Building Fair AI Models

Moninder Singh
IBM Research AI

PyData New York, 2018

Resources

- Tutorial (This document + Notebook):
  - https://github.com/monindersingh/pydata2018_fairAI_models_tutorial
- Toolkit Used – AI Fairness 360 (AIF 360):
  - https://github.com/IBM/AIF360
- AIF 360 Installation Instructions:
  - https://github.com/IBM/AIF360/blob/master/README.md
- AIF 360 Dataset Download Instructions:
  - https://github.com/IBM/AIF360/tree/master/aif360/data

Alternative Setup Instructions

- https://github.com/monindersingh/pydata2018_fairAI_models_tutorial/blob/master/README.md

Create and activate environment

```
conda create --name aif360 python=3.5
conda activate aif360
```

Clone AIF360 from GitHub:

```
git clone https://github.com/IBM/AIF360
```

Install R-essentials for downloading MEPS data

```
conda install -c r r-essentials
```

Alternative Setup Instructions (contd.)

Download datasets and place under appropriate folders under AIF360/aif360/data/raw by cloning this repository (NOTE: clone at same level as AIF360) and running the belowmentioned notebooks in the root folder

```
git clone https://github.com/monindersingh/pydata2018_fairAI_models_tutorial.git
```

Change to the root folder of just cloned repository and run

```
jupyter notebook pydata_datasets.ipynb
jupyter notebook pydata_meps_datasets.ipynb
```

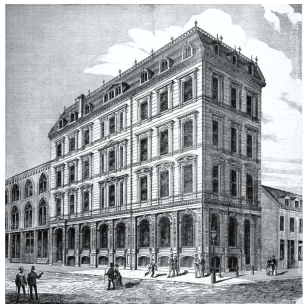
Then, navigate to the root directory of the cloned AIF360 project and run:

```
pip install .
```

Finally, install the additional requirements as follows:

```
conda install ecos
pip install -r requirements.txt
```

AI is now used in many high-stakes decision making applications



Credit

Employment

Admission

Sentencing

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

(Other than that it is 99% accurate)

Is it fair?

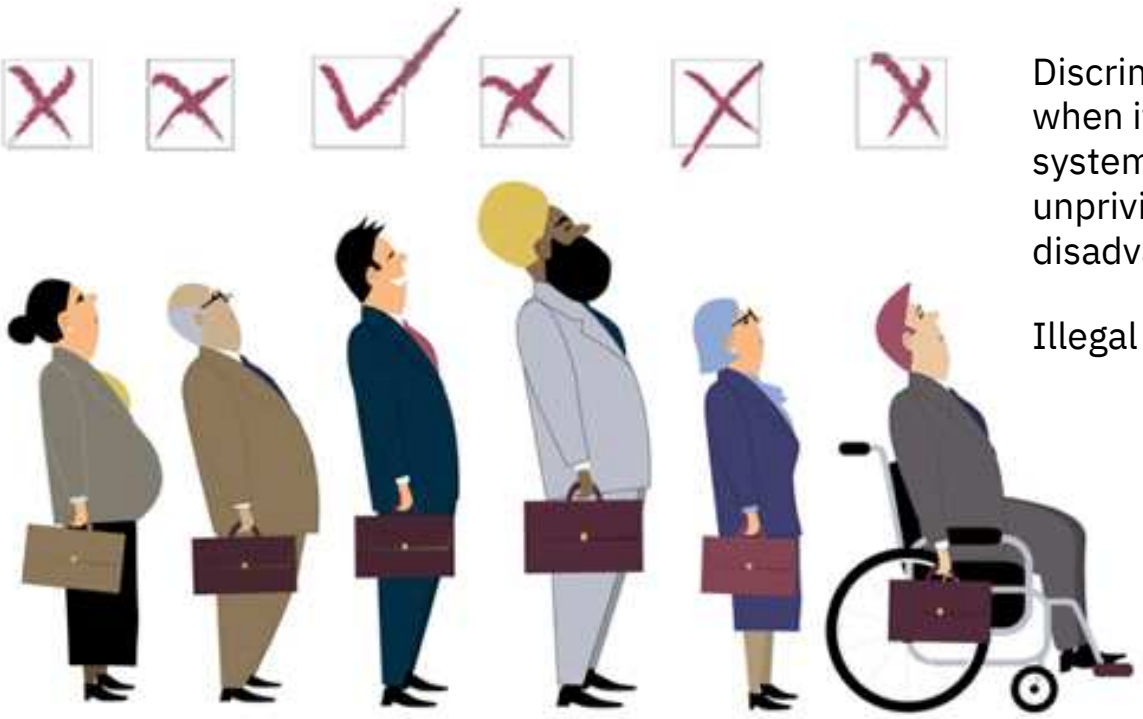
Is it easy to understand?

Did anyone tamper with it?

Is it accountable?

Unwanted bias and algorithmic fairness

Machine learning, by its very nature, is always a form of statistical discrimination



Discrimination becomes objectionable when it places certain privileged groups at systematic advantage and certain unprivileged groups at systematic disadvantage

Illegal in certain contexts

Some Examples

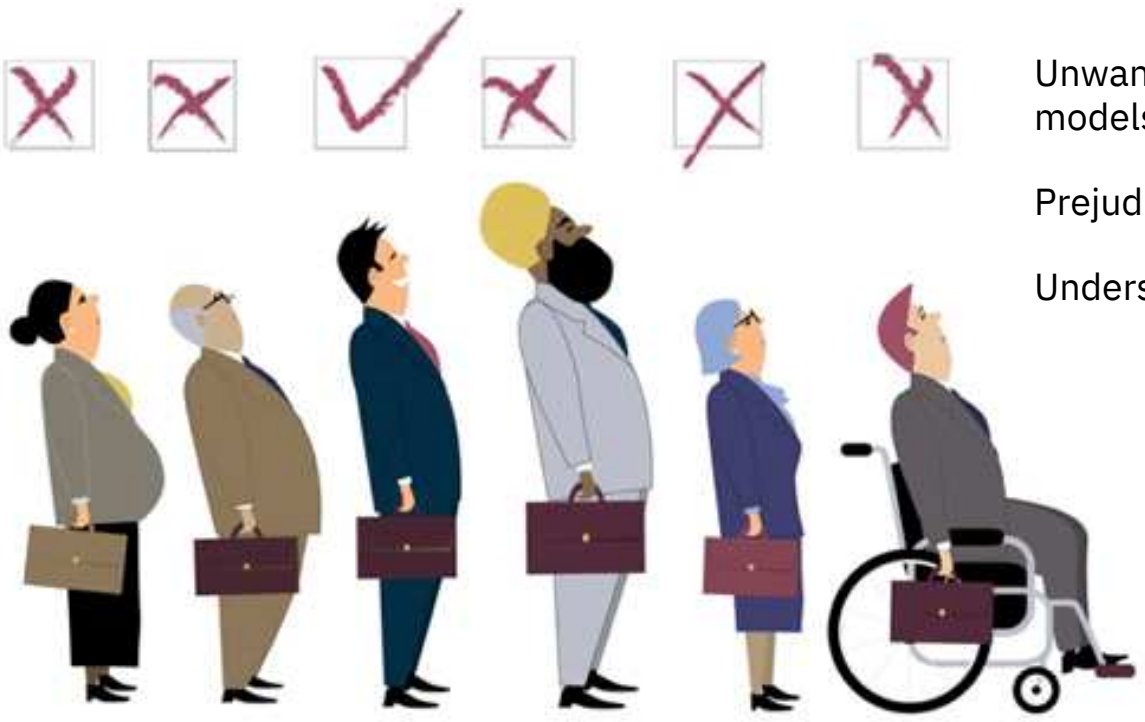
- Staples Online Pricing[1]
  - Adjust prices based on proximity to competitor stores
  - Higher prices for lower income people who generally live farther from such stores
- Compas Recidivism Data
  - Predict risk score for recidivism i.e. re-offend
  - Racial bias: Black defendants were often predicted to be at a higher risk of recidivism than they actually were; white defendants were often predicted to be less risky than they were
- Google Image Tagger
  - Offensive labels with images of black people

1. J. Valentino-Devries, J. Singer-Vine, and A. Soltani, "Websites vary prices, deals based on users' information," https://www.wsj.com/articles/SB10001424127887323777204578189391813881534, Dec 2012.
2. J. Larson, S. Mattu, L.Kirchner and J. Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, May 2016
3. J. Guynn, "Google photos labeled black people 'gorillas'," https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/, July 2015

AI Fairness 360 (AIF360)

- Open source, Python package - a comprehensive set of
  - metrics for datasets and models to test for biases,
  - explanations for these metrics
  - algorithms to mitigate bias in datasets and models
- Interactive demo:
  - http://aif360.mybluemix.net/data
- Additional tutorials/examples
  - http://aif360.mybluemix.net/resources#
  - https://github.com/IBM/AIF360/tree/master/examples
- Some guidance on choosing metrics/algorithms
  - http://aif360.mybluemix.net/resources#guidance

Unwanted bias and algorithmic fairness



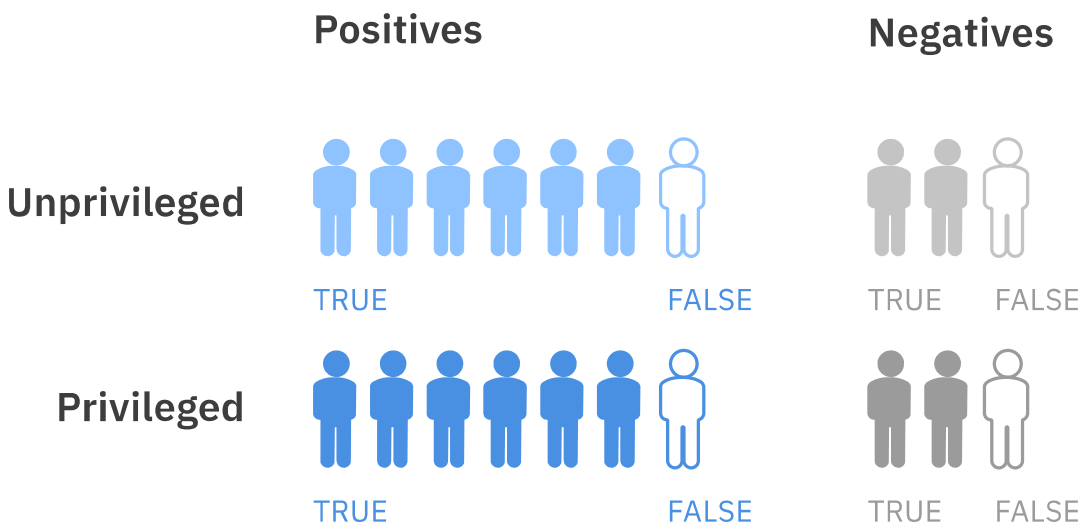
Unwanted bias in training data yields models with unwanted bias that scale out

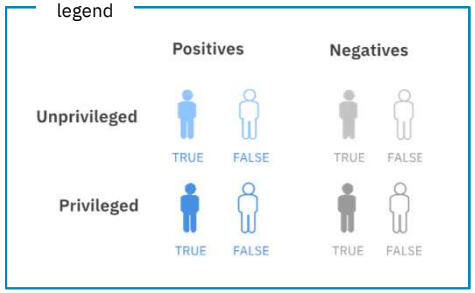
Prejudice in labels

Undersampling or oversampling

Statistical definitions of group fairness

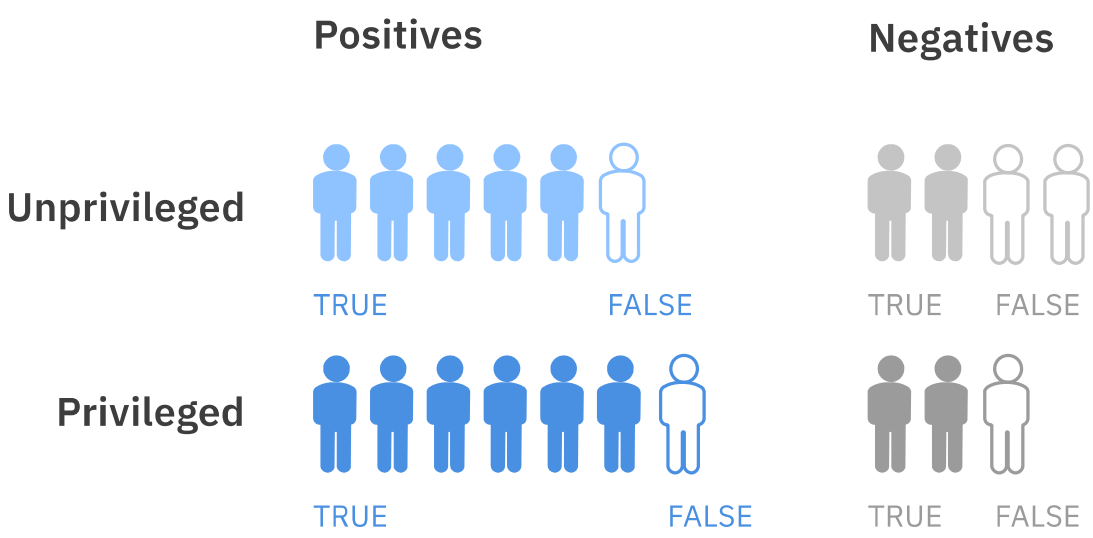
situation 1

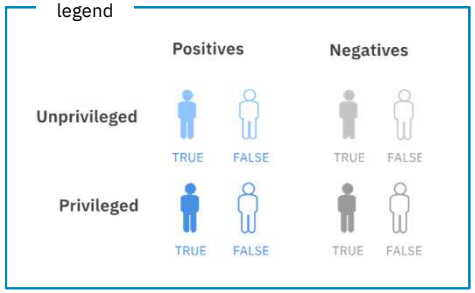




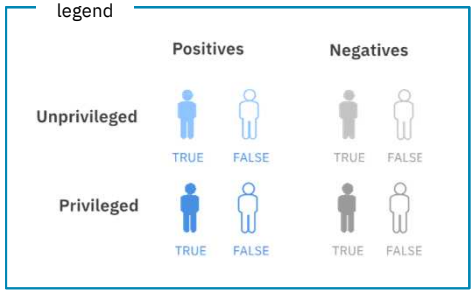
Statistical definitions of group fairness

situation 2

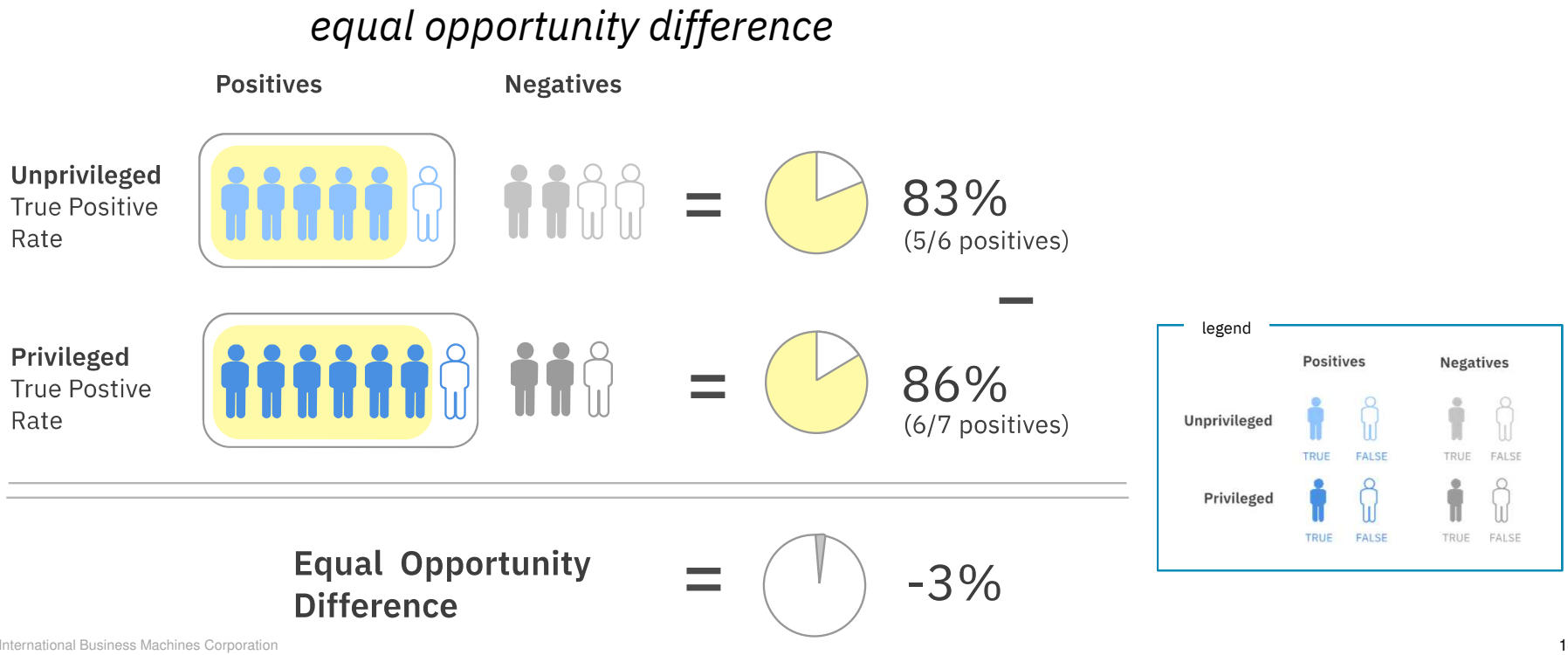




Statistical definitions of group fairness

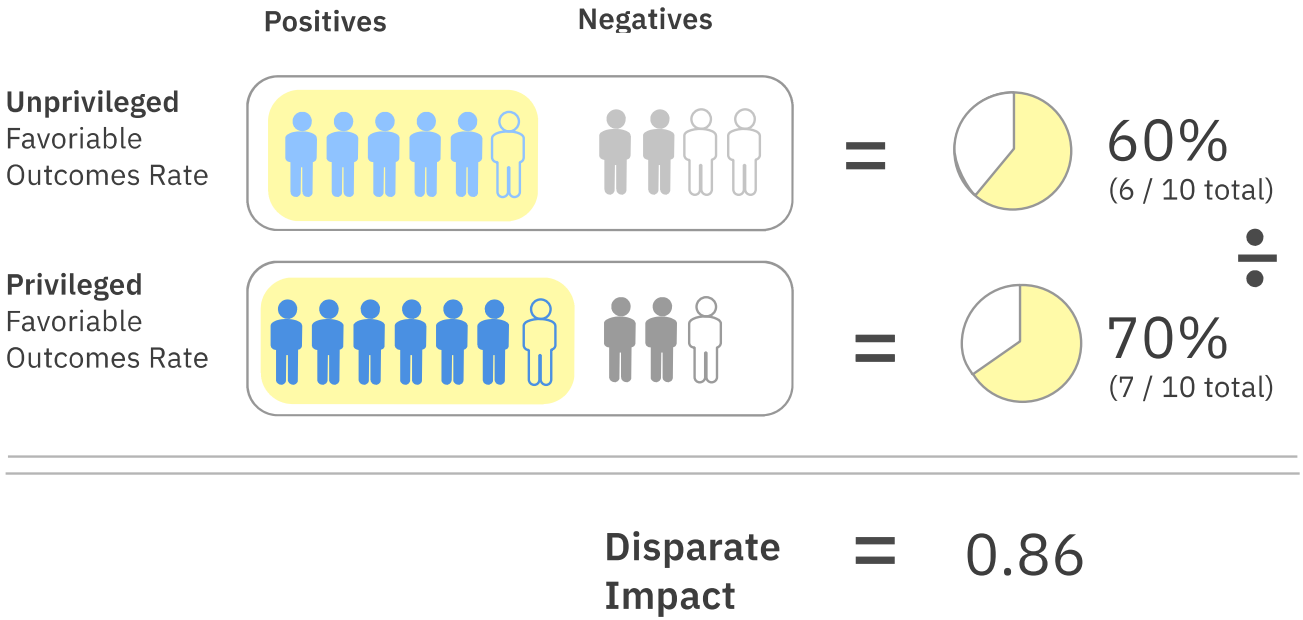


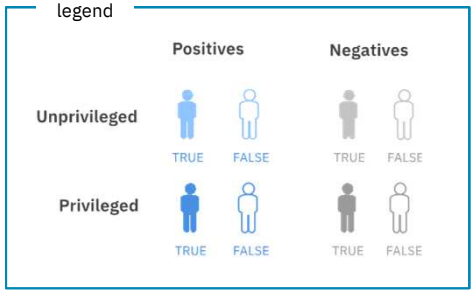
Statistical definitions of group fairness



Statistical definitions of group fairness

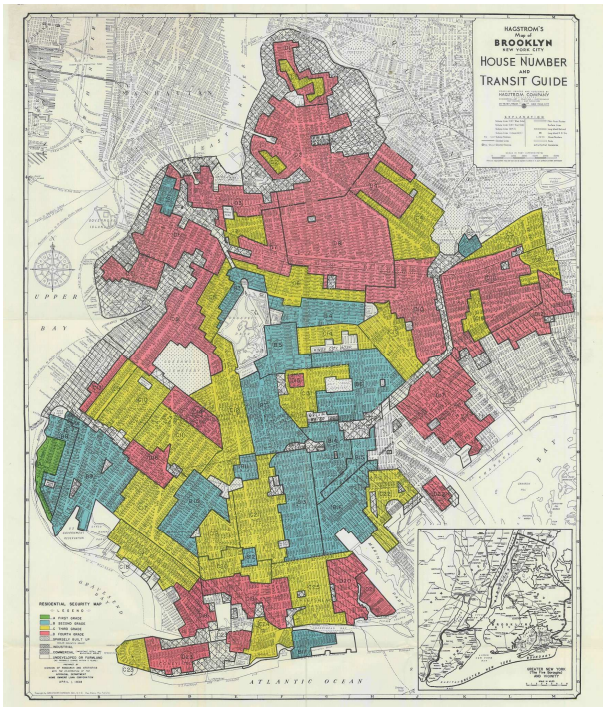
disparate impact



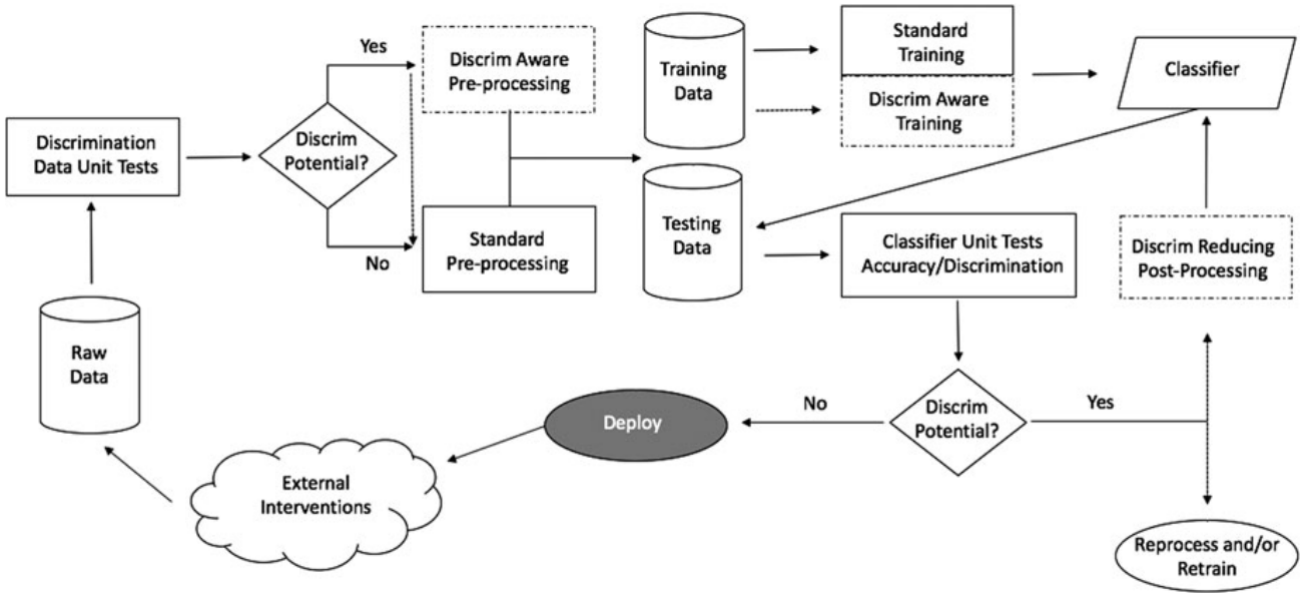


Bias mitigation is not easy

Cannot simply drop protected attributes because features are correlated with them

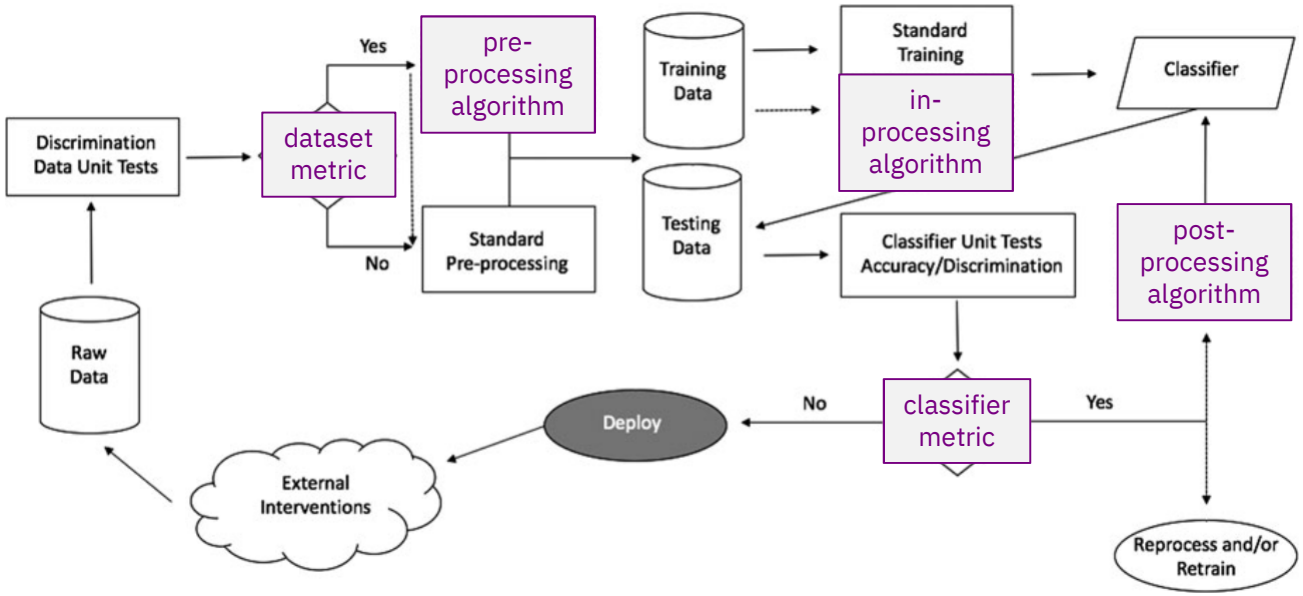


Fairness in building and deploying models



B. d'Alessandro, C. O'Neil and T. LaGatta``Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification'', Big Data, 5(2), 2017

Metrics, Algorithms



Some Pre-Processing Bias Mitigation Algorithms

- Pre-Processing techniques attempt to remove underlying bias by changing the input data itself by changing the labels, weights, and/or feature values

  – ReWeighing: assigns different weights to different training samples to reduce bias

    F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification without Discrimination," Knowledge and Information Systems, 2012.

  – Disparate Impact Remover: changes feature values to increase group fairness by trying to break dependence between sensitive features and other features while still trying to maintain model accuracy

    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.

  – Optimized Pre-Processing: learns a probabilistic transformation that edits the features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives

    F. P. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. "Optimized Pre-Processing for Discrimination Prevention." Conference on Neural Information Processing Systems, 2017

Some In-Processing Bias Mitigation Algorithms

- In-Processing techniques change the algorithm itself to reduce bias during model training

  - Prejudice Remover: adds a discrimination aware regularizer term that penalizes bias during learning

    T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-Aware Classifier with Prejudice Remover Regularizer," Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2012.

  - Adversarial Debiasing: learns a classifier to maximize prediction accuracy and simultaneously reduce an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit.

    B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, 2018.

Some Post-Processing Bias Mitigation Algorithms

- In-Processing techniques change the model output to reduce bias
  - Equalized Odds Post Processing: that solves a linear program to find probabilities with which to change output labels to optimize equalized odds

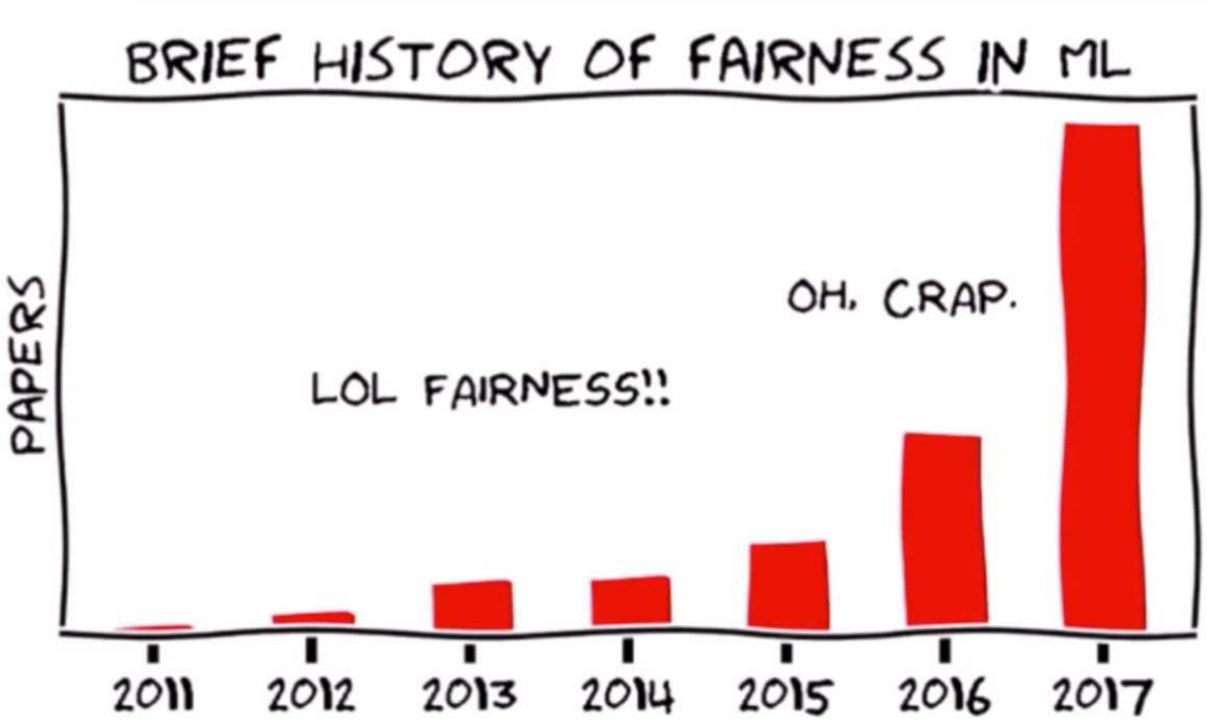
    G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On Fairness and Calibration," Conference on Neural Information Processing Systems, 2017.

  - Reject Option Classification: gives favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.

    F. Kamiran, A. Karim, and X. Zhang, "Decision Theory for Discrimination-Aware Classification," IEEE International Conference on Data Mining, 2012.

Research

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



(Hardt, 2017)

Many other open source fairness libraries...

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