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

IBM Data & AI / July 16, 2019 / © 2019 IBM Corporation

Why Do We Care About Fairness?

What is Fairness?

- There are 21 definitions of fairness
- Many of the definitions conflict
- The way you define fairness impacts bias

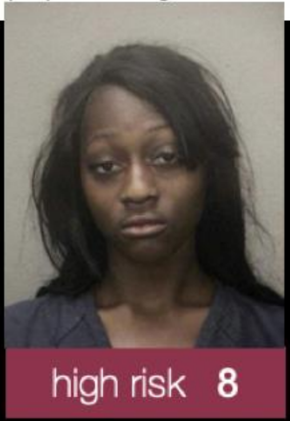
Fairness in Machine Learning Algorithms

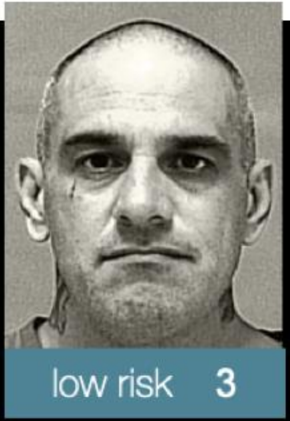
Prediction Fails Differently for Black Defendants

	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.”

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

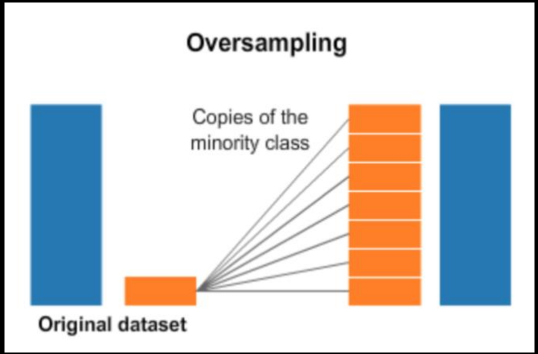
Extensible Toolkit for Detecting, Understanding, & Mitigating Unwanted **Algorithmic Bias**

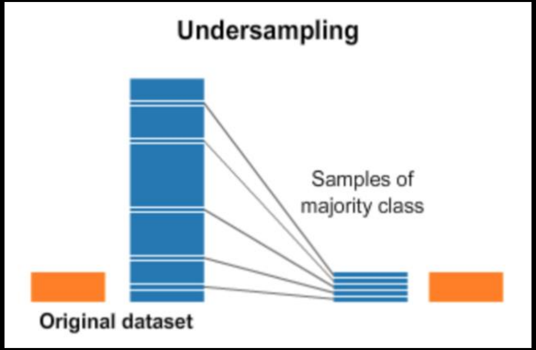
Leading Fairness Metrics and Algorithms from **Industry** & **Academia**

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





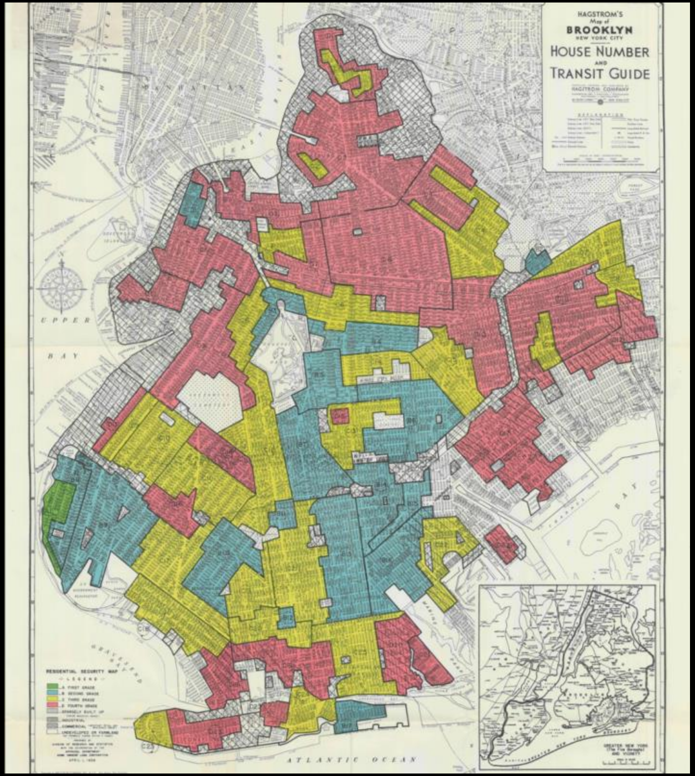
MIT Study of Top Face Recognition Services

99% accurate
for lighter-skinned males

65% accurate
for darker-skinned
females

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

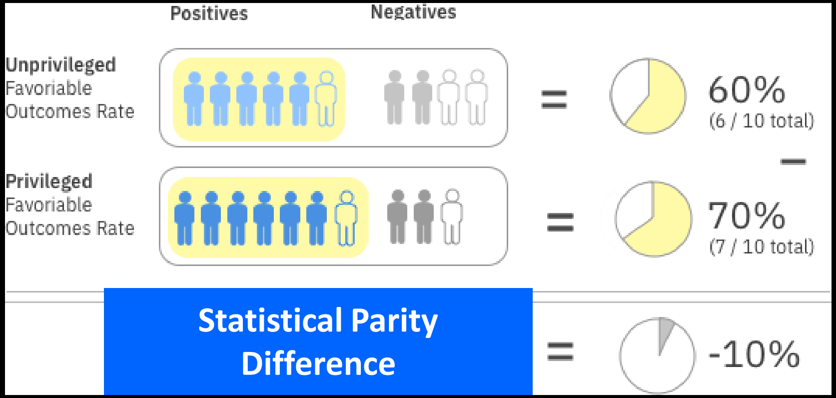
Group Fairness – Groups defined by protected attributes receiving similar treatments or outcomes

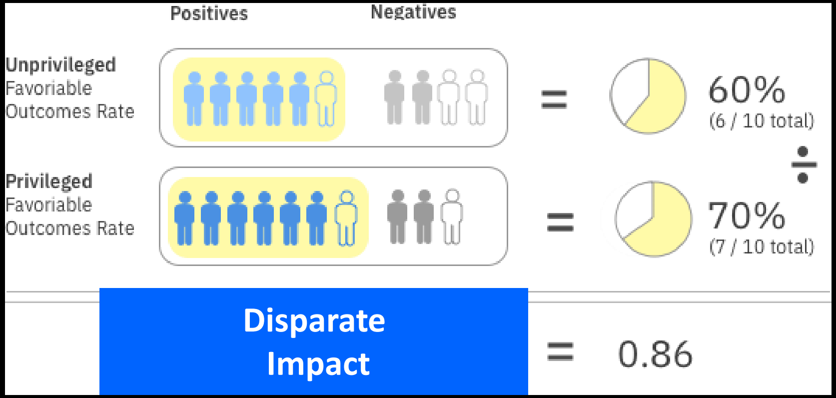
Individual Fairness –Similar individuals receiving similar treatments or outcomes

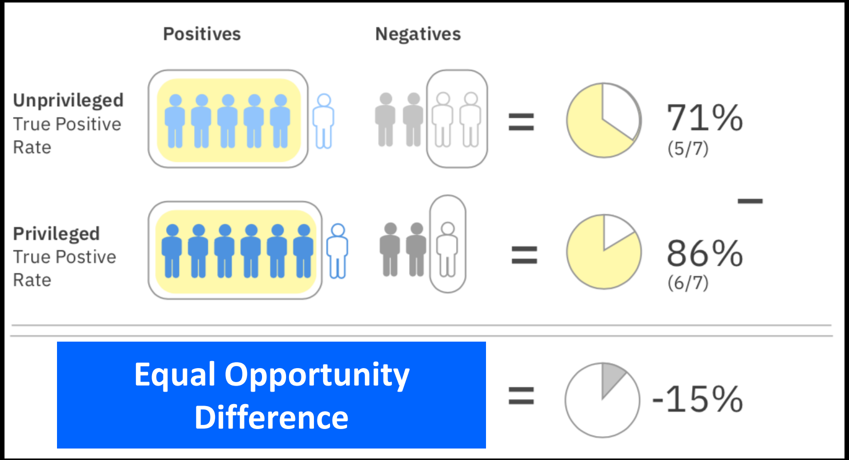
Fairness Metric – a measure of unwanted bias in training data or models

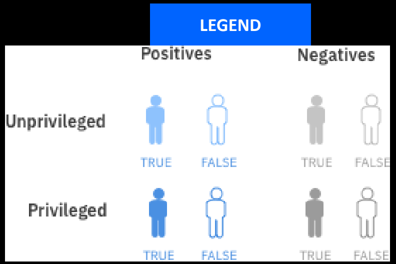
Favorable Label – a label whose value corresponds to an outcome that provides an advantage to the recipient

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

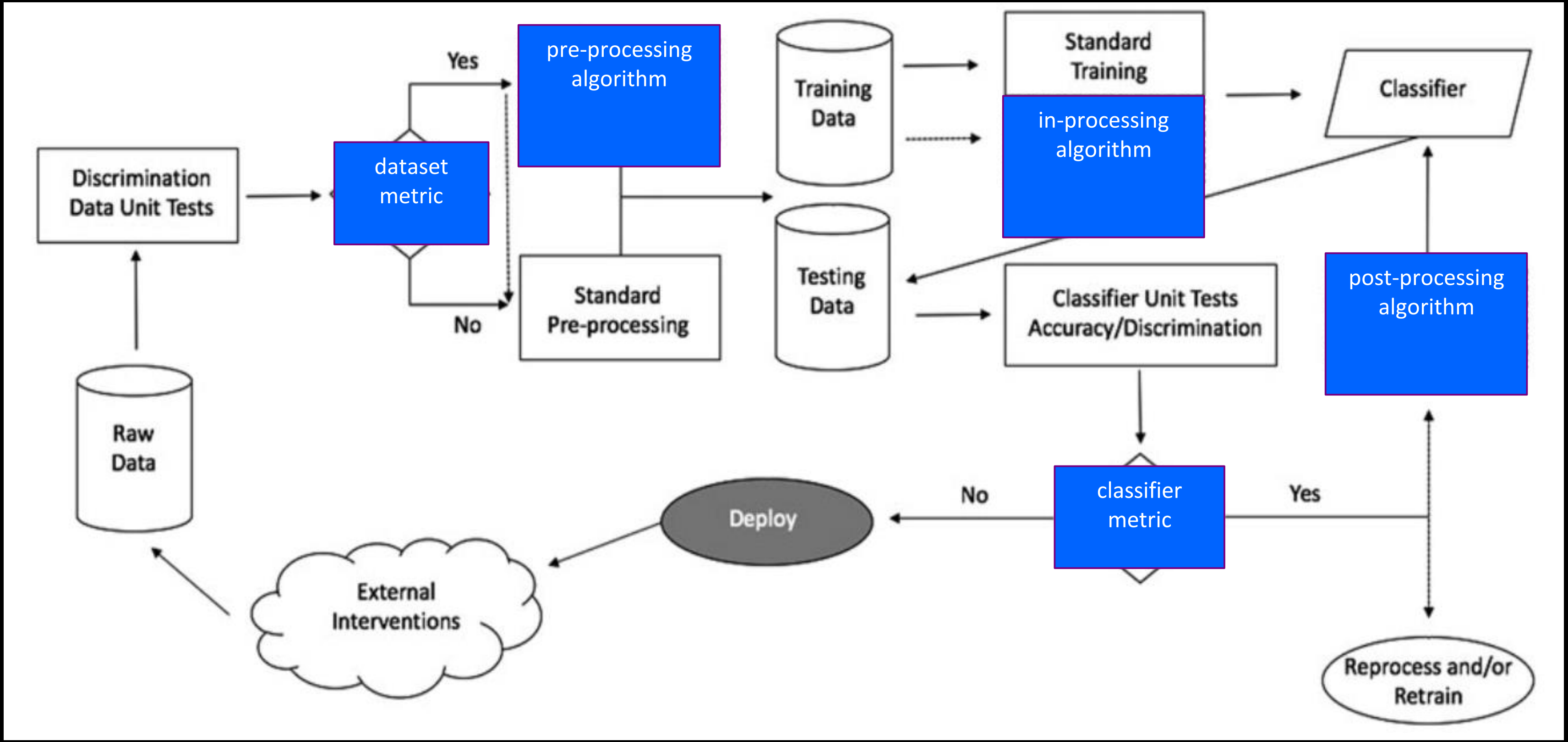
NO

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**

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

In-Processing Algorithms Mitigates Bias in **Classifiers**

Adversarial Debiasing

Uses adversarial techniques to maximize accuracy & reduce evidence of protected attributes in predictions

Prejudice Remover

Adds a discrimination-aware regularization term to the learning objective

Meta Fair Classifier

Takes the fairness metric as part of the input & returns a classifier optimized for the metric

Post-Processing Algorithms Mitigates Bias in **Predictions**

Reject Option Classification

Changes predictions from a classifier to make them fairer

Calibrated Equalized Odds

Optimizes over calibrated classifier score outputs that lead to fair output labels

Equalized Odds

Modifies the predicted label using an optimization scheme to make predictions fairer

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 (Penalties)

ETHICAL
What's your company's Ethics (Amazon Echo)

TRUST
Losing customer's Trust costly (Facebook)

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

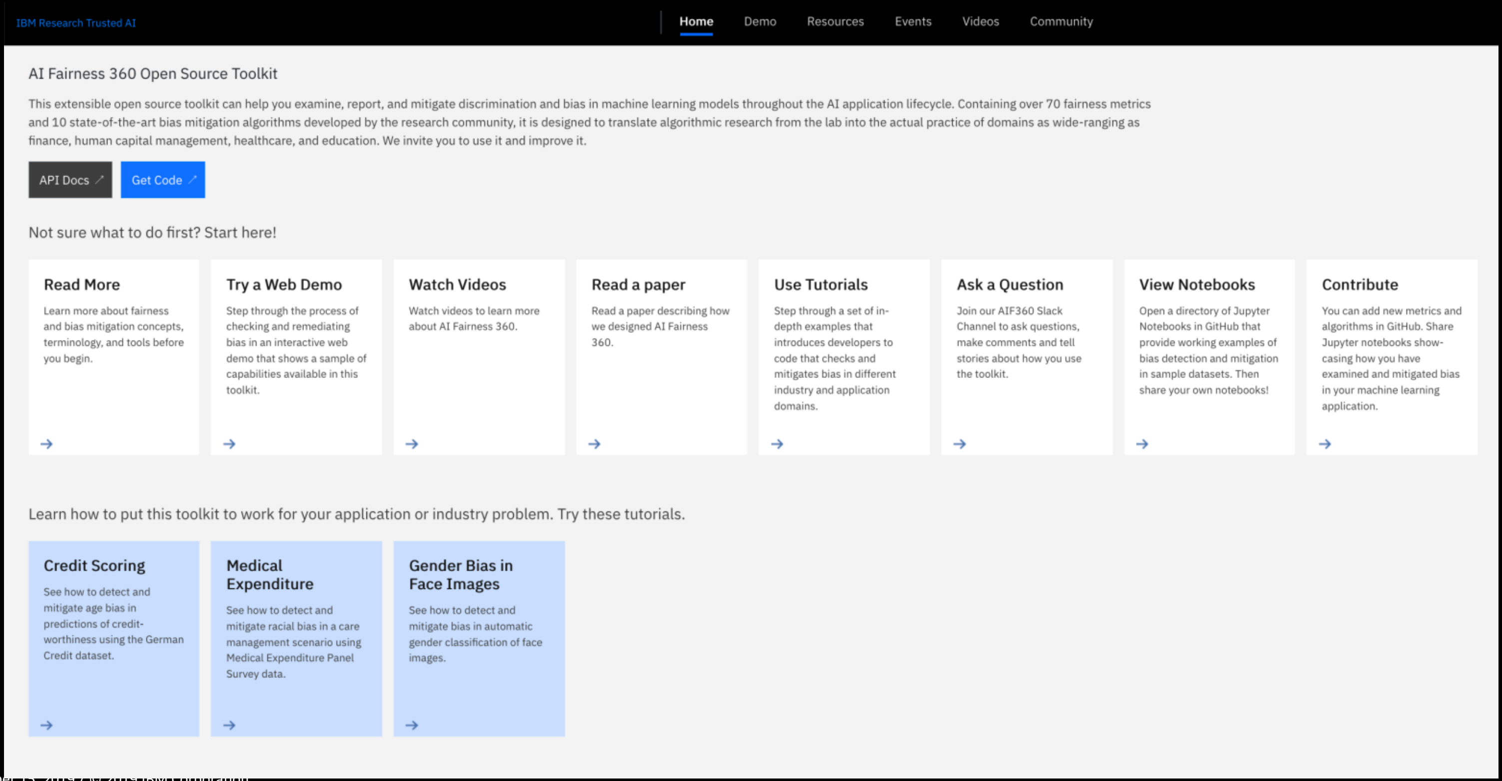
Check for bias as often as possible using any metrics that are applicable

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

Next Section:
Toolkit Overview
& Interactive Demo

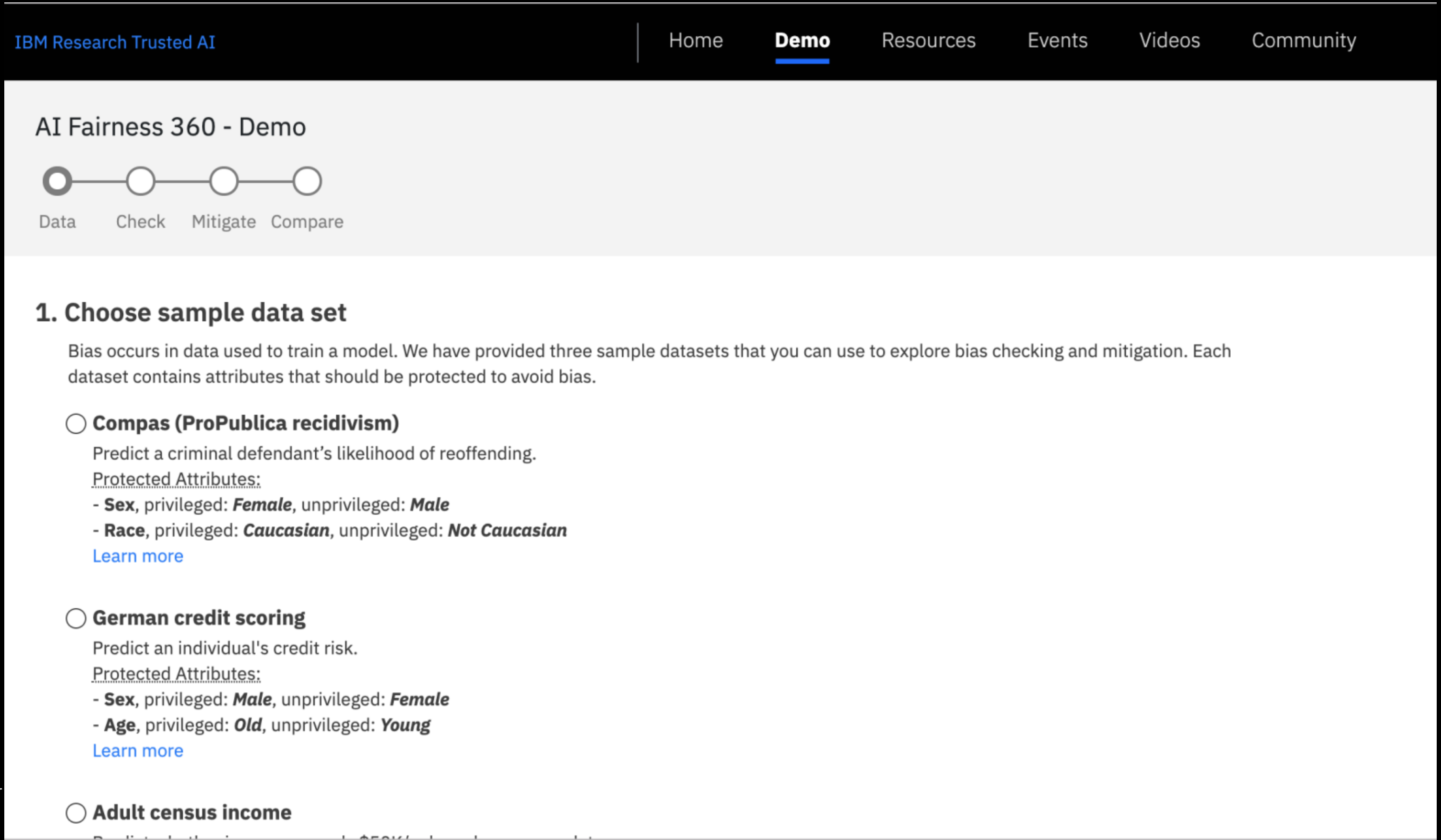
AI Fairness 360 Toolkit Overview

https://aif360.mybluemix.net/



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.

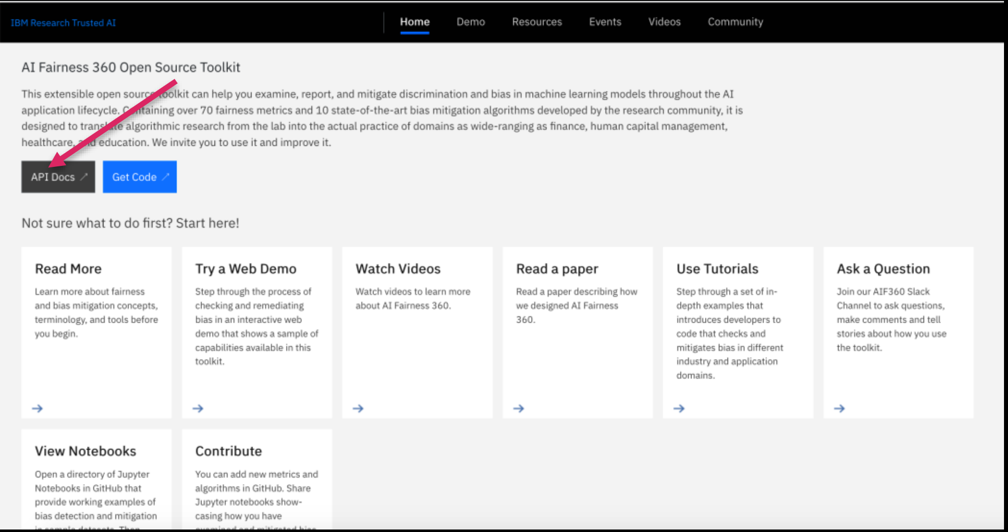
Protected Attributes:

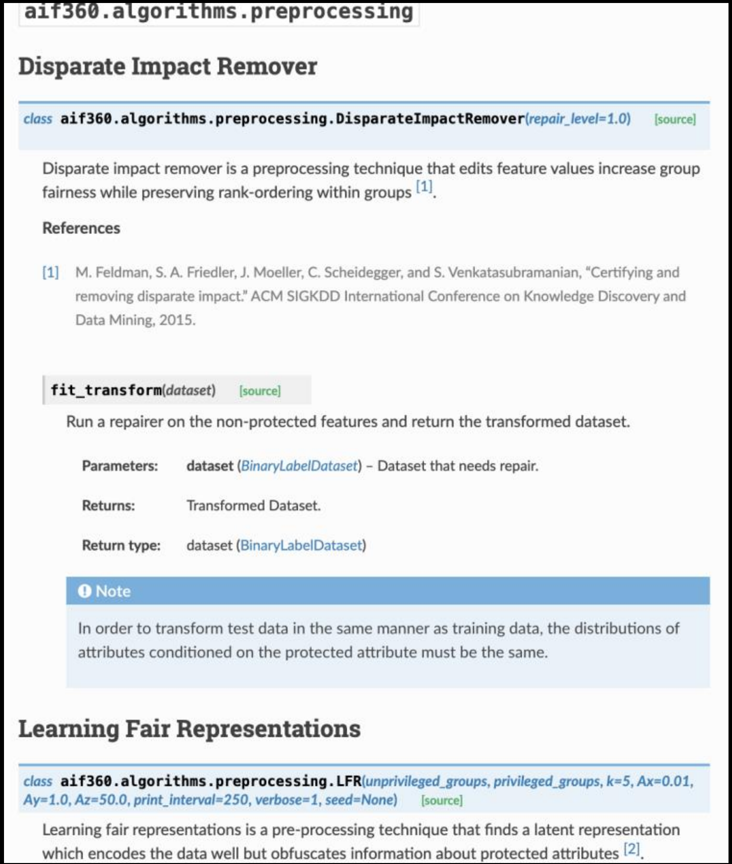
- **Sex**, privileged: ***Male***, unprivileged: ***Female***
- **Age**, privileged: ***Old***, unprivileged: ***Young***

Learn more

Adult census income

Toolkit API – Definitions, Formulas & References





Next Section:
Medical Use Case
Python Tutorial

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