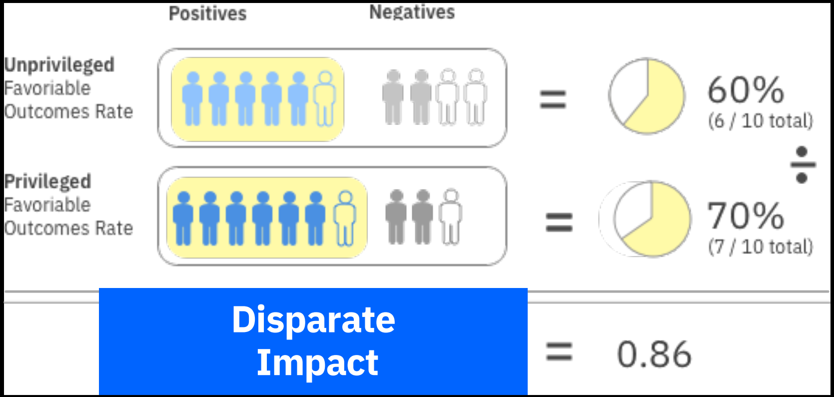
Individual Fairness – Similar individuals receiving similar treatments or outcomes

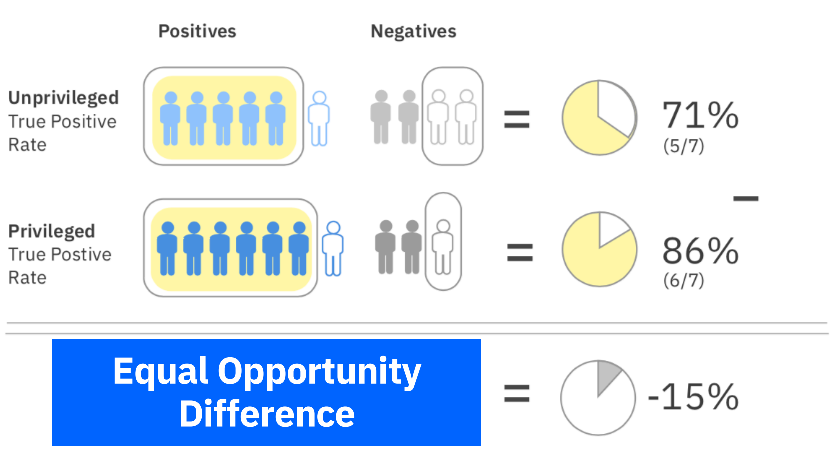
Fairness Metric – a measure of unwanted bias in training data or models

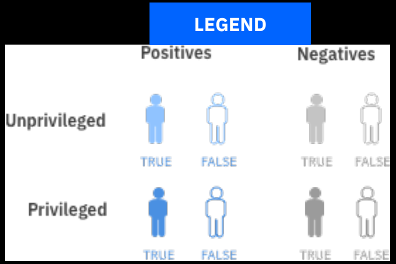
Favorable Label – a label whose value corresponds to an outcome that provides an advantage to the recipient

How To Measure Fairness – Some Group Fairness Metrics









How You Define Fairness Impacts How You Measure It

Do SAT Scores Correctly Compare The Abilities of Applicants?

YES

SAT score correlates well with future success and correctly compare the abilities of applicants

METRICS:

average_odds_difference & average_abs_odds_difference

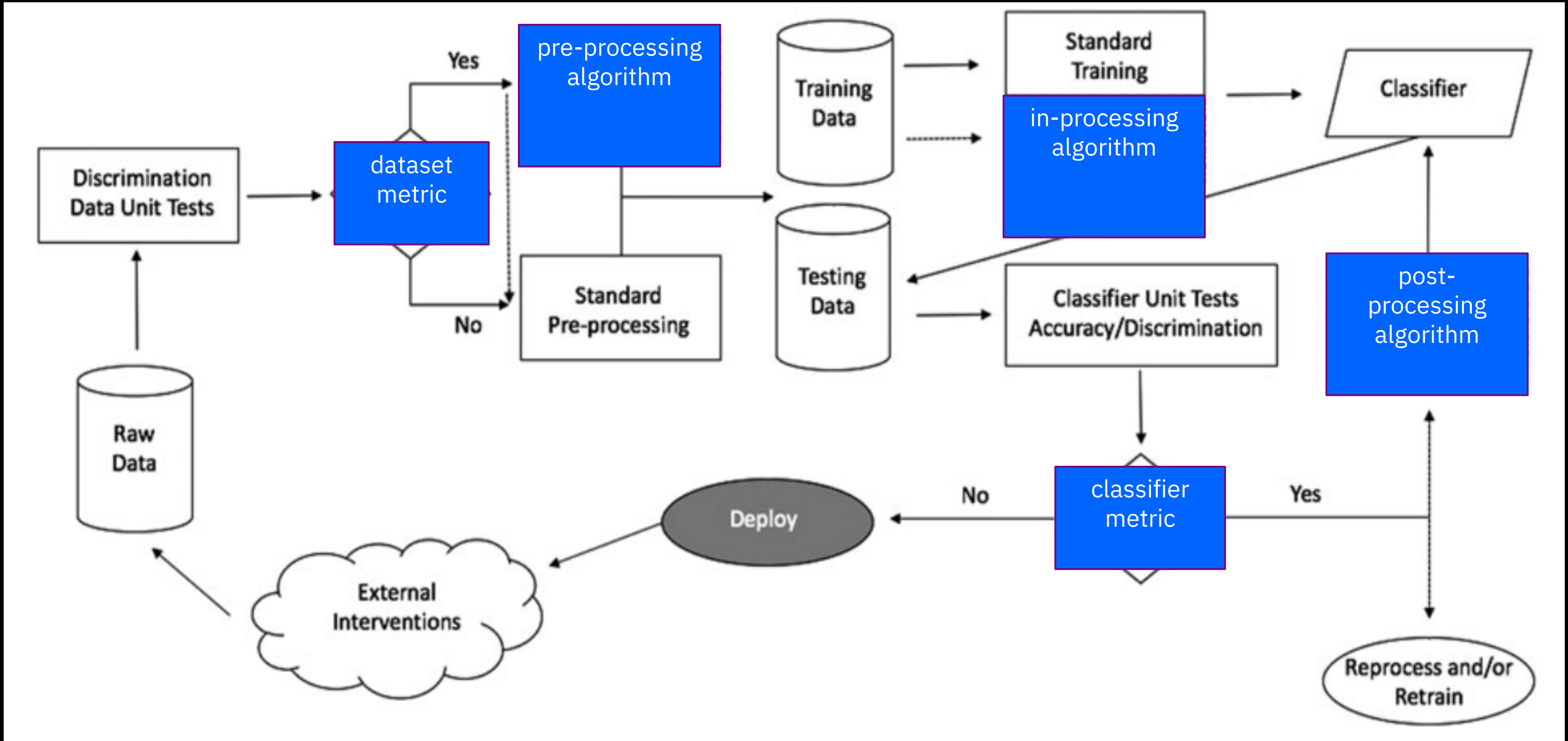
NO

SAT score may contain structural biases so its distribution is different across groups *(non-English speaking parents, single parents, low income, no SAT Prep)*

METRICS:

disparate_impact & statistical_parity_difference

Bias In the Machine Learning Pipeline



Where Can You Intervene in the Pipeline?

Pre-Processing Algorithm

Bias mitigation algorithms applied to Training Data

→

In-Processing Algorithm

Bias mitigation algorithm applied to a model during its training

→

Post-Processing Algorithm

Bias mitigation algorithm applied to predicted labels

- If you can modify the Training Data, then pre-processing can be used.
- If you can modify the Learning Algorithm, then in-processing can be used.
- If you can only treat the learned model as a black box and can't modify the training data or learning algorithm, then only post-processing can be used

Bias Mitigation Algorithms For Each Phase of the Pipeline

Pre-Processing Algorithms Mitigates Bias in **Training Data**	**In-Processing Algorithms** Mitigates Bias in **Classifiers**	**Post-Processing Algorithms** Mitigates Bias in **Predictions**
Reweighing Modifies the weights of different training examples	**Adversarial Debiasing** Uses adversarial techniques to maximize accuracy & reduce evidence of protected attributes in predictions	**Reject Option Classification** Changes predictions from a classifier to make them fairer
Disparate Impact Remover Edits feature values to improve group fairness	**Prejudice Remover** Adds a discrimination-aware regularization term to the learning objective	**Calibrated Equalized Odds** Optimizes over calibrated classifier score outputs that lead to fair output labels
Optimized Preprocessing Modifies training data features & labels	**Meta Fair Classifier** Takes the fairness metric as part of the input & returns a classifier optimized for the metric	**Equalized Odds** Modifies the predicted label using an optimization scheme to make predictions fairer
Learning Fair Representations Learns fair representations by obfuscating information about protected attributes		

AIF360 Includes The Top Algorithms In Industry/Academia

- Optimized Preprocessing (Calmon et al., NIPS 2017)
- Meta-Algorithm for Fair Classification (Celis et al., FAT* 2019)
- Disparate Impact Remover (Feldman et al., KDD 2015)
- Equalized Odds Postprocessing (Hardt et al., NIPS 2016)
- Reweighing (Kamiran and Calders, KIS 2012)
- Reject Option Classification (Kamiran et al., ICDM 2012)
- Prejudice Remover Regularizer (Kamishima et al., ECML PKDD 2012)
- Calibrated Equalized Odds Postprocessing (Pleiss et al., NIPS 2017)
- Learning Fair Representations (Zemel et al., ICML 2013)
- Adversarial Debiasing (Zhang et al., AIES 2018)

Pre-Processing is the Optimal Time to Mitigate Bias

Pre-Processing Algorithms
Mitigates Bias in **Training Data**

Reweighing

Modifies the weights of different training examples

Disparate Impact Remover

Edits feature values to improve group fairness

Optimized Preprocessing

Modifies training data features & labels

Learning Fair Representations

Learns fair representations by obfuscating information about protected attributes

Reweighing only changes **Weights** applied to training samples (no changes to feature/labels). Ideal if you cannot change values

Disparate Impact Remover and **Optimized Preprocessing** yield modified datasets in the same space as the input training data (provides transparency)

Learning Fair Representations yields modified datasets in the latent space

Tradeoffs - Bias vs. Accuracy

1. Is your model doing good things or bad things to people?
   - If your model is sending people to jail, may be better to have more false positives than false negatives
   - If your model is handing out loans, may be better to have more False Negatives than False Positives
2. Determine your threshold for accuracy vs. fairness based upon your legal, ethical and trust guidelines

LEGAL
Doing what is legal is top priority

ETHICAL
What's your company's Ethics

TRUST
Losing customer's Trust costly

Preventing Bias Is Hard!

Work with your stakeholders early to define fairness, protected attributes & thresholds

Apply the earliest mitigation in the pipeline that you have permission to apply

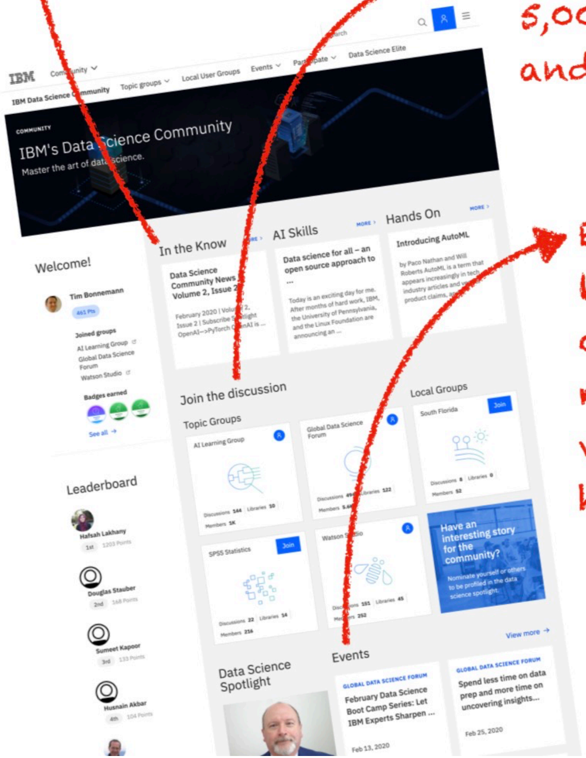
Check for bias as often as possible using any metrics that are applicable

Caveat: AIF360 should only be used with well defined data sets & well defined use cases

Learn:
Resources galore &
experts on tap (just ask)

Share:
5,000+ members
and counting!

Engage:
Webinars,
online
meetups &
virtual
hackathon*



*Details soon!

Join the IBM Data Science Community & get a complimentary month of Coursera!

http://ibm.biz/SF-Python-Coursera

Join the AI Fairness Slack Channel

Join the AIF360 Slack https://aif360.slack.com/

Ask questions and speak to AI Fairness 360 researchers, experts, and developers

Next Section:
Toolkit Overview & Interactive Demo

AI Fairness 360 Toolkit Overview

https://aif360.mybluemix.net/

IBM Research Trusted AI

Home Demo Resources Events Videos Community

AI Fairness 360 Open Source Toolkit

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. Containing over 70 fairness metrics and 10 state-of-the-art bias mitigation algorithms developed by the research community, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

API Docs ↗ Get Code ↗

Not sure what to do first? Start here!

Read More
Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.

→

Try a Web Demo
Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.

→

Watch Videos
Watch videos to learn more about AI Fairness 360.

→

Read a paper
Read a paper describing how we designed AI Fairness 360.

→

Use Tutorials
Step through a set of in-depth examples that introduces developers to code that checks and mitigates bias in different industry and application domains.

→

Ask a Question
Join our AIF360 Slack Channel to ask questions, make comments and tell stories about how you use the toolkit.

→

View Notebooks
Open a directory of Jupyter Notebooks in GitHub that provide working examples of bias detection and mitigation in sample datasets. Then share your own notebooks!

→

Contribute
You can add new metrics and algorithms in GitHub. Share Jupyter notebooks show-casing how you have examined and mitigated bias in your machine learning application.

→

Learn how to put this toolkit to work for your application or industry problem. Try these tutorials.

Credit Scoring
See how to detect and mitigate age bias in predictions of credit-worthiness using the German Credit dataset.

→

Medical Expenditure
See how to detect and mitigate racial bias in a care management scenario using Medical Expenditure Panel Survey data.

→

Gender Bias in Face Images
See how to detect and mitigate bias in automatic gender classification of face images.

→

Interactive Demo

https://aif360.mybluemix.net/data

IBM Research Trusted AI | Home | **Demo** | Resources | Events | Videos | Community

AI Fairness 360 - Demo

Data — Check — Mitigate — Compare

1. Choose sample data set

Bias occurs in data used to train a model. We have provided three sample datasets that you can use to explore bias checking and mitigation. Each dataset contains attributes that should be protected to avoid bias.

◯ **Compas (ProPublica recidivism)**

Predict a criminal defendant's likelihood of reoffending.
Protected Attributes:
- **Sex**, privileged: ***Female***, unprivileged: ***Male***
- **Race**, privileged: ***Caucasian***, unprivileged: ***Not Caucasian***

Learn more

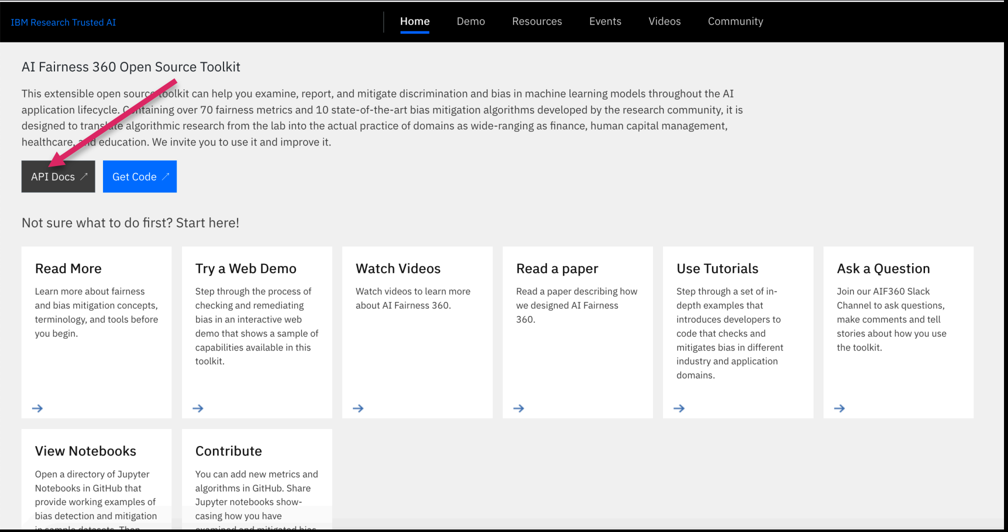
◯ **German credit scoring**

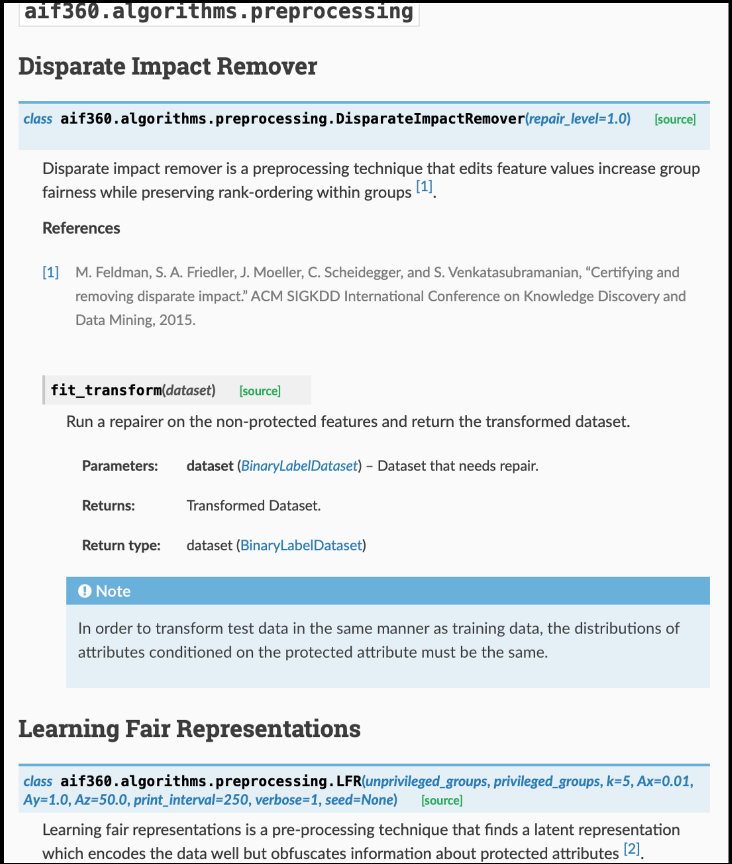
Predict an individual's credit risk.
Protected Attributes:
- **Sex**, privileged: ***Male***, unprivileged: ***Female***
- **Age**, privileged: ***Old***, unprivileged: ***Young***

Learn more

◯ **Adult census income**

Toolkit API – Definitions, Formulas & References




`aif360.algorithms.preprocessing`

Disparate Impact Remover

class `aif360.algorithms.preprocessing.DisparateImpactRemover`(*repair_level=1.0*) [source]

Disparate impact remover is a preprocessing technique that edits feature values increase group fairness while preserving rank-ordering within groups [1].

References

[1] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact." ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.

`fit_transform`(*dataset*) [source]

Run a repairer on the non-protected features and return the transformed dataset.

Parameters: **dataset** (*BinaryLabelDataset*) – Dataset that needs repair.

Returns: Transformed Dataset.

Return type: dataset (BinaryLabelDataset)

Note

In order to transform test data in the same manner as training data, the distributions of attributes conditioned on the protected attribute must be the same.

Learning Fair Representations

class `aif360.algorithms.preprocessing.LFR`(*unprivileged_groups, privileged_groups, k=5, Ax=0.01, Ay=1.0, Az=50.0, print_interval=250, verbose=1, seed=None*) [source]

Learning fair representations is a pre-processing technique that finds a latent representation which encodes the data well but obfuscates information about protected attributes [2].

Next Section:
Medical Use Case
Python Tutorial

Toolkit API – Definitions, Formulas & References

IBM Research Trusted AI

Home | Demo | Resources | Events | Videos | Community

AI Fairness 360 Open Source Toolkit

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. Containing over 70 fairness metrics and 10 state-of-the-art bias mitigation algorithms developed by the research community, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

API Docs ↗ | Get Code ↗

Not sure what to do first? Start here!

Read More
Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.

Try a Web Demo
Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.

Watch Videos
Watch videos to learn more about AI Fairness 360.

Read a paper
Read a paper describing how we designed AI Fairness 360.

Use Tutorials
Step through a set of in-depth examples that introduces developers to code that checks and mitigates bias in different industry and application domains.

Ask a Question
Join our AIF360 Slack Channel to ask questions, make comments and tell stories about how you use the toolkit.

View Notebooks
Open a directory of Jupyter Notebooks in GitHub that provide working examples of bias detection and mitigation

Contribute
You can add new metrics and algorithms in GitHub. Share Jupyter notebooks showcasing how you have

`aif360.algorithms.preprocessing`

Disparate Impact Remover

class `aif360.algorithms.preprocessing.DisparateImpactRemover`(*repair_level=1.0*) [source]

Disparate impact remover is a preprocessing technique that edits feature values increase group fairness while preserving rank-ordering within groups [1].

References

[1] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact." ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.

`fit_transform`(*dataset*) [source]

Run a repairer on the non-protected features and return the transformed dataset.

Parameters: dataset (*BinaryLabelDataset*) – Dataset that needs repair.

Returns: Transformed Dataset.

Return type: dataset (BinaryLabelDataset)

Note

In order to transform test data in the same manner as training data, the distributions of attributes conditioned on the protected attribute must be the same.

Learning Fair Representations

class `aif360.algorithms.preprocessing.LFR`(*unprivileged_groups, privileged_groups, k=5, Ax=0.01, Ay=1.0, Az=50.0, print_interval=250, verbose=1, seed=None*) [source]

Learning fair representations is a pre-processing technique that finds a latent representation which encodes the data well but obfuscates information about protected attributes [2].

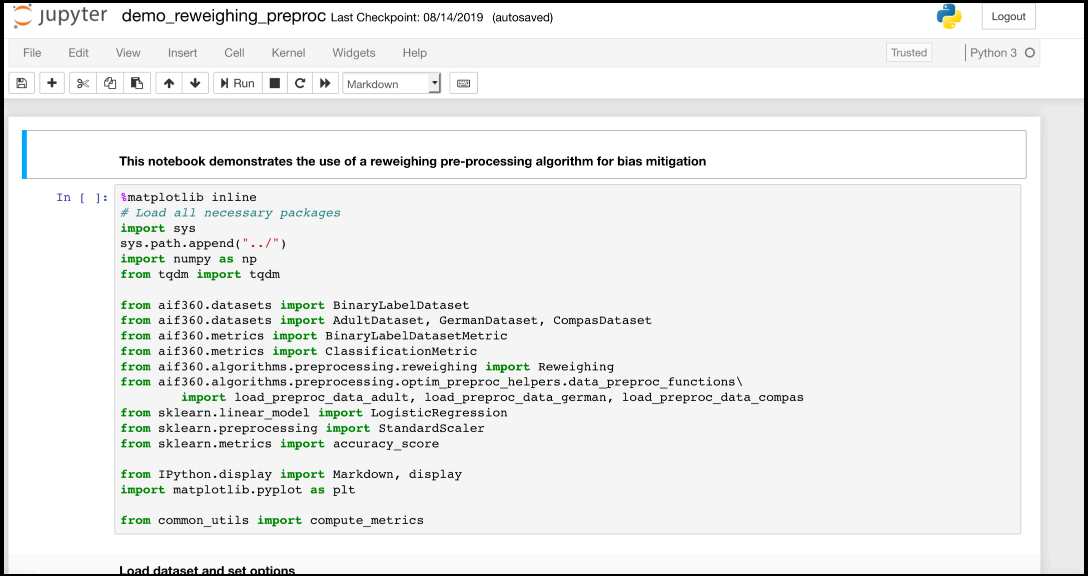
Reweighing Preprocessing Algorithm

- Demonstrate how AIF 360 Toolkit can be used to detect and reduce bias with a learning classifier using a fairness metrics and algorithms

- Classifiers are built using Logistic Regression

- Bias Detection Metrics Used:
    - Disparate Impact
    - Average Odds Difference

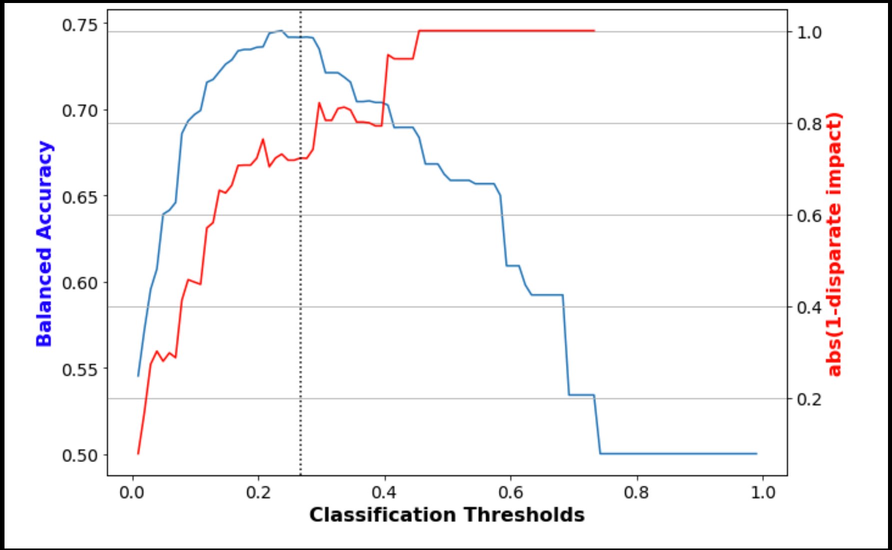
- Bias Mitigation Algorithm Used:
    - Reweighing (pre-processing algorithm)

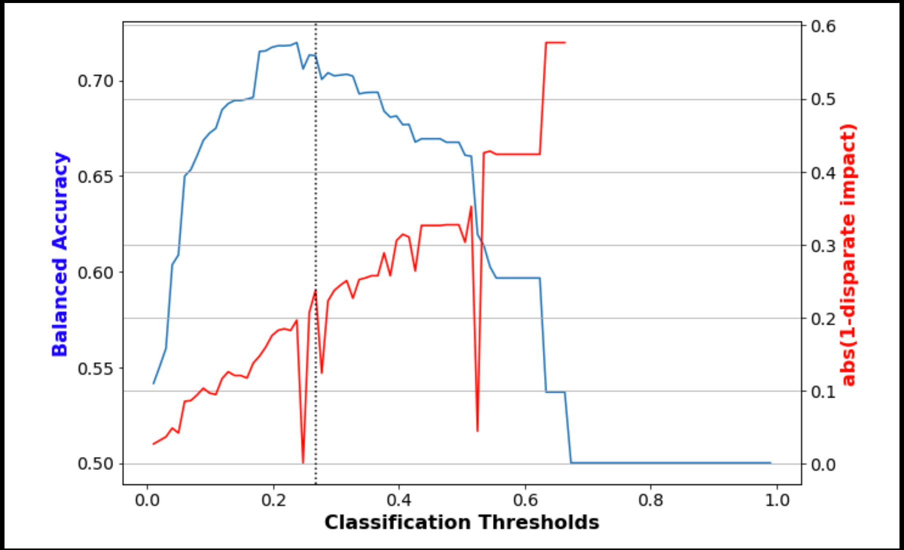



```python
%matplotlib inline
# Load all necessary packages
import sys
sys.path.append("../")
import numpy as np
from tqdm import tqdm

from aif360.datasets import BinaryLabelDataset
from aif360.datasets import AdultDataset, GermanDataset, CompasDataset
from aif360.metrics import BinaryLabelDatasetMetric
from aif360.metrics import ClassificationMetric
from aif360.algorithms.preprocessing.reweighing import Reweighing
from aif360.algorithms.preprocessing.optim_preproc_helpers.data_preproc_functions\
        import load_preproc_data_adult, load_preproc_data_german, load_preproc_data_compas
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score

from IPython.display import Markdown, display
import matplotlib.pyplot as plt

from common_utils import compute_metrics
```

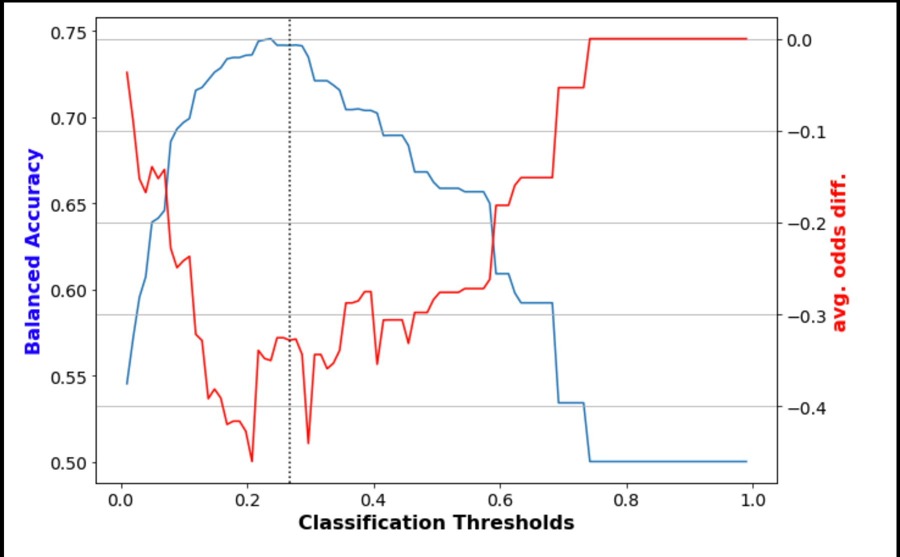
Load dataset and set options

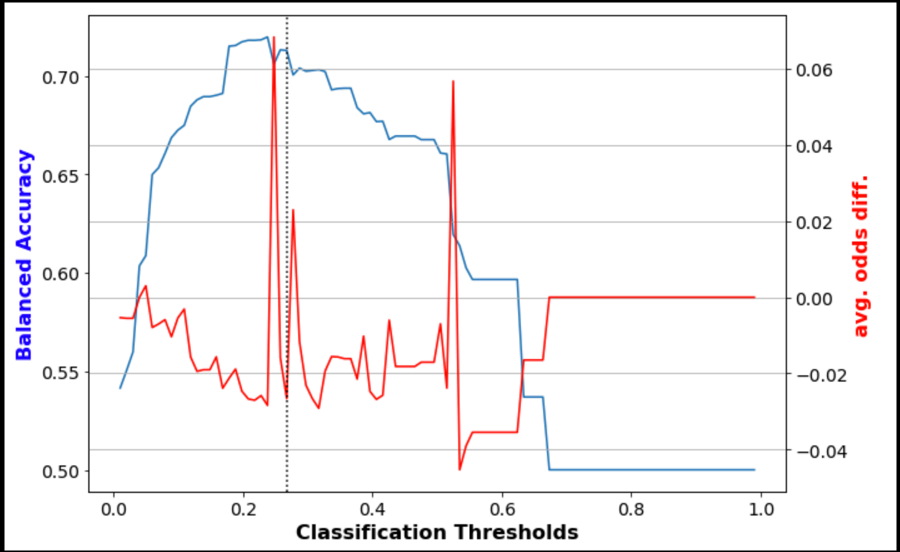
Bias Metric – Disparate Impact





Bias Metric - Average Odds Difference





Bias Metric - Average Odds Difference

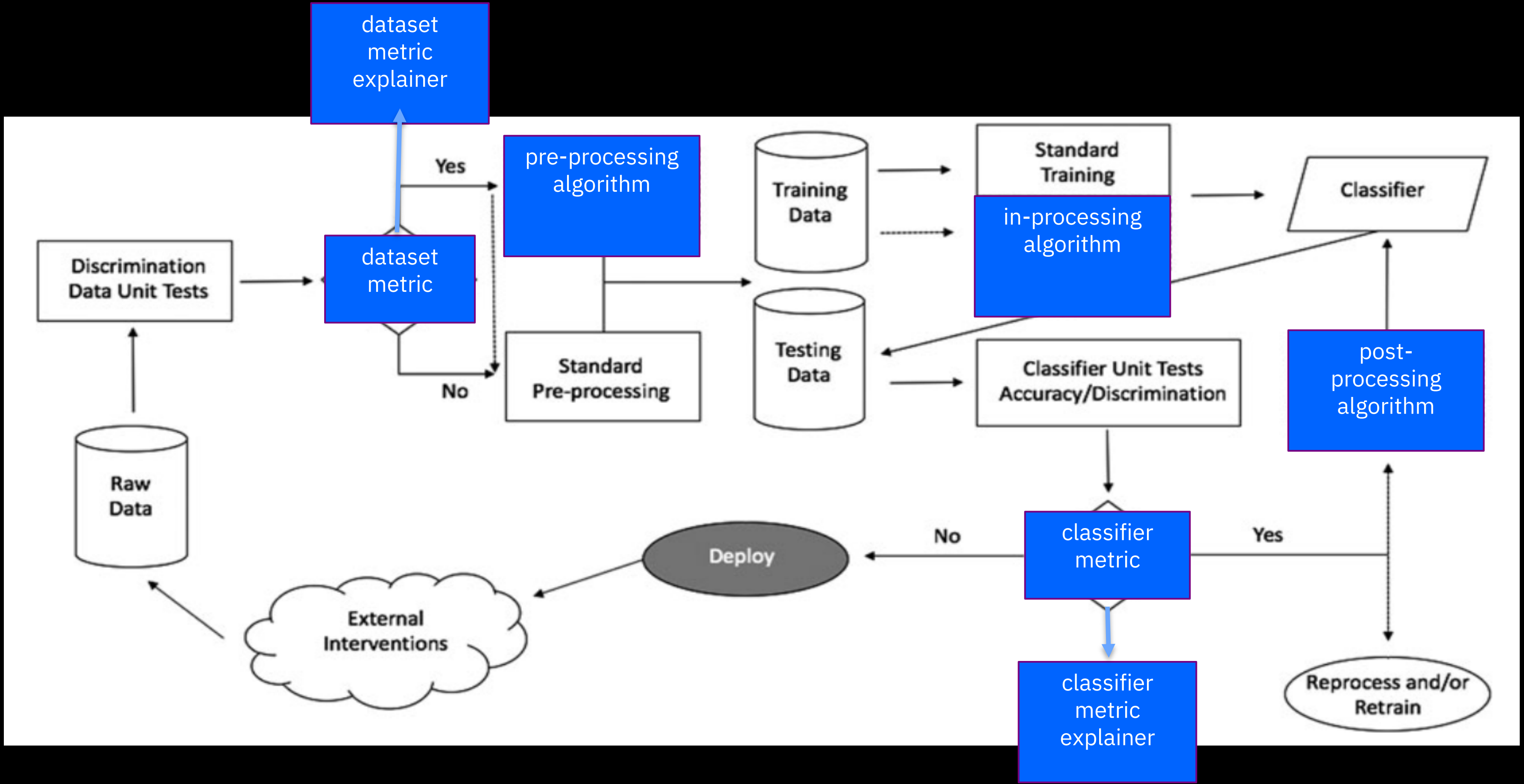
	Disparate Impact	
"Sex"	Before	After
Accuracy	0.7417	0.7128
Fairness	0.2774	0.7625

Disparate Impact
Fairness is Closer to 1 (Ratio)

	Average Odds Difference	
"Sex"	Before	After
Accuracy	0.7417	0.7128
Fairness	-0.3281	-0.0266

Average Odds Difference
Fairness is Closer to 0 (Difference)

Metrics, Algorithms & Explainers



Join the AI Fairness Slack Channel

Join the AIF360 Slack https://aif360.slack.com/

Ask questions and speak to AI Fairness 360 researchers, experts, and developers