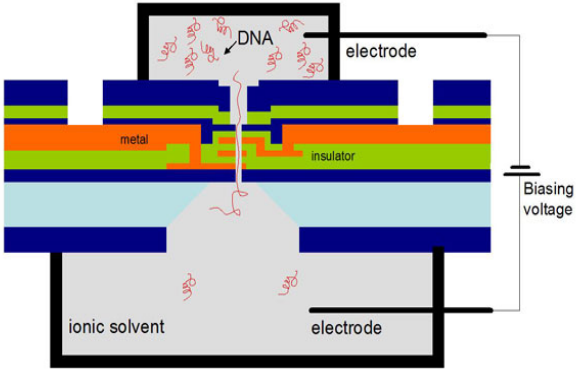
Trusted AI: Bringing Trust and Transparency into AI through Open Source

Animesh Singh
STSM and Chief Architect
Data and AI Open Source Platform

IBM has a a history of tech for social good…

The moon landing

Human genome sequencing



Tracking infectious diseases

Disaster Response

and in the same spirit, we are working to bring

Trust and Transparency into AI..

but why does it matter?

“What is vital is to make anything about AI **explainable**, **fair**, **secure**, and **with lineage** — meaning that anyone could very simply see how applications of AI are developed and why.”

Ginni Rometty
CES 2019 Opening Keynote

IBM News Room - IBM @ CES <https://newsroom.ibm.com/CES>

“Instrumenting trust into data sets and machine learning models will **accelerate the adoption of AI** and engender increased confidence in these general-purpose technologies.”

Aleksandra Mojsilovic
IBM Fellow
Head of Foundations of Trusted AI

Building trust into AI <https://www.ibm.org/responsibility/2018/trusted-ai#story>

“If we fail to make **ethical** and **inclusive** artificial intelligence we risk losing gains made in civil rights and gender equity under the guise of machine neutrality.”

Joy Buolamwini
Gender Shades
MIT Media Lab

So what does it take to trust a decision made by a machine?

(Other than that it is 99% accurate)?

Did anyone tamper with it?

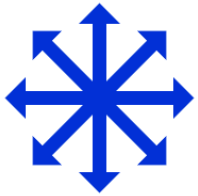
Is it fair?

Is it easy to understand?

Is it accountable?

Our vision for Trusted AI

Pillars of trust, woven into the lifecycle of an AI application



ROBUSTNESS

Did anyone tamper with it?

FAIRNESS

Is it fair?

EXPLAINABILITY

Is it easy to understand?

LINEAGE

Is it accountable?

IBM has a long history in the open source ecosystem

and

We are leveraging this to bring Trust and Transparency into AI though Open Source..

THE APACHE™ SOFTWARE FOUNDATION

THE LINUX FOUNDATION

ECLIPSE™ FOUNDATION

CLOUD NATIVE COMPUTING FOUNDATION

Let’s talk about Robustness, why is it important?

ROBUSTNESS

Did anyone tamper with it?

FAIRNESS

Is it fair?

EXPLAINABILITY

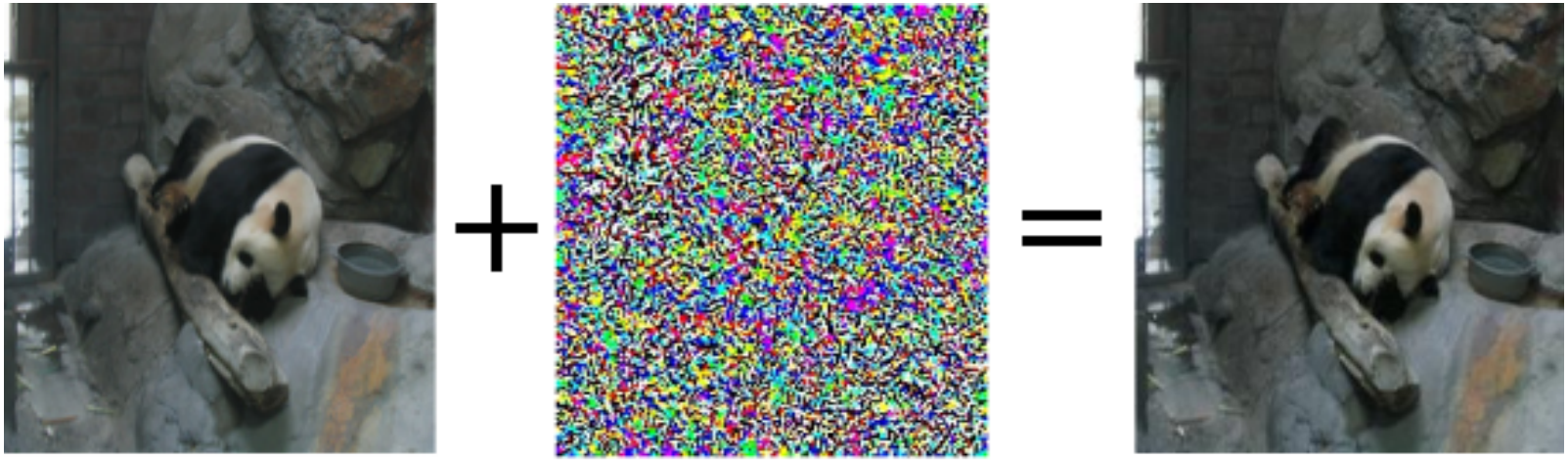
Is it easy to understand?

LINEAGE

Is it accountable?

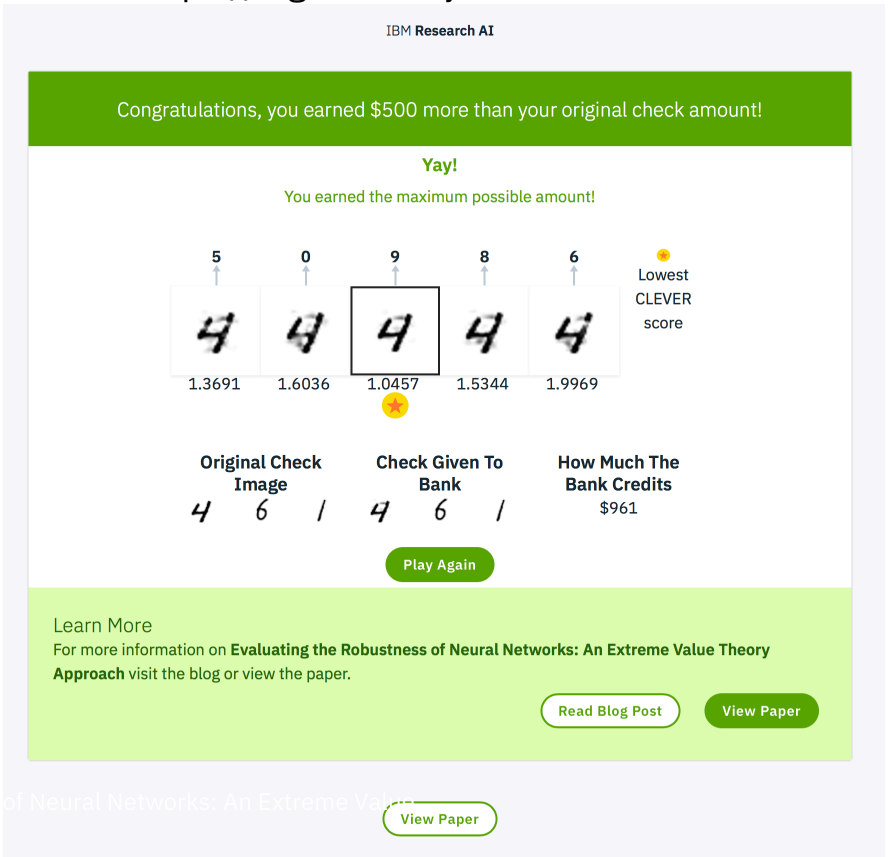
Deep learning and adversarial attacks

Deep Learning models are now used in many areas - Can we trust them?

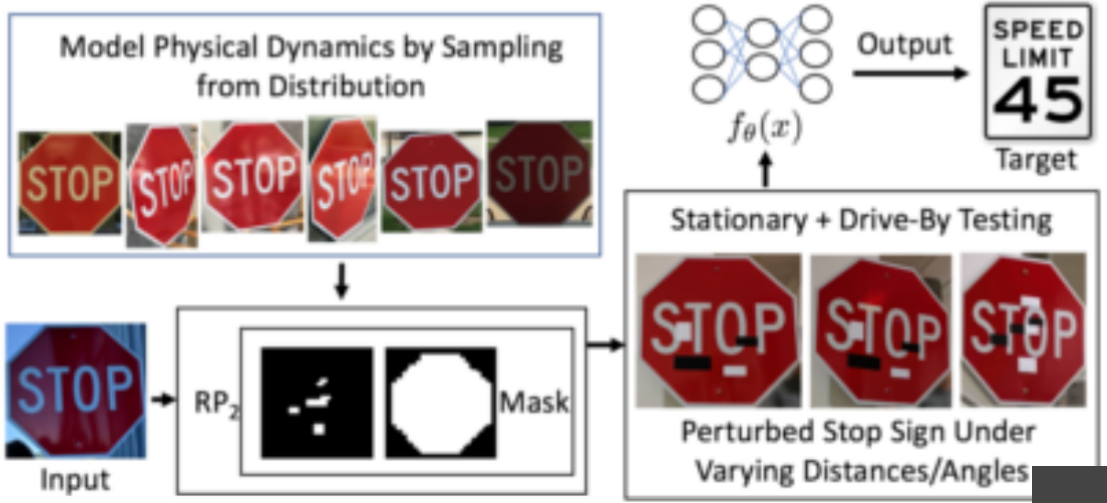


How does it impact us?

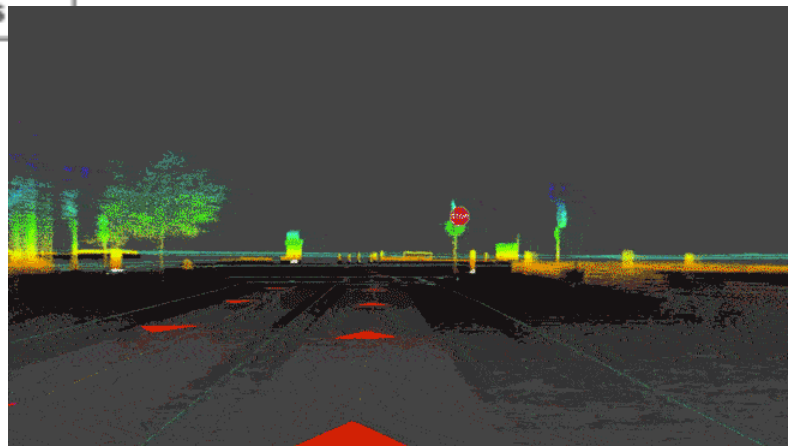
https://bigcheck.mybluemix.net



Scarier example..



1707.08945.pdf



Another scarier example..

https://arxiv.org/abs/1704.05712

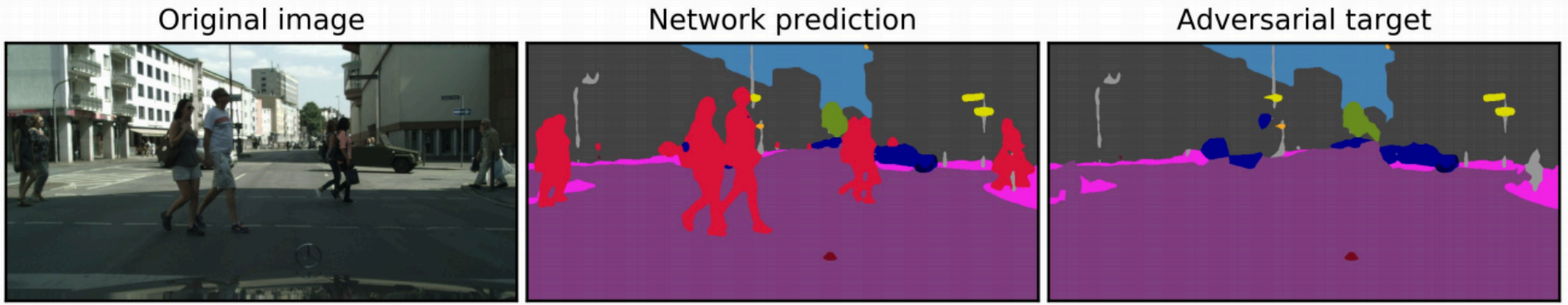
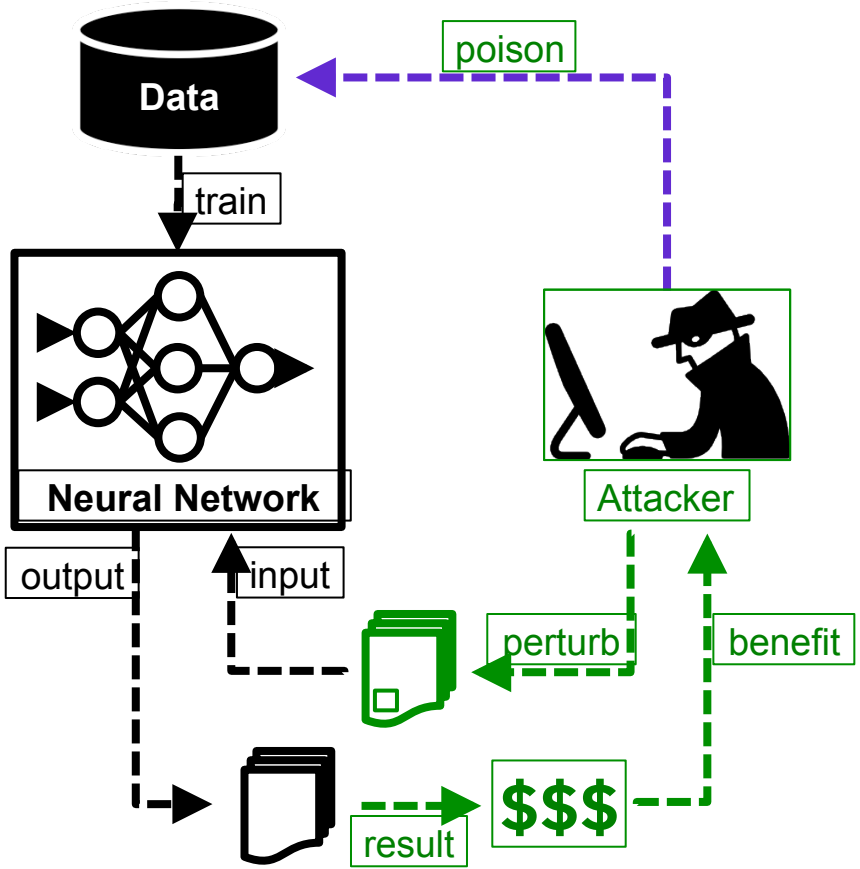


Figure 2. Illustration of an adversary generating a dynamic target segmentation for hiding pedestrians.

Adversarial Threats to AI



Evasion attacks

- Performed at test time
- Perturb inputs with crafted noise
- Model fails to predict correctly
- Undetectable by humans

Poisoning attacks

- Performed at training time
- Insert poisoned sample in training data
- Use backdoor later

Adversarial Robustness Toolbox ↳ (ART)

https://github.com/IBM/adversarial-robustness-toolbox

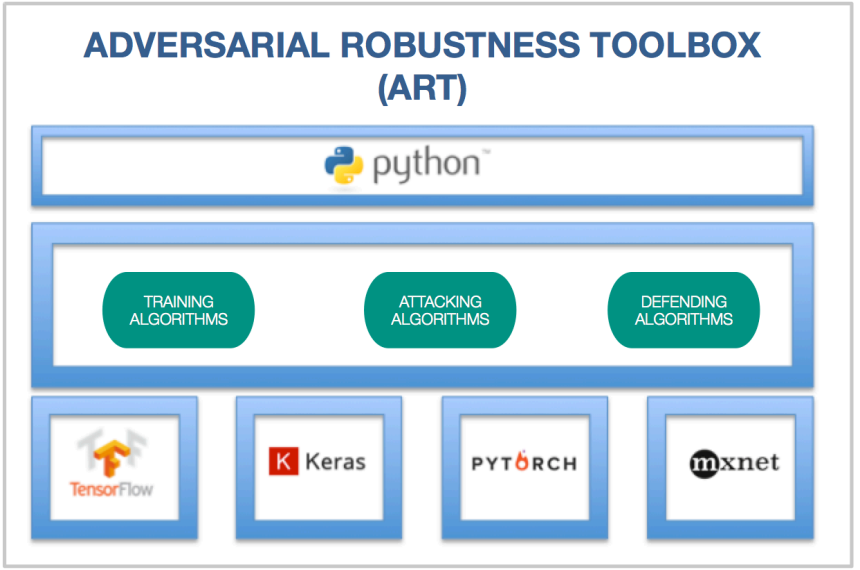
ART is a library dedicated to adversarial machine learning. Its purpose is to allow rapid crafting and analysis of **attack, defense and detection methods** for machine learning models. Applicable domains include finance, self driving vehicles etc.

The Adversarial Robustness Toolbox provides an implementation for many state-of-the-art methods for attacking and defending classifiers.

Toolbox: Attacks, defenses, and metrics
Evasion attacks
Defense methods
Detection methods
Robustness metrics

https://art-demo.mybluemix.net/

ART



Implementation for state-of-the-art methods for attacking and defending classifiers.

Evasion attacks

- FGSM
- JSMA
- BIM
- PGD
- Carlini & Wagner
- DeepFool
- NewtonFool
- Universal perturbation

Evasion defenses

- Feature squeezing
- Spatial smoothing
- Label smoothing
- Adversarial training
- Virtual adversarial training
- Thermometer encoding
- Gaussian data augmentation

Poisoning detection

- Detection based on clustering activations
- Proof of attack strategy

Evasion detection

- Detector based on inputs
- Detector based on activations

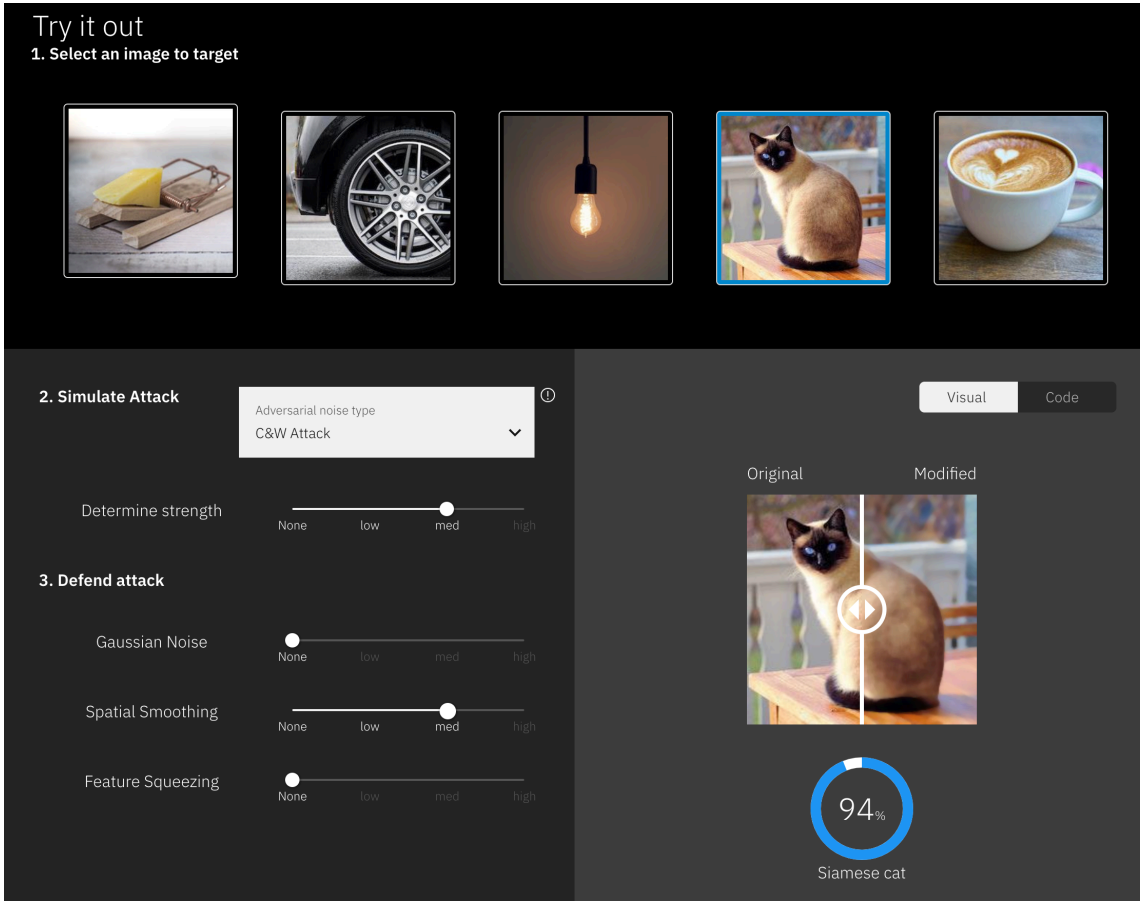
Robustness metrics

- CLEVER
- Empirical robustness
- Loss sensitivity

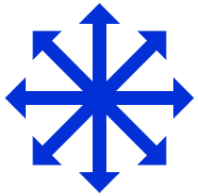
Unified model API

- Training
- Prediction
- Access to loss and prediction gradients

ART Demo: https://art-demo.mybluemix.net/



Now let’s discuss Fairness, why bias in AI is such an important topic?



ROBUSTNESS

Did anyone tamper with it?

FAIRNESS

Is it fair?

EXPLAINABILITY

Is it easy to understand?

LINEAGE

Is it accountable?

What is bias?

*A **cognitive bias** is a systematic pattern of deviation from norm or rationality in judgment. Individuals create their own "subjective social reality" from their perception of the input.”*

- Wikipedia

Bias in AI Example: Criminal Justice System

Since 2008, nearly every arrestee in Broward County, Florida has been assigned a risk score using Northpointe's COMPAS algorithm.

Defendants with low risk scores are released on bail.

It falsely flagged black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants

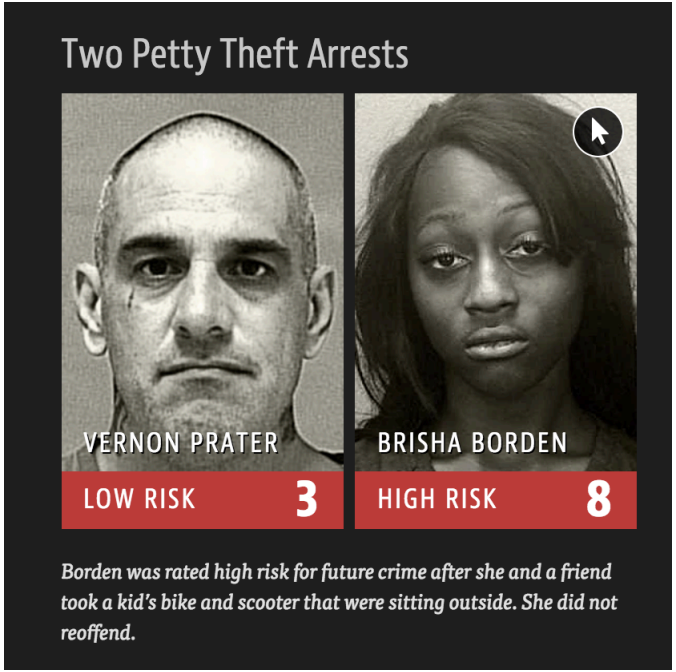
Machine Bias — ProPublica <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

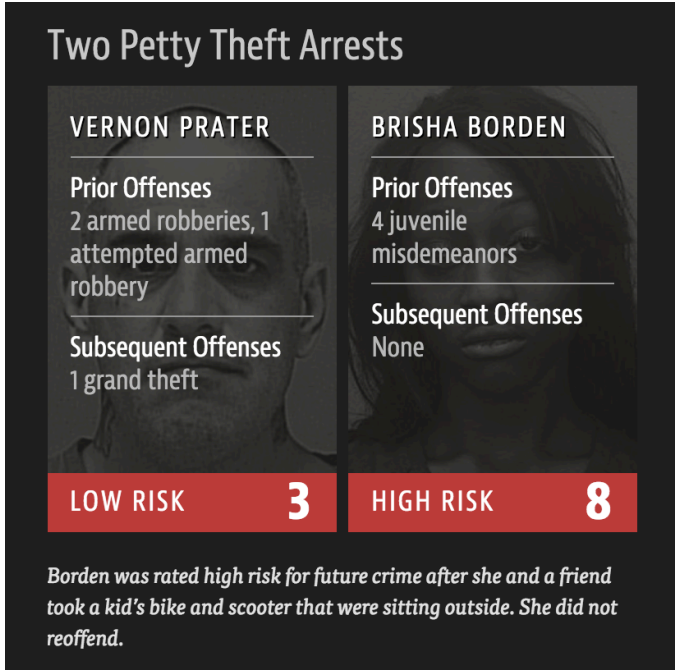
Bias in AI Example: Criminal Justice System

Since 2008, nearly every arrestee in Broward County, Florida has been assigned a risk score using Northpointe's COMPAS algorithm.

Defendants with low risk scores are released on bail.

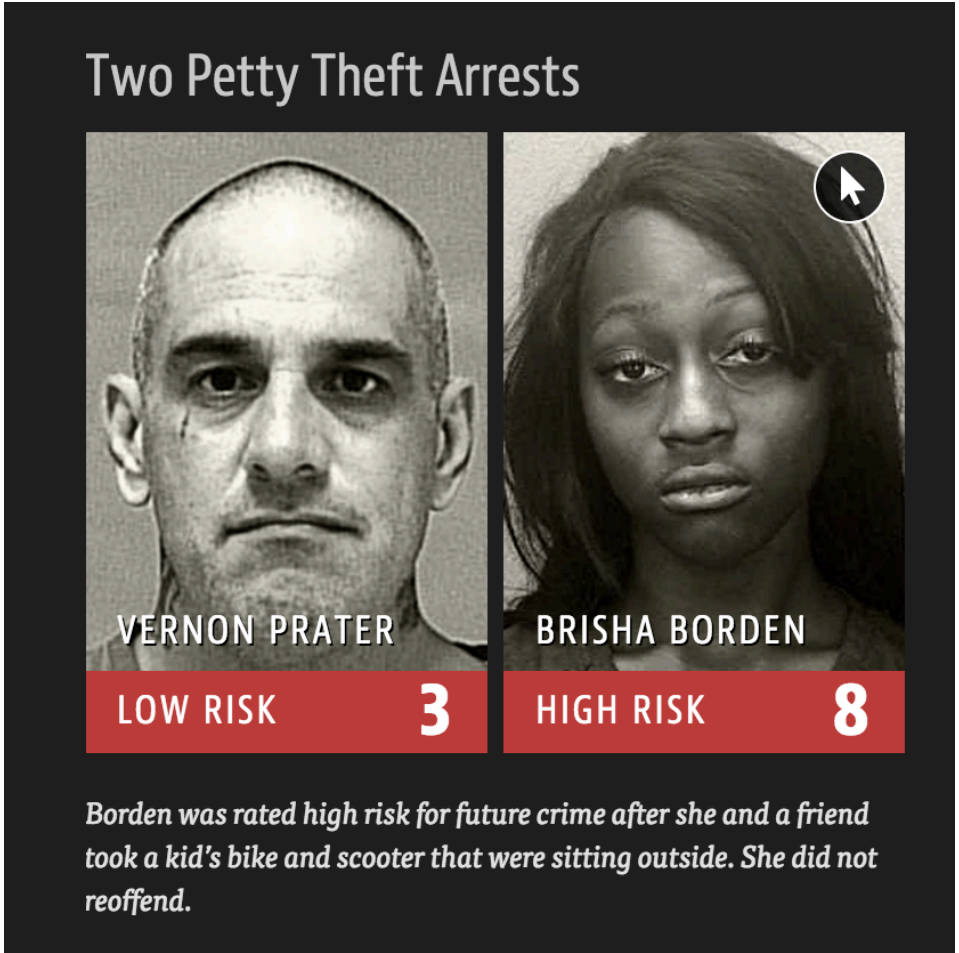
It falsely flagged black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants



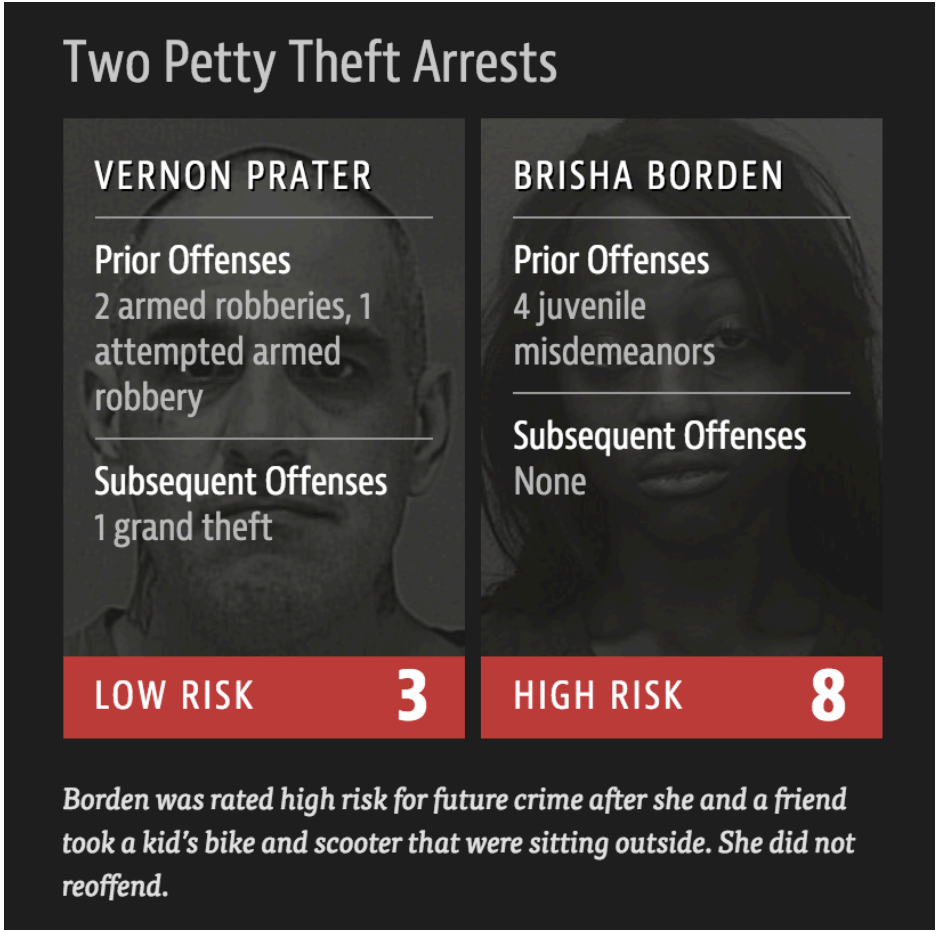


Machine Bias — ProPublica <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

Bias in Recidivism Assessment (Propublica, May 2016)



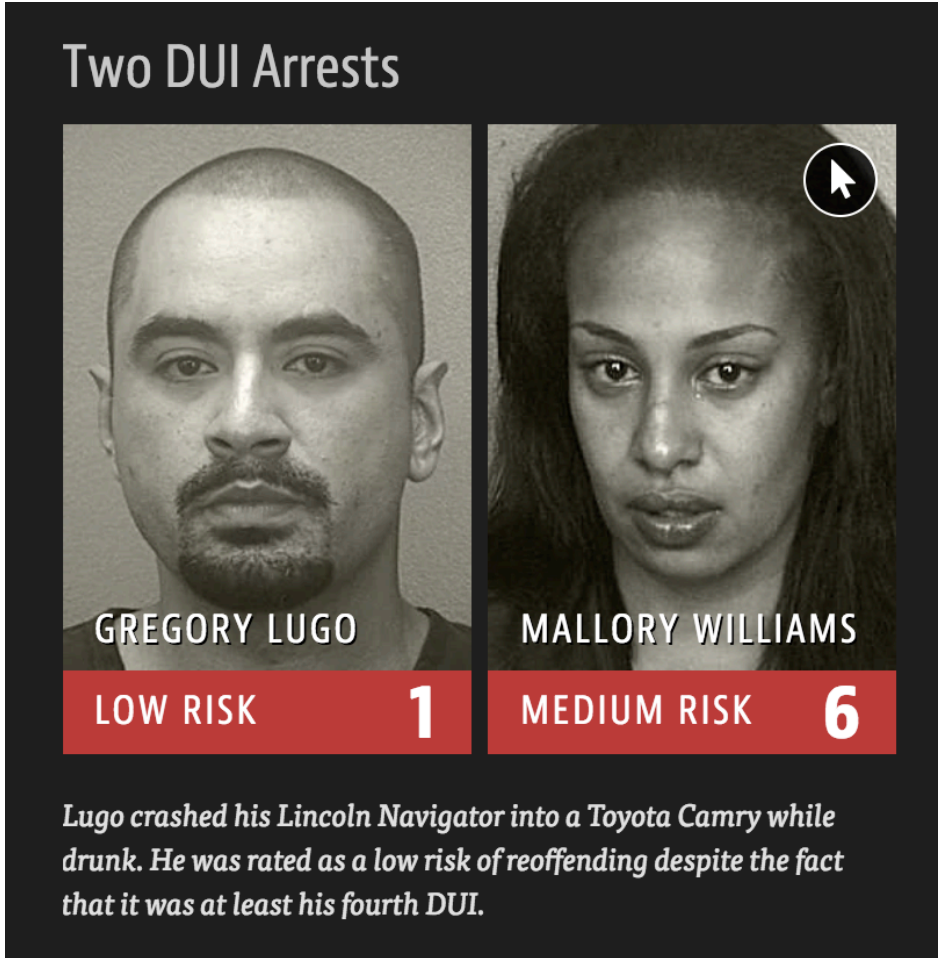
Bias in Recidivism Assessment (Propublica, May 2016)



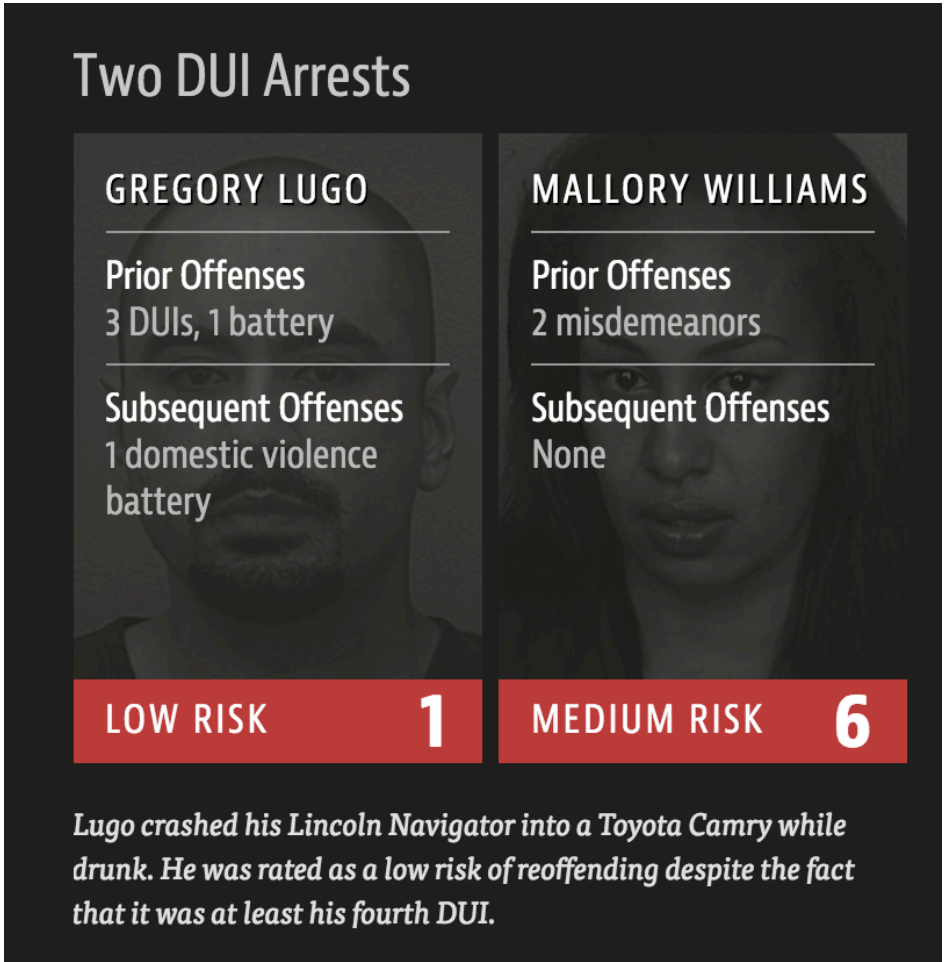
Bias in Recidivism Assessment (Propublica, May 2016)

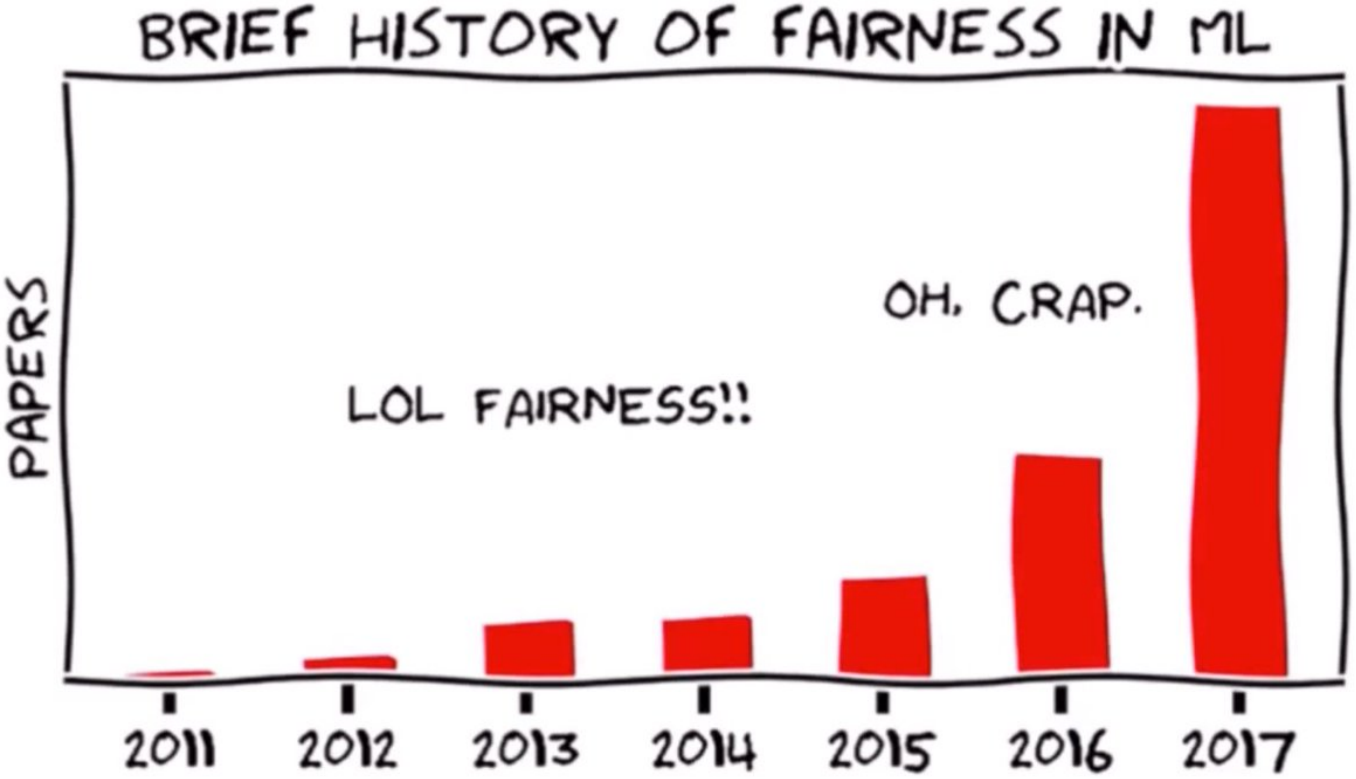
Bias in Recidivism Assessment (Propublica, May 2016)

Bias in Recidivism Assessment (Propublica, May 2016)



Bias in Recidivism Assessment (Propublica, May 2016)




BRIEF HISTORY OF FAIRNESS IN ML
PAPERS
LOL FAIRNESS!!
OH. CRAP.
2011
2012
2013
2014
2015
2016
2017

(Hardt, 2017)

AI Fairness 360
↳ (AIF360)

https://github.com/IBM/AIF360

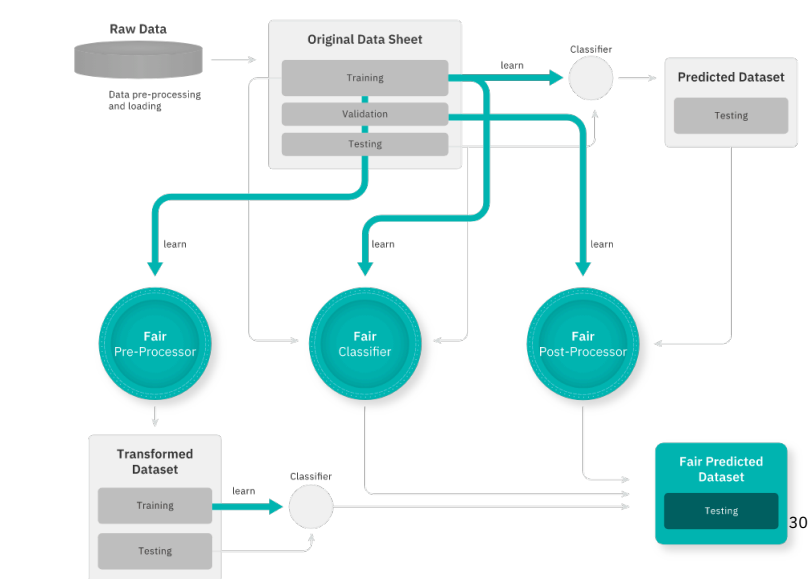
AIF360 toolkit is an open-source library to help detect and remove bias in machine learning models. **AIF360 translates algorithmic research from the lab into practice**. Applicable domains include finance, human capital management, healthcare, and education.

The AI Fairness 360 Python package includes a comprehensive set of metrics for datasets and models to test for biases, explanations for these metrics, and algorithms to mitigate bias in datasets and models.

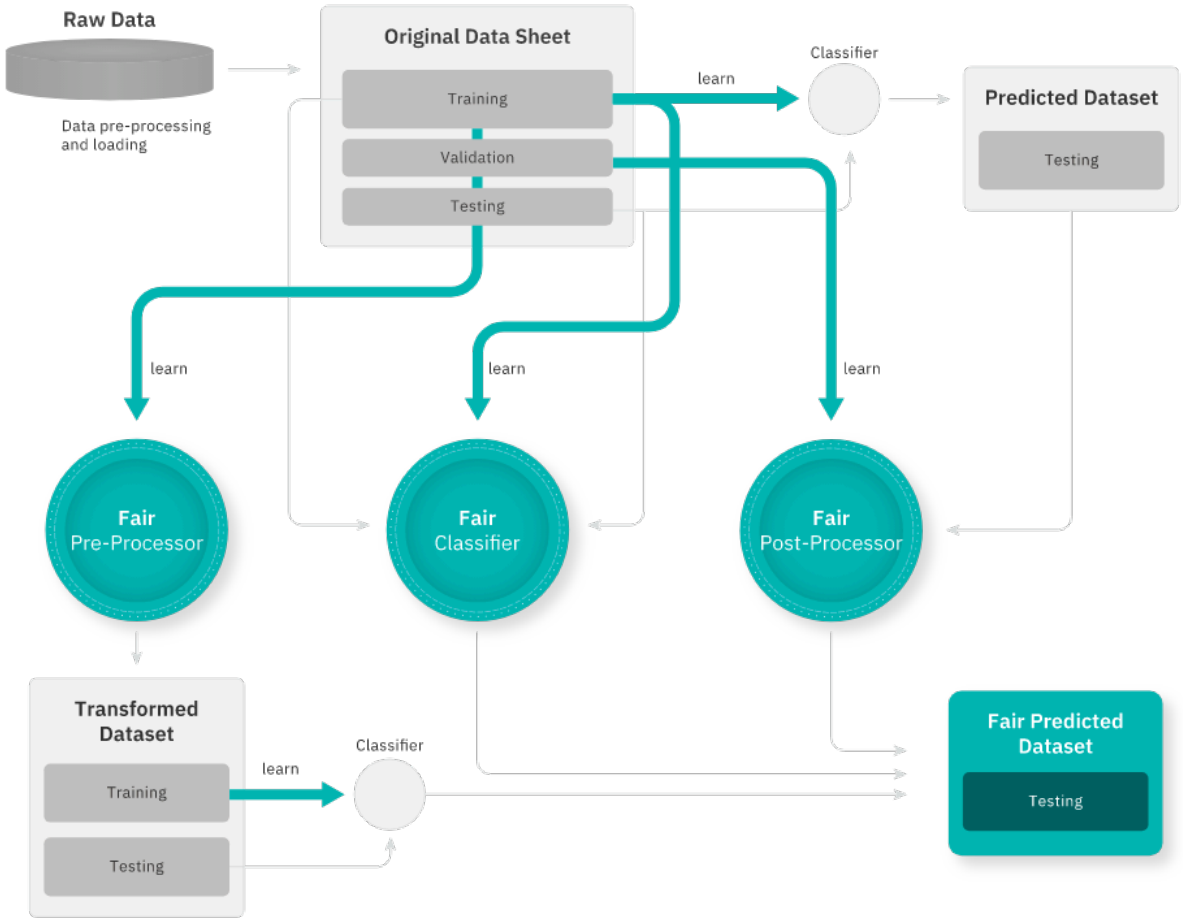
Toolbox
Fairness metrics (70+)
Fairness metric explanations
Bias mitigation algorithms (10+)

http://aif360.mybluemix.net/

AIF360



AIF 360 detects for fairness in building and deploying models throughout AI Lifecycle



Metrics (70+)

Statistical Parity Difference

The difference of the rate of favorable outcomes received by the unprivileged group to the privileged group.

Equal Opportunity Difference

The difference of true positive rates between the unprivileged and the privileged groups.

Average Odds Difference

The average difference of false positive rate (false positives/negatives) and true positive rate (true positives/positives) between unprivileged and privileged groups.

Disparate Impact

The ratio of rate of favorable outcome for the unprivileged group to that of the privileged group.

Theil Index

Measures the inequality in benefit allocation for individuals.

Euclidean Distance

The average Euclidean distance between the samples from the two datasets.

Mahalanobis Distance

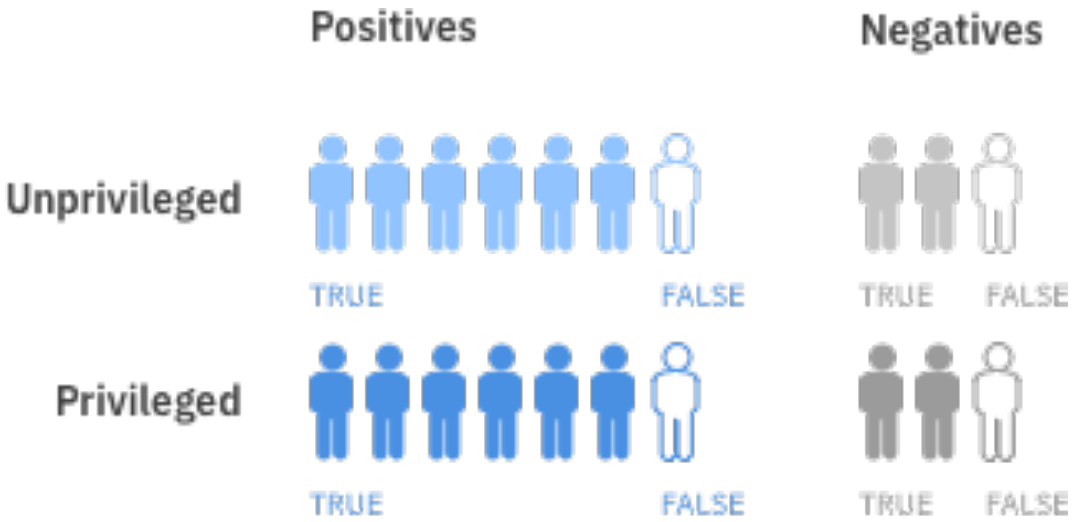
The average Mahalanobis distance between the samples from the two datasets.

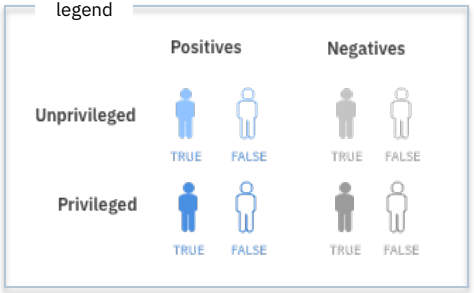
Manhattan Distance

The average Manhattan distance between the samples from the two datasets.

Group fairness metrics

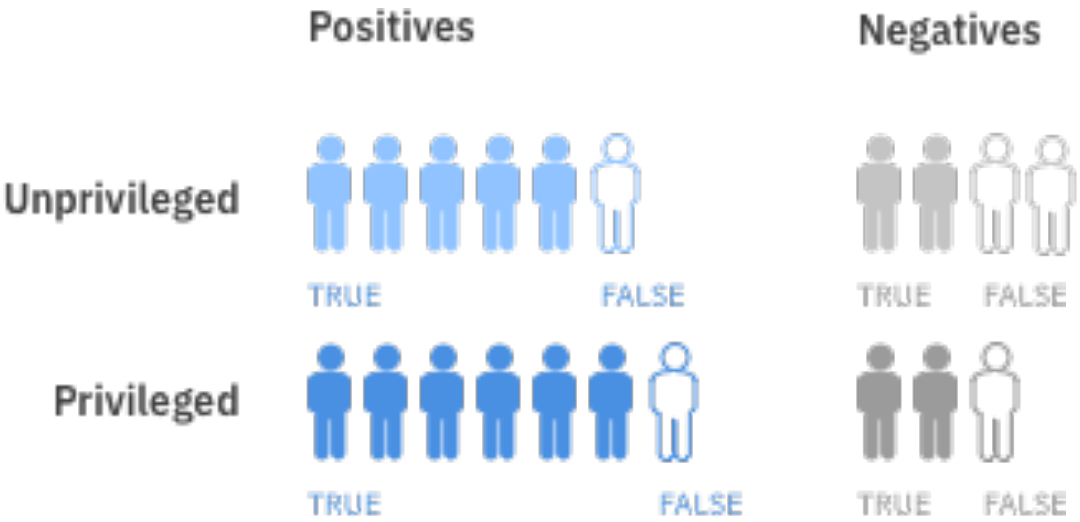
situation 1

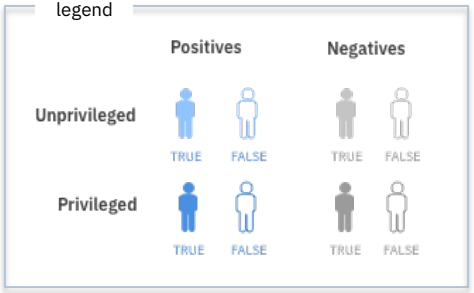




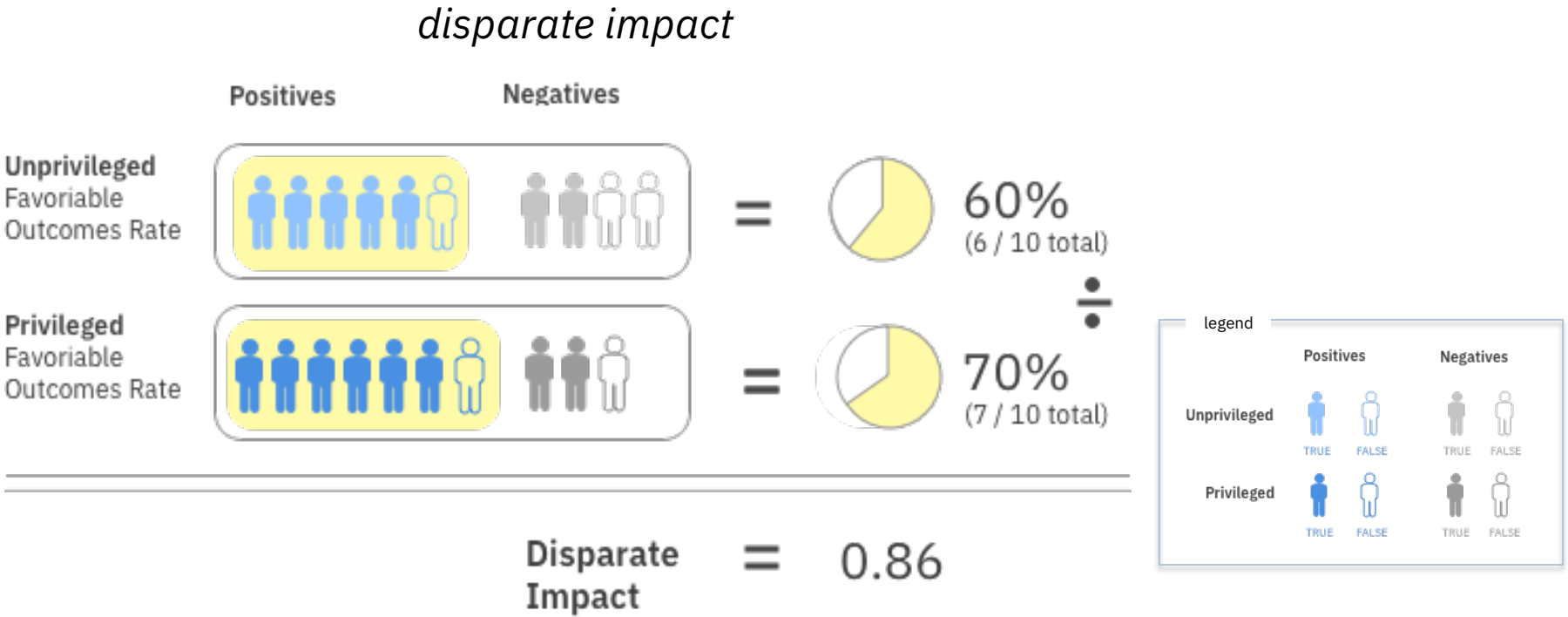
Group fairness metrics

situation 2

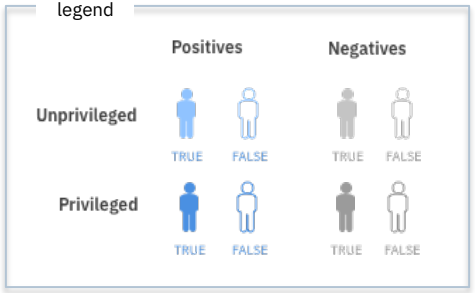




Group fairness metrics

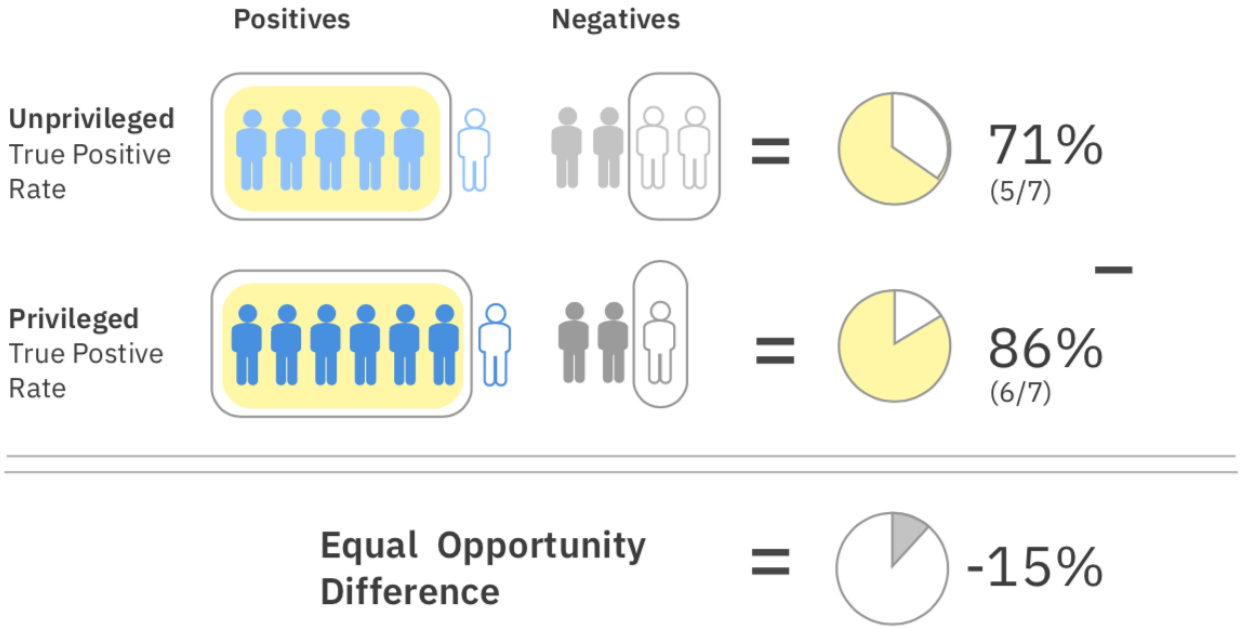


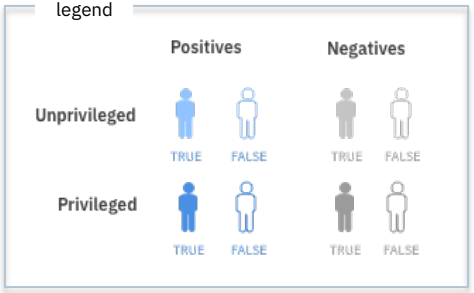
Group fairness metrics



Group fairness metrics

equal opportunity difference





Algorithms (10)

Optimized Pre-processing

Use to mitigate bias in training data. Modifies training data features and labels.

Reweighing

Use to mitgate bias in training data. Modifies the weights of different training examples.

Adversarial Debiasing

Use to mitigate bias in classifiers. Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions.

Reject Option Classification

Use to mitigate bias in predictions. Changes predictions from a classifier to make them fairer.

Disparate Impact Remover

Use to mitigate bias in training data. Edits feature values to improve group fairness.

Learning Fair Representations

Use to mitigate bias in training data. Learns fair representations by obfuscating information about protected attributes.

Prejudice Remover

Use to mitigate bias in classifiers. Adds a discrimination-aware regularization term to the learning objective.

Calibrated Equalized Odds Post-processing

Use to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels.

Equalized Odds Post-processing

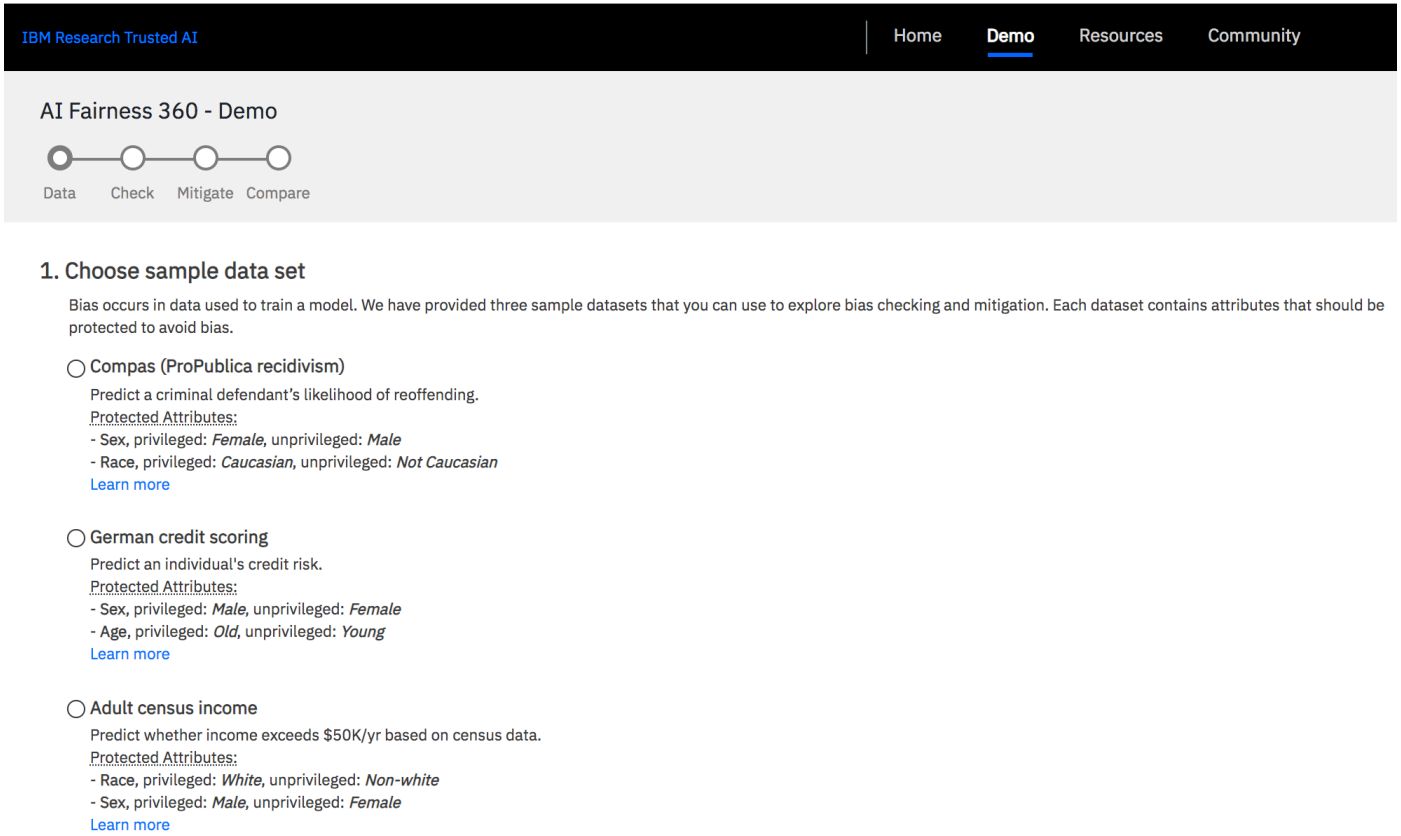
Use to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer.

Meta Fair Classifier

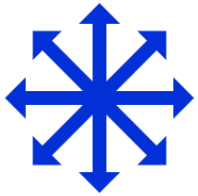
Use to mitigate bias in classifier. Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized for that metric.

Demo Application: AI Fairness 360 Web Application

http://aif360.mybluemix.net/



Shifting the focus towards Explainability…



ROBUSTNESS

Did anyone tamper with it?

FAIRNESS

Is it fair?

EXPLAINABILITY

Is it easy to understand?

LINEAGE

Is it accountable?

AI needs to explain its decision, and there are different ways to explain

One explanation does not fit all

Different stakeholders require explanations for different purposes and with different objectives, and explanations will have to be tailored to their needs.

End users/customers (trust)

Doctors: *Why did you recommend this treatment?*

Customers: *Why was my loan denied?*

Teachers: *Why was my teaching evaluated in this way?*

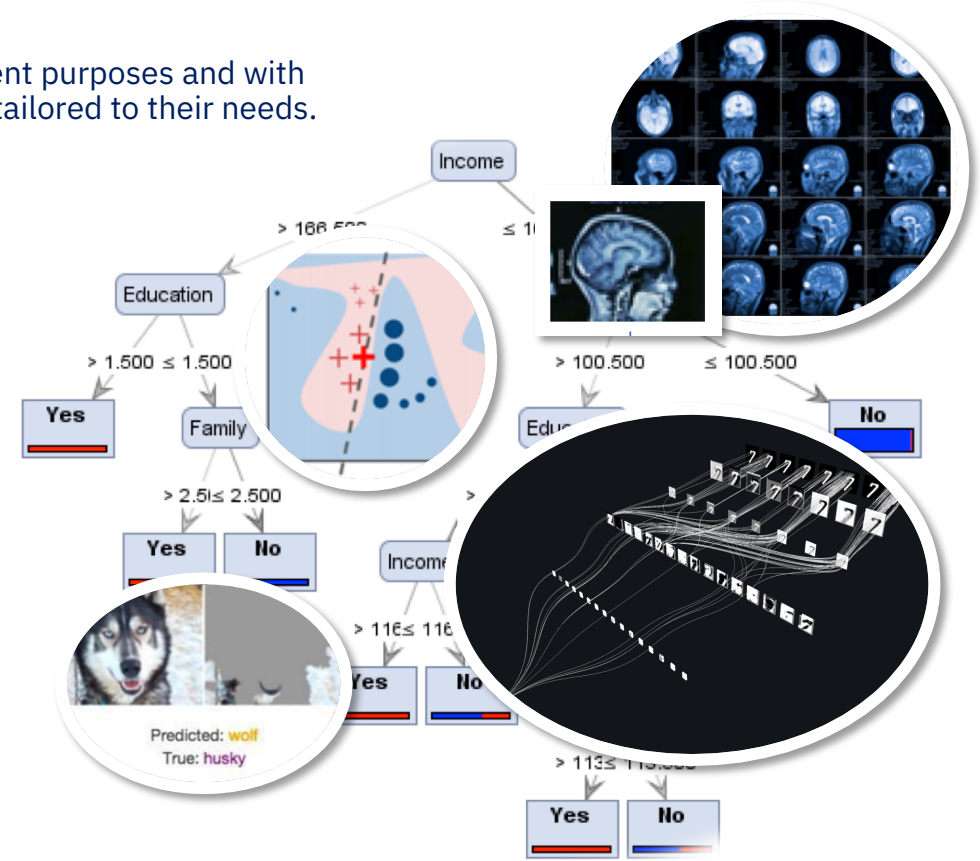
Gov't/regulators (compliance, safety)

Prove to me that you didn't discriminate.

Developers (quality, "debuggability")

Is our system performing well?

How can we improve it?



AI Explainability 360
↳ (AIX360)

https://github.com/IBM/AIX360

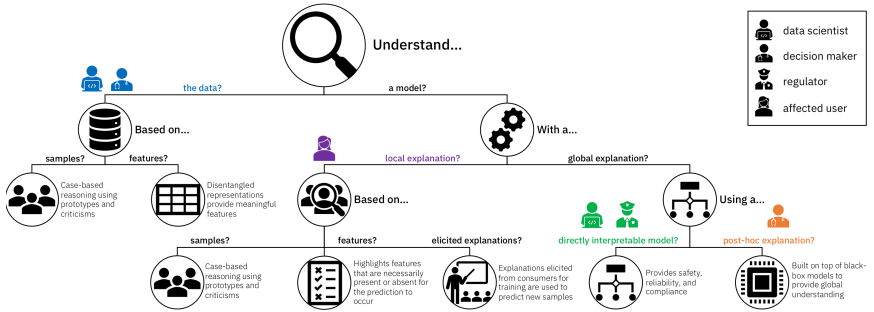
AIX360

AIX360 toolkit is an open-source library to help explain AI and machine learning models and their predictions. This includes three classes of algorithms: local post-hoc, global post-hoc, and directly interpretable explainers for models that use image, text, and structured/tabular data.

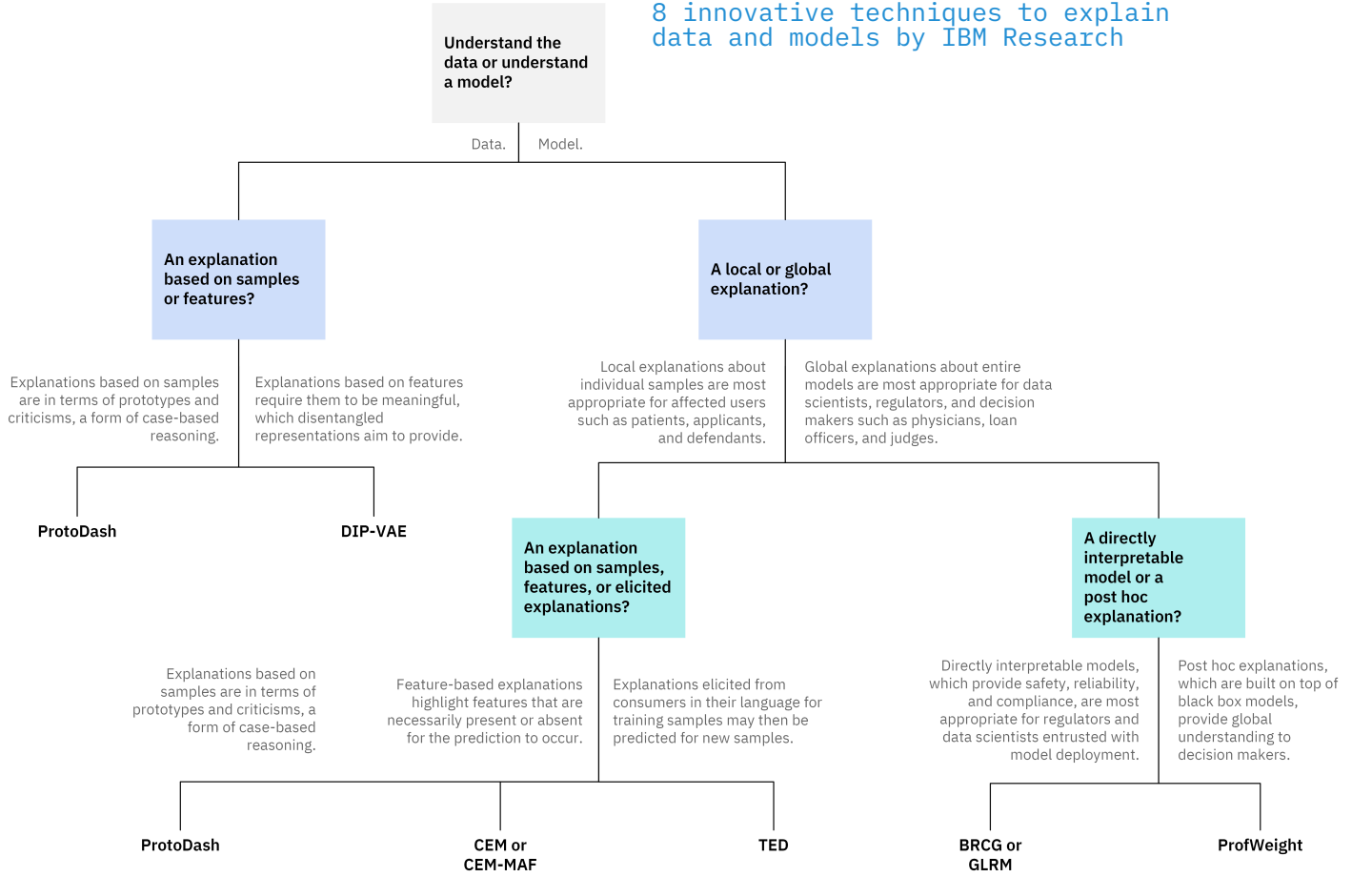
The AI Explainability360 Python package includes a comprehensive set of explainers, both at global and local level.

Toolbox
Local post-hoc
Global post-hoc
Directly interpretable

http://aix360.mybluemix.net



AIX360: Multiple dimensions of explainability



8 innovative techniques to explain data and models by IBM Research

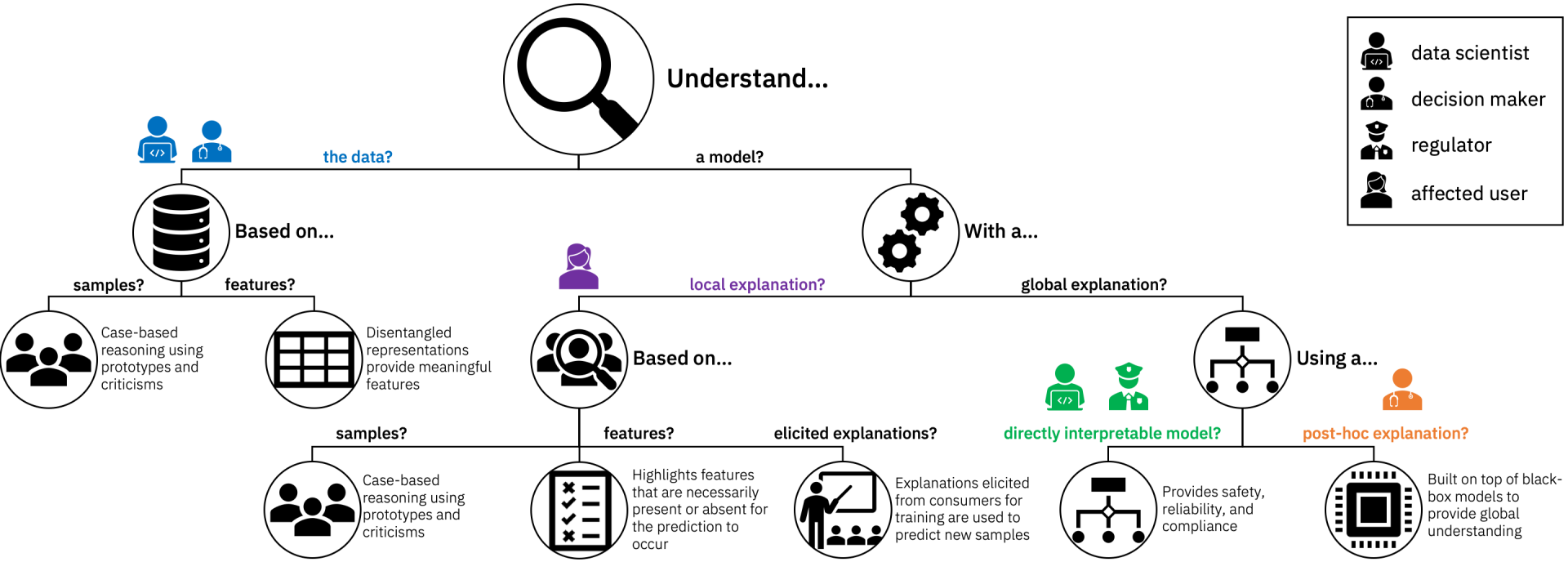
Data vs. Model

directly interpretable
vs.
Post-hoc interpretation

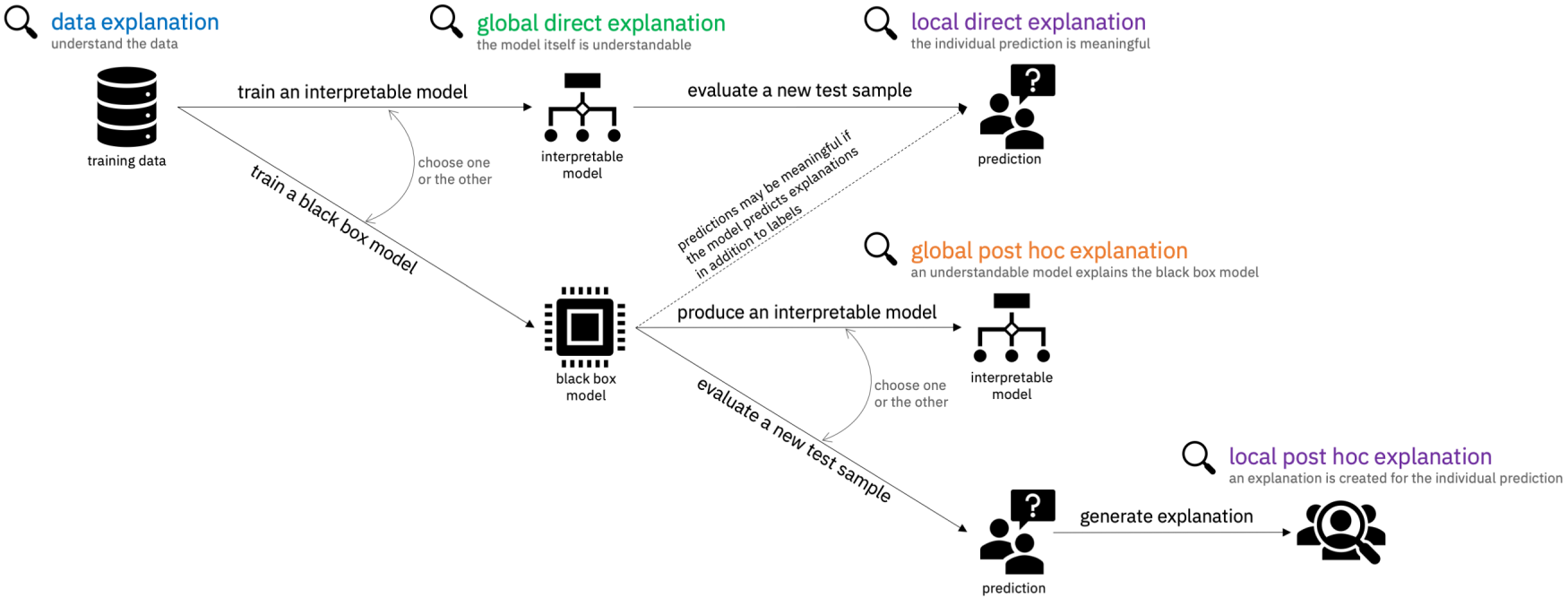
Local vs. Global

source: IBM Research AI Explainability 360

AIX360: Multiple dimensions of explainability

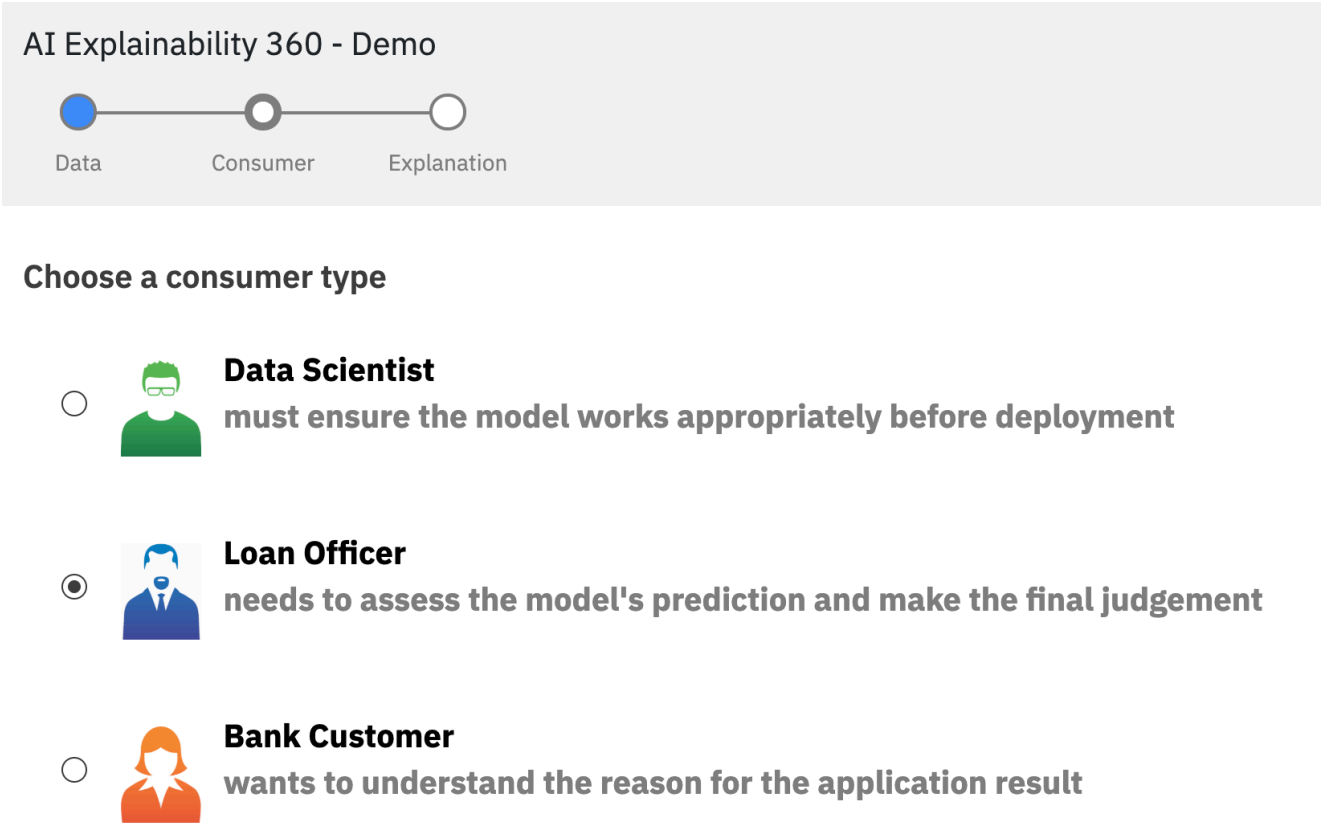


AIX360: Multiple dimensions of explainability



Demo Application: AI Explainability 360 Web Application

http://aix360.mybluemix.net/



We are also making these capabilities around Trusted AI available to businesses through

Watson OpenScale

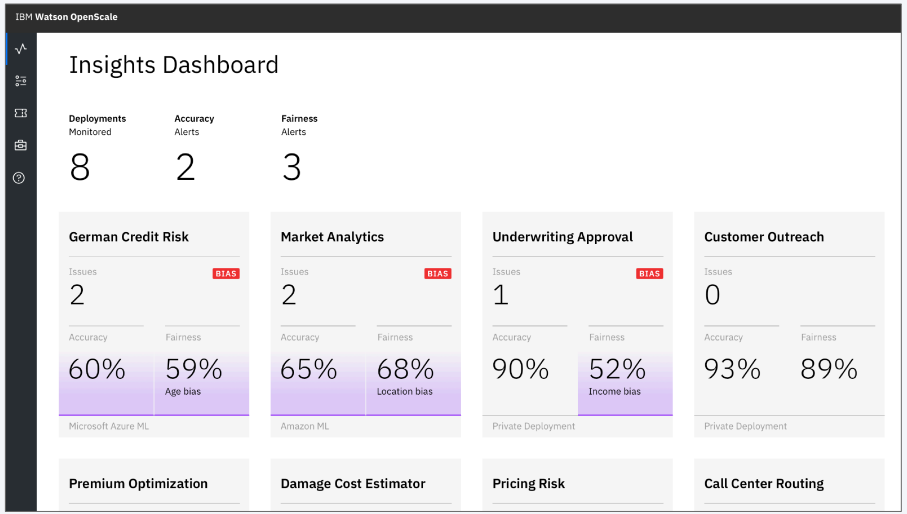
Watson OpenScale **tracks and measures trusted AI outcomes across its lifecycle**, and adapts and governs AI to changing business situations — for models built and running anywhere.

Measure and track AI outcomes

Track performance of production AI and its impact on business goals, with actionable metrics in a single console.

Govern, detect bias and explain AI

Maintain **regulatory compliance** by **tracing and explaining AI decisions** across workflows, and intelligently **detect and correct bias** to improve outcomes.



We would like to partner with community to build Trusted and Transparent AI

To collaborate, look at the corresponding projects here

codait.org

or

https://github.com/topics/trusted-ai

and reach out via github or send an email to

singhan@us.ibm.com

CODAIT

Center for Open Source Data and AI Technologies