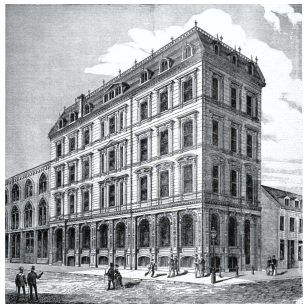
Building Fair AI Models

Moninder Singh
IBM Research AI

PyData New York, 2018

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?

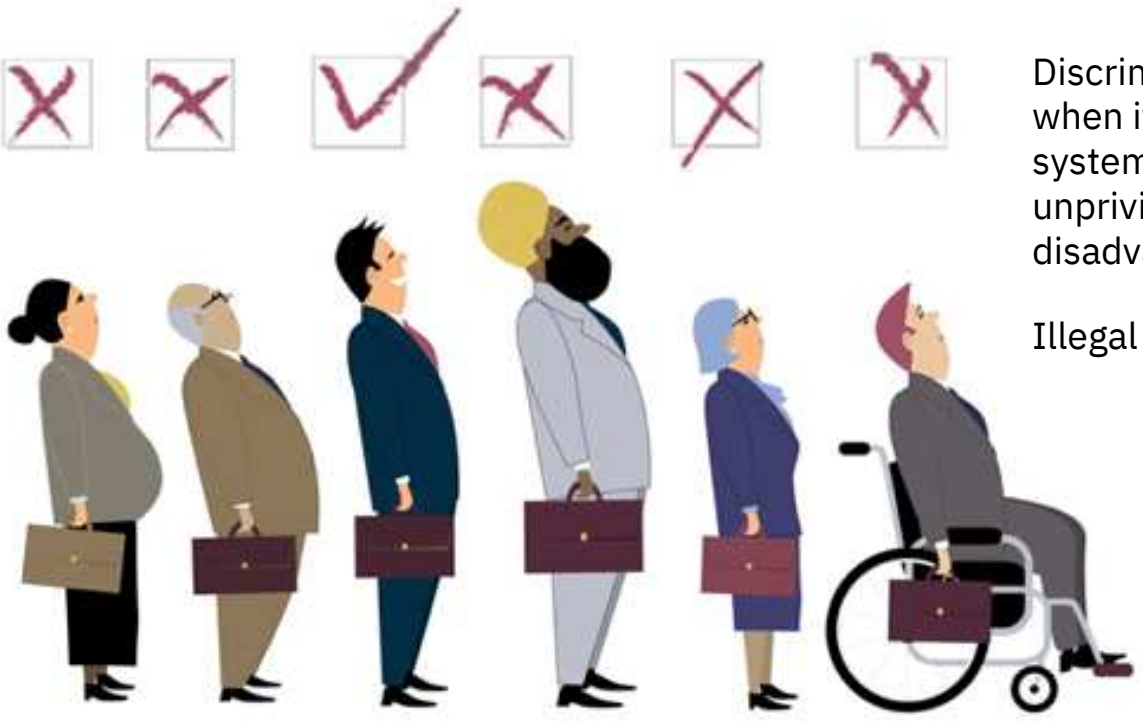
Is it easy to understand?

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

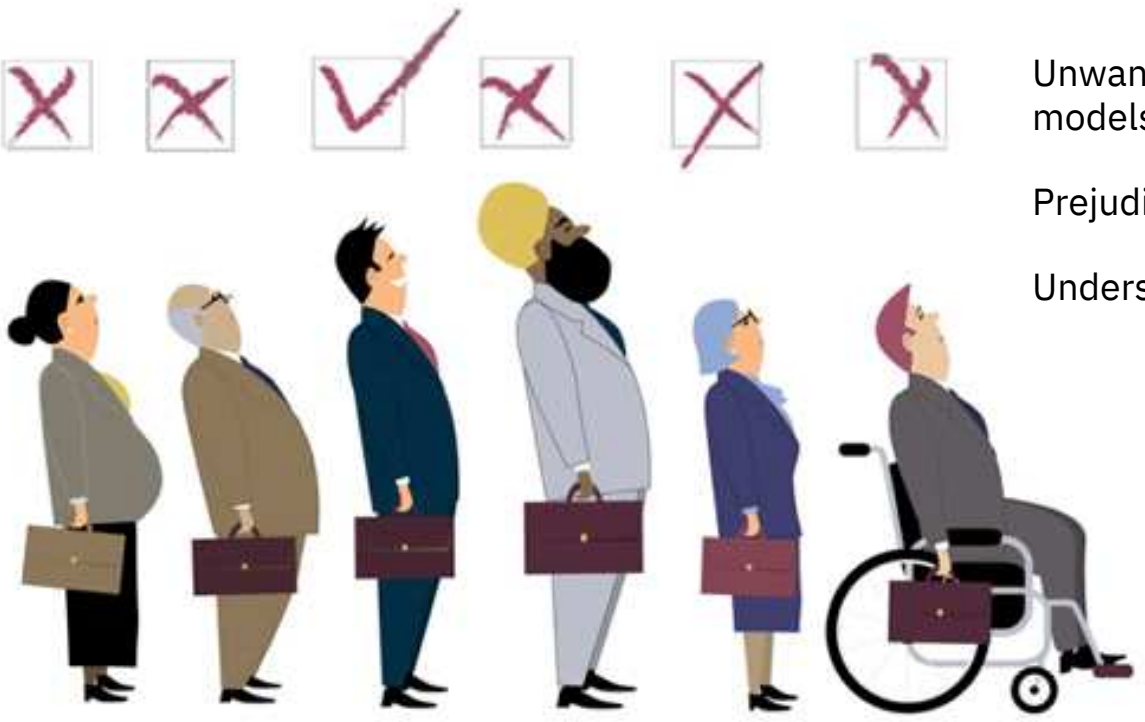
Illegal in certain contexts

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

Unwanted bias and algorithmic fairness



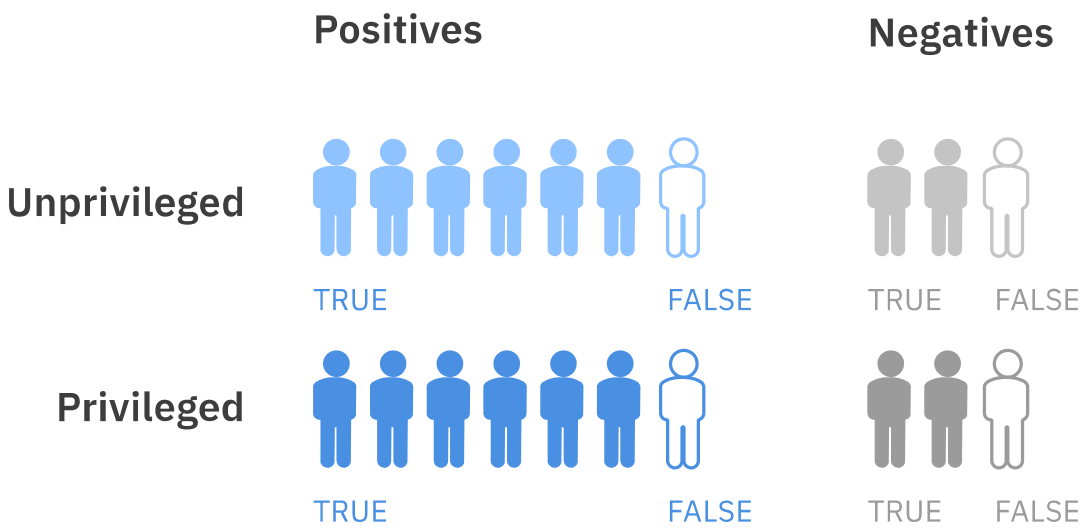
Unwanted bias in training data yields models with unwanted bias that scale out

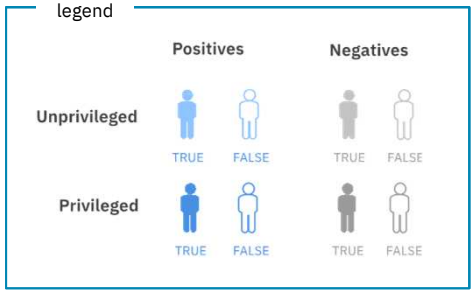
Prejudice in labels

Undersampling or oversampling

Statistical definitions of group fairness

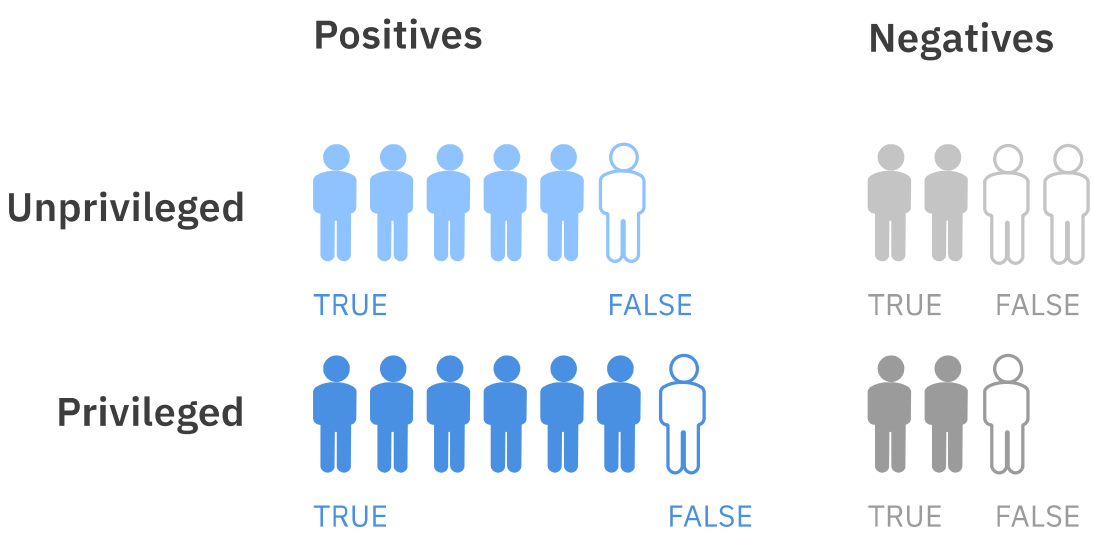
situation 1

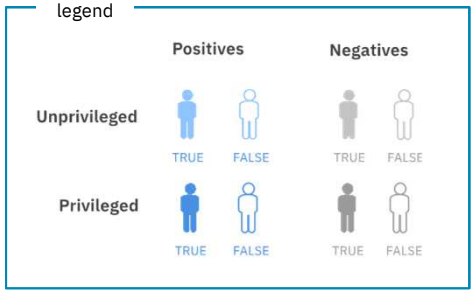




Statistical definitions of group fairness

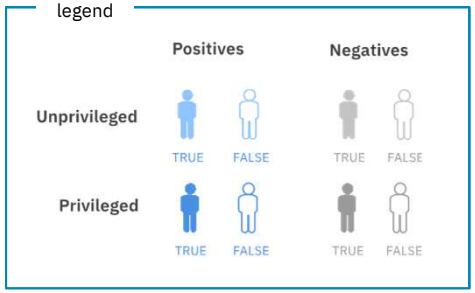
situation 2



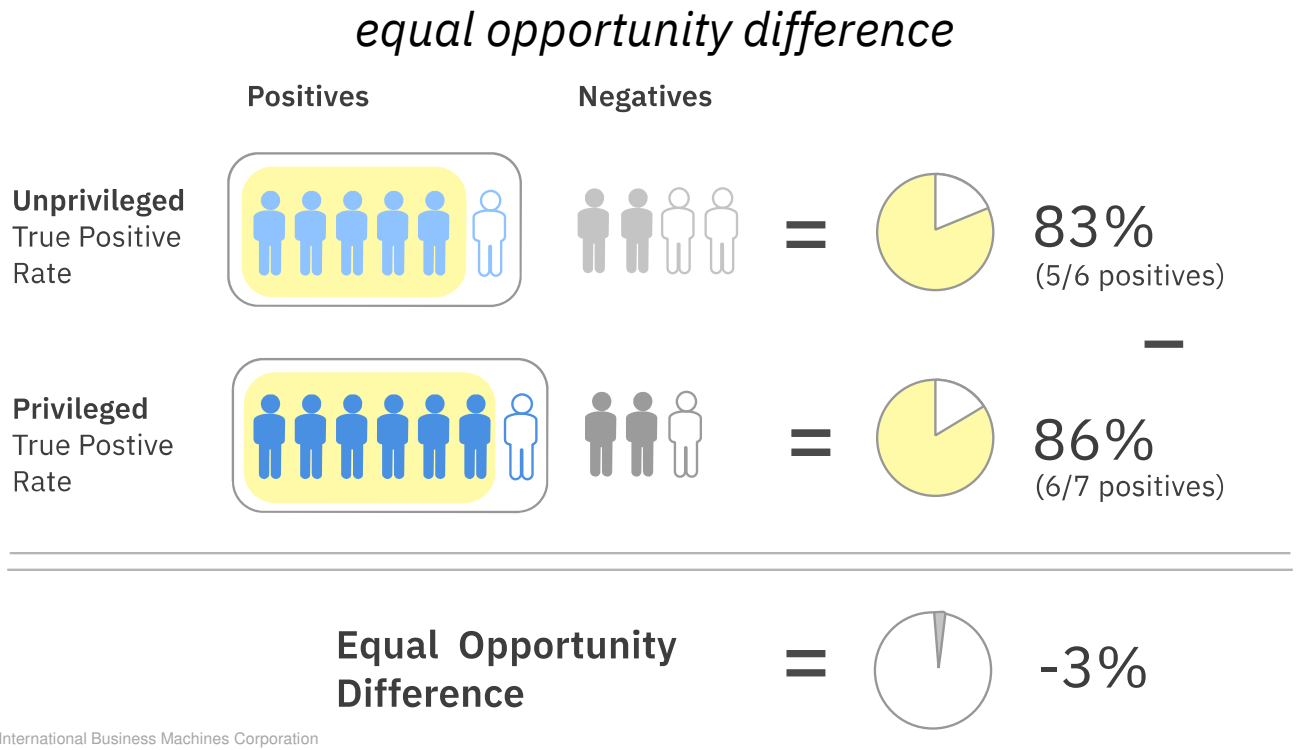


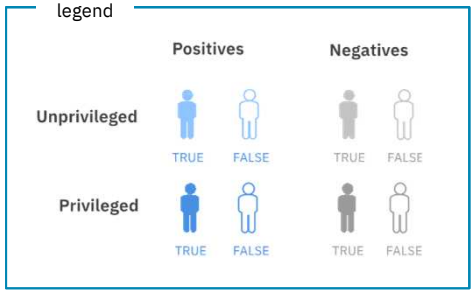
Statistical definitions of group fairness

statistical parity difference



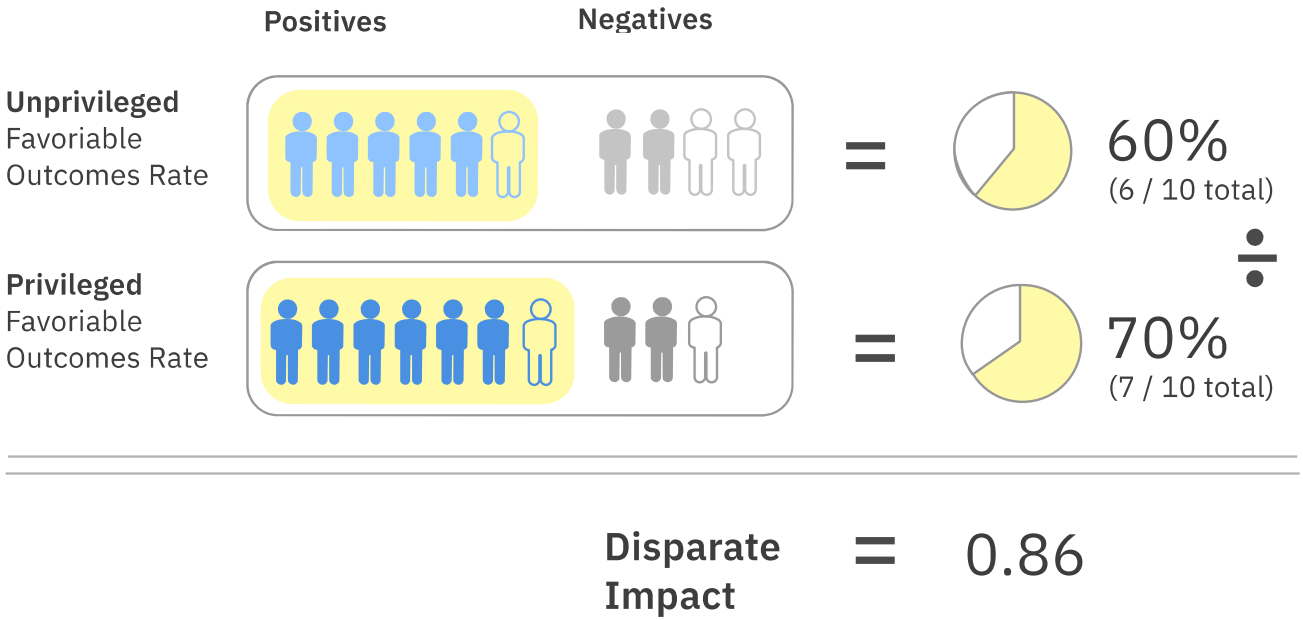
Statistical definitions of group fairness

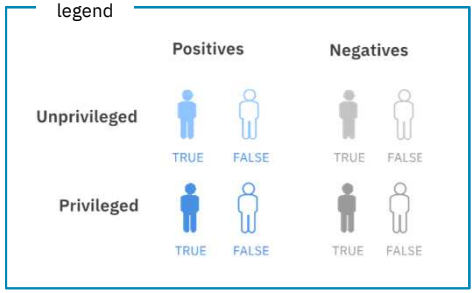




Statistical definitions of group fairness

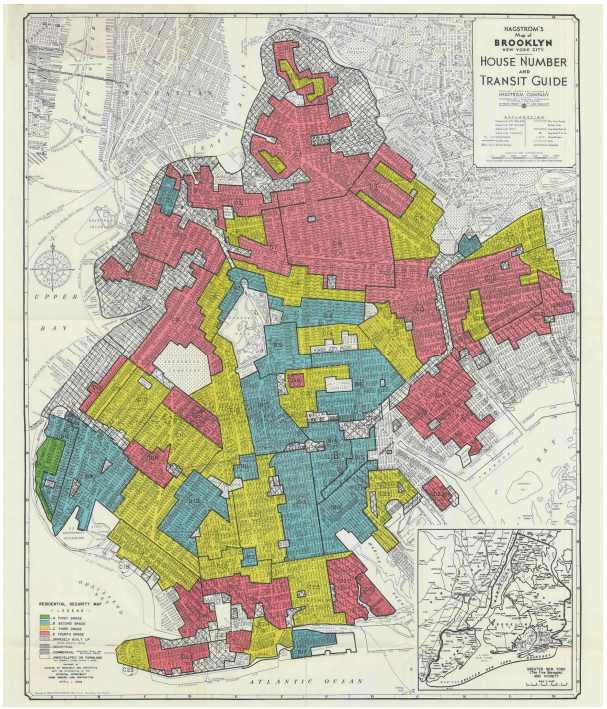
disparate impact



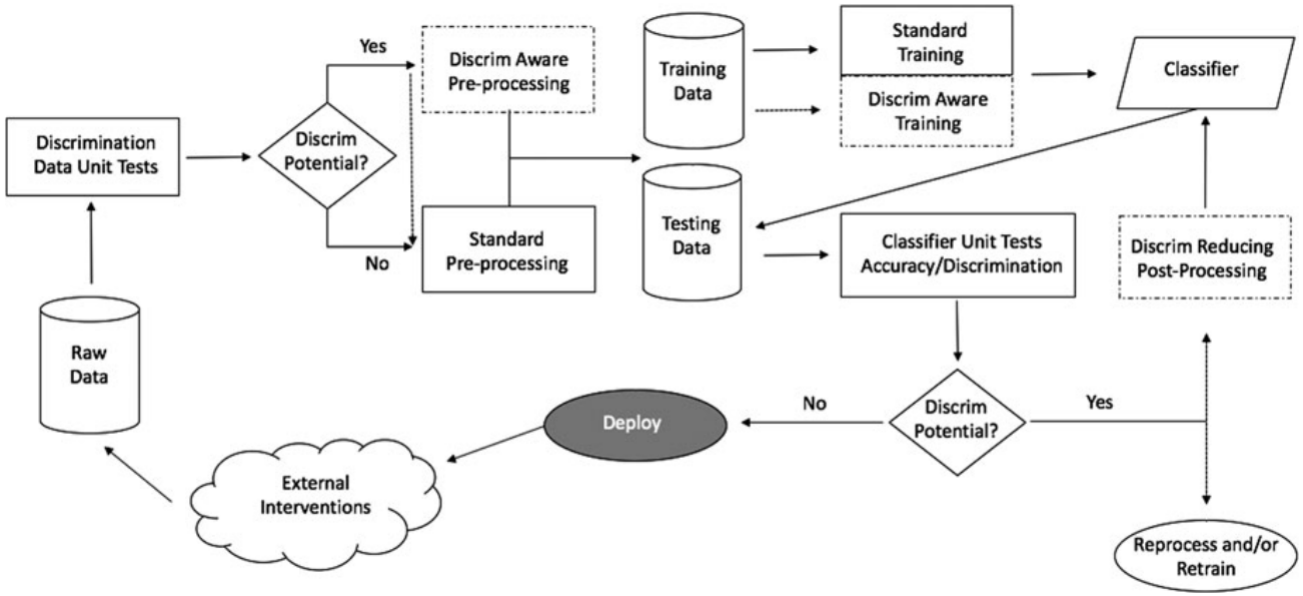


Bias mitigation is not easy

Cannot simply drop protected attributes because features are correlated with them

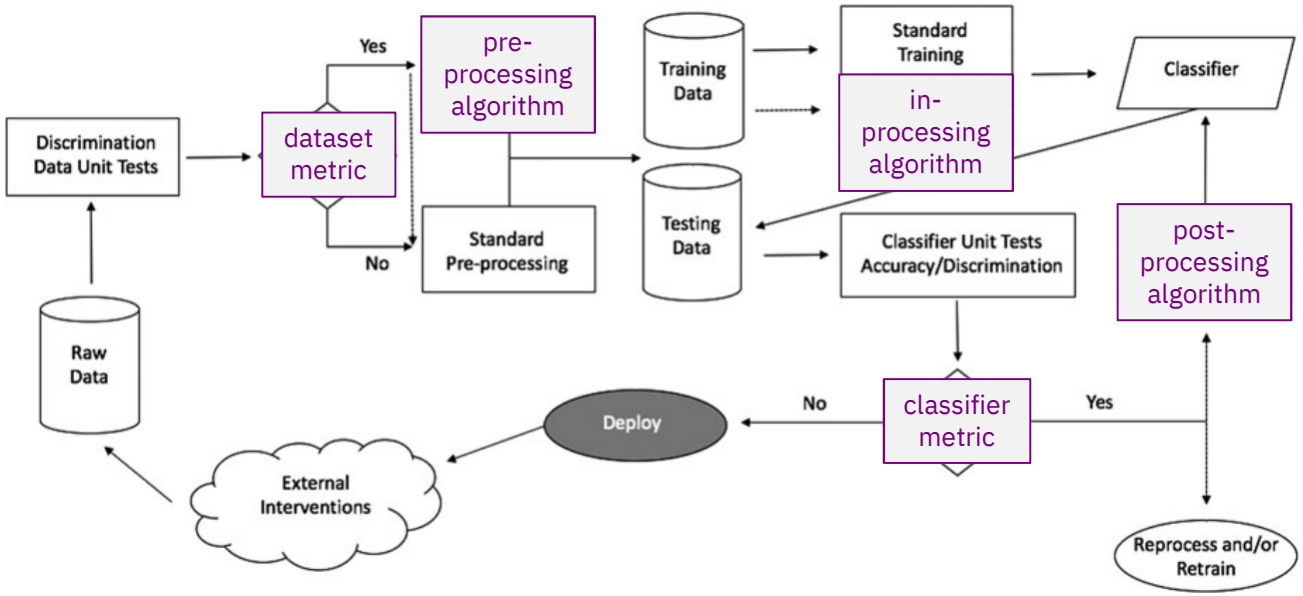


Fairness in building and deploying models



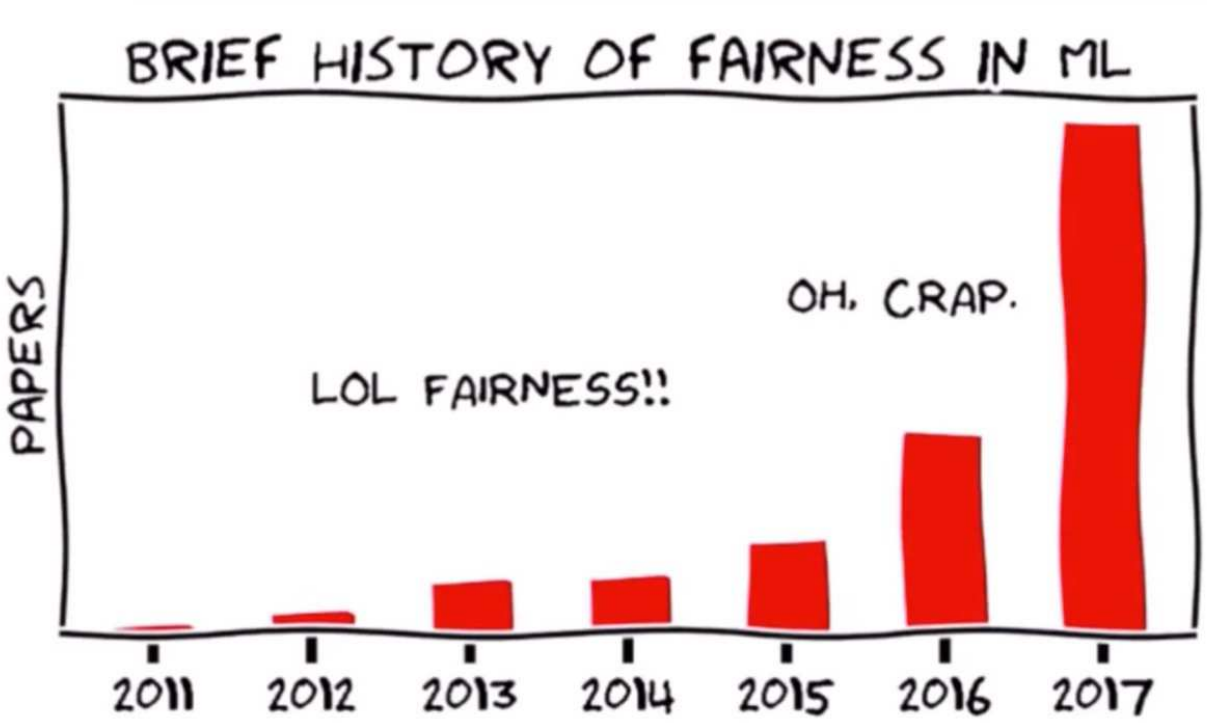
(d'Alessandro et al., 2017)

Metrics, Algorithms



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