

Toolbox to Mitigate Bias in AI

Presented By:

Gabriela de Queiroz
Saishruthi Swaminathan
Stacey Ronaghan



Agenda



Responsible AI



AI Fairness 360



Demo



Q&A

Responsible AI

- *AI Opportunities*
 - Increased Revenue
 - Efficiencies
- *AI Risks*
 - Harm to Users
 - Harm to Business
- *A Solution*
 - Regulation
 - Ethical & Moral Practices



Unwanted
Bias

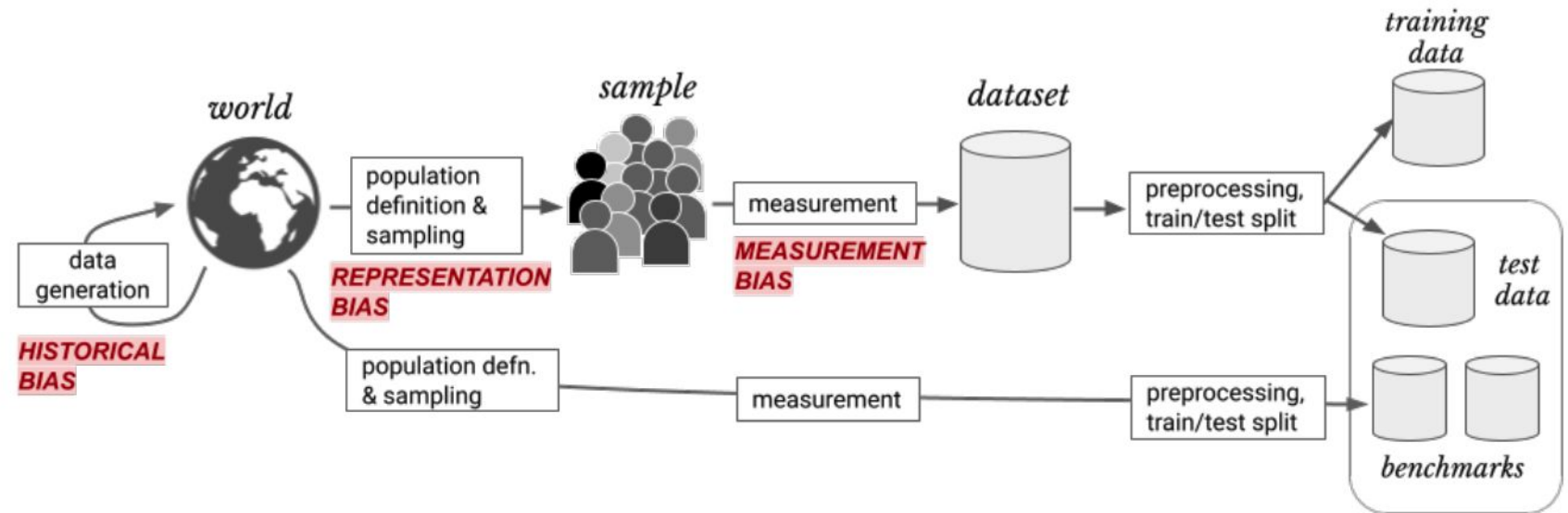
Places

privileged groups at
systematic **advantage**

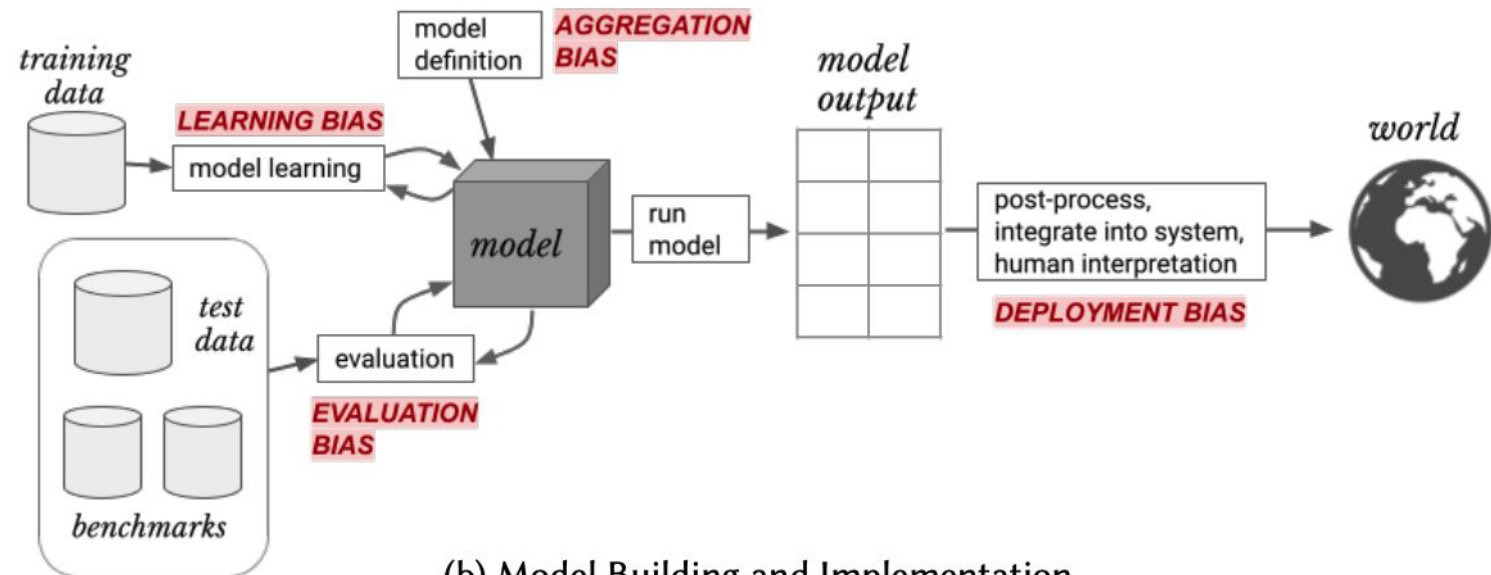
and

unprivileged groups at
systematic **disadvantage.**

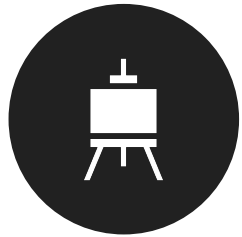
Types of Bias



(a) Data Generation



(b) Model Building and Implementation



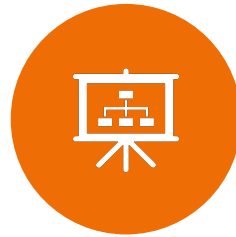
DESIGN

- Human Centric
- Optimization Metrics



DATA

- Representative
- Protected



MODEL

- Interpretable
- Fair



MONITORING

- Staged rollout
- Feedback loop



ACCOUNTABILITY

- Transparency
- Responsibilities

Responsible AI Pipeline

Responsible AI Benefits

- Prevent harm
- Build an inclusive product
- Delightful customer experiences
- Responsible branding

A Step Towards Building Trustworthy AI system

AI Fairness 360 Toolkit (AIF360)



AI Fairness 360 (AIF360)

An extensible, open source toolkit for measuring, understanding, and reducing AI bias. It combines the top bias metrics, bias mitigation algorithms, and metric explainers from fairness researchers across industry and academia

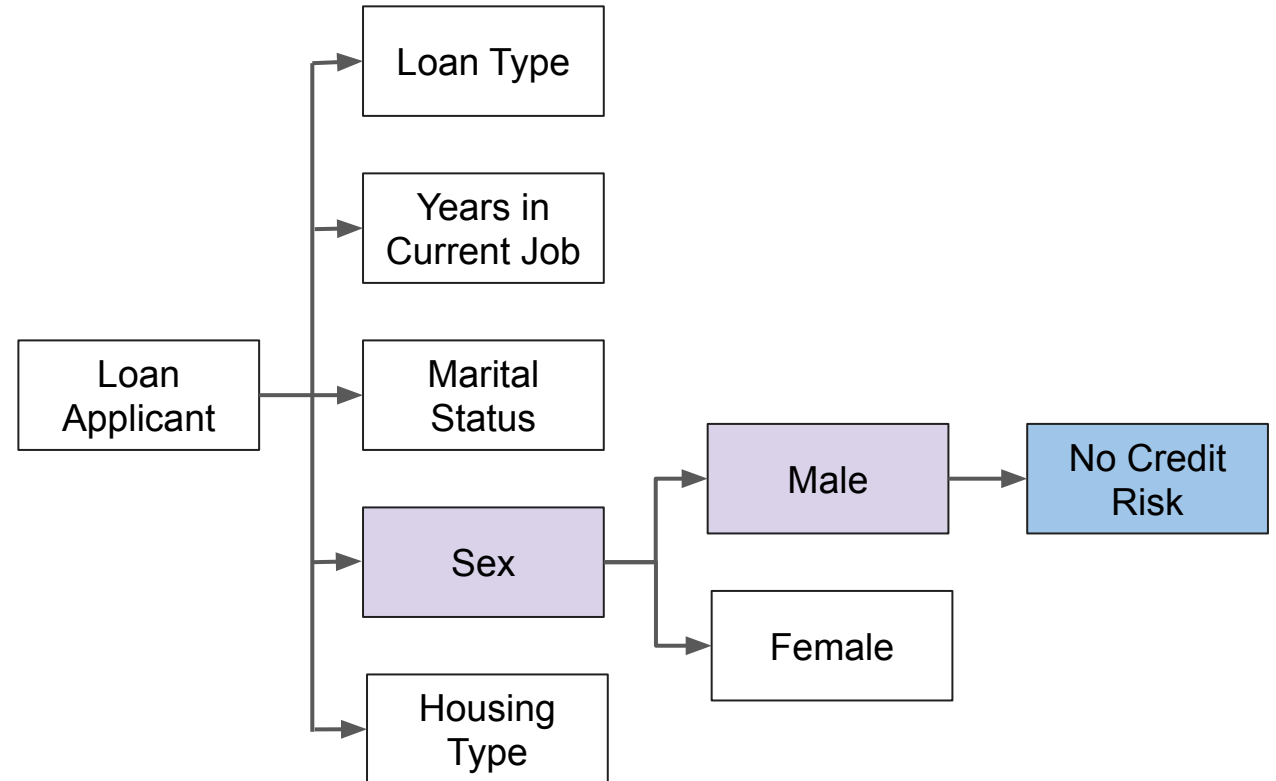
- Implement techniques from eight published papers across the greater AI fairness research community
- Available in Python and R

Terminology

- **Favorable label:** A label whose value corresponds to an outcome that provides an advantage to the recipient (such as receiving a loan, being hired for a job, not being arrested)
- **Protected attribute:** An attribute that partitions a population into groups whose outcomes should have parity (such as race, sex, caste, and religion)
- **Privileged value** (of a protected attribute): A protected attribute value indicating a group that has historically been at a systemic advantage
- **Discrimination/unwanted bias:** When specific privileged groups are placed at a systematic advantage and specific unprivileged groups are placed at a systematic disadvantage. This relates to attributes such as race, sex, age, and sexual orientation.

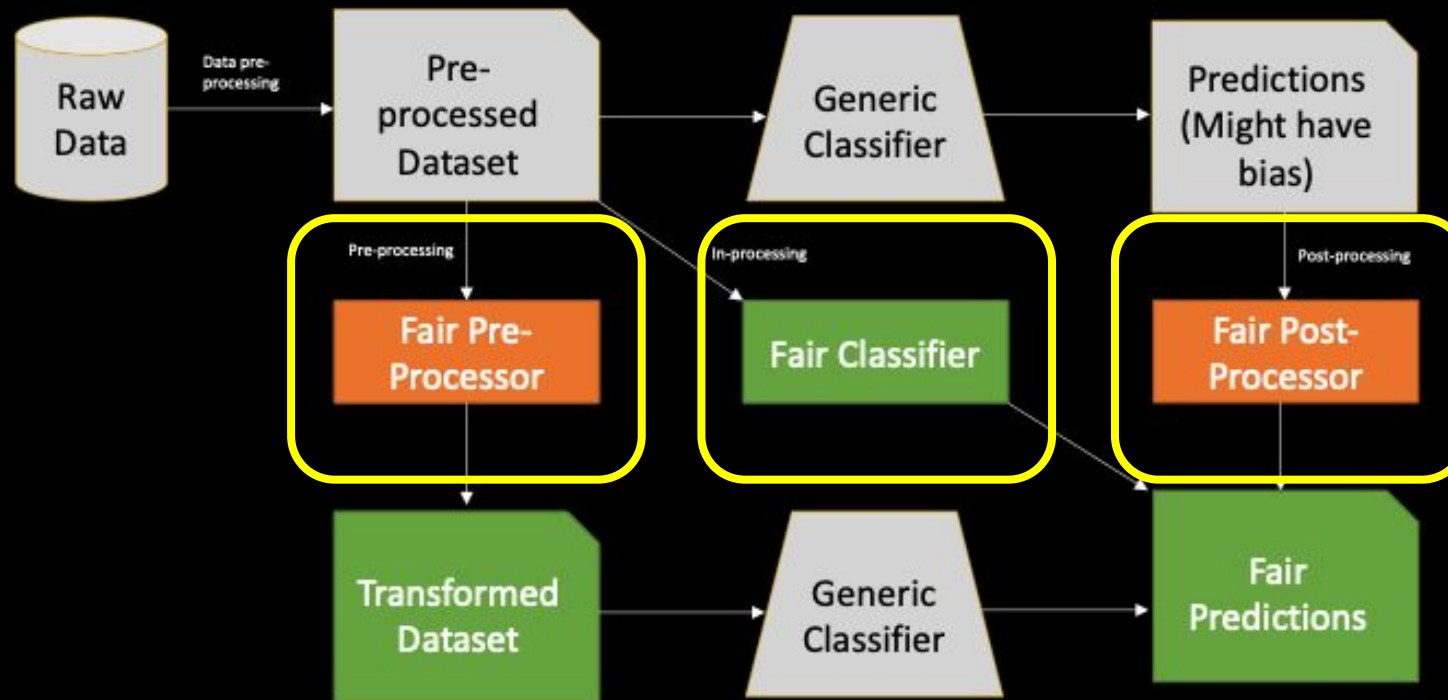
Terminologies

- Favorable label: No Credit Risk
- Unfavorable label: Credit Risk
- Protected Attribute: Sex
- Privileged Protected Attribute: Male



Mitigating bias throughout the AI application lifecycle

Usage



Metrics

A quantification of unwanted bias in training data or models.

Group fairness

Partitions a population into groups defined by protected attributes & seeks for some statistical measure to be equal across groups.

Individual fairness

Seeks for similar individuals to be treated similarly.

Metrics

Group Fairness

Data Vs Model

Measure fairness on the training data

Vs

Measure fairness on the learned model

Metrics

Group Fairness

We are all Equal

All groups have similar abilities with respect to the task
(even if we cannot observe this properly)

Vs

What You See is What You Get

Observations reflect ability with respect to the task

Group Fairness Metrics

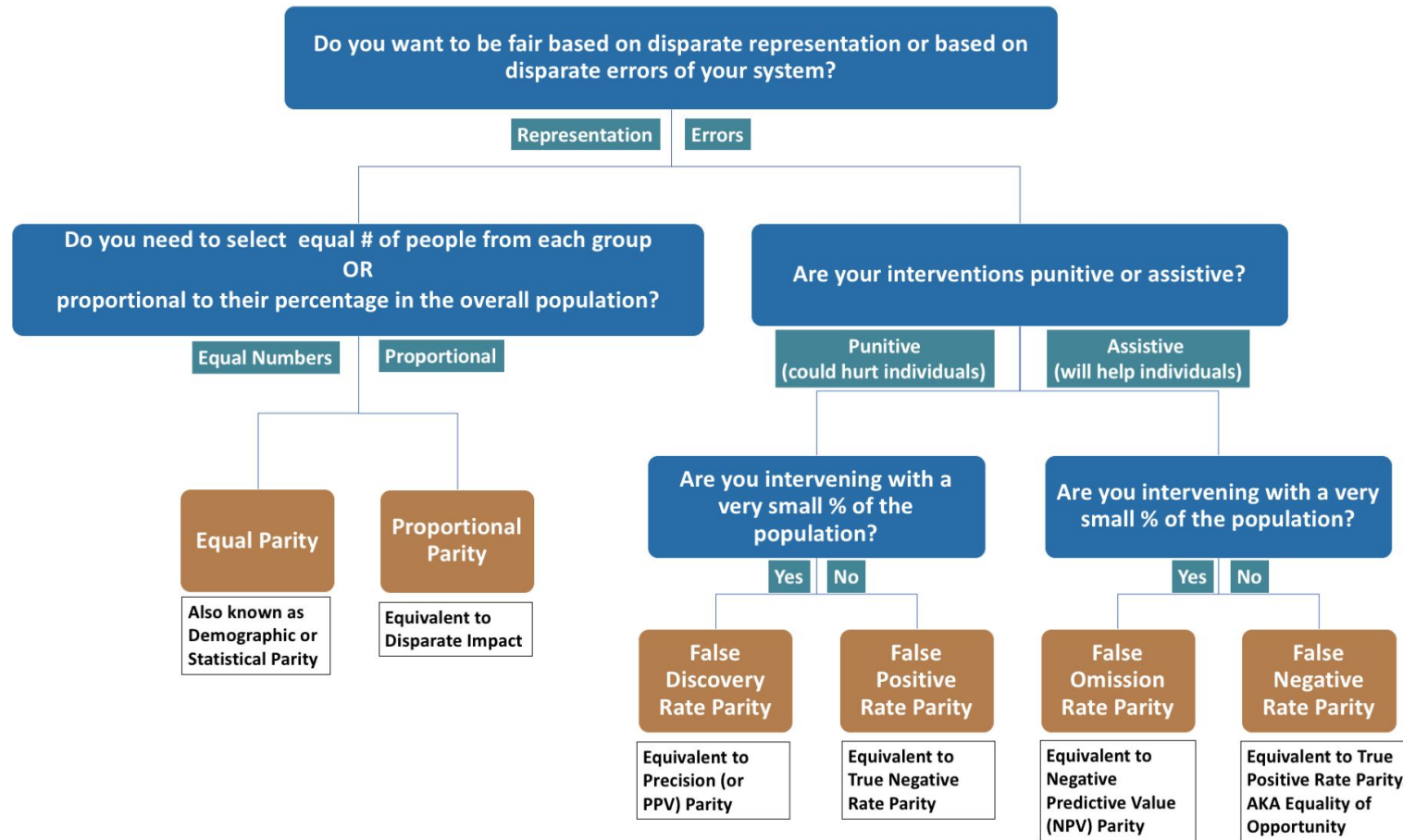
- Difference
- Number of instance
- Ratio
- Base Rate
- Consistency
- Difference
- Disparate Impact
- Mean Difference
- Number of negatives
- Number of Positives
- Ratio
- Smoothed empirical differential fairness
- Statistical Parity Difference
- Rich Subgroup
- `false_negative_rate`
- `false_negative_rate`

Individual Fairness Metrics

- average_euclidean_distance
- average_mahalanobis_distance
- average_manhattan_distance
- difference
- euclidean_distance
- mahalanobis_distance
- manhattan_distance
- mean_euclidean_distance_difference
- mean_euclidean_distance_ratio
- mean_mahalanobis_distance_difference
- mean_mahalanobis_distance_ratio
- mean_manhattan_distance_difference
- mean_manhattan_distance_ratio

Full Guidance Available Here: <https://aif360.mybluemix.net/resources#guidance>

Fairness Metrics Tree



Full Guidance Available Here: <https://aif360.mybluemix.net/resources#guidance>

Algorithms

- Bias mitigation algorithms attempt to improve the fairness metrics by modifying the training data, the learning algorithm, or the predictions.
- These algorithm categories are known as pre-processing, in-processing, and post-processing, respectively.

Algorithms

Pre-Processing Algorithms Mitigate bias in training data	In-Processing Algorithms Mitigate bias in classifiers	Post-Processing Algorithms Mitigate bias in predictions
Reweighting Modifies the weights of different training examples	Adversarial Debiasing Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions	Reject Option Classification Changes predictions from a classifier to make them more fair
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 and labels	Meta Fair Classifier Takes the fairness metric as part of the input and returns a classifier optimized for the metric	Equalized Odds Modifies the predicted label using an optimization scheme to make predictions more fair
Learning Fair Representations Learns fair representations by obfuscating information about protected attributes		

Using AIF360 in R

R Package Installation

You can install the **aif360** R package in your machine

Or you can use **Docker** for example and install the package

Example Use-Case

Business use-case

Select customers who are likely to buy our new product.

Target Audience

Those whose income is over \$50,000.

Dataset

<https://archive.ics.uci.edu/ml/datasets/adult>

Example Attributes

- Age
- Work Classification (e.g. Private, Self-Employed, Never worked, etc.)
- Education (e.g. Bachelors, Some college, High School graduate, etc.)
- Years of Education
- Marital Status
- Occupation
- Race
- Sex
- Hours-per-week
- Native country

Live Demo

```
## 3) Calculate the mean difference
metric_train <- binary_label_dataset_metric(data_aif_train,
                                             privileged_groups = privileged_groups,
                                             unprivileged_groups = unprivileged_groups)

metric_train$mean_difference()
# [1] -0.1932321
# The difference between the proportion of positive outcomes for the unprivileged vs
# the privileged group
#  $P(Y=1|D=unprivileged) - P(Y=1|D=privileged)$ 

## 4) Apply Adversarial debiasing is an in-processing technique that learns a classifier
## to maximize prediction accuracy and simultaneously reduce an adversary's ability to determine
## the protected attribute from the predictions
sess <- tf$compat.v1$Session()

debiased_model <- adversarial_debiasing(privileged_groups = privileged_groups,
                                       unprivileged_groups = unprivileged_groups,
                                       scope_name = 'debiased_classifier',
                                       debias = TRUE,
                                       sess = sess)

debiased_model$fit(data_aif_train)
# predictions
data_aif_train_debiasing <- debiased_model$predict(data_aif_train)

# Right now we are just caring about fairness
metric_preds <- binary_label_dataset_metric(data_aif_train_debiasing,
                                             privileged_groups = privileged_groups,
                                             unprivileged_groups = unprivileged_groups)

metric_preds$mean_difference()
# [1] -0.08583602 after
# [1] -0.1932321 before
```

```
adult_dataset.R
1 ## Load the library
2 library(aif360)
3 load_aif360_lib()
4
5 ## Load the data
6 original_data <- readr::read_csv(
7   "https://www.dropbox.com/s/ga8tr1glj17nrgk/adult_data_preprocessed.csv?dl=1"
8 )
9 original_data <- original_data[, -1]
10 head(original_data)
11 str(original_data)
12
13 # Predict whether income exceeds $50K/yr based on census data.
14 # Variables:
15 ## sex: 1 male, 0 female
16 ## income binary: 1 > $50k, 0 <= $50k
17
18 privileged_groups <- list("sex", 1)
19 unprivileged_groups <- list("sex", 0)
20
21 ## 1) Convert the dataframe into the aif360 format -----
22 data_aif <- aif_dataset(data_path = original_data,
23                        favor_label = 1,
24                        unfavor_label = 0,
25                        privileged_protected_attribute = 1,
26                        unprivileged_protected_attribute = 0,
27                        target_column = "Income Binary",
28                        protected_attribute = "sex")
29
30 ## 2) Let's split in train and test -----
31 # train should be 70%
32 # test should be 30%
33 set.seed(1234)
34 data_aif_split <- data_aif$split(num_on_size_splits = list(0.70))
35 data_aif_train <- data_aif_split[[1]]
36 data_aif_test <- data_aif_split[[2]]
37
```


Call To Action

Label issues and pull requests for new contributors [Dismiss](#)

Now, GitHub will help potential first-time contributors [discover issues](#) labeled with **good first issue**

Filters ▾



is:open is:issue author:SSaishruthi label:R

Labels 12

Milestones 1

New issue

✕ Clear current search query, filters, and sorts

<input type="checkbox"/>	5 Open ✓ 0 Closed	Author ▾	Label ▾	Projects ▾	Milestones ▾	Assignee ▾	Sort ▾
<input type="checkbox"/>	<div><div>Update R package dependencies</div><div>enhancement R</div><div>#261 opened on Aug 12 by SSaishruthi III New Issues</div></div>						
<input type="checkbox"/>	<div><div>[AIF360-R] Post-processing algorithms</div><div>contribution welcome R</div><div>#211 opened on Oct 26, 2020 by SSaishruthi III New Issues 2 tasks</div></div>						
<input type="checkbox"/>	<div><div>[AIF360-R] In-processing algorithms</div><div>contribution welcome R</div><div>#210 opened on Oct 26, 2020 by SSaishruthi III New Issues 3 tasks</div></div>						
<input type="checkbox"/>	<div><div>[AIF360-R] Pre-processing Algorithms</div><div>contribution welcome R</div><div>#209 opened on Oct 26, 2020 by SSaishruthi III New Issues 2 tasks</div></div>						
<input type="checkbox"/>	<div><div>R - Next Batch Updates</div><div>R</div><div>#198 opened on Aug 10, 2020 by SSaishruthi III New Issues 2 of 6 tasks</div></div>						

AIF360 - Interactive Demo

aif360.mybluemix.net

AI Fairness 360

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it.

[Python API Docs ↗](#)

[Get Python Code ↗](#)

[Get R Code ↗](#)

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

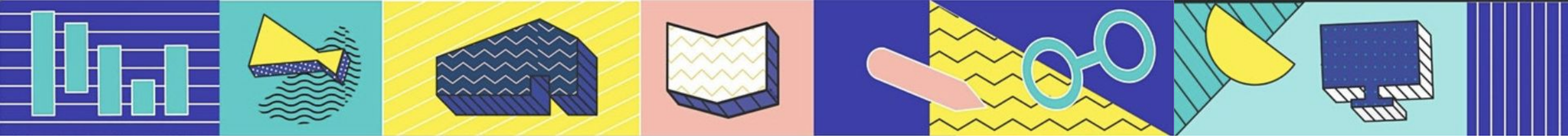
Open a directory of Jupyter Notebooks in GitHub that provide working examples of bias detection and mitigation in sample datasets. Then share your own notebooks!



Contribute

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





Thank You!



[linkedin.com/in/
staceyronaghan](https://www.linkedin.com/in/staceyronaghan)



[linkedin.com/in/
gabrieladequeiroz](https://www.linkedin.com/in/gabrieladequeiroz)



[linkedin.com/in/
saishruthi-swaminathan](https://www.linkedin.com/in/saishruthi-swaminathan)