AI Fairness 360

IBM AI Research

Presented by:
(Ethan) Yuqiang Heng, Dave Van Veen

AI Fairness is Important

- AI used to make decisions in increasingly more and higher-stake aspects of our life: credit, employment, admission, sentencing
  - Objectionable when places privileged groups at systematic advantage and unprivileged groups at systematic disadvantage
- What is fairness and how to make models fair?
  - AI fairness research has produced dozens of metrics and algorithms
  - **Confusing and overwhelming for practitioners!**

Image from
https://dumielauxepices.net/sites/default/files/injustice-clipart-religion-discrimination-648162-3448199.jpg

Existing Tools

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
Themis-ML	Python library built on scikit-learn that implements fairness-aware machine learning algorithms	https://github.com/cosmicBboy/themis-ml
FairML	Looks at significance of model inputs to quantify prediction dependence on inputs	https://github.com/adebayoj/fairml
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
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

Screenshot from IBM presentation https://www.youtube.com/watch?v=X1NsrcaRQTE

There isn’t an all-in-one solution!

AI Fairness 360: Is it fair and how do I make it fair?

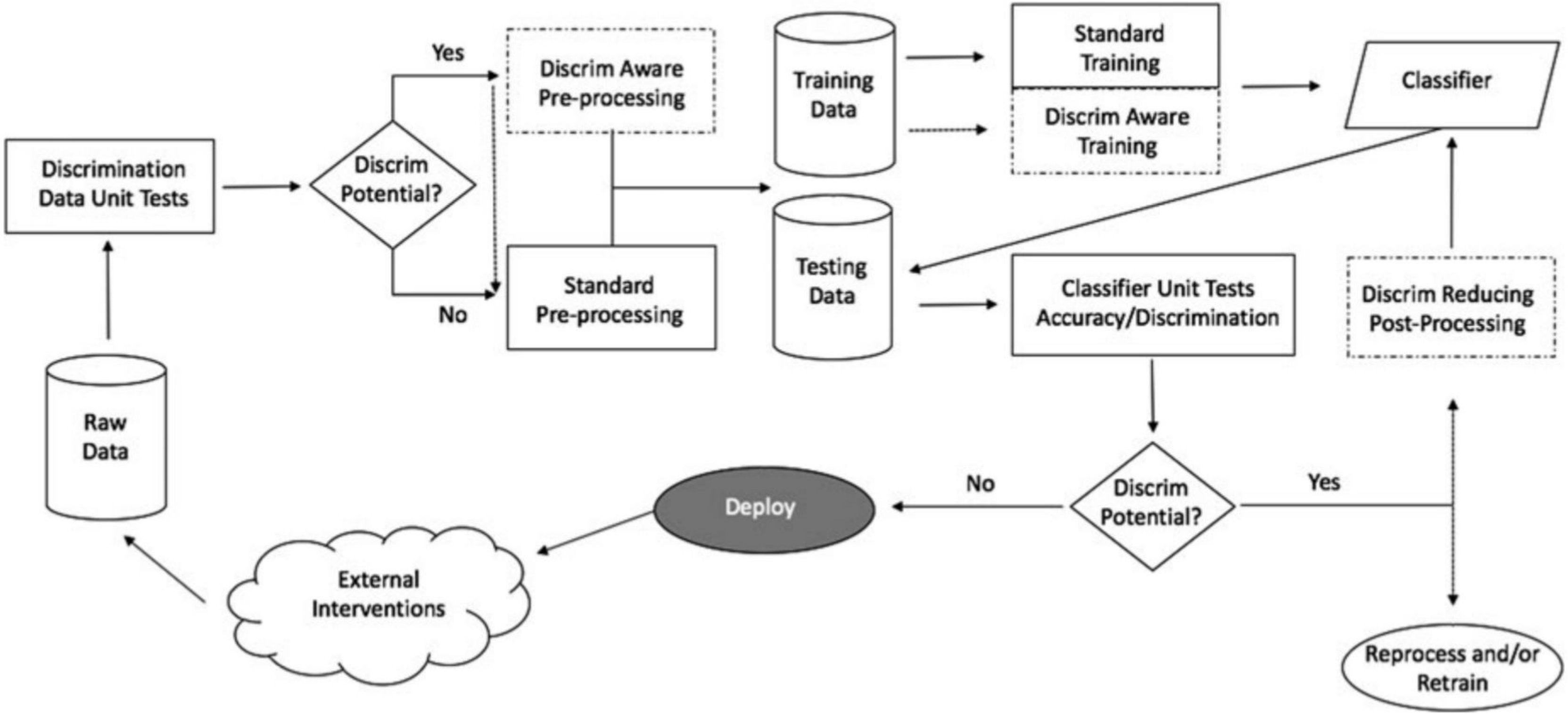
What does it offer?

- Datasets
- Fairness Toolbox
  - 30+ fairness metrics
  - Fairness metric explanations
  - 9+ bias mitigation algorithms
- Guidance
  - Which metric and algorithm to consider based on your scenario
- Industry-specific tutorials

What differentiates it from competition?

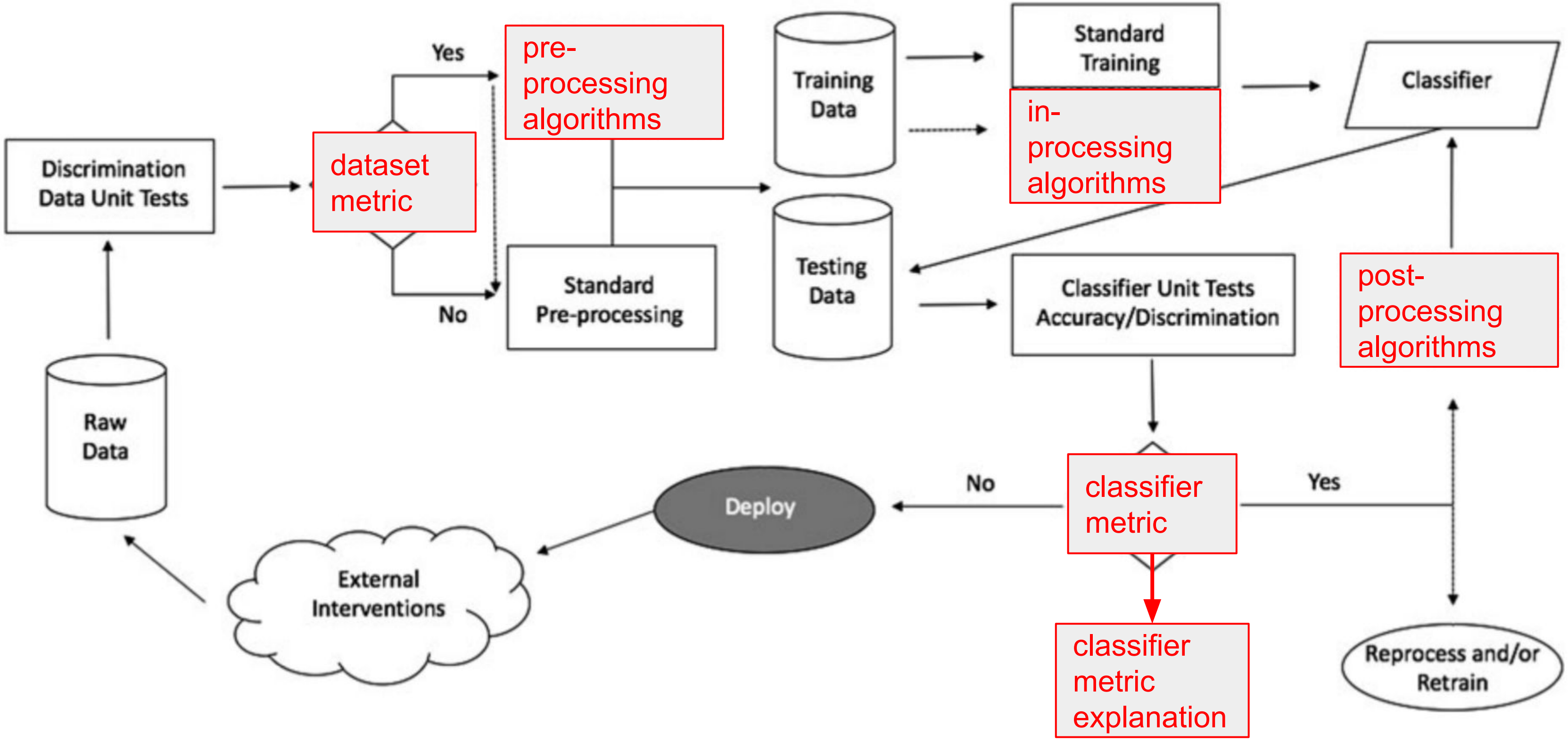
- Comprehensive set of both metrics and bias mitigation algorithms (some unique from IBM research and exclusive to this toolbox)
- Designed to be extensible and easily adopted (scikit-learn style)
- Translate results from research labs to industry practitioners

Workflow for Building Fair Models



d'Alessandro et al., 2017

Workflow for AI Fairness 360



Dataset Class

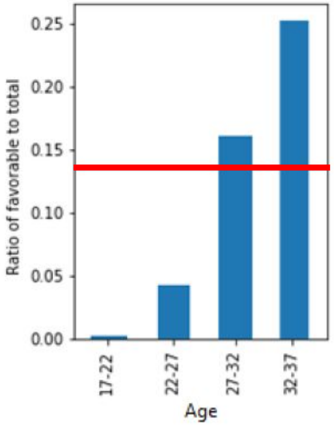
- Standardize loading raw dataset from CSV format
  - Provides interface for data ‘cleaning’: converting categorical features, specifying features, labels, protected attributes, privileged status, favorable status, etc.
- Includes common datasets
  - Adult Census Income (Kohavi, 1996), German Credit (Dheeru & Karra Taniskidou, 2017), ProPublica Recidivism (COMPAS) (Angwin et al., 2016), Bank Marketing (Moro et al., 2014), and three versions of Medical Expenditure Panel Surveys (AHRQ, 2015; 2016)
- Includes common functions
  - split, compare, converting to Pandas DF, tracking previous versions

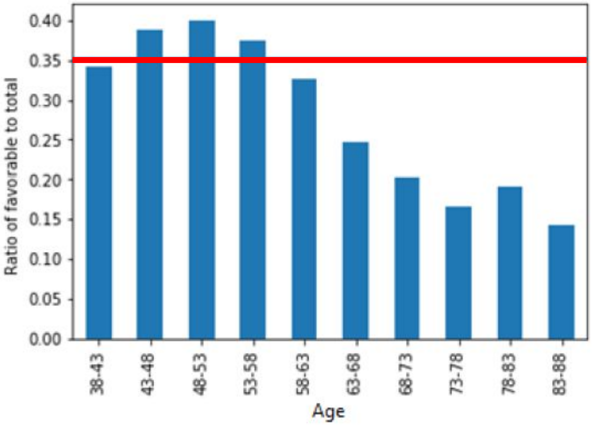
Metric Class

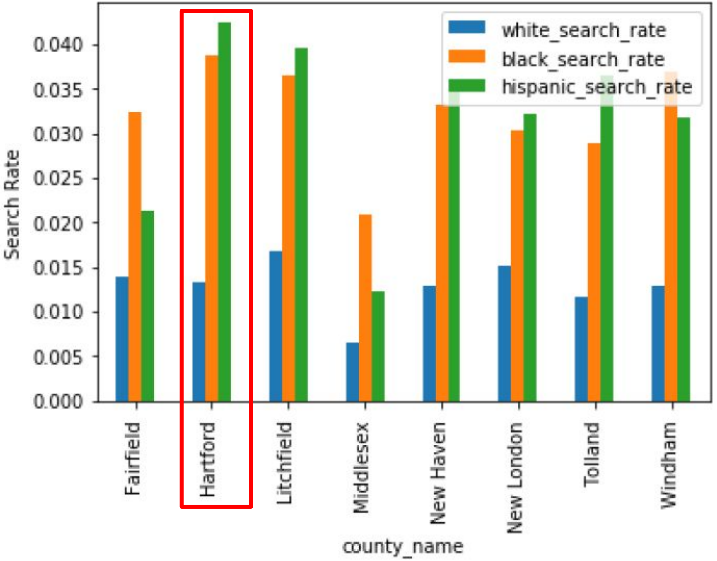
- Metric either applies to single dataset to get group or individual fairness measures or compares original and transformed datasets
- Individual vs. Group Fairness, or both
- Fairness in Data vs. Model
- We’re All Equal vs. What You See is What You Get
- Ratios vs. Differences

Explainer Class

- Explainer associated with metric class, provides:
  - Text description and explanation of metric
  - Fine-grained localization
    - finds critical values in protected attributes for
      - privileged vs unprivileged groups
    - compare fairness measure across attributes





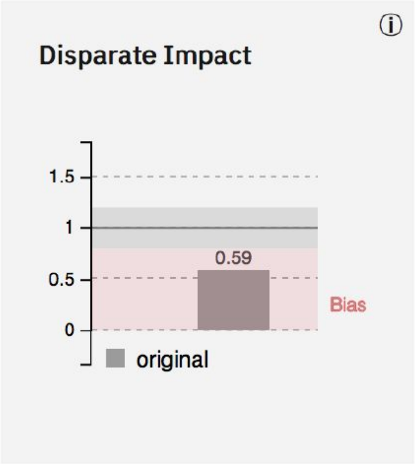


Algorithms Class

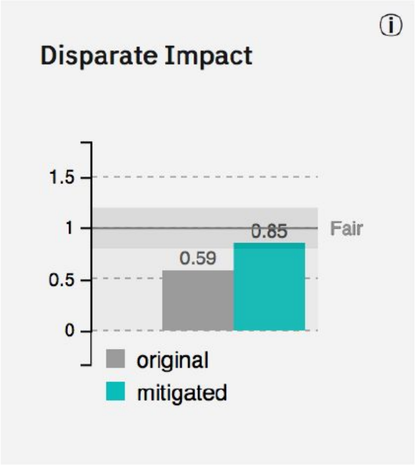
- Pre-Processing: allowed to modify training data
  - Reweighing (Kamiran & Calders, 2012), Optimized preprocessing (Calmon et al., 2017), Learning fair representations (Zemel et al., 2013), Disparate impact remover (Feldman et al., 2015)
- In-Processing: allowed to change the learning procedure
  - Adversarial debiasing (Zhang et al., 2018), Prejudice remover (Kamishima et al., 2012)
- Post-Processing: treat learned model as black box without modifying training data or algorithm
  - Equalized odds postprocessing (Hardt et al., 2016), Calibrated equalized odds postprocessing (Pleiss et al., 2017), Reject option classification (Kamiran et al., 2012)

Adoption and Maintenance

- Adoption
  - Web interactive demo with intuitive visualizations for consumers without programming background
  - Notebook tutorials, guidance and community forum for new developers
- Maintain Quality of Code
  - Unit and integration tests to ensure code quality and API compliance while allowing contributions and extensions



Before mitigation



After adversarial debiasing mitigation

AIF360 Demo

Two Philosophies

- <u>WAE</u>: We're All Equal
    - All groups have similar abilities w.r.t. the task
- <u>WYSIWYG</u>: What You See Is What You Get
    - Observations reflect abilities w.r.t. the task

- Example - SAT scores
    - <u>WYSIWYG</u>: Score correlates well with success → score can be used to compare abilities across applicants
    - <u>WAE</u>: SAT may contain structural biases → different distribution across groups should not be mistaken for difference in ability

Metrics Examples

disparity_impact

% classified as favorable, ratio of unprivileged:privileged [fair = 1]

statistical_parity_difference

% classified as favorable, difference of unprivileged minus privileged [fair = 0]

equal_opportunity_difference

TP / (TP + FN), difference of unprivileged minus privileged [fair = 0]

FAIRNESS TREE