IBM Research AI

AI Fairness 360

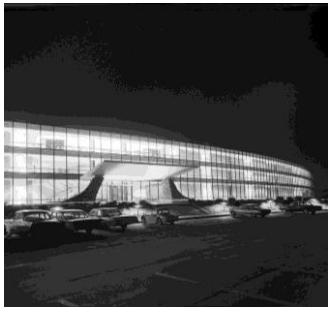
Kush R. Varshney

krvarshn@us.ibm.com
http://krvarshney.github.io
@krvarshney

http://aif360.mybluemix.net
https://github.com/ibm/aif360
https://pypi.org/project/aif360

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

Unwanted bias and algorithmic fairness

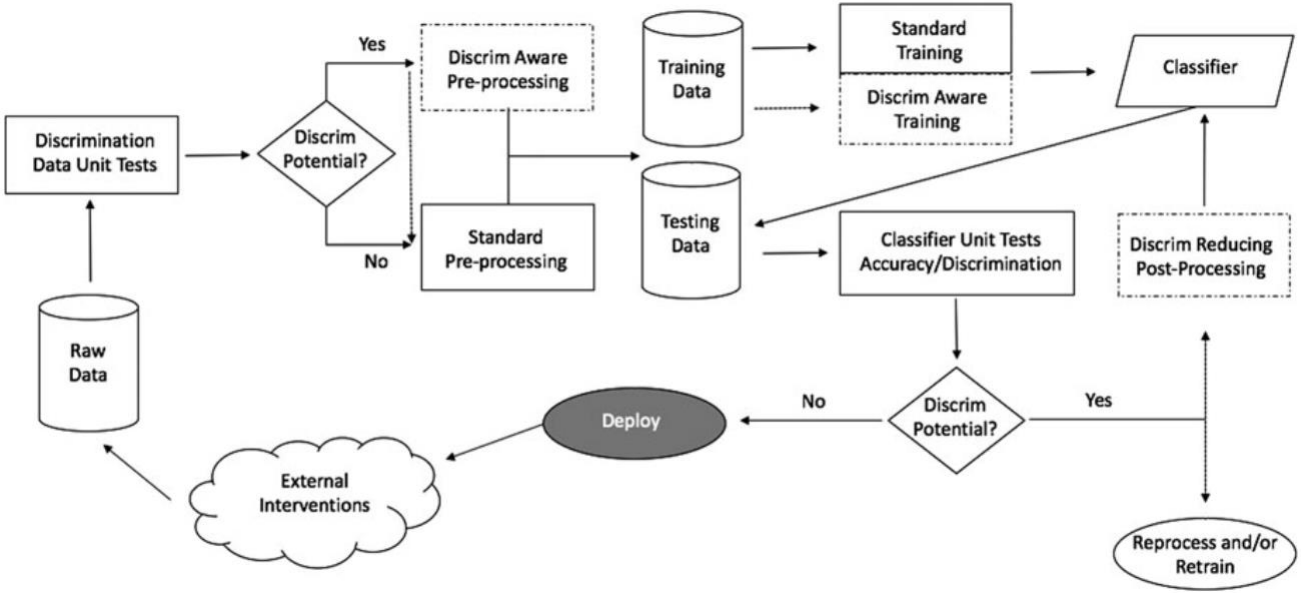
Machine learning, by its very nature, is always a form of statistical discrimination

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

Prejudice in labels

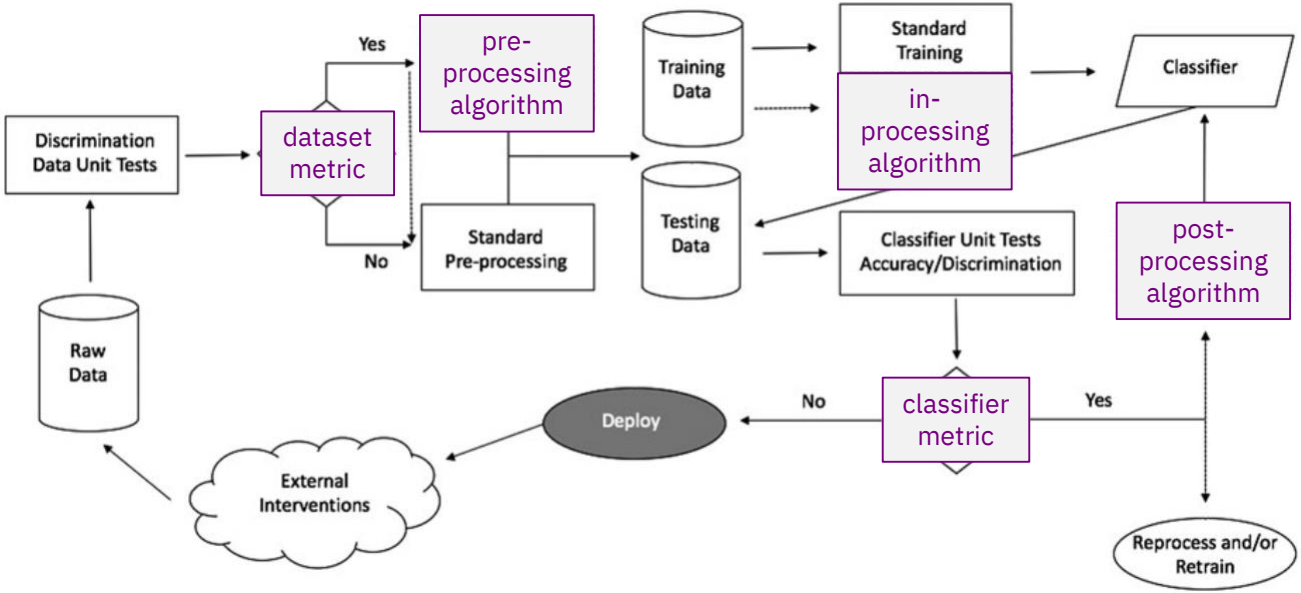
Undersampling or oversampling

Fairness in building and deploying models

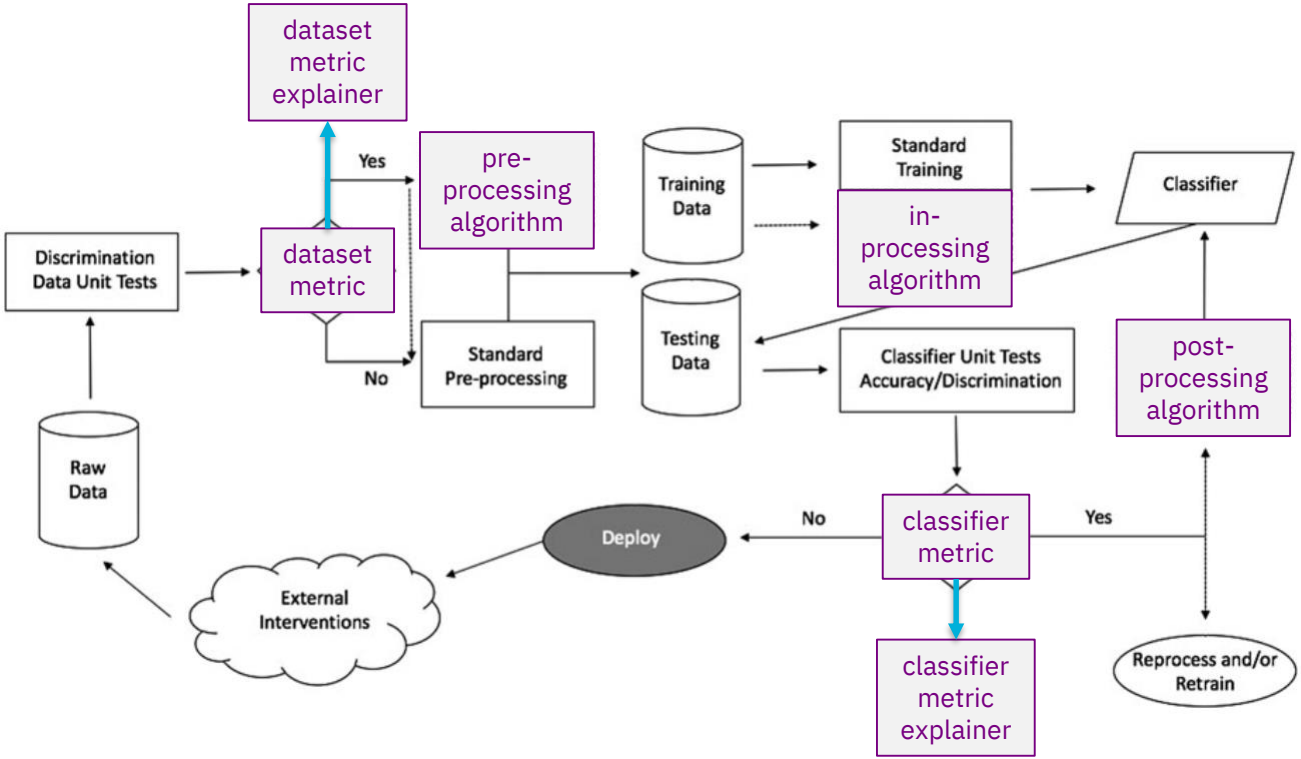


(d'Alessandro et al., 2017)

Metrics, Algorithms



Metrics, Algorithms, and Explainers



21 (or more) definitions of fairness

and the need for a toolbox with guidance

There is no one definition of fairness applicable in all contexts

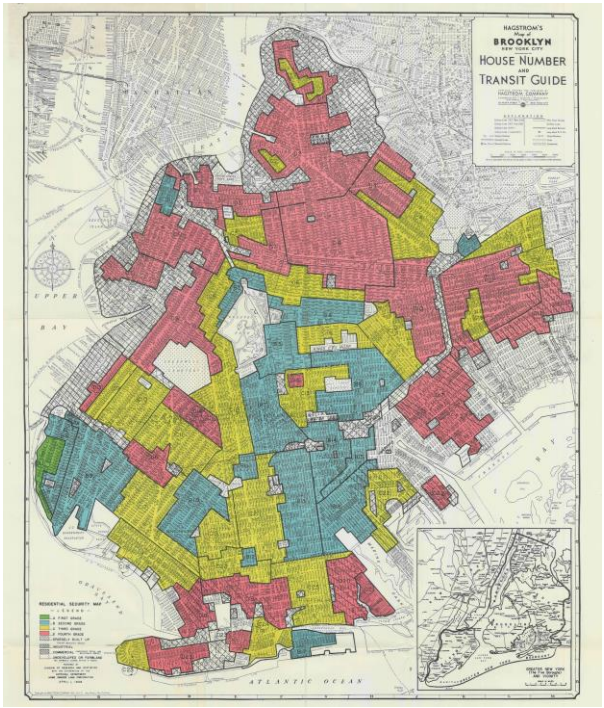
Some definitions even conflict

Requires a comprehensive set of fairness metrics and bias mitigation algorithms

Also requires some guidance to industry practitioners

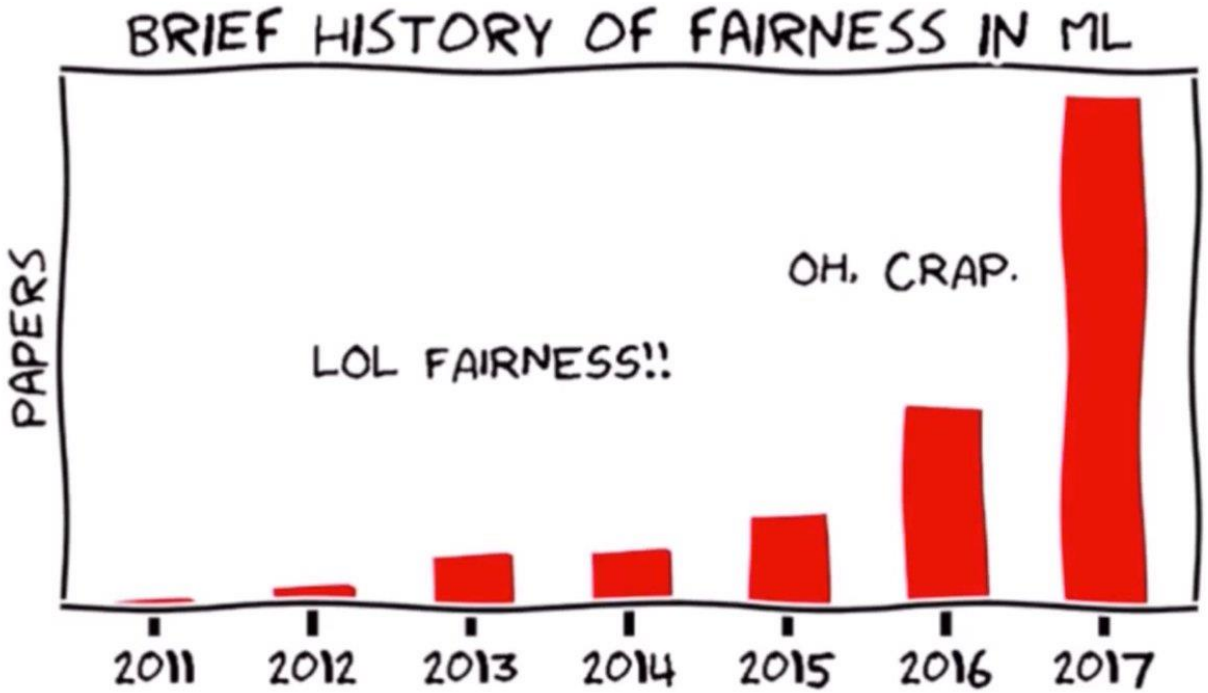
Bias mitigation is not easy

Cannot simply drop protected attributes because features are correlated with them



Research

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



(Hardt, 2017)

05/03/18	facebook	Facebook says it has a tool to detect bias in its artificial intelligence	Quartz
05/25/18	Microsoft	Microsoft is creating an oracle for catching biased AI algorithms	MIT Technology Review
05/31/18	pymetrics	Pymetrics open-sources Audit AI, an algorithm bias detection tool	VentureBeat
06/07/18	Google	Google Education Guide to Responsible AI Practices – Fairness	Google
06/09/18	accenture	Accenture wants to beat unfair AI with a professional toolkit	TechCrunch

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

AI Fairness 360

Datasets

Toolbox

- Fairness metrics (30+)
- Fairness metric explanations
- Bias mitigation algorithms (9+)

Guidance

Industry-specific tutorials

Differentiation

Comprehensive bias mitigation toolbox
(including unique algorithms from IBM Research)

Several metrics and algorithms that have **no available implementations** elsewhere

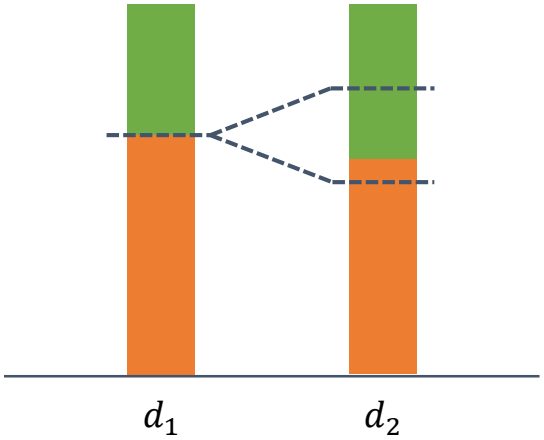
Extensible

Designed to translate new research from the lab to industry practitioners
(e.g. scikit-learn's fit/predict paradigm)

Optimized Preprocessing (NIPS 2017)

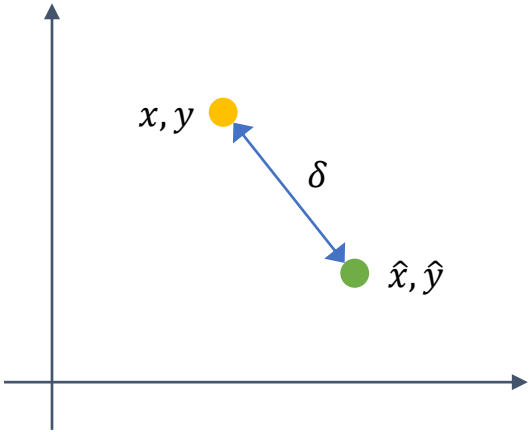
1. Group discrimination

Control dependence $p_{\hat{Y}|D}$ of transformed outcome $\hat{Y}$ on D



2. Individual distortion

Avoid large changes in individual features



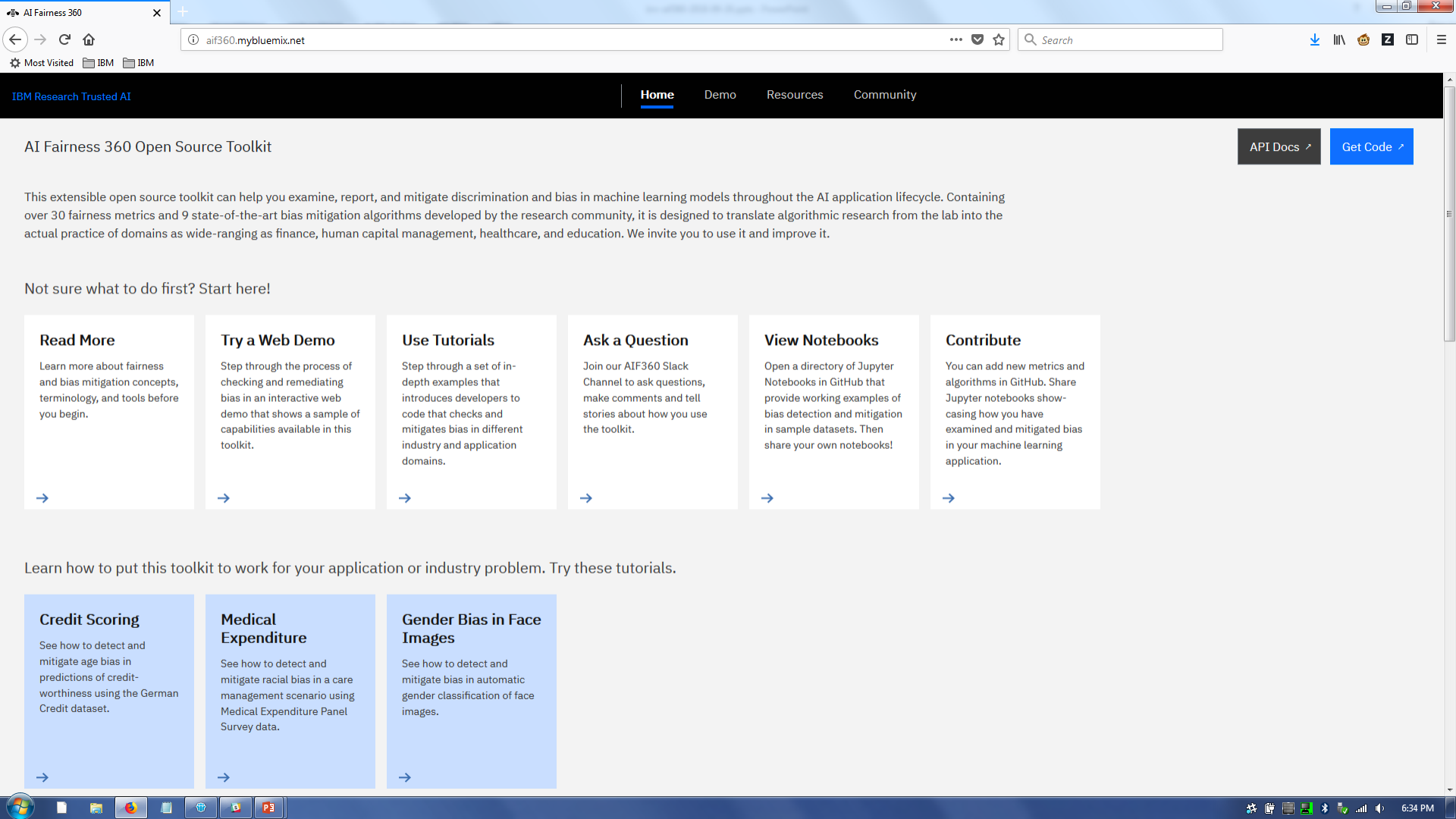
3. Utility preservation

Retain joint distribution $p_{X,Y}$ so model can still learn task

$$\min \quad \Delta(p_{\hat{X},\hat{Y}}, p_{X,Y})$$

$$\text{s.t.} \quad J\left(p_{\hat{Y}|D}(\hat{y}|d_1), p_{\hat{Y}|D}(\hat{y}|d_1)\right) \leq \epsilon$$

$$\mathbf{E}\left[\delta\left((x,y),(\hat{X},\hat{Y})\right) \mid d, x, y\right] \leq c$$


AI Fairness 360
aif360.mybluemix.net
Search
Most Visited
IBM
IBM
IBM Research Trusted AI
Home
Demo
Resources
Community
AI Fairness 360 Open Source Toolkit
API Docs
Get Code
This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. Containing over 30 fairness metrics and 9 state-of-the-art bias mitigation algorithms developed by the research community, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.
Not sure what to do first? Start here!
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.
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.
Learn how to put this toolkit to work for your application or industry problem. Try these tutorials.
Credit Scoring
See how to detect and mitigate age bias in predictions of credit-worthiness using the German Credit dataset.
Medical Expenditure
See how to detect and mitigate racial bias in a care management scenario using Medical Expenditure Panel Survey data.
Gender Bias in Face Images
See how to detect and mitigate bias in automatic gender classification of face images.
6:34 PM