Ethics in AI:

Detecting and Remediating Bias in AI

by Creating Ethical AI Practices

Romeo Kienzler, Animesh Sign

IBM Center for Open Source Data and AI Technologies CODAIT

The .. singularity .. is a hypothetical point in the future when technological growth becomes uncontrollable and irreversible, resulting in unfathomable changes to human civilization.

source: wikipedia

…resulting in **unfathomable** changes to human civilization.

…resulting in **unfavourable** changes to human civilization.

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?

Did anyone tamper with it?

Robustness…

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

https://hackernoon.com/dogs-wolves-data-science-and-why-machines-must-learn-like-humans-do-41c43bc7f982

Adversarial Examples

- Perturb model inputs with crafted noise
- Model fails to recognize input correctly
- Attack undetectable by humans
- Random noise does not work.

Attack noise hides pedestrians from the detection system.

Metzen et al., Universal Adversarial Perturbations against Semantic Image Segmentation.
https://arxiv.org/abs/1704.05712

a)

b)

c)

d)

e)

f)

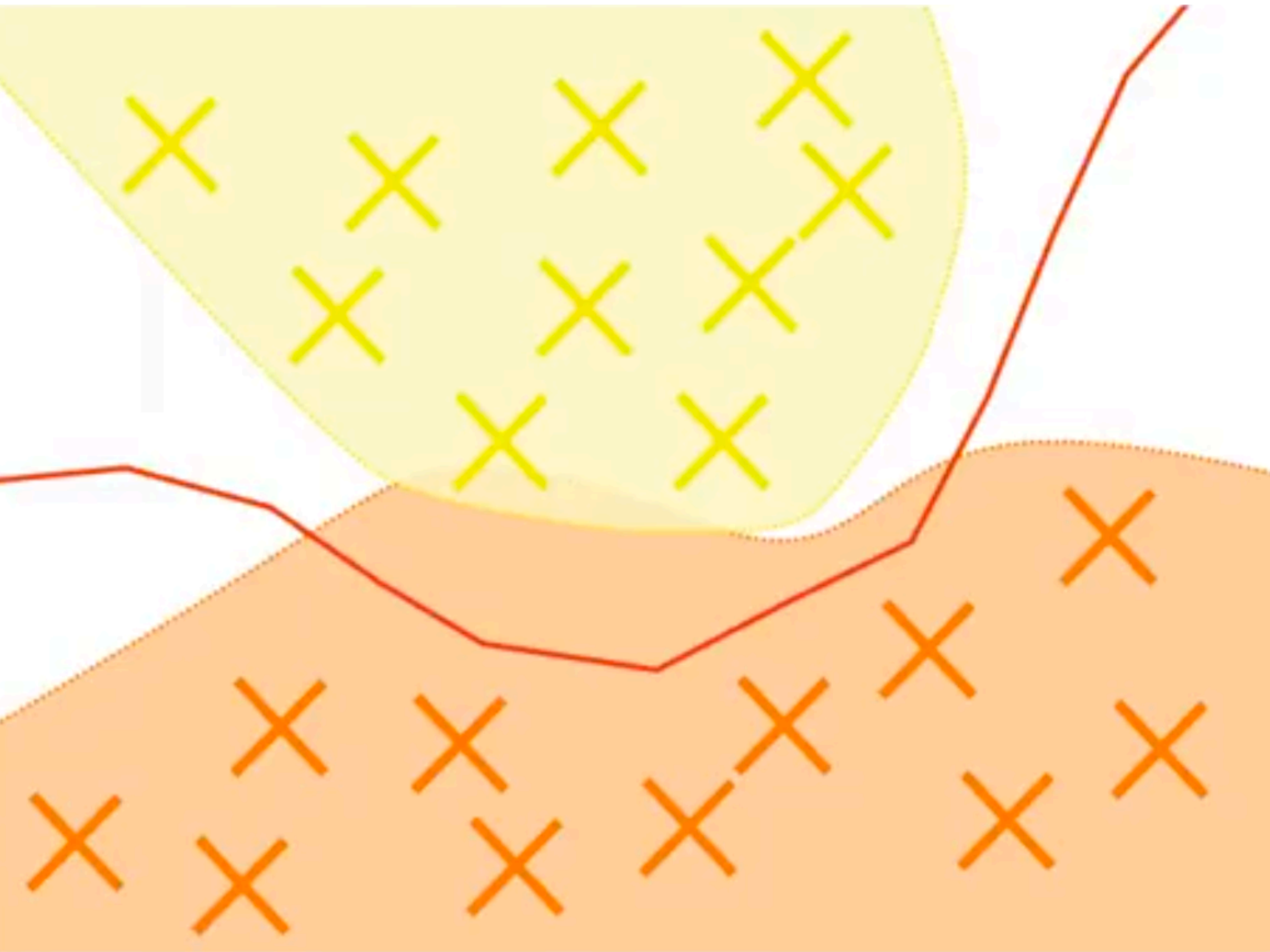
Figure 8: A sample from the adversarial signs that were tested on the test field. Each sign has its own adversarial target $\tilde{y} =$: a) 120 km/h, b) 60 km/h, c) 50 km/h, d) 30 km/h, e) 60 km/h, f) 80 km/h

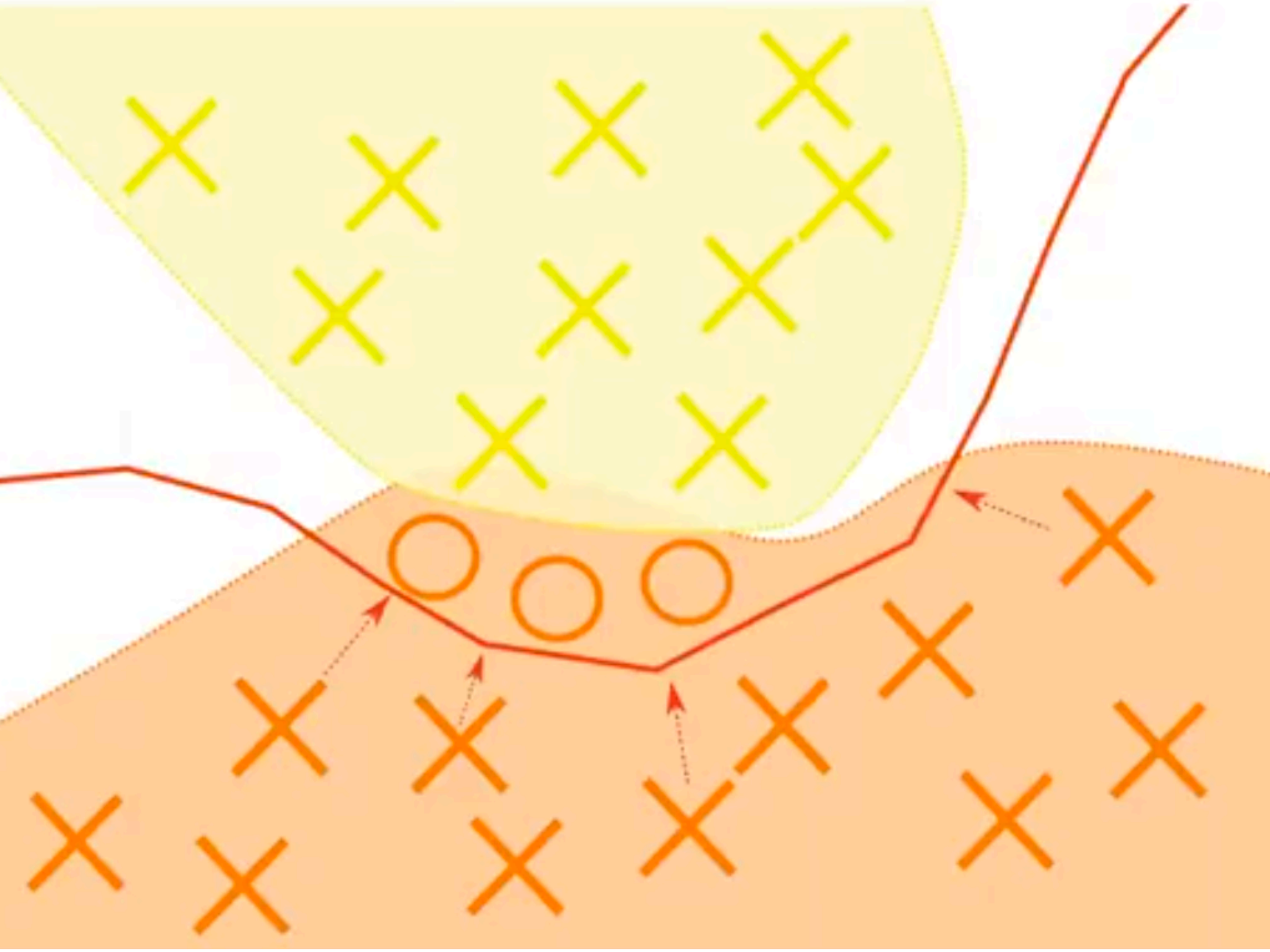
Source: https://arxiv.org/abs/1907.00374

Fooling a Real Car with Adversarial Traffic Signs

Nir Morgulis, Alexander Kreines, Shachar Mendelowitz, Yuval Weisglass

(Submitted on 30 Jun 2019)





Adversarial Robustness Toolbox

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

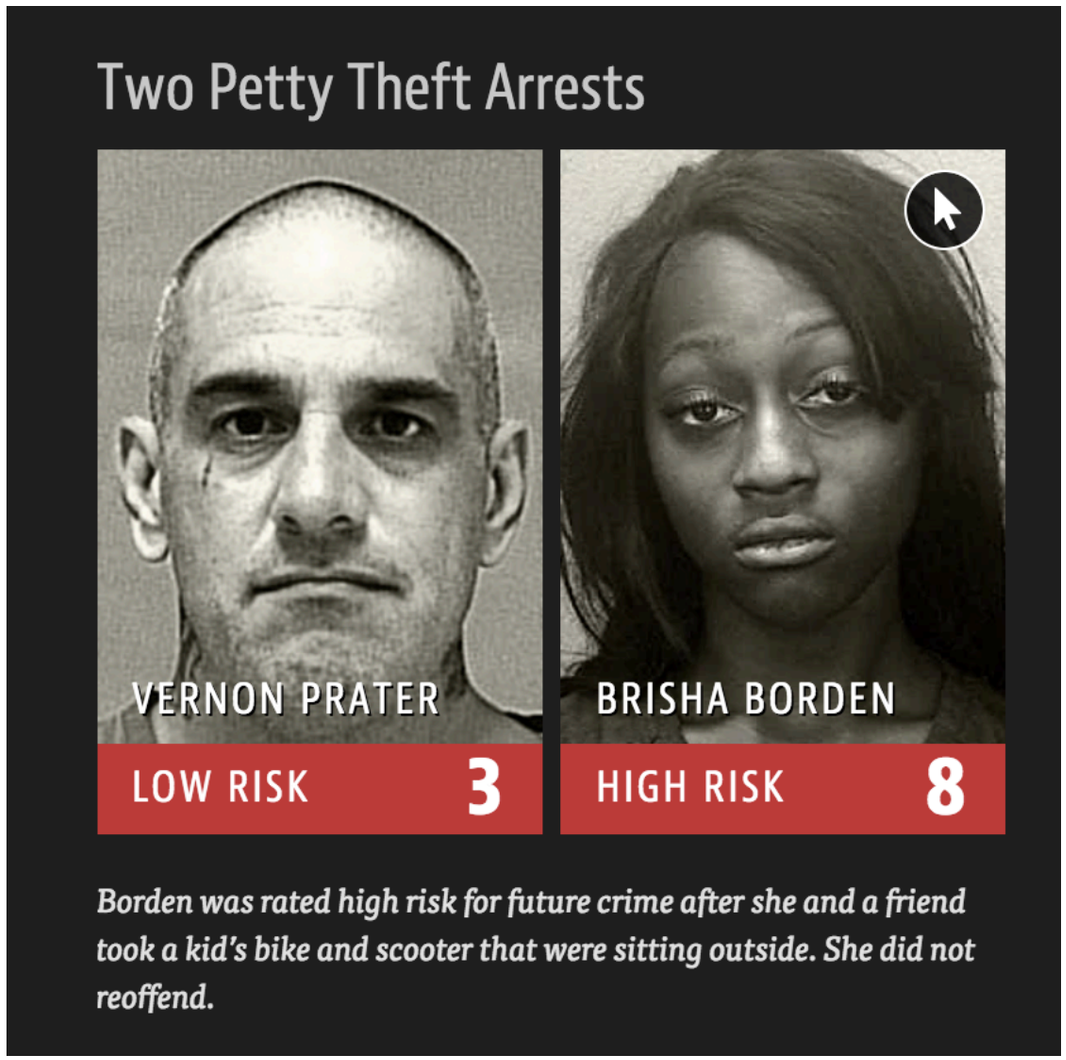
Attacks	Defenses
DeepFool	Feature Squeezing
Fast Gradient Method	Spatial Smoothing
Jacobian Saliency Map	Label Smoothing
NewtonFool	Adversarial Training
Universal Perturbation	Virtual Adversarial Training
C&W Attack	Gaussian Augmentation
Virtual Adversarial Method	

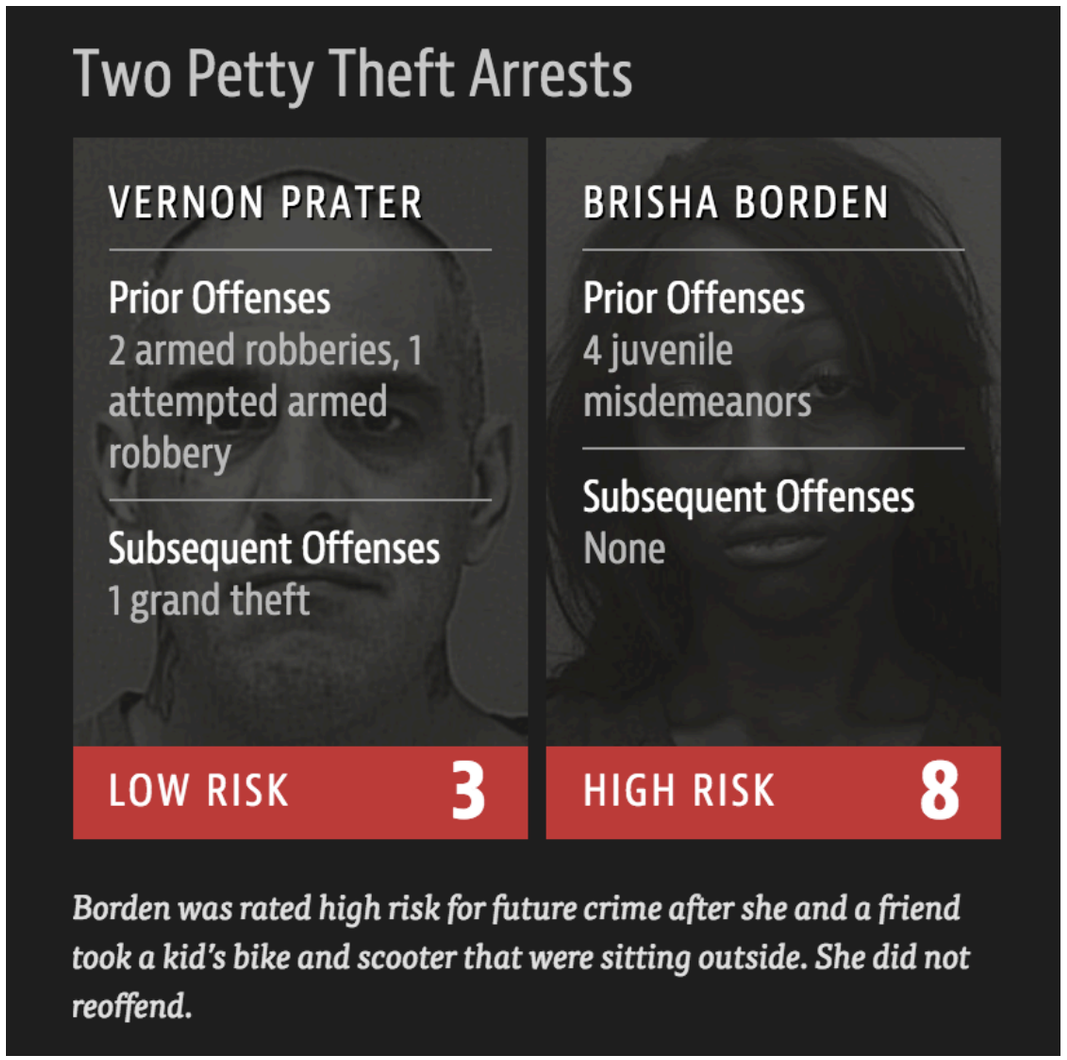
Frameworks	Metrics
TensorFlow	Loss sensitivity
Keras	Empirical robustness
PyTorch (soon)	CLEVER
MXNet (soon)	

Is it fair?

Bias

Northpointe's COMPAS algorithm widely used since 2008 in Broward County, Florida is racially biased





flagging black people 45% vs. white people 24% for risk for future crime

The problem of racist AI is not always a problem of the AI. It is the problem of a racist world (moral-robots.com)

AI Fairness 360 - Demo

Data — Check — Mitigate — Compare

Back Next

2. Check bias metrics

Dataset: German credit scoring
Mitigation: none

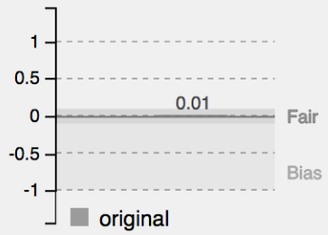
Protected Attribute: Sex

Privileged Group: ***Male***, Unprivileged Group: ***Female***

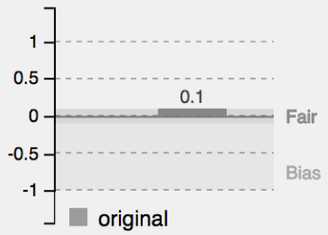
Accuracy with no mitigation applied is 76%

With default thresholds, bias against unprivileged group detected in 0 out of 5 metrics

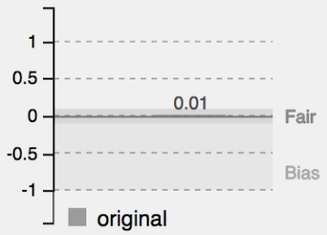
Statistical Parity Difference



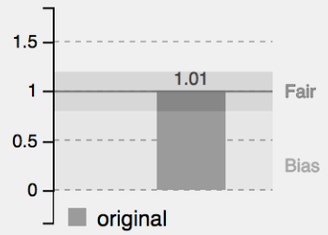
Equal Opportunity Difference



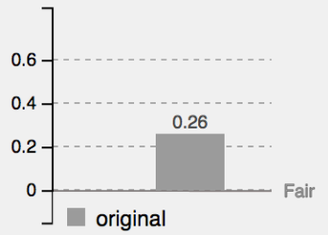
Average Odds Difference

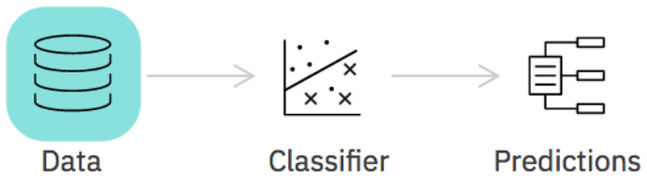


Disparate Impact



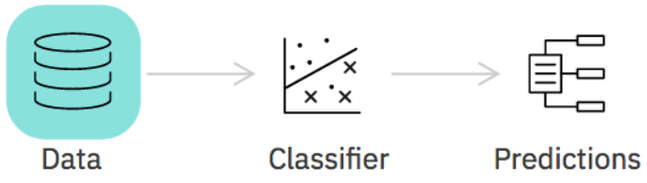
Theil Index





○ **Reweighing**

Weights the examples in each (group, label) combination differently to ensure fairness before classification.

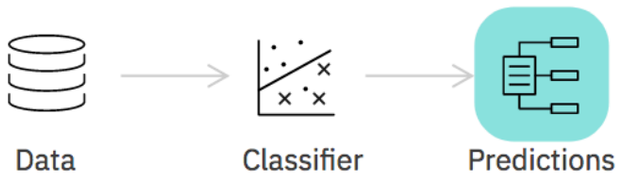


◉ **Adversarial Debiasing**

Learns a classifier that maximizes prediction accuracy and simultaneously reduces 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.

○ **Reject Option Based Classification**

Changes predictions from a classifier to make them fairer. Provides favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.



4. Compare original vs. mitigated results

Dataset: German credit scoring

Mitigation: Adversarial Debiasing algorithm applied

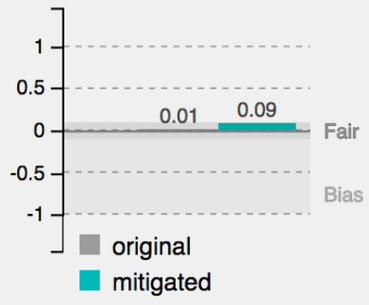
Protected Attribute: Sex

Privileged Group: ***Male***, Unprivileged Group: ***Female***

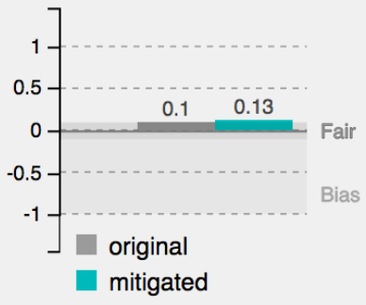
Accuracy after mitigation changed from 76% to 62%

Bias against unprivileged group unchanged after mitigation (0 of 5 metrics indicate bias)

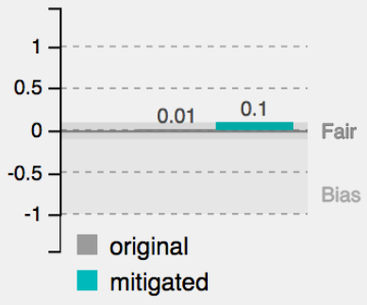
Statistical Parity Difference



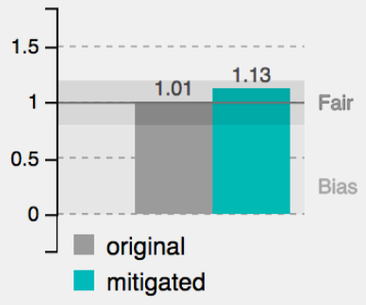
Equal Opportunity Difference



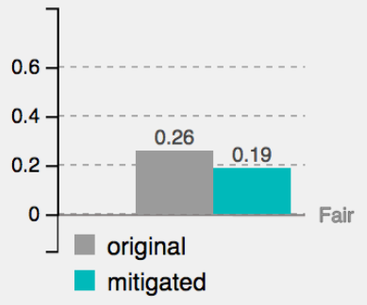
Average Odds Difference

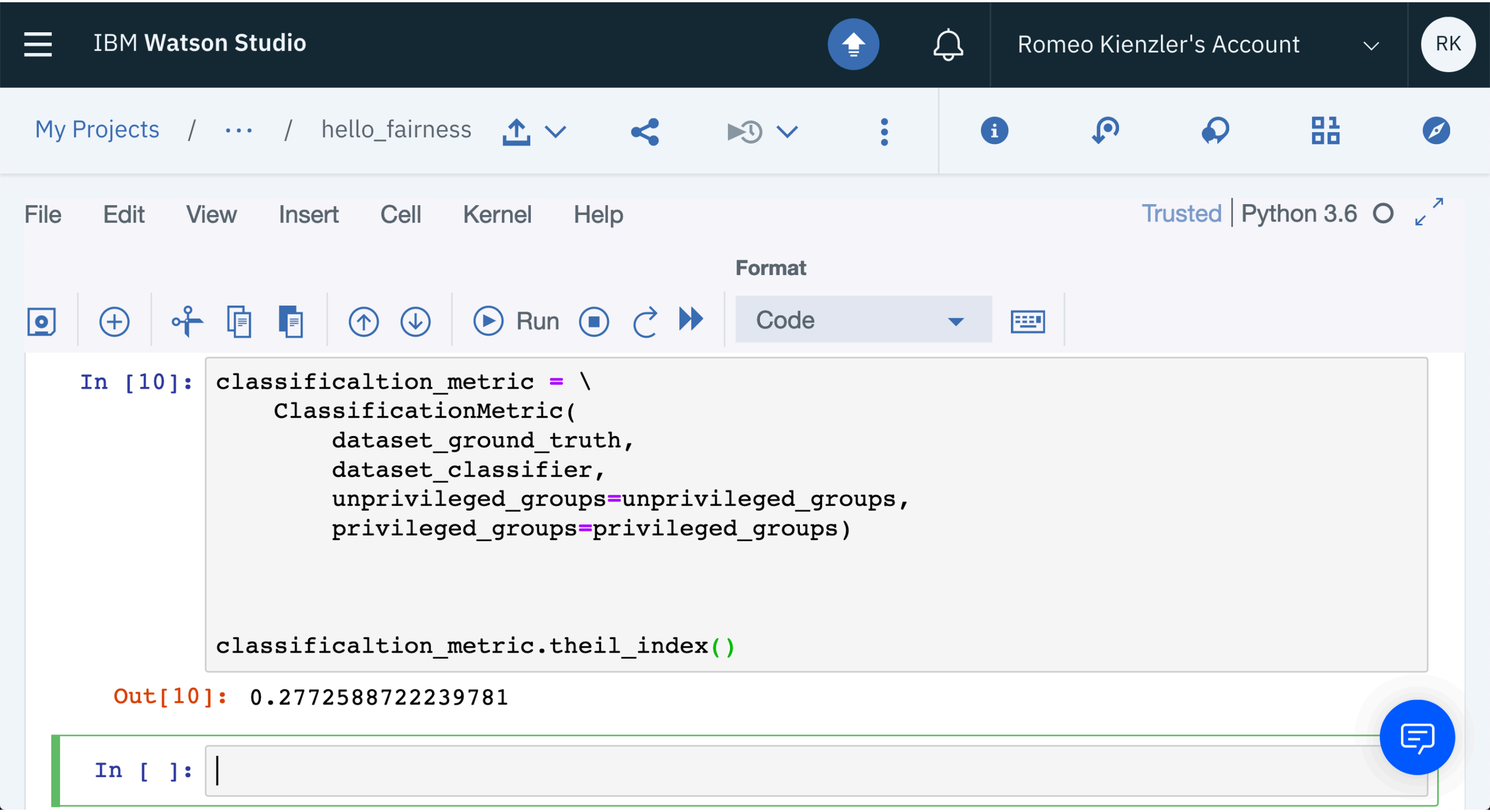


Disparate Impact



Theil Index




IBM Watson Studio
Romeo Kienzler's Account
RK
My Projects / ··· / hello_fairness
File Edit View Insert Cell Kernel Help
Trusted | Python 3.6
Format
Run
Code
In [10]:
classificaltion_metric = \
ClassificationMetric(
dataset_ground_truth,
dataset_classifier,
unprivileged_groups=unprivileged_groups,
privileged_groups=privileged_groups)
classificaltion_metric.theil_index()
Out[10]: 0.2772588722239781
In [ ]:

AI Fairness 360 Toolbox

https://github.com/IBM/AIF360

Is it easy to understand?

Explainability...

Hi can you please tell my why my number 004179 is blocked?

request #38815

Question about WhatsApp for Android

Inbox ×

support@support.whatsapp.com Tue, Jun 25, 9:18 AM

to me

##- WhatsApp Support -##

Hi,

Thanks for your message.

We understand you're currently unable to access WhatsApp and are working diligently to answer your request. We appreciate your patience and will get back to you as soon as possible. For more information, please read this article.

support@support.whatsapp.com Jun 28, 2019, 7:06 PM

to me

##- WhatsApp Support -##

Hi,

Thanks for your message.

Your WhatsApp account has been banned because your activity violated our Terms of Service.

Be aware that we ban accounts if we believe the account activity is in violation of our Terms of Service. Please review the “Acceptable use of our services” section in our Terms of Service carefully to learn more about the appropriate uses of WhatsApp and the activities that violate our Terms of Service.

We might not issue a warning before banning your account. If you think your account was banned by mistake, please respond to this email and we'll look into your case.

Note: WhatsApp reserves the right to modify, suspend or terminate service for any reason without prior notice, at our sole discretion.

WhatsApp Support Team

Sun, Jun 30, 10:39 AM

to support

Hi

I can't find a reason why my number has been banned regarding your terms and services

Please explain

Thanks a lot!

support@support.whatsapp.com Jun 30, 2019, 10:52 AM

to me

##- WhatsApp Support -##

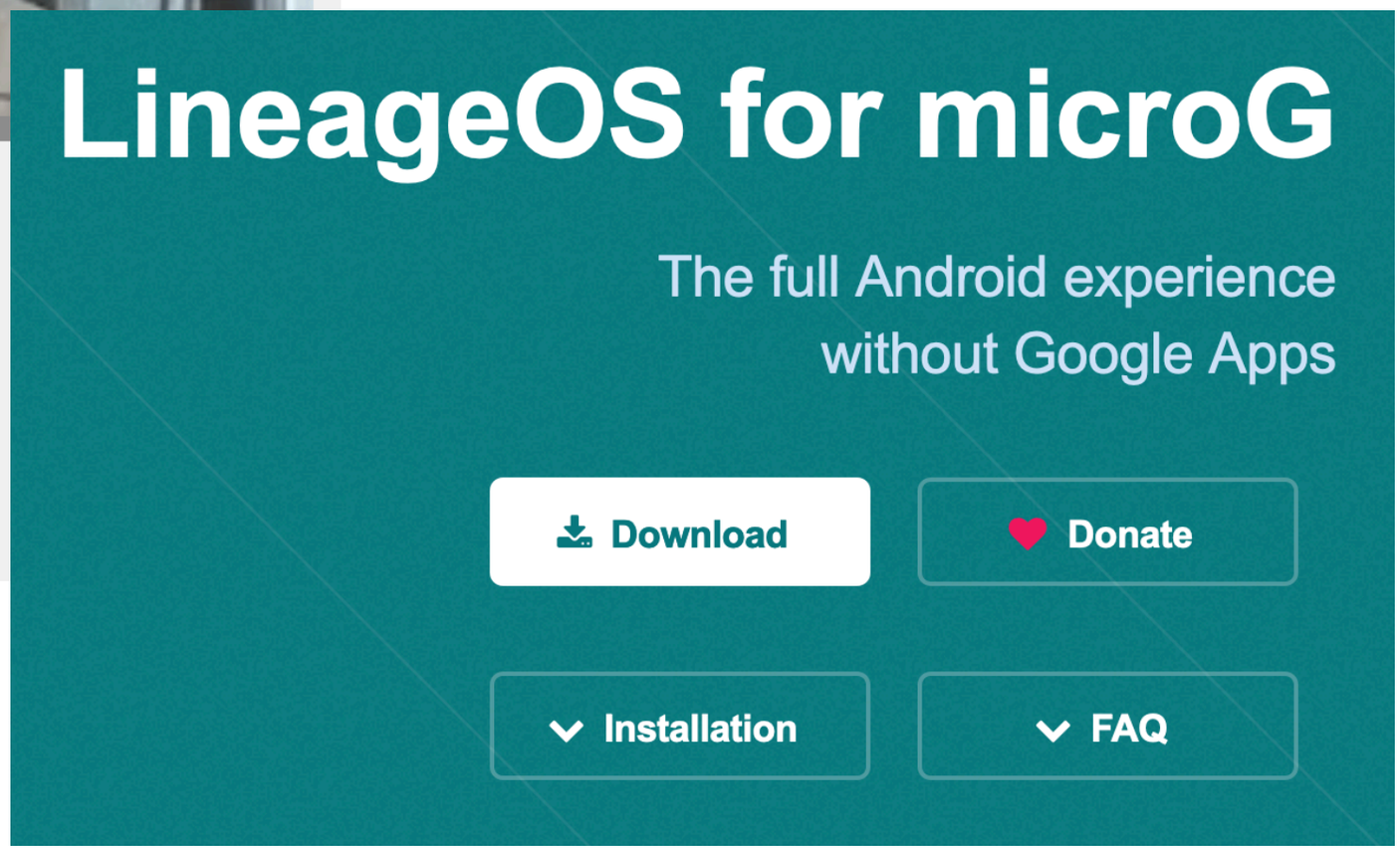
Hi,

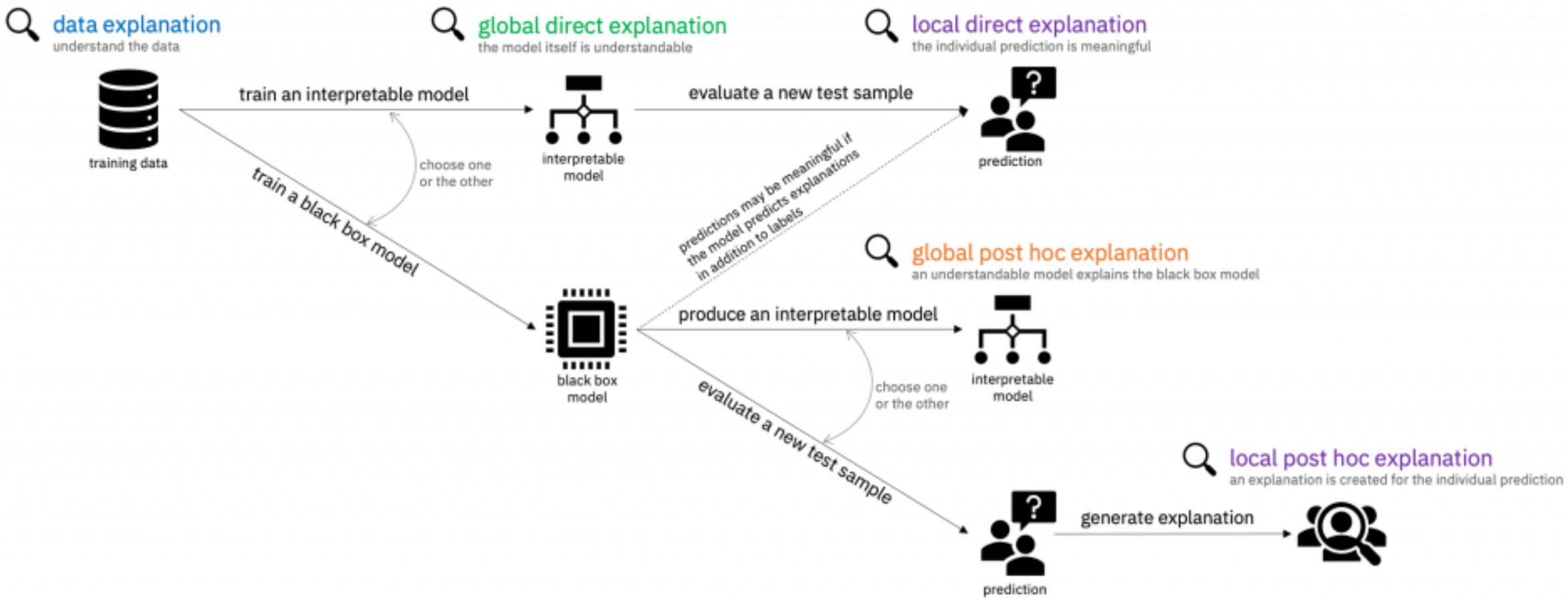
We have reason to believe your account activity has violated our Terms of Service and decided to keep your account banned. We received a large number of complaints about your account and in order to protect our users' privacy, we won't disclose the nature of the complaints.

Responses to this email thread won't be read.

FairPhone2 Adventures: Replacing the internal...

48 views • 6 months ago





 romeo kienzler

AI Explainability 360

https://github.com/IBM/AIX360

 romeo kienzler

Is it accountable?

Data Lineage...

Kubeflow

Pipelines

Experiments

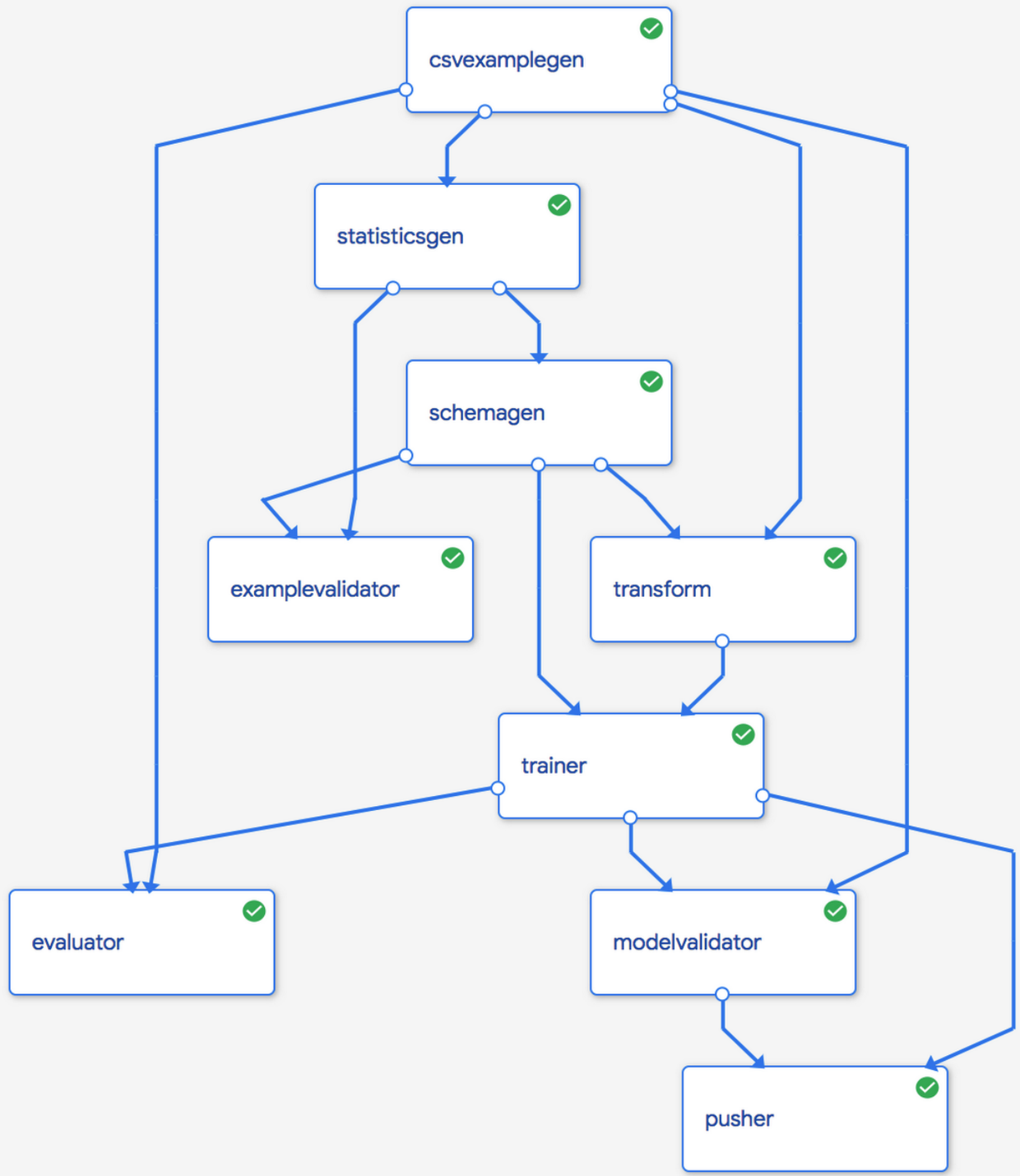
Notebooks

All runs

Chicago Taxi Pipeline

Clone run Refresh

Graph Run output Config



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?

Trusted AI Lifecycle through Open Source

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

Did anyone tamper with it?

ROBUSTNESS

Adversarial Robustness 360

↳ (ART)

github.com/IBM/adversarial-robustness-toolbox

art-demo.mybluemix.net

Is it fair?

FAIRNESS

AI Fairness 360

↳ (AIF360)

github.com/IBM/AIF360

aif360.mybluemix.net

Is it easy to understand?

EXPLAINABILITY

AI Explainability 360

↳ (AIX360)

github.com/IBM/AIX360

aix360.mybluemix.net

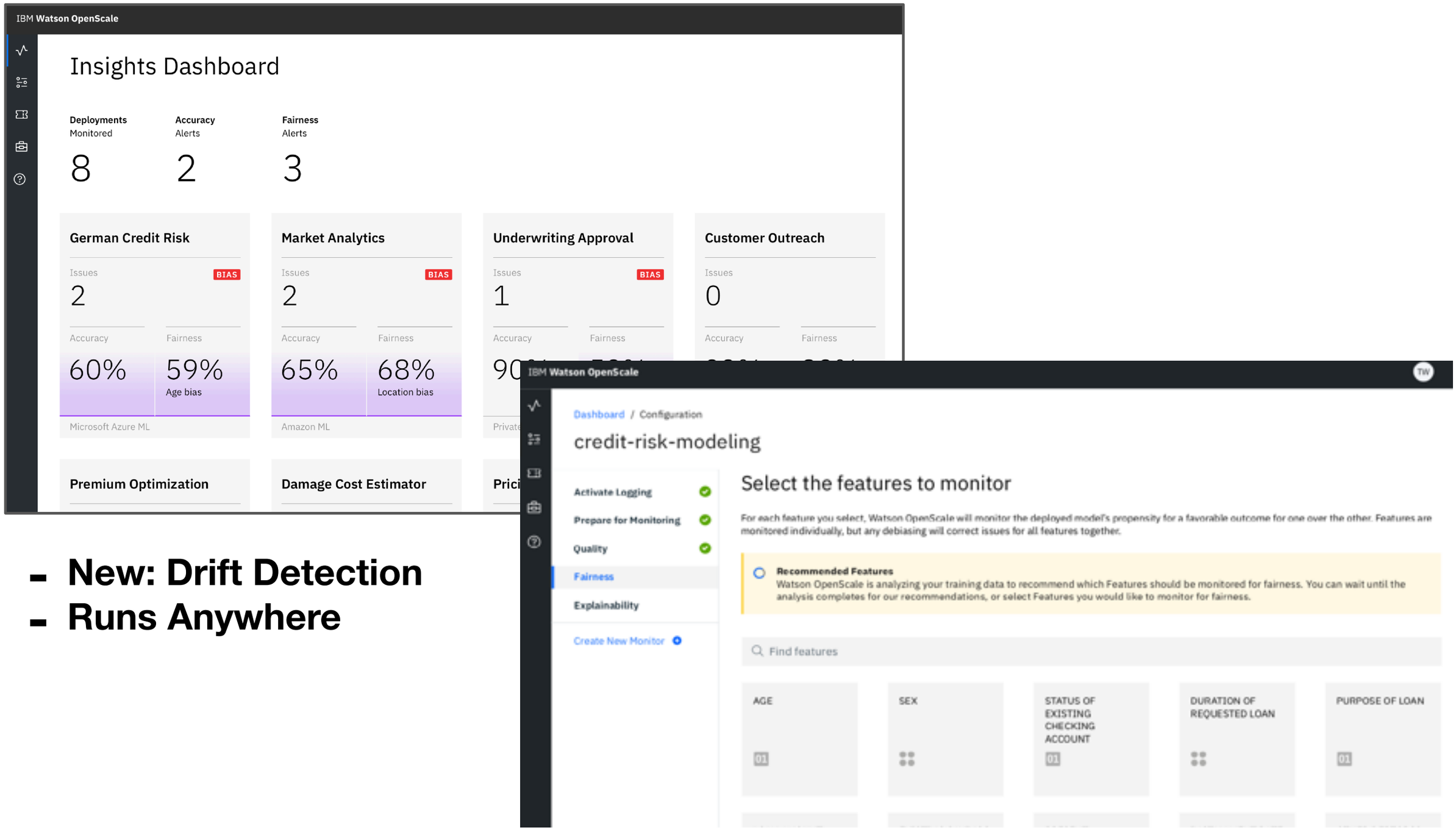
Is it accountable?

LINEAGE

In the works!

romeo kienzler

Watson OpenScale



- **New: Drift Detection**
- **Runs Anywhere**

romeo kienzler

ANNOUNCING

LFAI Trusted AI Committee

↳ bit.ly/trusted-ai

Bring Trust, Transparency and Responsibility into AI

✓ Principles Working Group

✓ Usecases Working Group

Chairs	Region	Company
Animesh Singh	*North America*	*IBM*
Souad Ouali	*Europe*	*Orange*
Jeff Cao	*Asia*	*Tencent*

IBM orange™ AT&T ERICSSON Tech Mahindra Tencent HUAWEI amdocs ETHICAL ML INSTITUTE

LF AI TRUSTED AI

IBM CODAIT

romeo kienzler

“Hello Fairness” script

In [7]:
```
# initialize list of lists
# 0 black
# 1 white
# 0 reject
# 1 accept

df_ground_truth = pd.DataFrame([[0,0], [1,1], [0,1], [1,0], [1,0]], columns = ['protected','label'])
df_classifier = pd.DataFrame([[0,0], [1,1], [0,0], [1,1], [1,0]], columns = ['protected','label'])

privileged_groups=[{'protected': 0}]
unprivileged_groups=[{'protected': 1}]
```

In [8]:
```
dataset_ground_truth = BinaryLabelDataset(
    favorable_label=1,
    unfavorable_label=0,
    df=df_ground_truth,
    label_names=['label'],
    protected_attribute_names=['protected'],
    unprivileged_protected_attributes=unprivileged_groups)
```

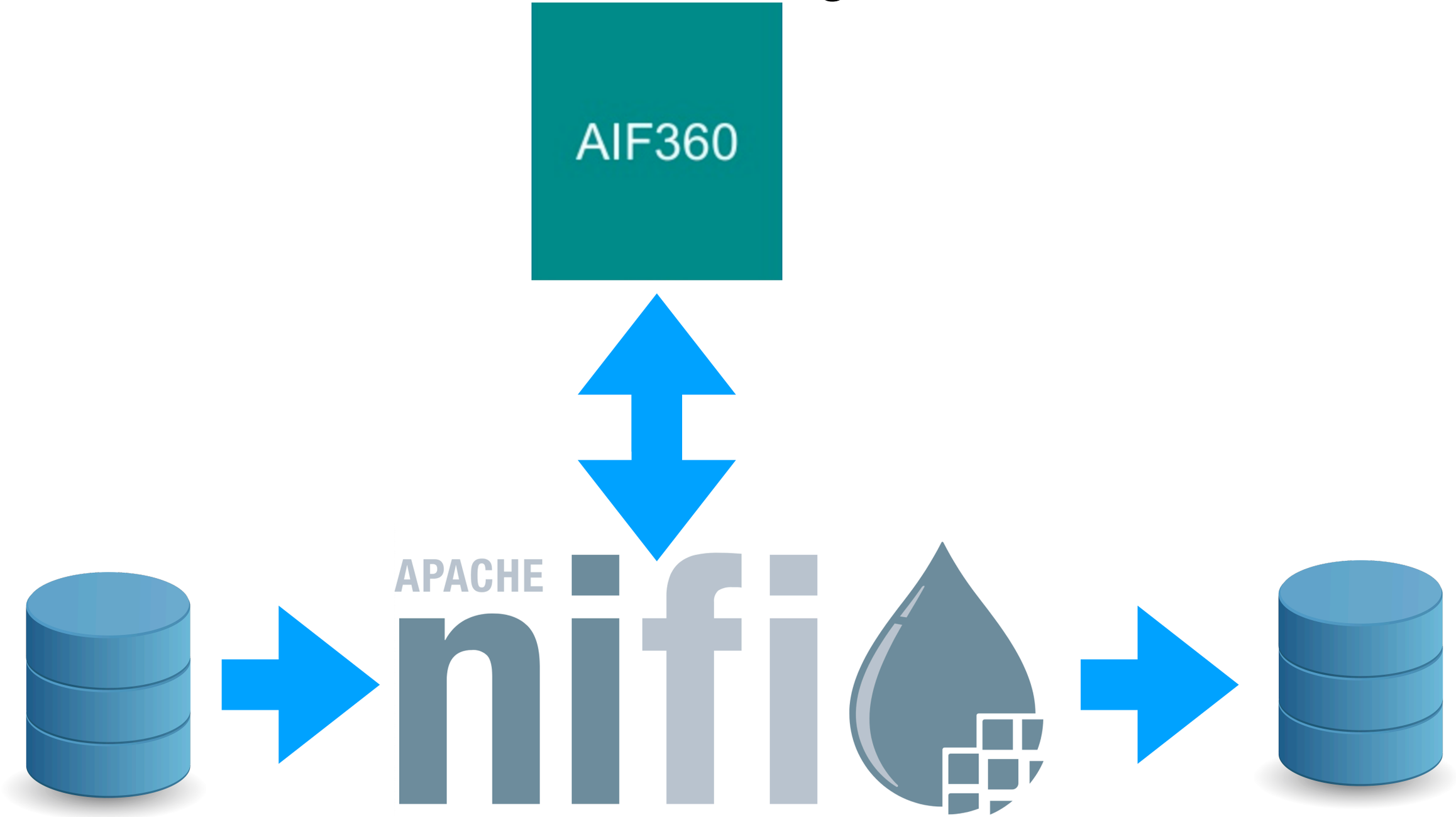
In [9]:
```
dataset_classifier = BinaryLabelDataset(
    favorable_label=1,
    unfavorable_label=0,
    df=df_classifier,
    label_names=['label'],
    protected_attribute_names=['protected'],
    unprivileged_protected_attributes=unprivileged_groups)
```

In [10]:
```
classificaltion_metric = \
    ClassificationMetric(
        dataset_ground_truth,
        dataset_classifier,
        unprivileged_groups=unprivileged_groups,
        privileged_groups=privileged_groups)

TPR = classificaltion_metric.true_positive_rate()
TNR = classificaltion_metric.true_negative_rate()
bal_acc_nodebiasing_test = 0.5*(TPR+TNR)
```

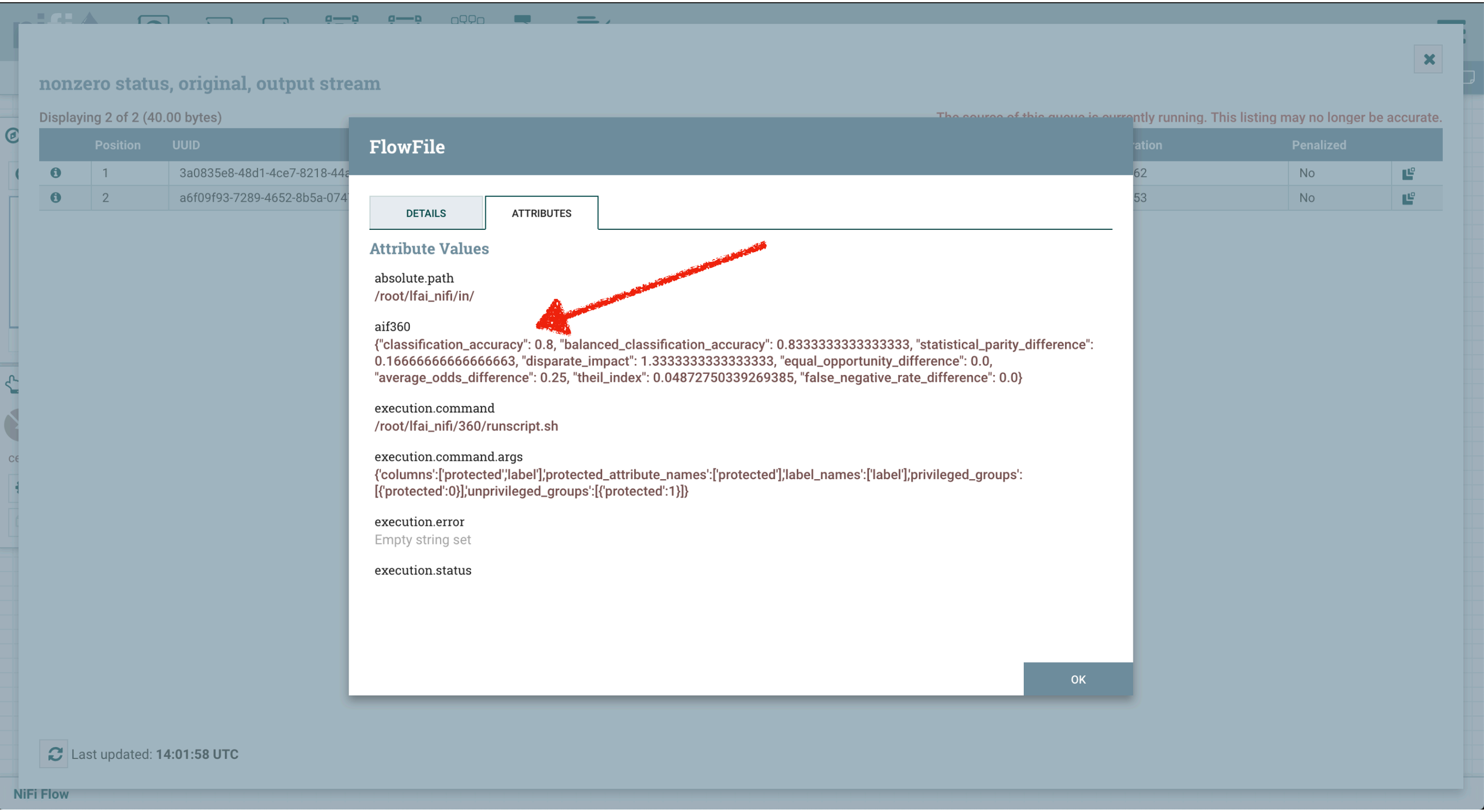
AIF360 Nifi Processor

data driven integration



romeo kienzler

Computed Bias Metrics



romeo kienzler

Wimbledon AI Highlights With Watson™

IBM

All Events | All Rounds | All Statistics | Day 13 Sun 14 J

Type Player Name

COVERAGE		HIGHLIGHTS		TOP 5 MAIN EVENTS HIGHL		
65745 Points	254 Hours	11 Produced	2 Published	2 Published	610 Secs Published	Secs

Shintaro Mochizuki vs Carlos Gimeno Valero

Set 2 : 40-AD : Match Point; Mochizuki wins the point with a backhand winner.

Most Excitement | Most Recent

Set Excitement Threshold

0.00 — 0.3 — 1.00

0.95 ● Sunday, 14 July 2019, 15:09

Shintaro Mochizuki vs Carlos Gimeno Valero

Set 2 : 40-AD : Match Point; Mochizuki wins the point with a backhand winner.

0.89 ● Sunday, 14 July 2019, 15:05

Shintaro Mochizuki vs Carlos Gimeno Valero

Set 2 : AD-40 : Mochizuki wins the point with a backhand winner.

IBM CODAIT

romeo kienzler

thank you