Trusted AI

Building Reproducible, Unbiased and Robust AI Pipelines using the python OpenSource Stack

Romeo Kienzler, Animesh Sign
IBM Center for Open Source Data and AI Technologies CODAIT

Center for Open-Source Data & AI Technologies

Improving the Enterprise AI Lifecycle in Open Source

- **#2 contributor to KubeFlow**
- **#4 contributor to TensorFlow**
    - **Google, DeepMind & nVidia**
- **#2 contributor to Apache Spark (50,000 lines of code)**
- **Keras**
- **#1 contributor to Apache SystemML (65,000 lines of code)**
- **Apache Arrow**
- **Apache Bahir**
- **Apache Toree**
- **Apache Zeppelin**
- **Apache Livy**
- **Fabric for DeepLearning (FfDL)**

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

… **intelligence explosion**, an upgradable intelligent agent **..** would enter a "runaway reaction" of **self-improvement cycles**, **..** surpass all human intelligence**.**

source: wikipedia

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

request #38815

…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


STOP
© Washington University


STOP
HATE
© Washington University


SPEED
LIMIT
45

a)

b)

c)

d)

e)

f)

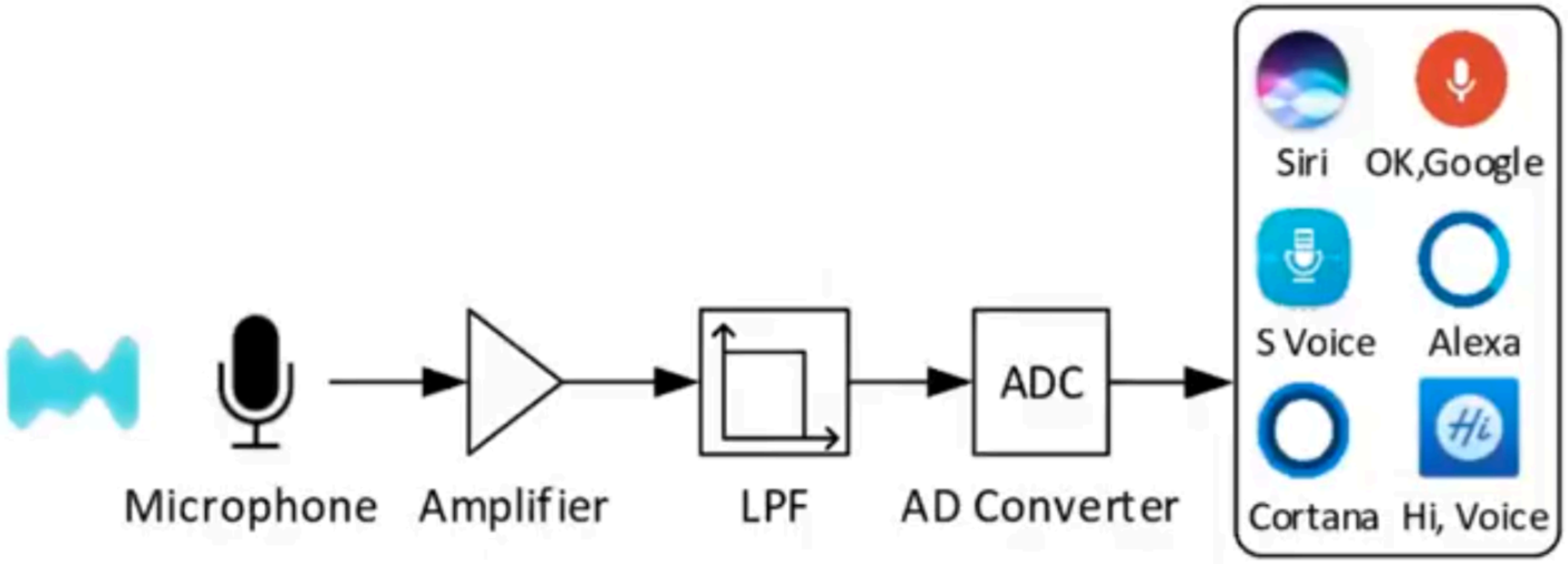
Figure 8: A sample from the adversarial signs that were tested on the test field. Each sign has its own adversarial target $\tilde{y} =$: a) 120 km/h, b) 60 km/h, c) 50 km/h, d) 30 km/h, e) 60 km/h, f) 80 km/h

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

Fooling a Real Car with Adversarial Traffic Signs

Nir Morgulis, Alexander Kreines, Shachar Mendelowitz, Yuval Weisglass

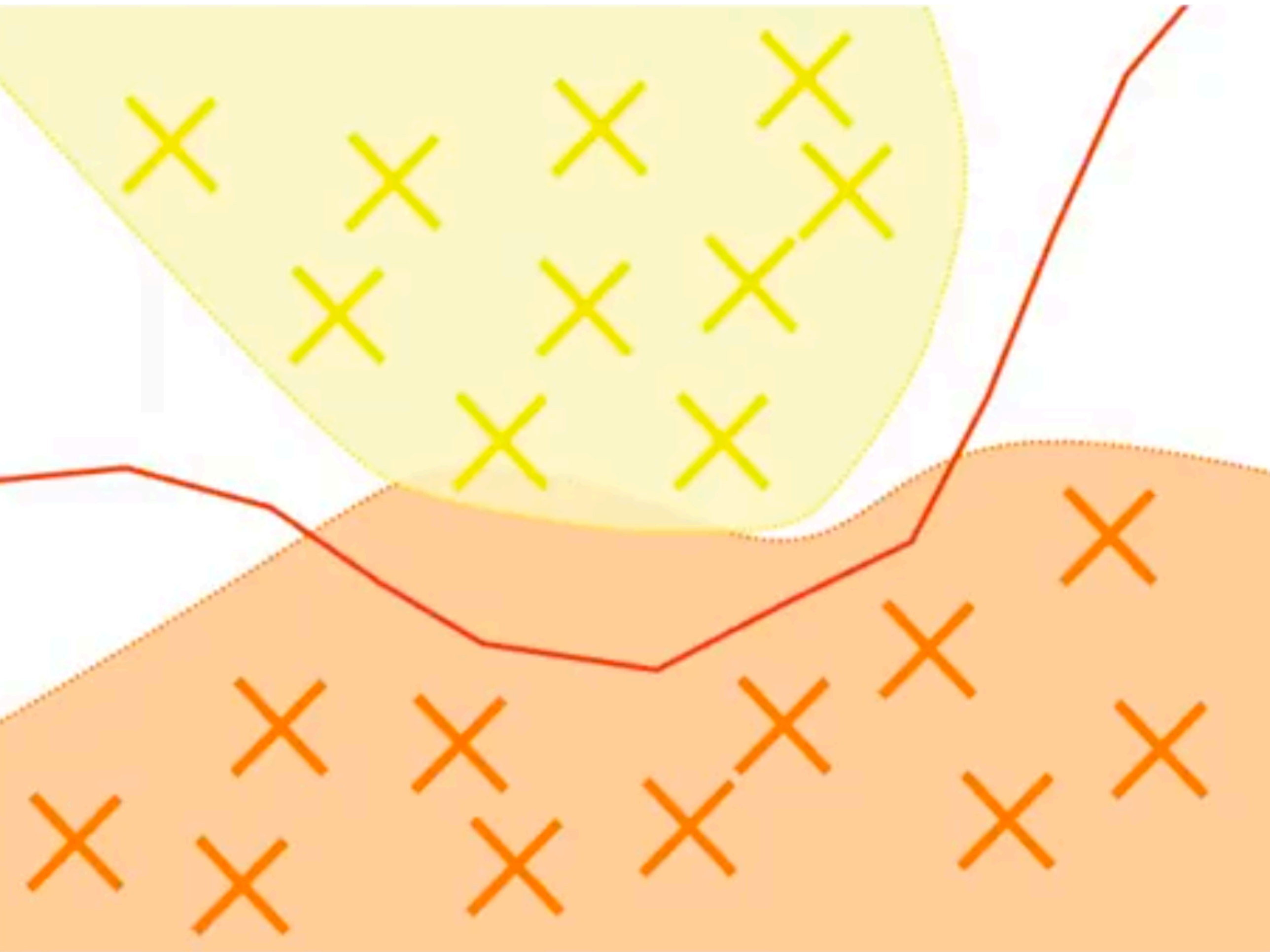
(Submitted on 30 Jun 2019)

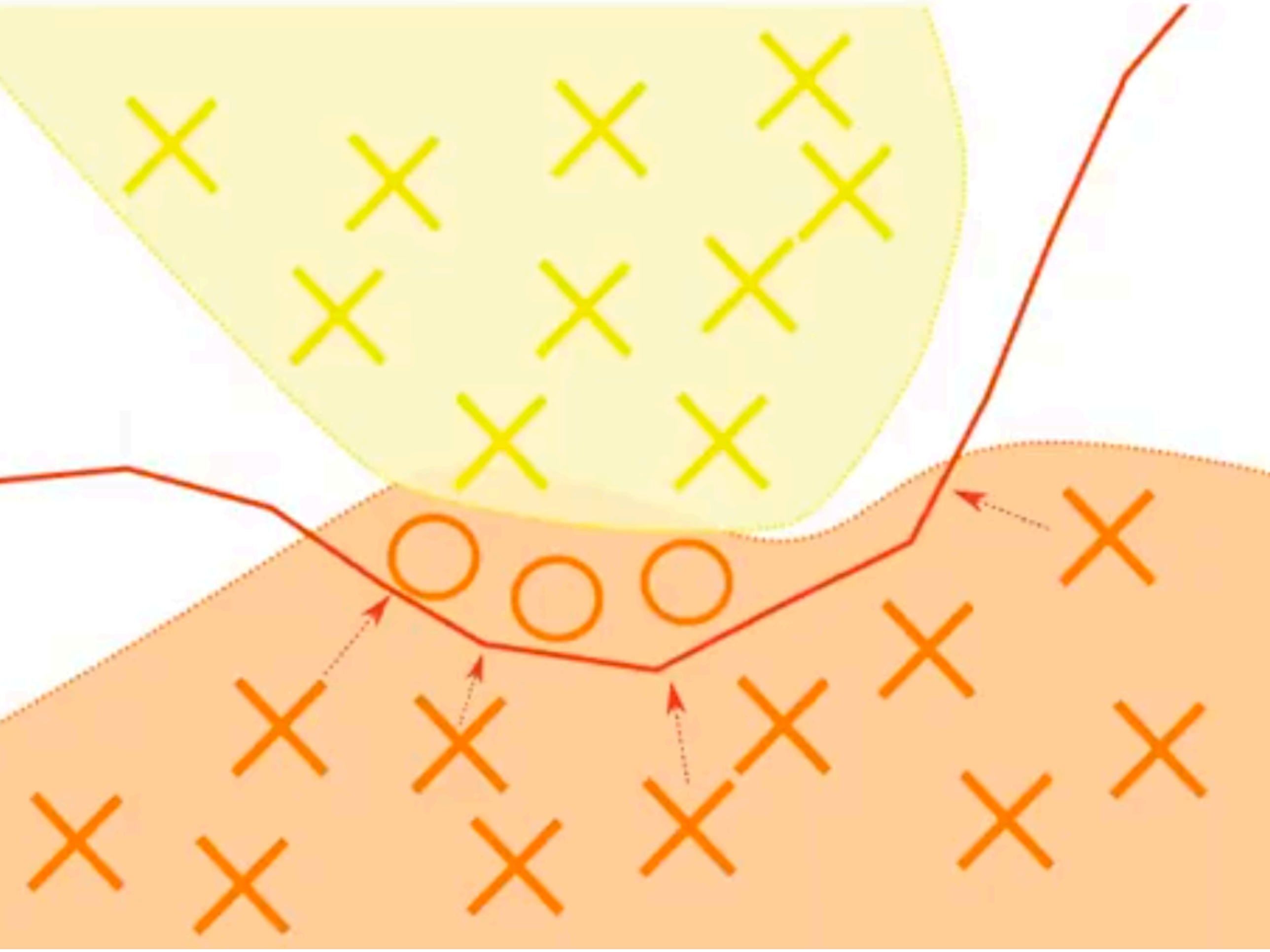


Okay Google, text John!

- Stealthy voice commands recognized by devices
- Humans cannot detect it.

Zang et al., Dolphin Attack: Inaudible Voice Commands, ACM C 2017





Adversarial Robustness Toolbox

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

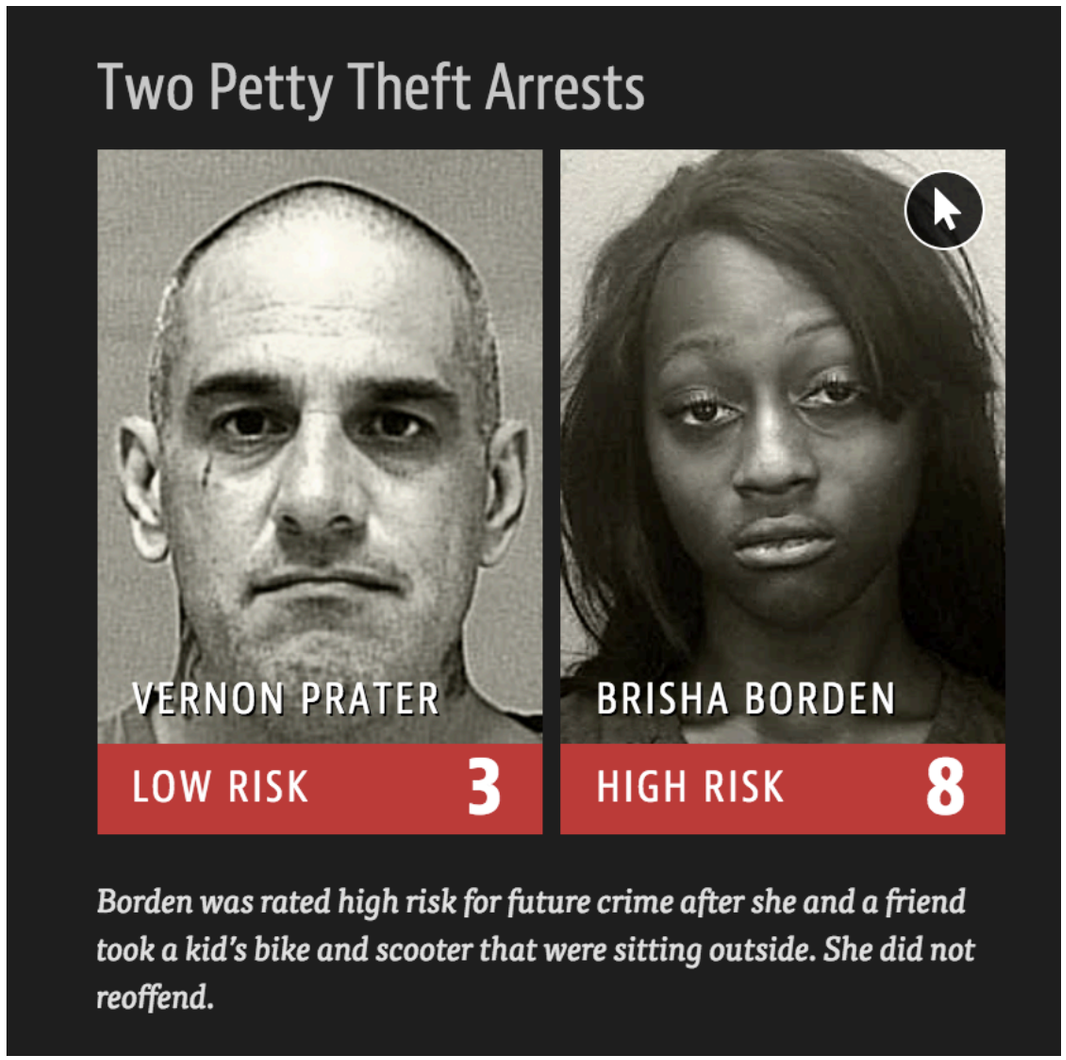
Attacks	Defenses
DeepFool	Feature Squeezing
Fast Gradient Method	Spatial Smoothing
Jacobian Saliency Map	Label Smoothing
NewtonFool	Adversarial Training
Universal Perturbation	Virtual Adversarial Training
C&W Attack	Gaussian Augmentation
Virtual Adversarial Method	

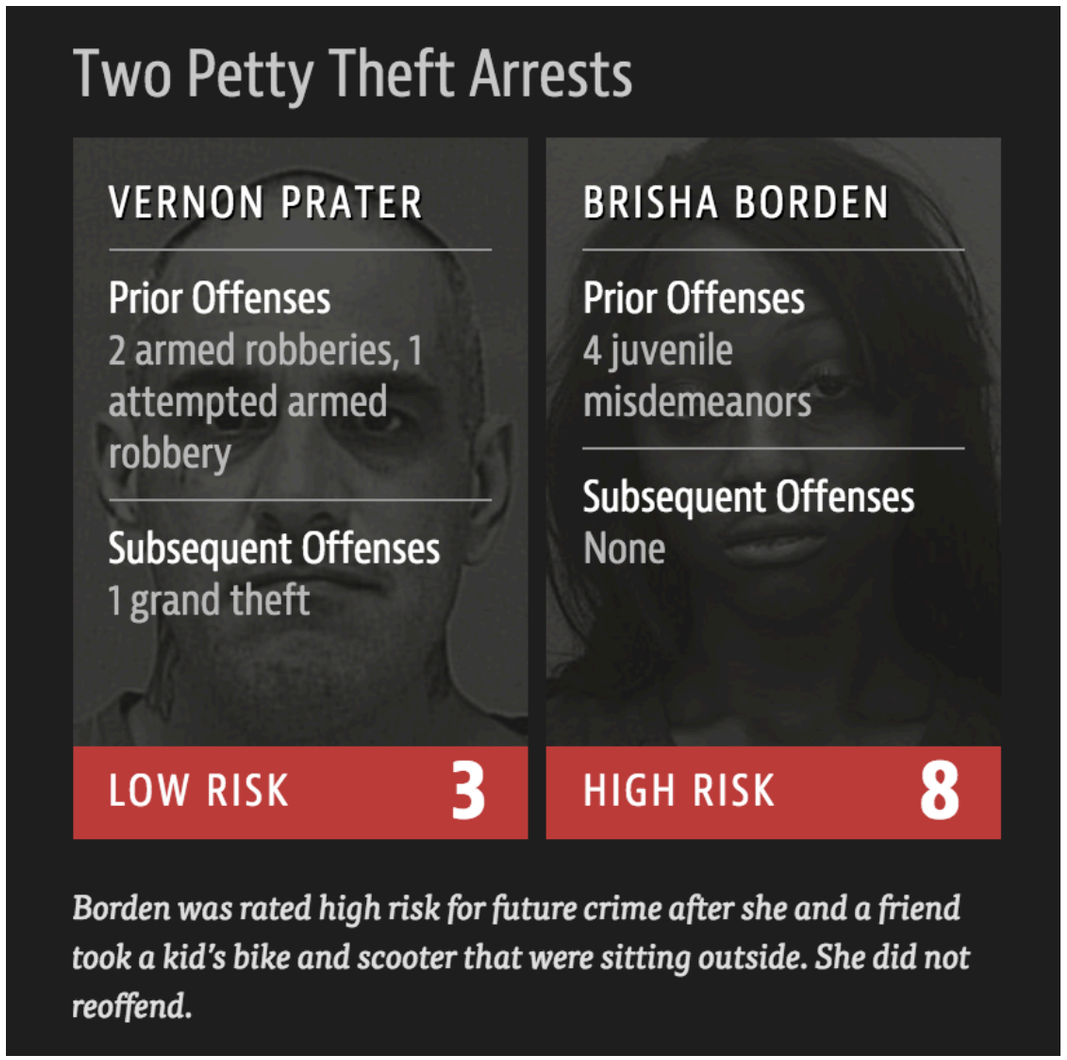
Frameworks	Metrics
TensorFlow	Loss sensitivity
Keras	Empirical robustness
PyTorch (soon)	CLEVER
MXNet (soon)	

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

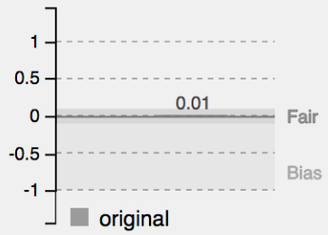
Protected Attribute: Sex

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

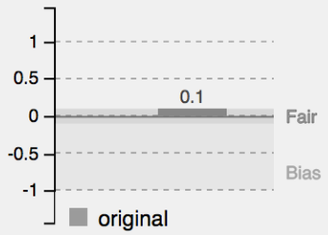
Accuracy with no mitigation applied is 76%

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

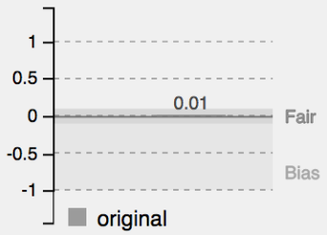
Statistical Parity Difference



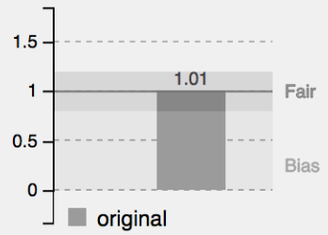
Equal Opportunity Difference



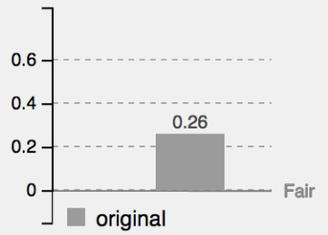
Average Odds Difference

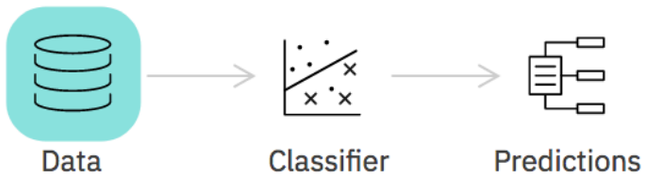


Disparate Impact



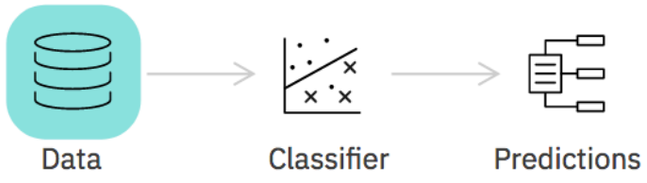
Theil Index





○ **Reweighing**

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

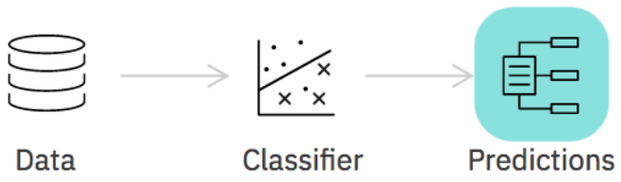


◉ **Adversarial Debiasing**

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

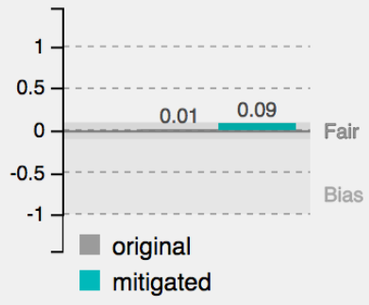
Protected Attribute: Sex

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

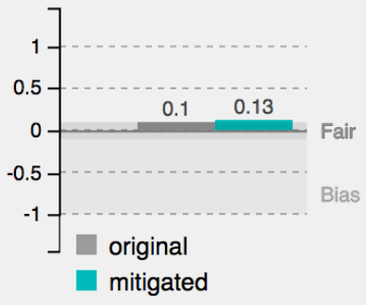
Accuracy after mitigation changed from 76% to 62%

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

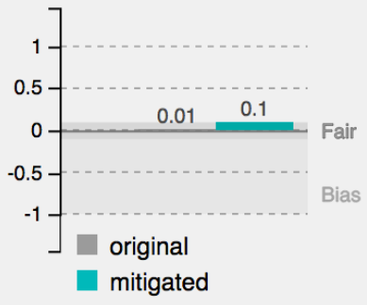
Statistical Parity Difference



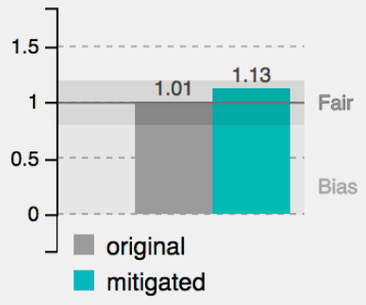
Equal Opportunity Difference



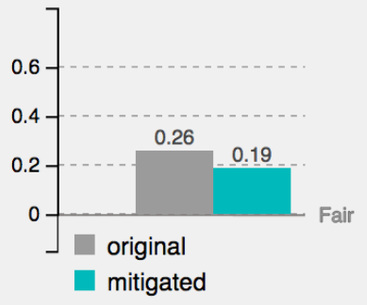
Average Odds Difference

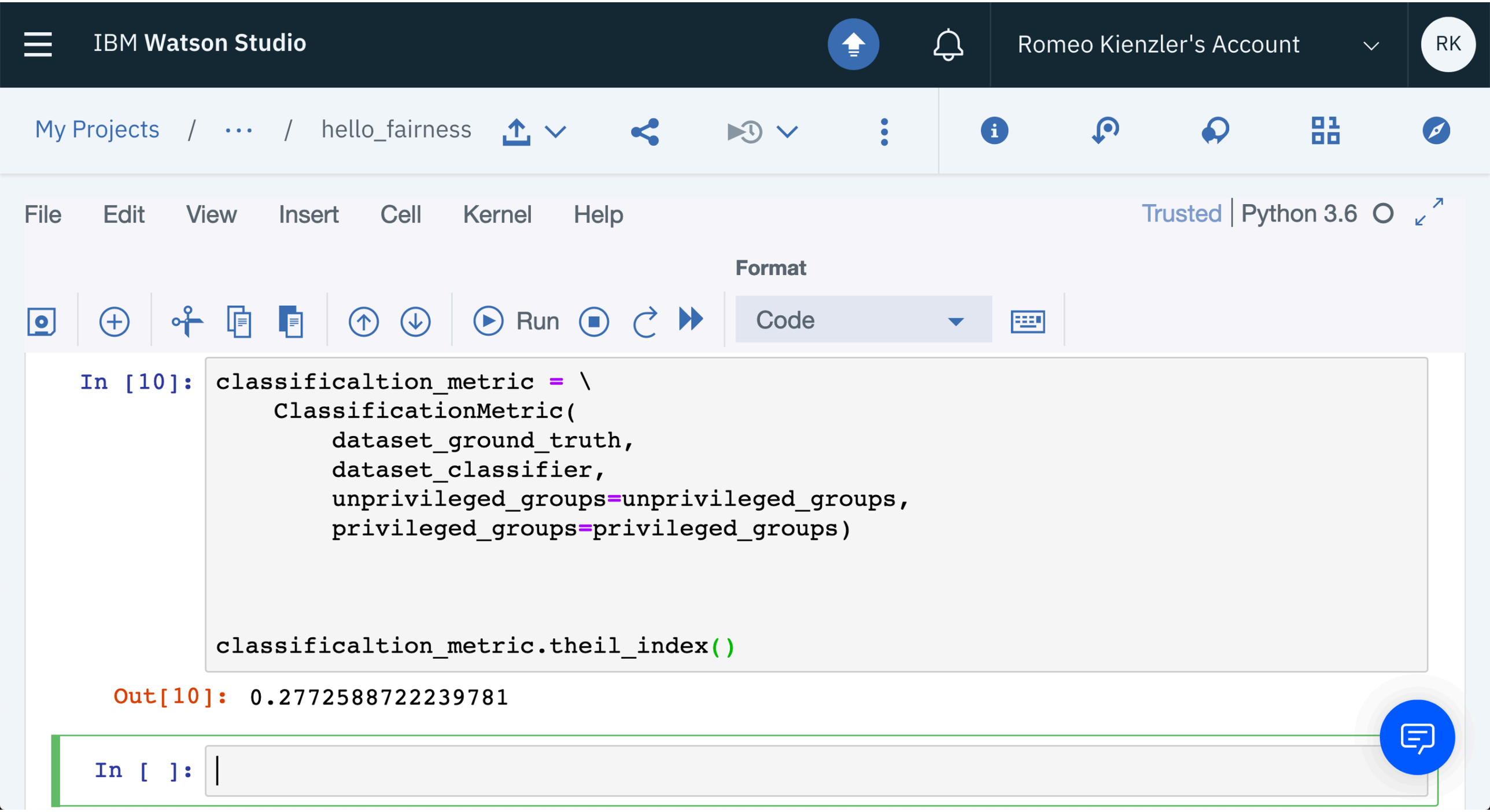


Disparate Impact



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 -##

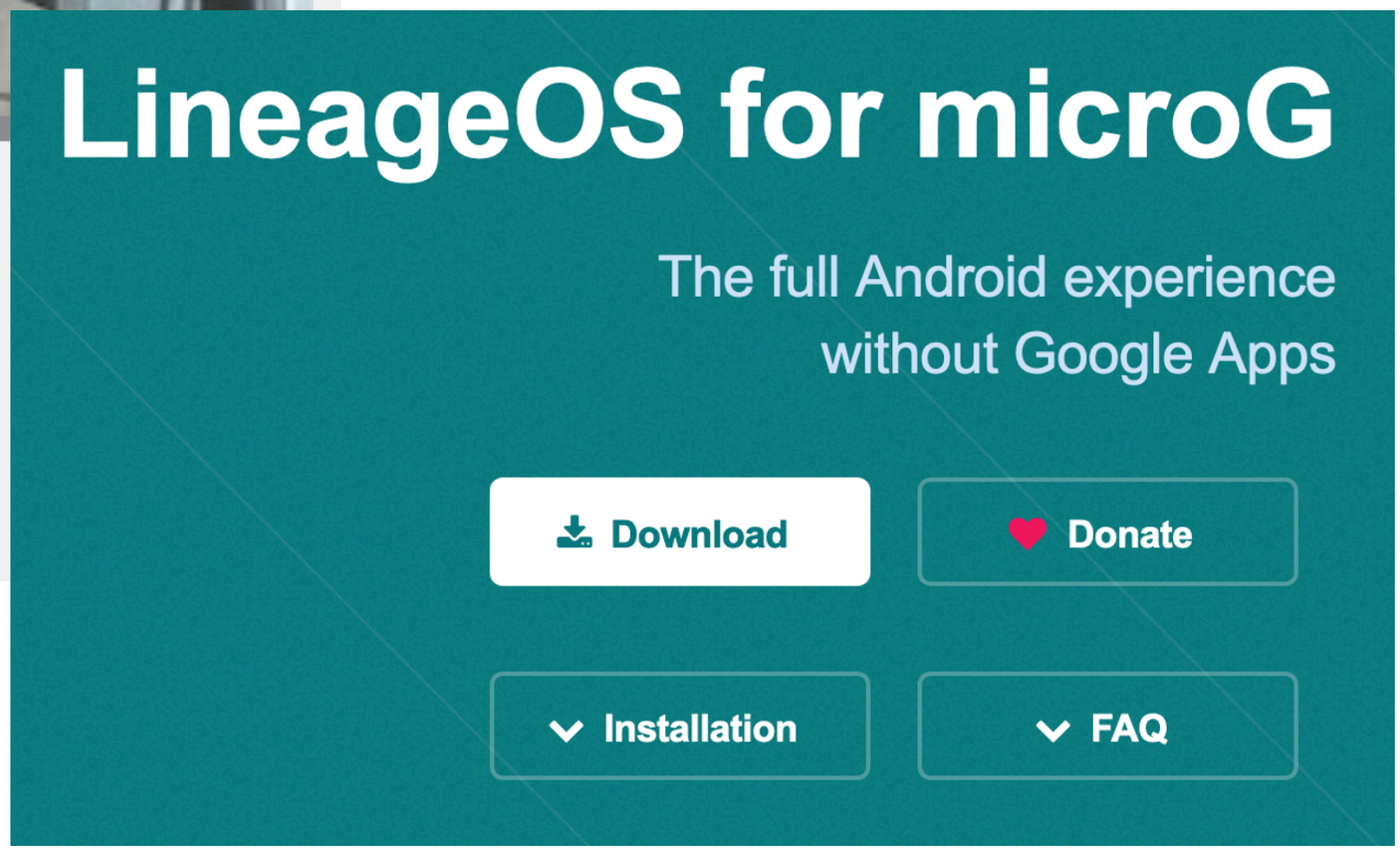
Hi,

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



Signal

SUPPORT BLOG DEVELOPERS CAREERS DONATE EN

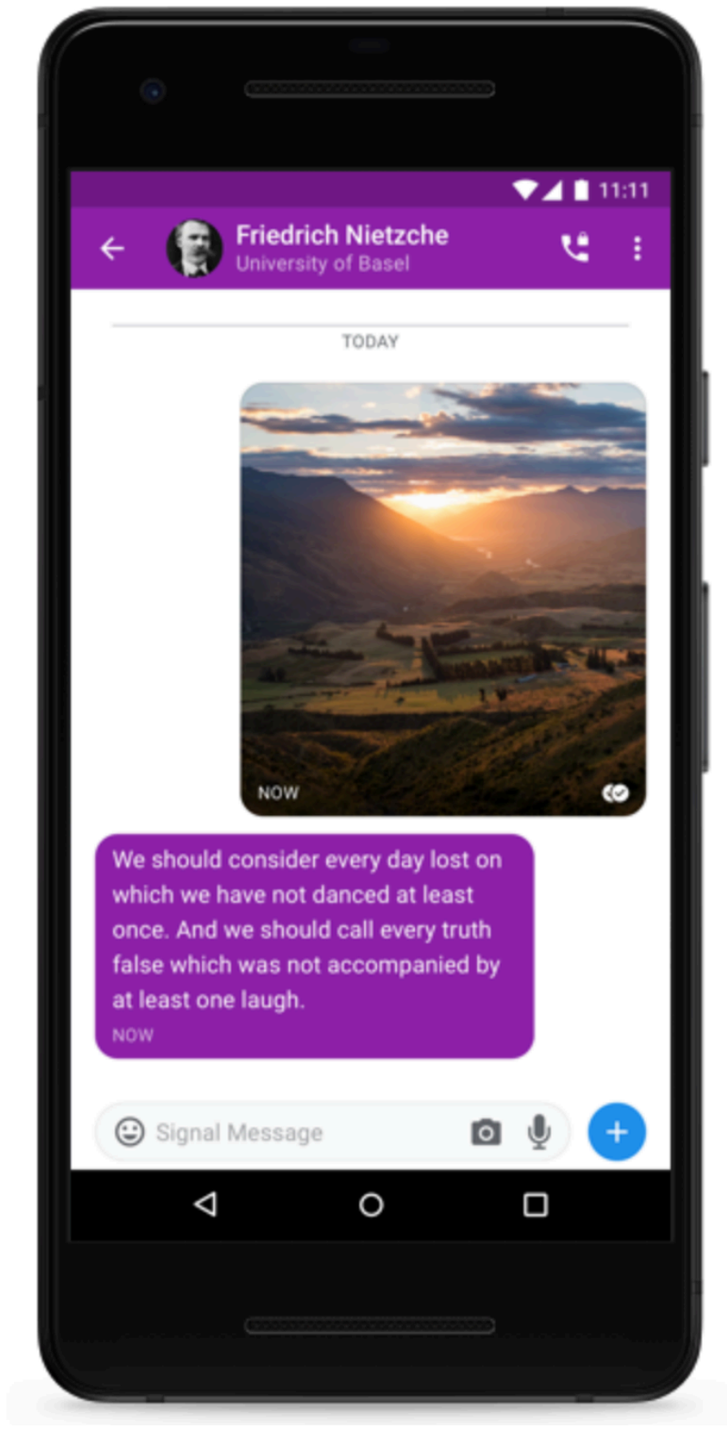
Fast, simple, secure.

Privacy that fits in your pocket.

Android

iPhone

Desktop



RIOT.IM

Liberate your communication

Communicate the way you want with Riot - a universal secure chat app entirely under your control.

Get started

LAUNCH NOW ON YOUR BROWSER

DOWNLOAD APP FOR DESKTOP

Download on the App Store

GET IT ON Google Play

GET IT ON F-Droid

Riot is for everyone, from casual chat to high powered collaboration

IBM CODAIT

 romeo kienzler

Keybase

Login Install

The Big Stellar Space Drop

2 Billion Lumens for Everyone

Just about $122 million USD

98,595 registered so far. The next round is **Nov 15, 2019**.

You'll get **1,014** Lumens (XLM) each month if no one else registers.

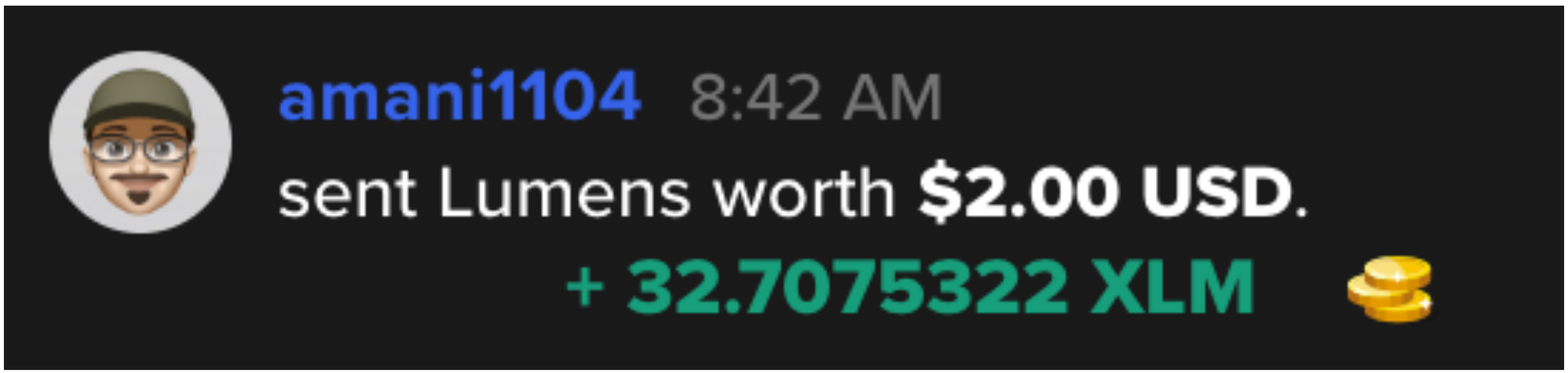
September 9, 2019

UPDATE #3 October 8: 💖 Good news, everyone! The Space Drop is back on, and the requirements to register are now looser.

At Keybase, we've **blogged before** why the Stellar network is our favorite cryptocurrency technology.

- Transactions take only 5 seconds and cost under 1¢
- Stellar does not burn mountains of fossil fuels
- Stellar has a "path payments" system for magical currency and token conversions.

The last point is special. I've got US dollars, and you want Japanese yen? No problem - 5 seconds later I've sent you yen, around the world.


amani1104 8:42 AM
sent Lumens worth $2.00 USD.
+ 32.7075322 XLM

Bitmessage File Help 100% Fri 18:31

Bitmessage

Messages | Send | Subscriptions | Chans | Blacklist | Network Status

Identities

- All accounts
  - inbox
  - new
  - sent
  - trash
- romeo.kienzler@mailchuck.co…
  - inbox
  - sent
  - trash

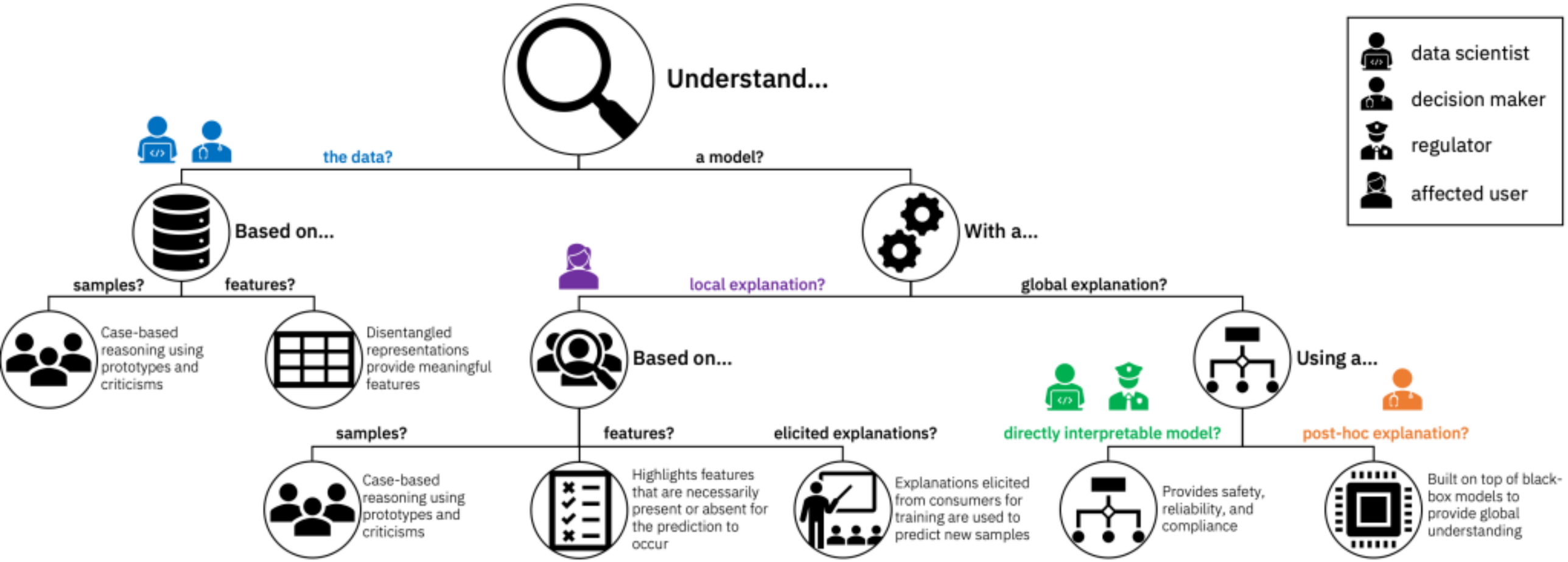
Search All

To	Subject	Sent
BM-2cTG3G92Sd4bwBHjkgUzZUnSicmw9Yivch	Re: `` ; ; : .. .,: :`,' : `, ` '	Message sent. Waiting for acknowledgement. Sent at Sun Aug 11 11:13:…
BM-2cTG3G92Sd4bwBHjkgUzZUnSicmw9Yivch	Re: `` ; ; : .. .,: :`,' : `, ` '	Acknowledgement of the message received Sun May 5 10:54:43 2019
BM-2cWvUgkPGKLw8tDe6tYCZMktKVv3KNiSRE	Re: test	Acknowledgement of the message received Sat Apr 13 23:56:06 2019
BM-2cWvUgkPGKLw8tDe6tYCZMktKVv3KNiSRE	Re: test	Acknowledgement of the message received Sat Apr 13 22:42:33 2019
BM-2cWvUgkPGKLw8tDe6tYCZMktKVv3KNiSRE	test	Acknowledgement of the message received Sat Apr 13 22:37:20 2019
romeo.kienzler@mailchuck.com	RE: selftest mac	Message sent. Sent at Tue Mar 12 13:48:54 2019
romeo.kienzler@mailchuck.com	selftest mac	Message sent. Sent at Tue Mar 5 07:50:26 2019
BM-2cVYYrhaY5Gbi3KqrX9Eae2NRNrkfrhCSA	romeo.kienzler	Acknowledgement of the message received Tue Feb 19 23:26:39 2019

und wie die ankommt !!!!!!!!!! voll geil

unstoppable!

nix gegen federated, aber true peer 2 peer is wieklich der oberhammer



 romeo kienzler

Computer Science > Artificial Intelligence

Boolean Decision Rules via Column Generation

Sanjeeb Dash, Oktay Günlük, Dennis Wei

(Submitted on 24 May 2018)

This paper considers the learning of Boolean rules in either disjunctive normal form (DNF, OR-of-ANDs, equivalent to decision rule sets) or conjunctive normal form (CNF, AND-of-ORs) as an interpretable model for classification. An integer program is formulated to optimally trade classification accuracy for rule simplicity. Column generation (CG) is used to efficiently search over an exponential number of candidate clauses (conjunctions or disjunctions) without the need for heuristic rule mining. This approach also bounds the gap between the selected rule set and the best possible rule set on the training data. To handle large datasets, we propose an approximate CG algorithm using randomization. Compared to three recently proposed alternatives, the CG algorithm dominates the accuracy-simplicity trade-off in 7 out of 15 datasets. When maximized for accuracy, CG is competitive with rule learners designed for this purpose, sometimes finding significantly simpler solutions that are no less accurate.

Subjects: **Artificial Intelligence (cs.AI)**

Cite as: arXiv:1805.09901 [cs.AI]

(or arXiv:1805.09901v1 [cs.AI] for this version)

Bibliographic data

[Enable Bibex(What is Bibex?)]

Submission history

From: Oktay Gunluk [view email]

[v1] Thu, 24 May 2018 21:12:26 UTC (116 KB)

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

Computer Science > Artificial Intelligence

Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, Payel Das

(Submitted on 21 Feb 2018 (v1), last revised 29 Oct 2018 (this version, v2))

In this paper we propose a novel method that provides contrastive explanations justifying the classification of an input by a black box classifier such as a deep neural network. Given an input we find what should be %necessarily and minimally and sufficiently present (viz. important object pixels in an image) to justify its classification and analogously what should be minimally and necessarily \emph{absent} (viz. certain background pixels). We argue that such explanations are natural for humans and are used commonly in domains such as health care and criminology. What is minimally but critically \emph{absent} is an important part of an explanation, which to the best of our knowledge, has not been explicitly identified by current explanation methods that explain predictions of neural networks. We validate our approach on three real datasets obtained from diverse domains; namely, a handwritten digits dataset MNIST, a large procurement fraud dataset and a brain activity strength dataset. In all three cases, we witness the power of our approach in generating precise explanations that are also easy for human experts to understand and evaluate.

Subjects: **Artificial Intelligence (cs.AI)**; Computer Vision and Pattern Recognition (cs.CV); Machine Learning (cs.LG)
Report number: accepted to NIPS 2018
Cite as: arXiv:1802.07623 [cs.AI]
(or arXiv:1802.07623v2 [cs.AI] for this version)

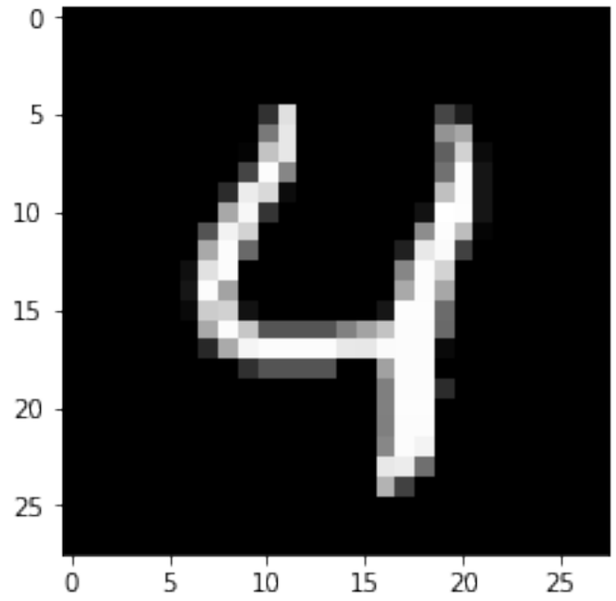
Bibliographic data

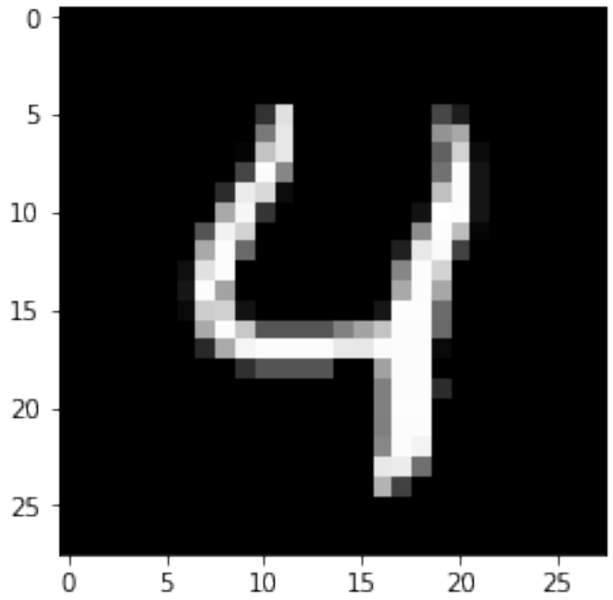
[Enable Bibex(What is Bibex?)]

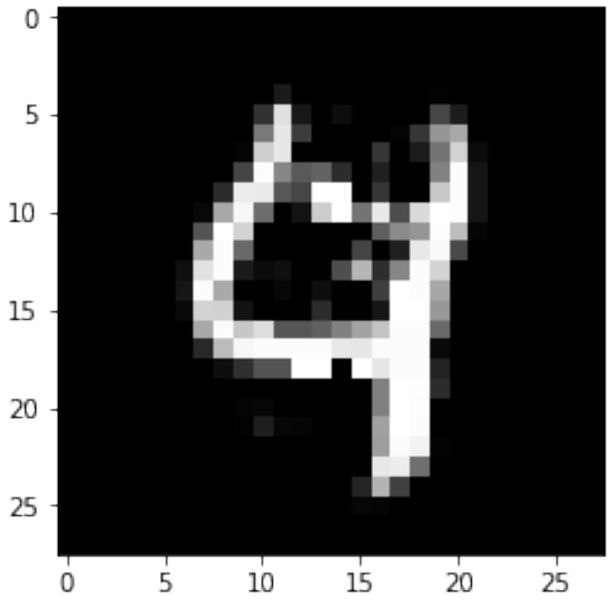
Submission history

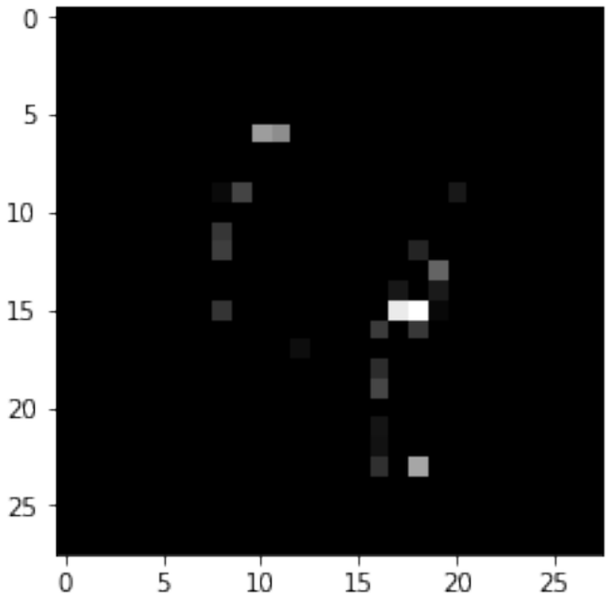
From: Amit Dhurandhar [view email]
[v1] Wed, 21 Feb 2018 15:51:38 UTC (539 KB)
[v2] Mon, 29 Oct 2018 16:08:36 UTC (975 KB)

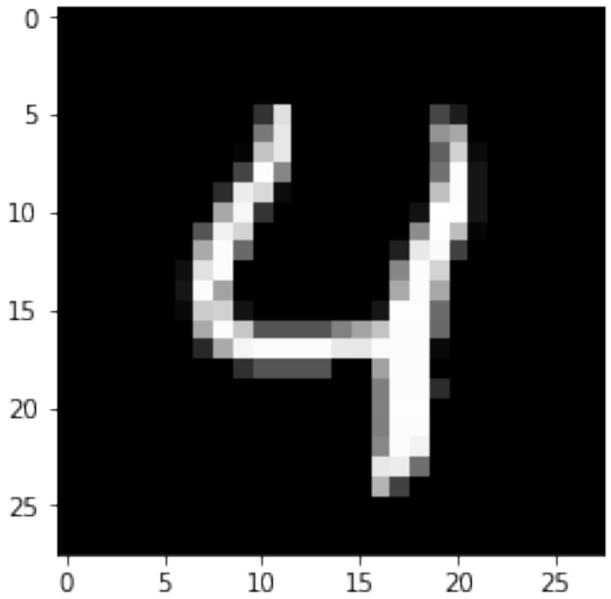
Which authors of this paper are endorsers? / *Disable MathJax (What is MathJax?)*

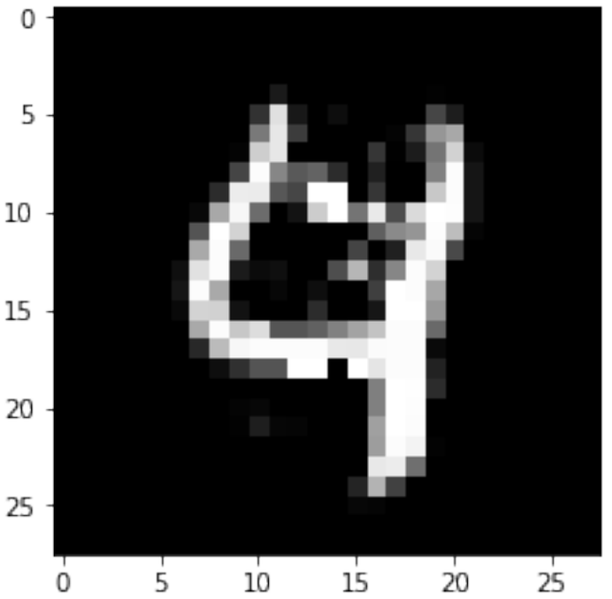










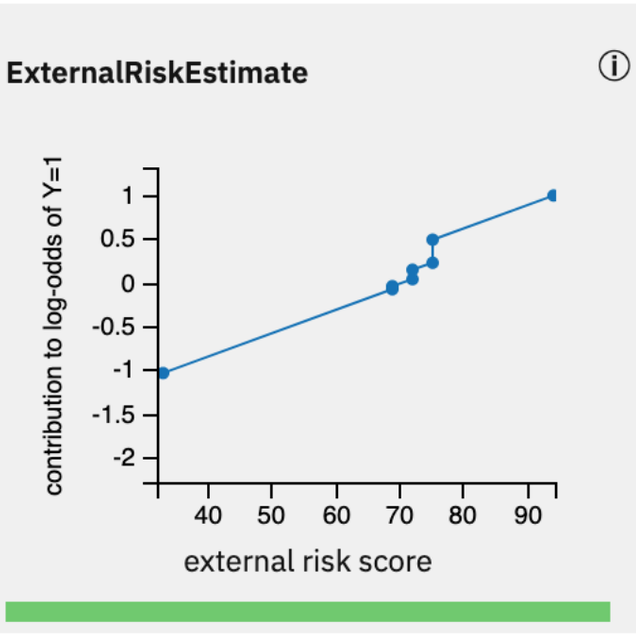


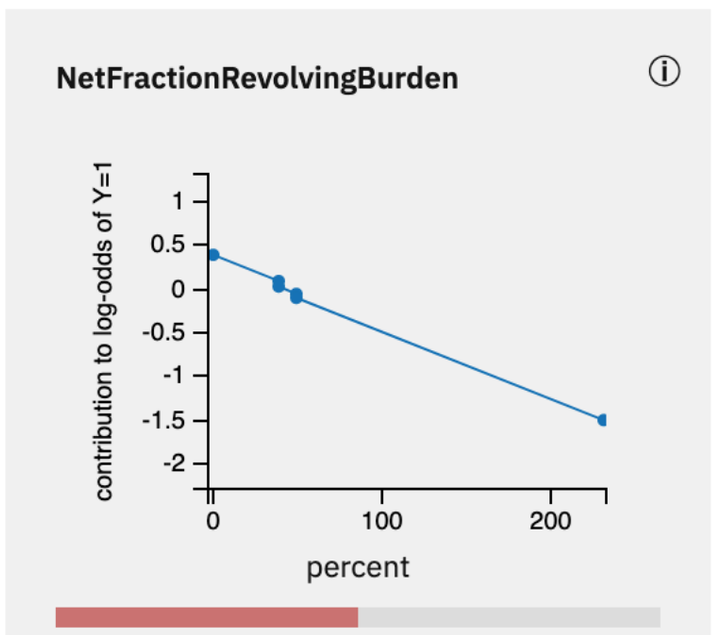
AI Explainability 360 - Demo

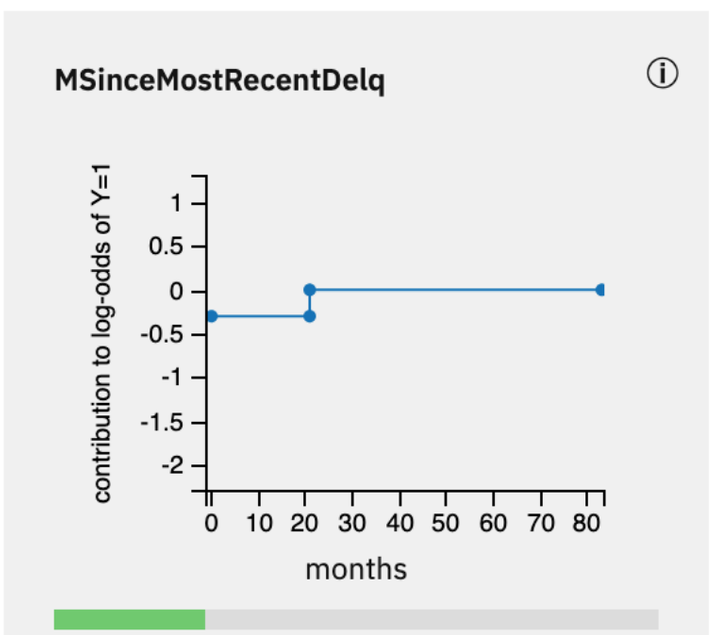
Data — Consumer — Explanation

A Data Scientist wants to understand:

What is the overall logic of the model in making decisions?
Is the logic reasonable, so that we can deploy the model with confidence?







 romeo kienzler

AI Explainability 360 - Demo

Data — Consumer — Explanation

A Loan Officer wants to understand:

Why is the model recommending this person's credit be approved or denied?
How can I inform my decision to accept or reject a line of credit by looking at similar individuals?

Using Similar Examples to Inform a Loan Decision

A Loan Officer typically makes the final decision when accepting or rejecting a customer's loan request. When using a predictive model, a Loan Officer wants to understand how and why the model came to that prediction in order to make an informed and trusted decision. One algorithm within AI Explainability 360—ProtoDash—works with an existing predictive model to show how the customer compares to others who have similar profiles and had similar repayment records to the model's prediction for the current customer, which helps to evaluate and predict the applicant's risk. Based on the model's prediction and the explanation for how it came to that recommendation, the Loan Officer can make a more informed decision.

Select a customer the Loan Officer wants to understand

Alice

Approved

Robert

Denied

AI Explainability 360 - Demo

Data — Consumer — Explanation

A Bank Customer wants to understand:

Why was my application rejected?
What can I improve to increase the likelihood my application is accepted?

Providing Contrastive Explanations for Insight into Loan Application Outcomes

The Bank Customer wants to know how and why the decision was made to accept or reject their loan application. The explanation given will help them understand if they've been treated fairly, and also provide insight into what – if their application was rejected – they can improve in order to increase the likelihood it will be accepted in the future. To help provide that insight and suggest avenues for improvement, we will use the Contrastive Explanations Method (CEM) algorithm available in AI Explainability 360. This algorithm sits on top of an existing predictive model and helps detect both the features that a bank customer could improve (e.g., amount of time since last credit inquiry, average age of accounts), and also further detects the features that will increase the likelihood of approval and those that are within reach for the customer. See examples below.

Select a customer asking for explanations

Jason	Ann	Julia
Denied	**Denied**	**Denied**

 romeo kienzler

AI Explainability 360

https://github.com/IBM/AIX360

 romeo kienzler

Is it accountable?

Data Lineage...

Kubeflow

Pipelines

Experiments

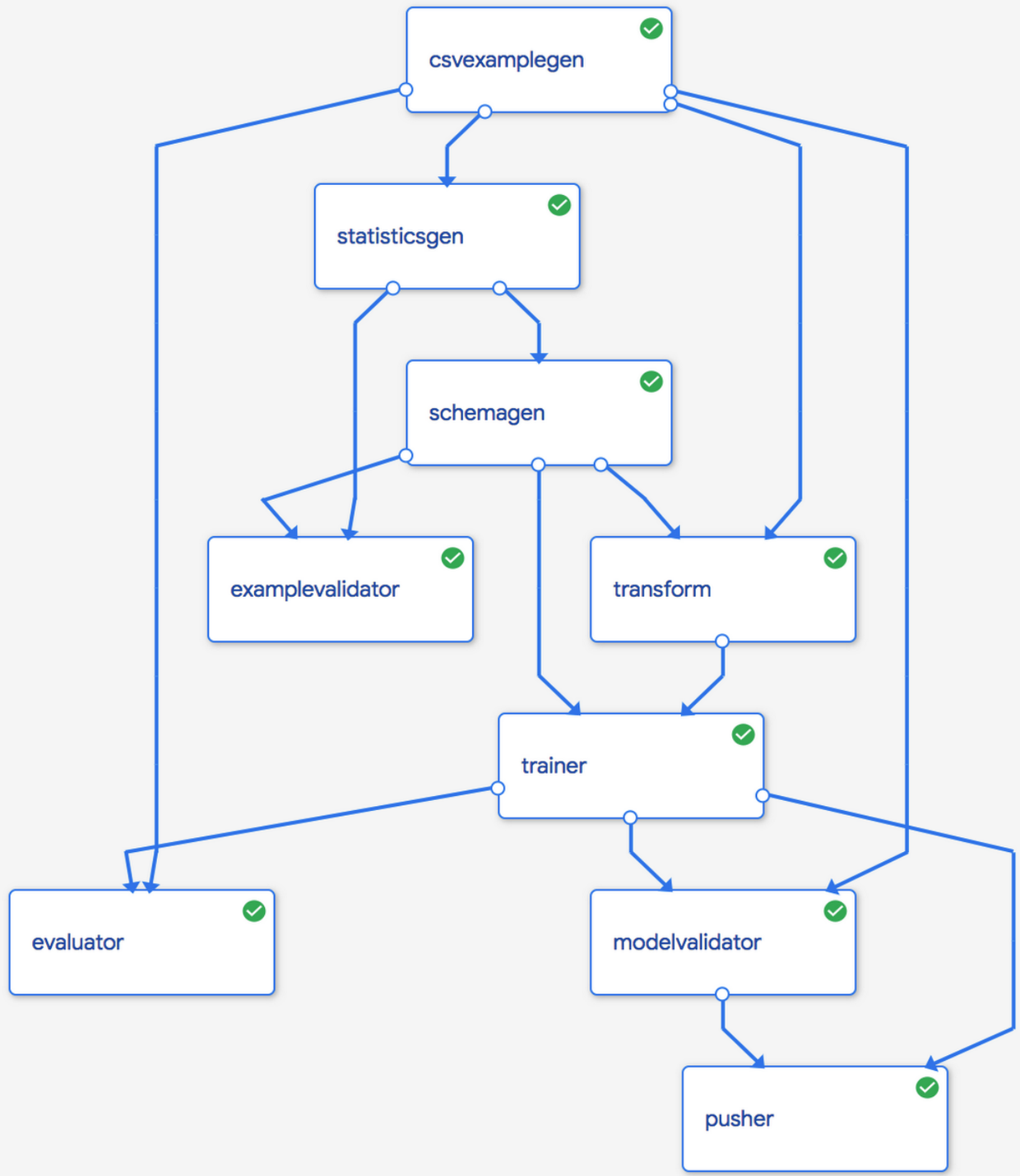
Notebooks

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

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

Watson OpenScale

Announced recently:

- ✓ A new capability called **Drift Detection** which detects when and how far a model "drifts" from its original parameters
- ✓ In line with our **Watson Anywhere** vision, the latest release of Cloud Pak for Data makes the world-class **Watson OpenScale** available in private, hybrid, and multi-cloud environments.

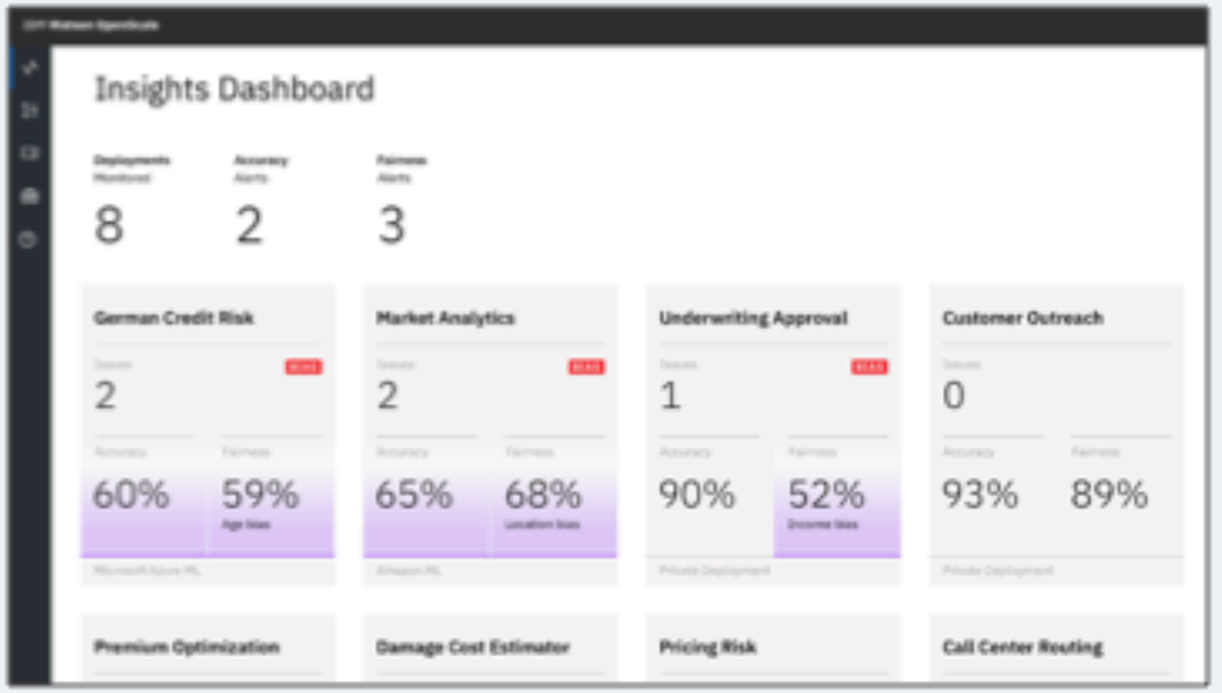
Watson OpenScale **tracks and measures trusted AI outcomes across its lifecycle**, and adapts and governs AI to changing business situations — for models built and running anywhere.

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.



IBM CODAIT

Vehicle Repair Estimator

Neural Network Synthesis

Deploy | Download Model

STATUS

IBM AI OpenScale has successfully created a synthetic deep learning model with **NeuNetS**. Your NeuNetS model was trained with your data set and tested for accuracy, precision, and recall.

Prepare — Synthesize — Test — Deploy

ACCURACY	PRECISION	RECALL
93% .46	94% .47	92% .48

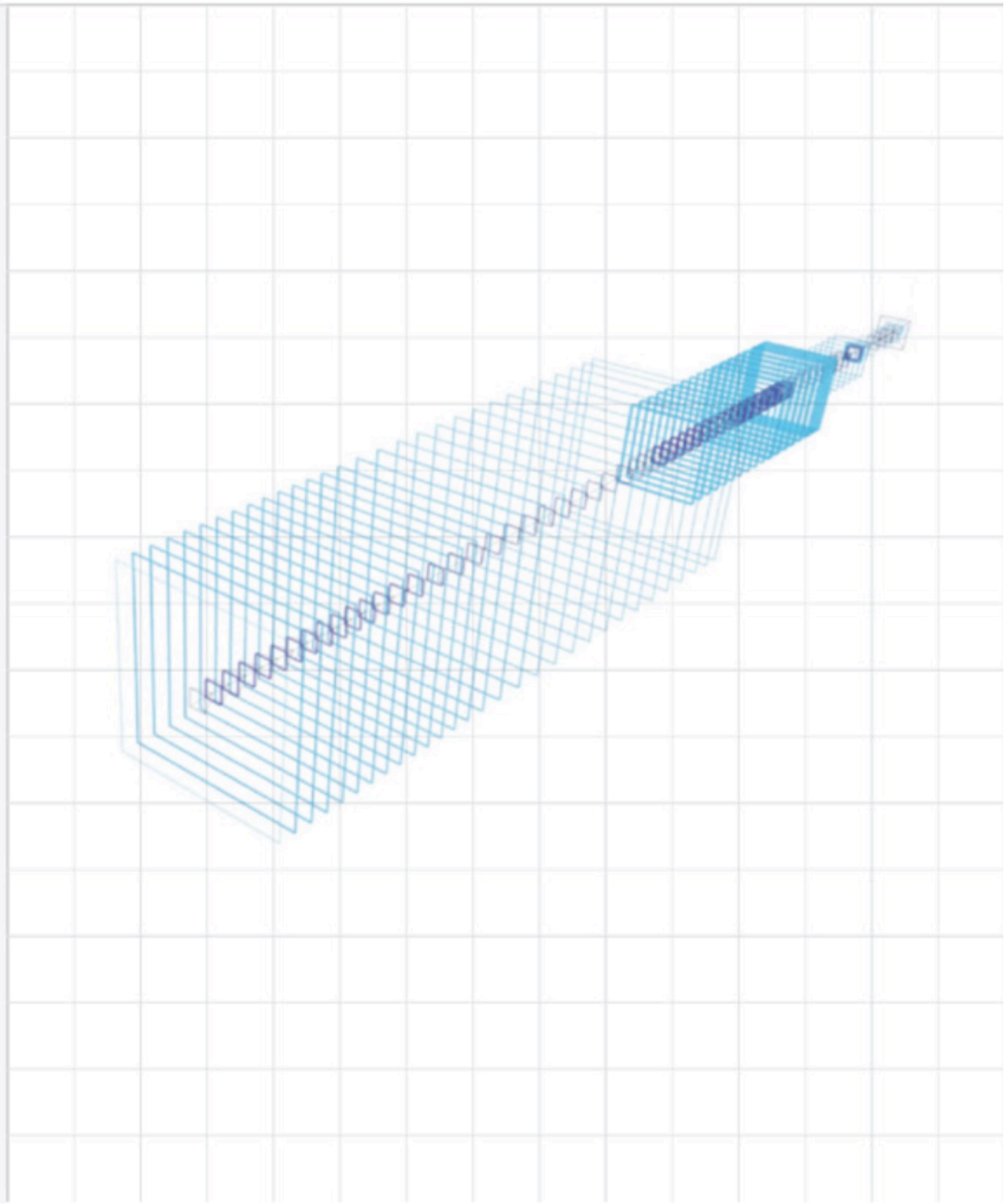
TRAINING DATA

Source	Test Sample
TrainingSet.zip	25%

Size	Features	Labels
10,245	256	765

DEPLOYMENT

AIOS Instance	WML Instance	Model Type	Framework
aios-tx	wml-ws	wml-1.1	Tensorflow 1.5

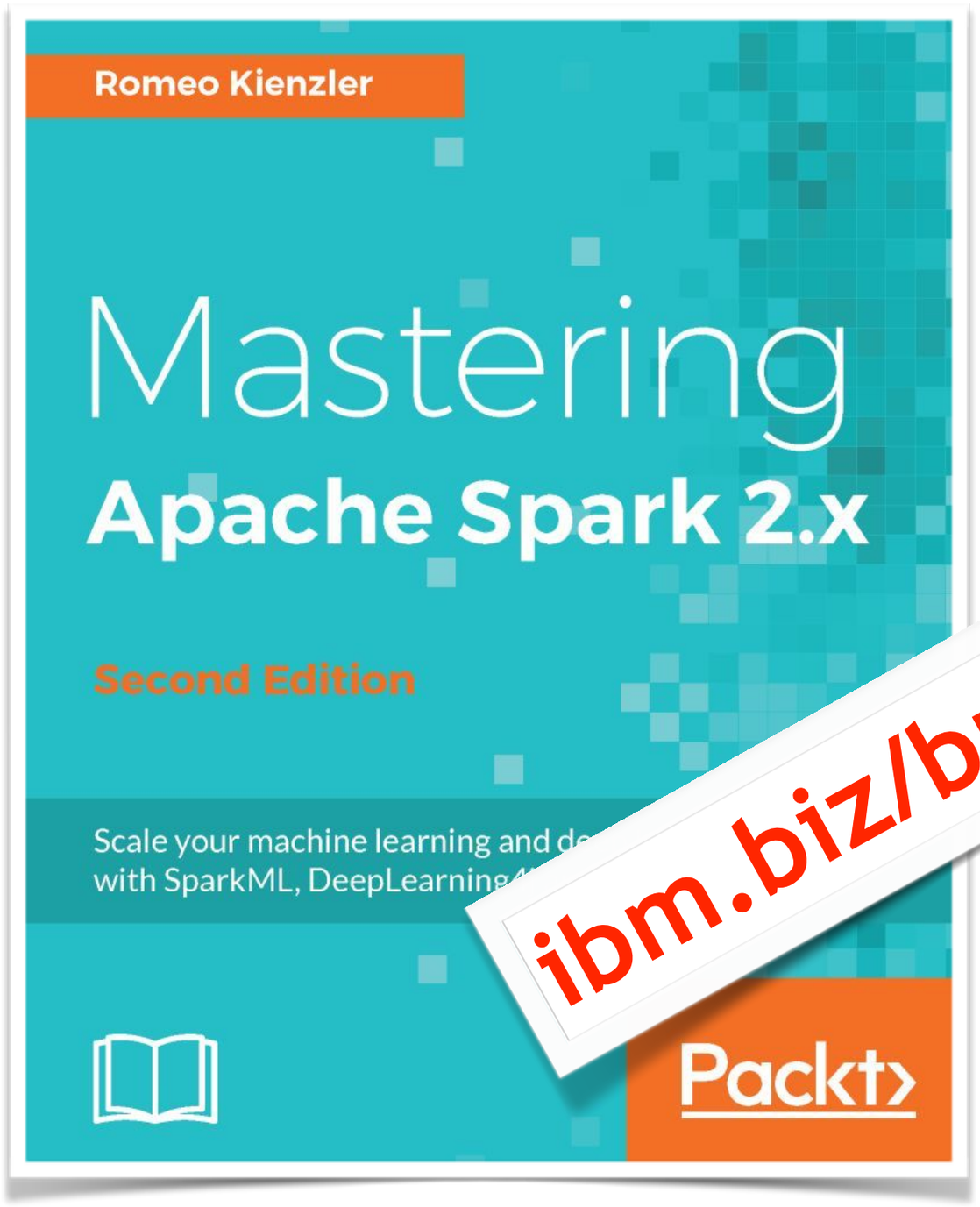


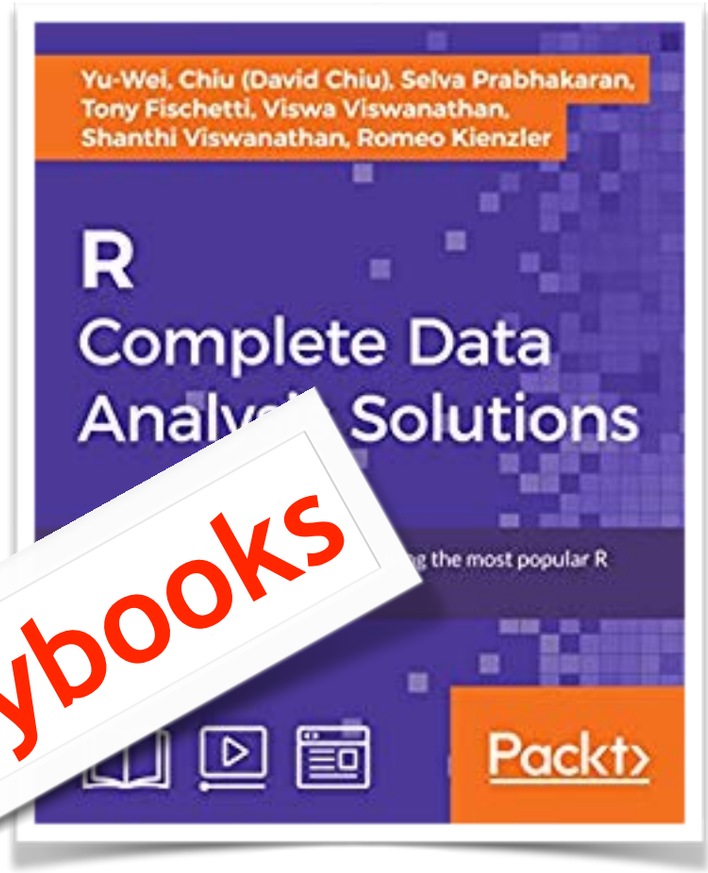
NeuNetS: An Automated Synthesis Engine for Neural Network Design

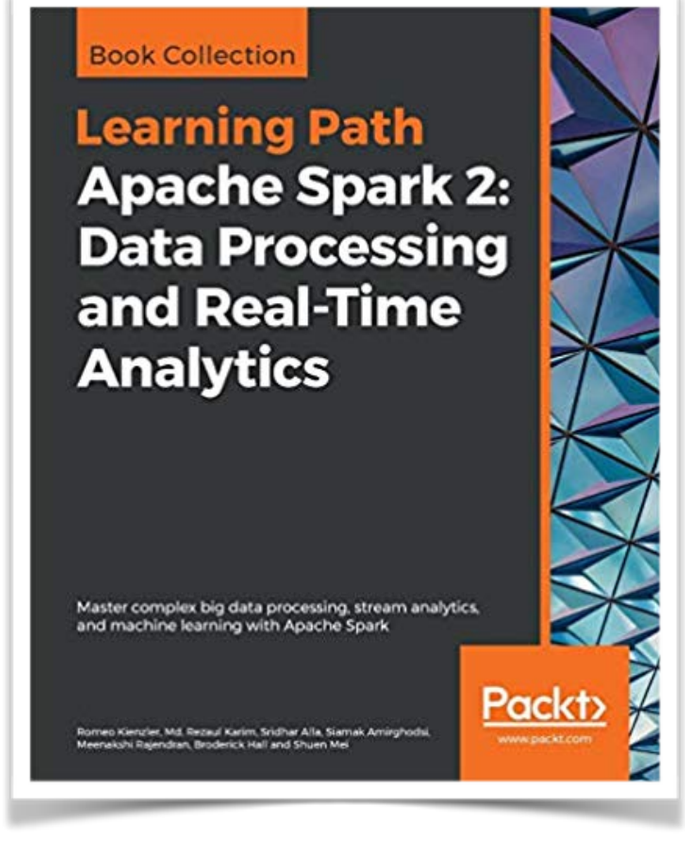
https://arxiv.org/abs/1901.06261

https://en.wikipedia.org/wiki/Neural_architecture_search

romeo kienzler







IBM Advanced Data Science Specialization Certificate on C sera

Fundamentals of Scalable Data
www.coursera.org/le

Advanced g and Signal Processing
www.coursera ed-machine-learning-signal ssing

Applied AI wit
www.co

Ac Data Science Capstone Project
https://www.coursera.org/learn/advanced-data-science-capstone

IBM Cloud Free Tier which enables all the open source presented here:

ibm.biz/freecloudregistration*

*includes a tracking URL in favour of myself