Explaining Machine Learning Models

Dr. Margriet Groenendijk

Data Science & AI Developer Advocate

IBM

@MargrietGr

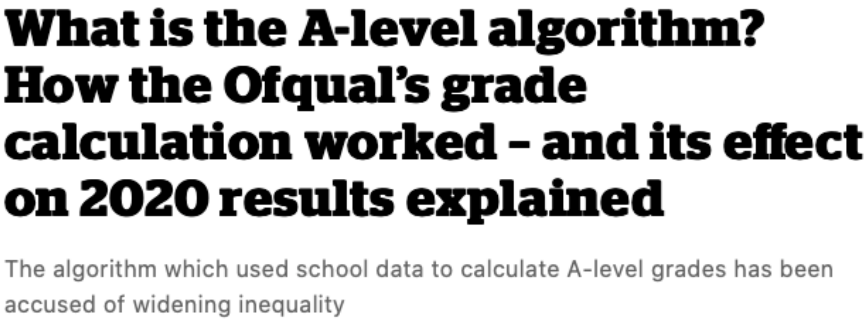
Getting started

Go to your IBM Cloud account or sign up

http://ibm.biz/explainai_mg

All instructions

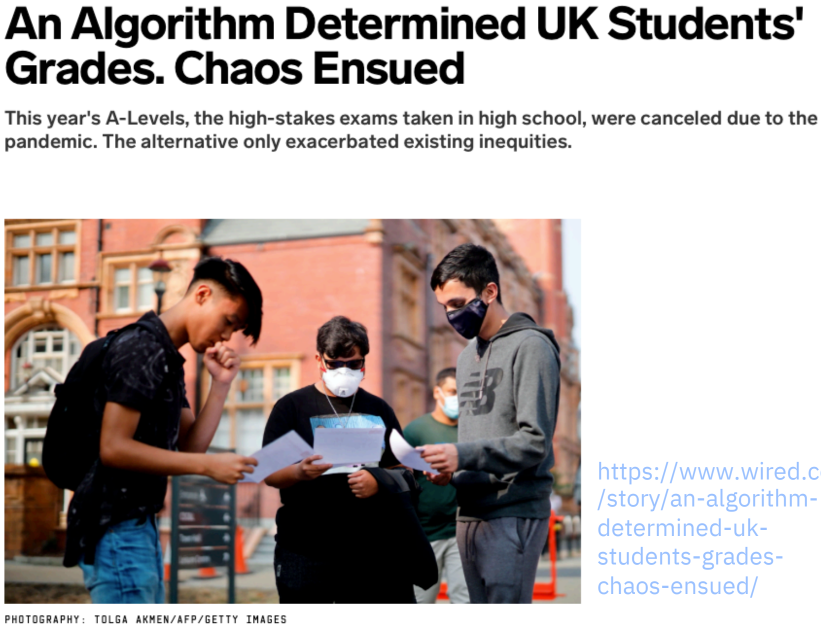
https://github.com/IBMDeveloperUK/AIX360-workshop



What is the A-level algorithm? How the Ofqual's grade calculation worked – and its effect on 2020 results explained

The algorithm which used school data to calculate A-level grades has been accused of widening inequality

https://inews.co.uk/news/education/a-level-algorithm-what-ofqual-grades-how-work-results-2020-explained-581250

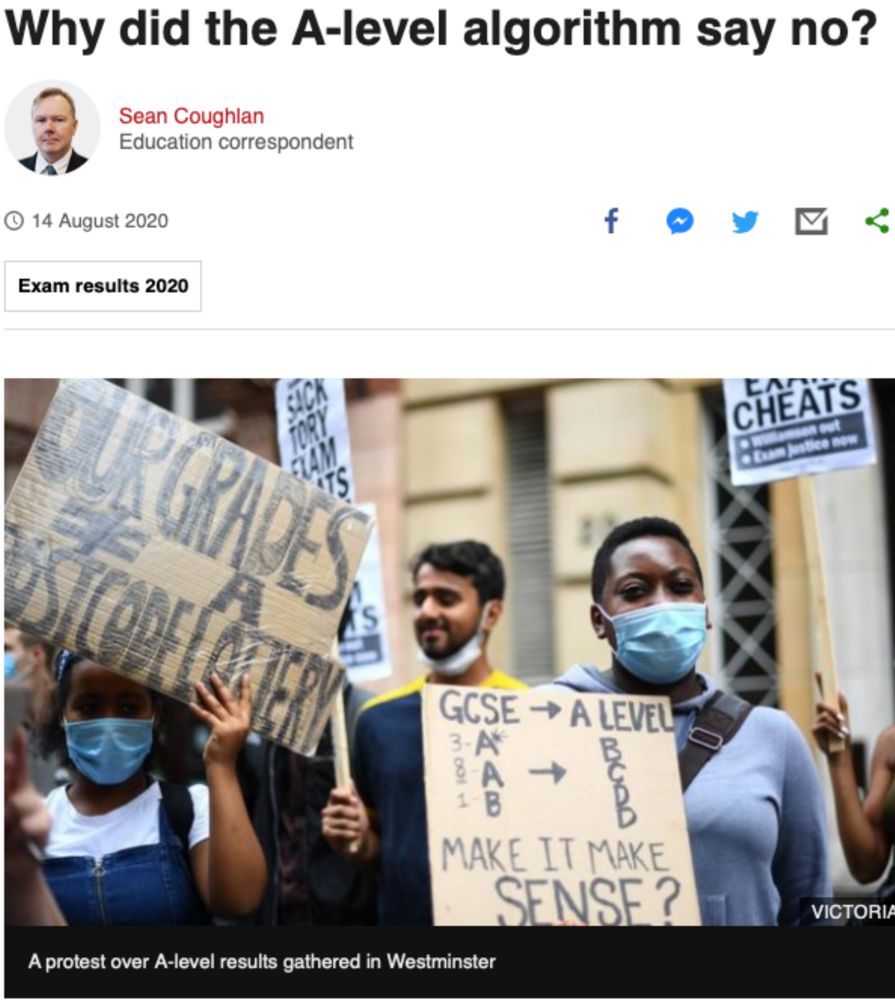


An Algorithm Determined UK Students' Grades. Chaos Ensued

This year's A-Levels, the high-stakes exams taken in high school, were canceled due to the pandemic. The alternative only exacerbated existing inequities.

PHOTOGRAPHY: TOLGA AKMEN/AFP/GETTY IMAGES

https://www.wired.com/story/an-algorithm-determined-uk-students-grades-chaos-ensued/



Why did the A-level algorithm say no?

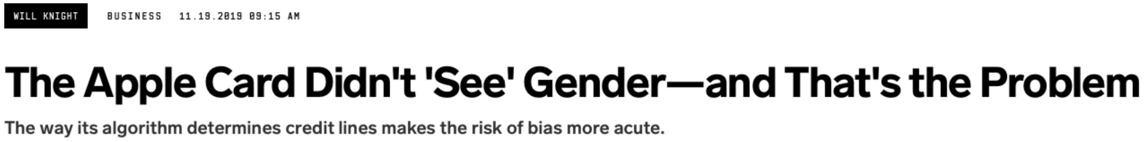
Sean Coughlan
Education correspondent

14 August 2020

Exam results 2020

A protest over A-level results gathered in Westminster

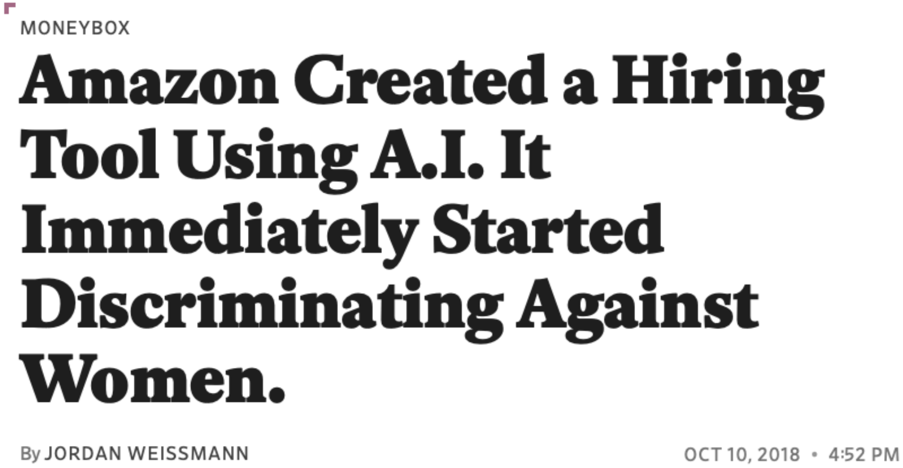
https://www.bbc.co.uk/news/education-53787203



WILL KNIGHT BUSINESS 11.19.2019 09:15 AM

The Apple Card Didn't 'See' Gender—and That's the Problem

The way its algorithm determines credit lines makes the risk of bias more acute.

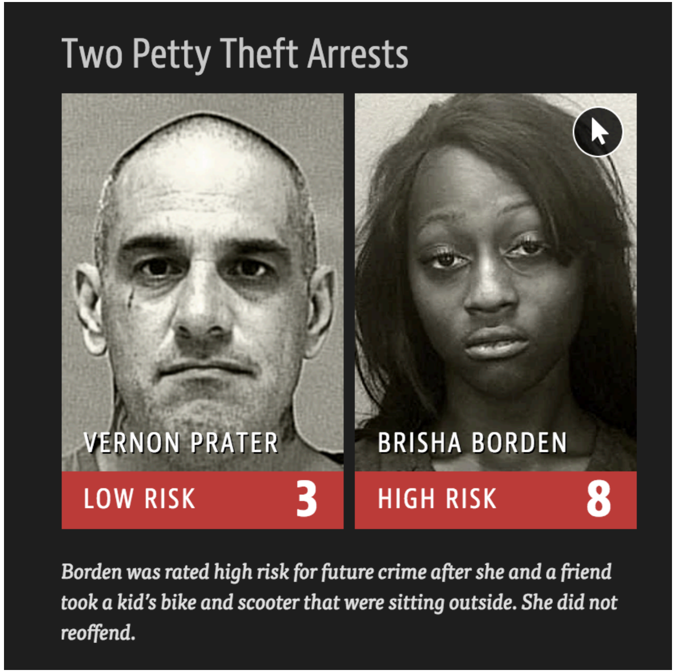


MONEYBOX

Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women.

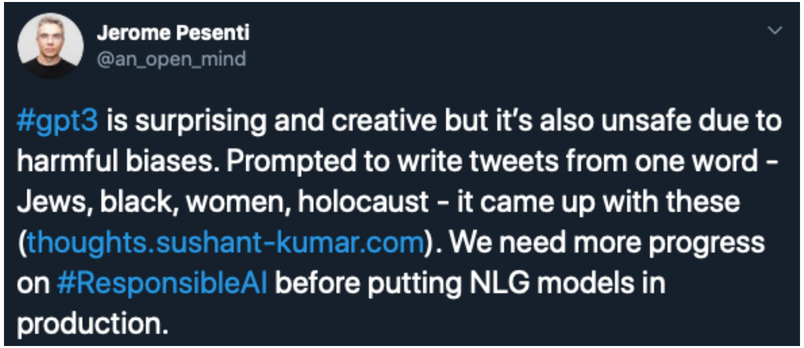
By JORDAN WEISSMANN

OCT 10, 2018 • 4:52 PM



Two Petty Theft Arrests

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.



Jerome Pesenti
@an_open_mind

#gpt3 is surprising and creative but it's also unsafe due to harmful biases. Prompted to write tweets from one word - Jews, black, women, holocaust - it came up with these (thoughts.sushant-kumar.com). We need more progress on #ResponsibleAI before putting NLG models in production.

@MargrietGr

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?

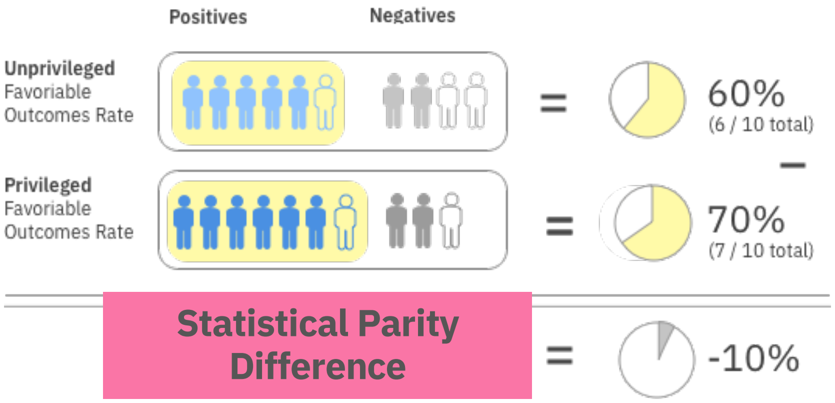
LINEAGE

AI FactSheets 360

↳ (AIFS360)

aifs360.mybluemix.net

How To Measure Fairness – Some Group Fairness Metrics



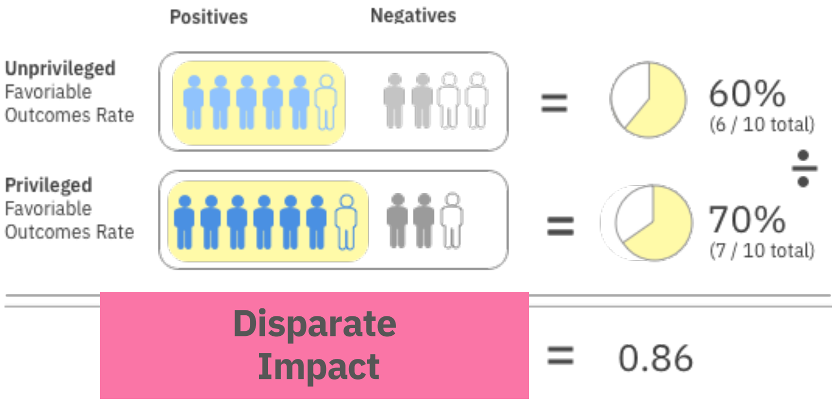
Positives Negatives

Unprivileged Favoriable Outcomes Rate = 60% (6 / 10 total)

–

Privileged Favoriable Outcomes Rate = 70% (7 / 10 total)

Statistical Parity Difference = -10%



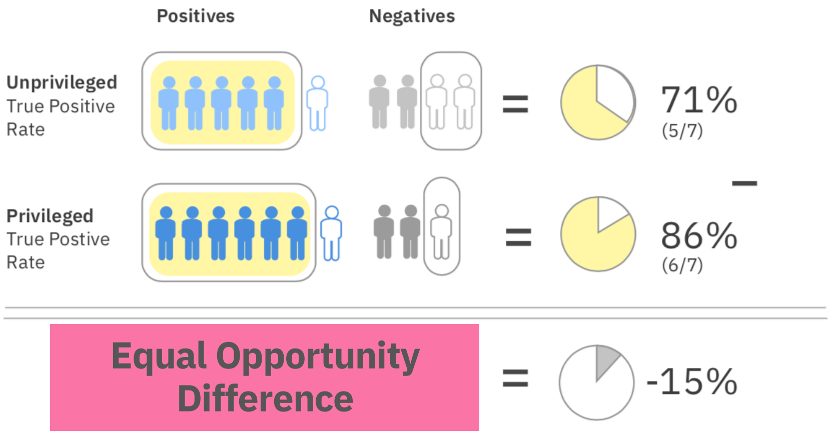
Positives Negatives

Unprivileged Favoriable Outcomes Rate = 60% (6 / 10 total)

÷

Privileged Favoriable Outcomes Rate = 70% (7 / 10 total)

Disparate Impact = 0.86



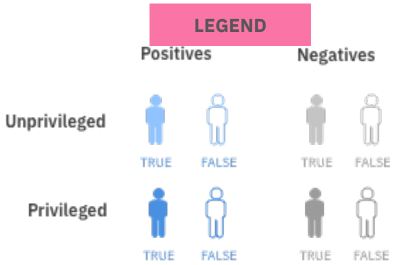
Positives Negatives

Unprivileged True Positive Rate = 71% (5/7)

–

Privileged True Postive Rate = 86% (6/7)

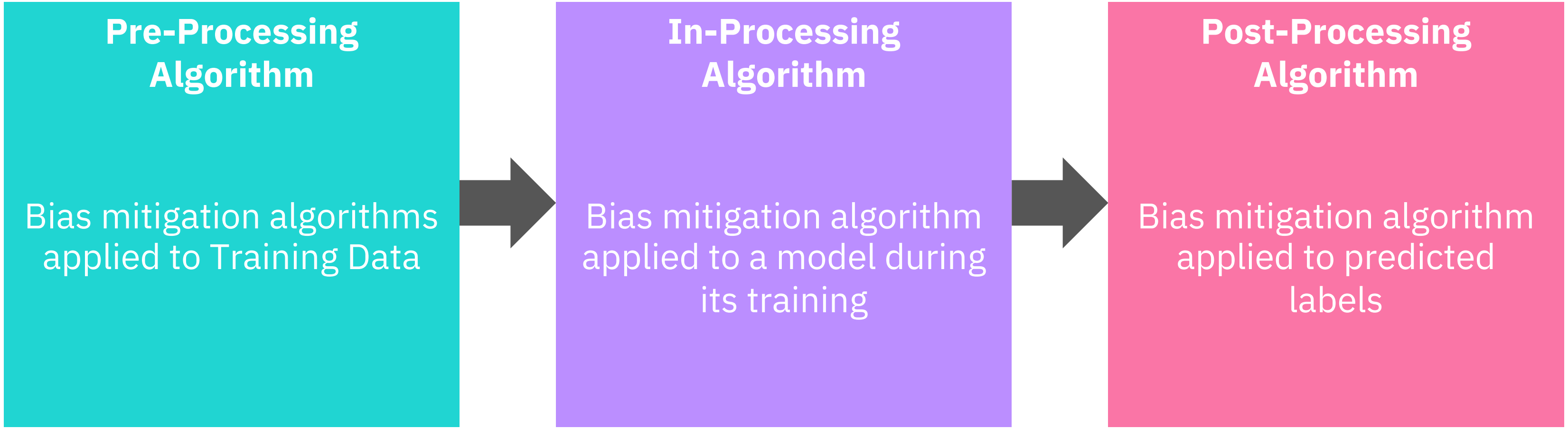
Equal Opportunity Difference = -15%



LEGEND

	Positives		Negatives	
Unprivileged	TRUE	FALSE	TRUE	FALSE
Privileged	TRUE	FALSE	TRUE	FALSE

Where Can You Intervene in the Pipeline?



Pre-Processing Algorithm

Bias mitigation algorithms applied to Training Data

In-Processing Algorithm

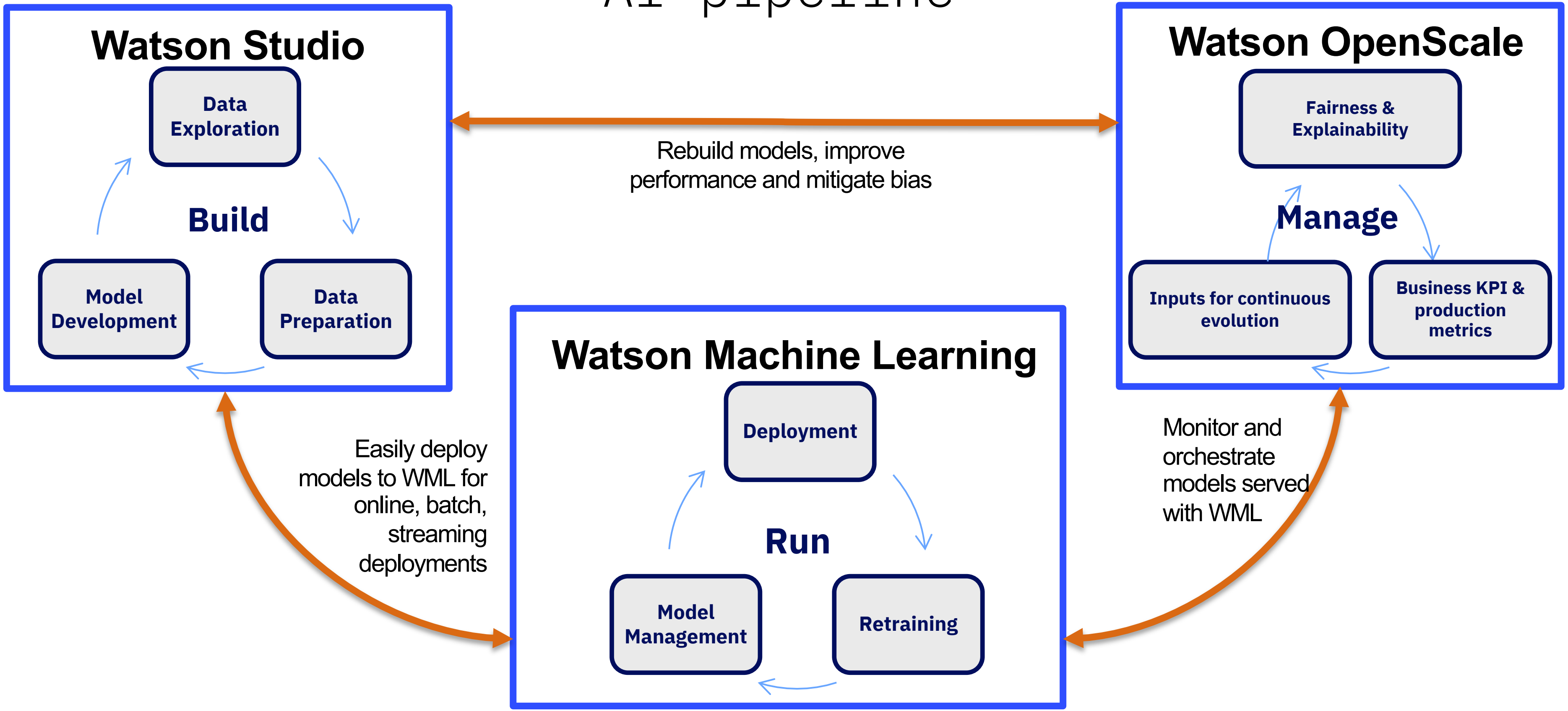
Bias mitigation algorithm applied to a model during its training

Post-Processing Algorithm

Bias mitigation algorithm applied to predicted labels

- If you can modify the Training Data, then pre-processing can be used.
- If you can modify the Learning Algorithm, then in-processing can be used.
- If you can only treat the learned model as a black box and can’t modify the training data or learning algorithm, then only post-processing can be used

AI pipeline



Watson Studio

Data Exploration

Build

Model Development

Data Preparation

Rebuild models, improve performance and mitigate bias

Watson OpenScale

Fairness & Explainability

Manage

Inputs for continuous evolution

Business KPI & production metrics

Easily deploy models to WML for online, batch, streaming deployments

Watson Machine Learning

Deployment

Run

Model Management

Retraining

Monitor and orchestrate models served with WML

IBM Cloud Pak for Data

Fully-integrated data and AI platform

Cloud Pak for Data…

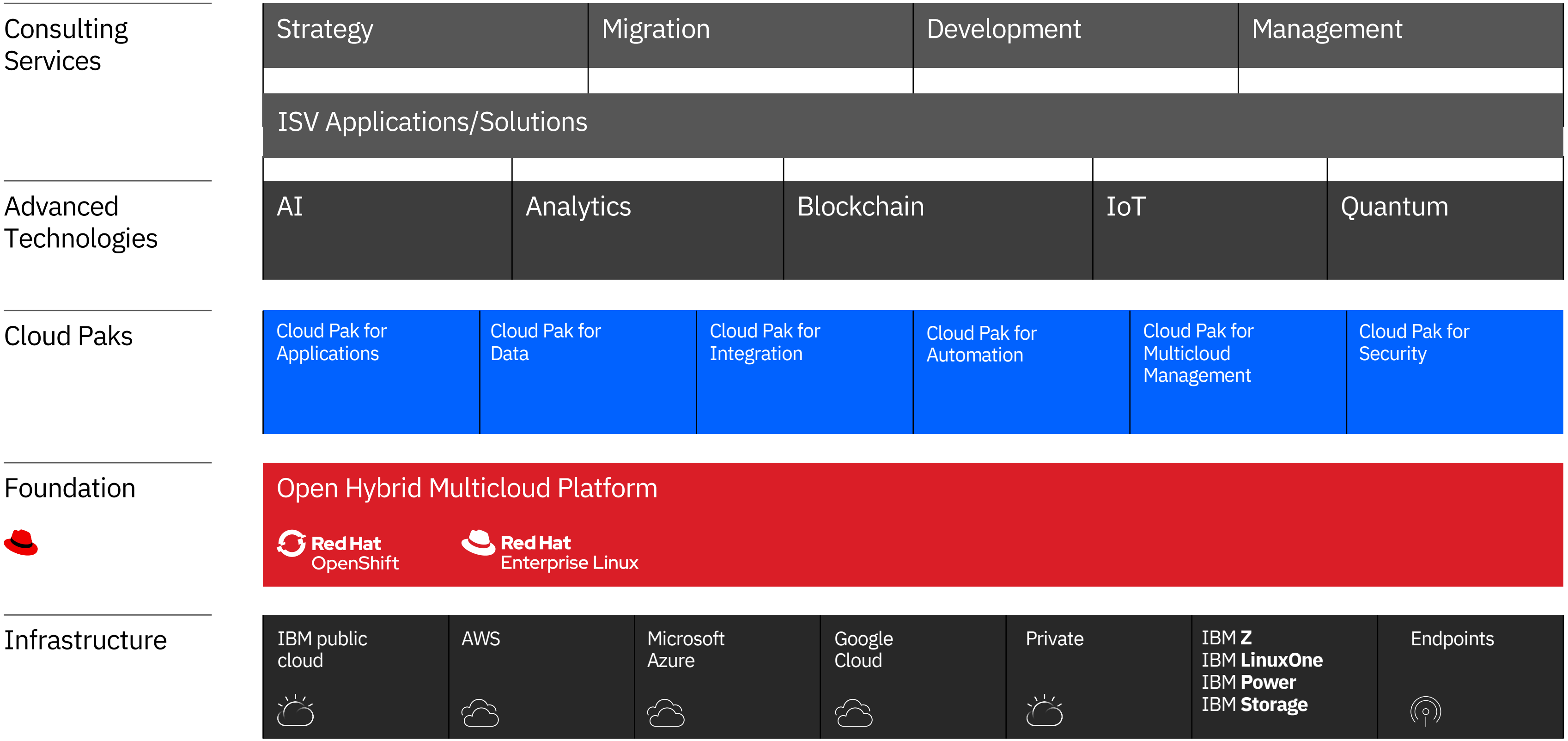
- Runs on Red Hat OpenShift and is a fully-integrated data and AI platform
- Supports multi-cloud environments such as AWS, Azure, Google Cloud, IBM Cloud, and private clouds
- Allows you to build, deploy, and manage ML models that scale throughout the organization and automates the AI lifecycle
- Enables integrations to popular open source and cloud native tools, as well as IBM application middleware and development services

Developer benefits…

- Full control over your data and its privacy
- Seamless integration of developer tools -- streamlines work by creating a pipeline for collecting, organizing, analyzing, and consuming data
- Single platform for data management and analysis, allowing developers to easily manage data connections and access to analysis tools
- Core operational services provided, including logging, monitoring, and security

https://ibm.biz/cpd-experiences

Build once. Deploy anywhere.



Consulting Services

Strategy | Migration | Development | Management

ISV Applications/Solutions

Advanced Technologies

AI | Analytics | Blockchain | IoT | Quantum

Cloud Paks

Cloud Pak for Applications | Cloud Pak for Data | Cloud Pak for Integration | Cloud Pak for Automation | Cloud Pak for Multicloud Management | Cloud Pak for Security

Foundation

Open Hybrid Multicloud Platform

Red Hat OpenShift

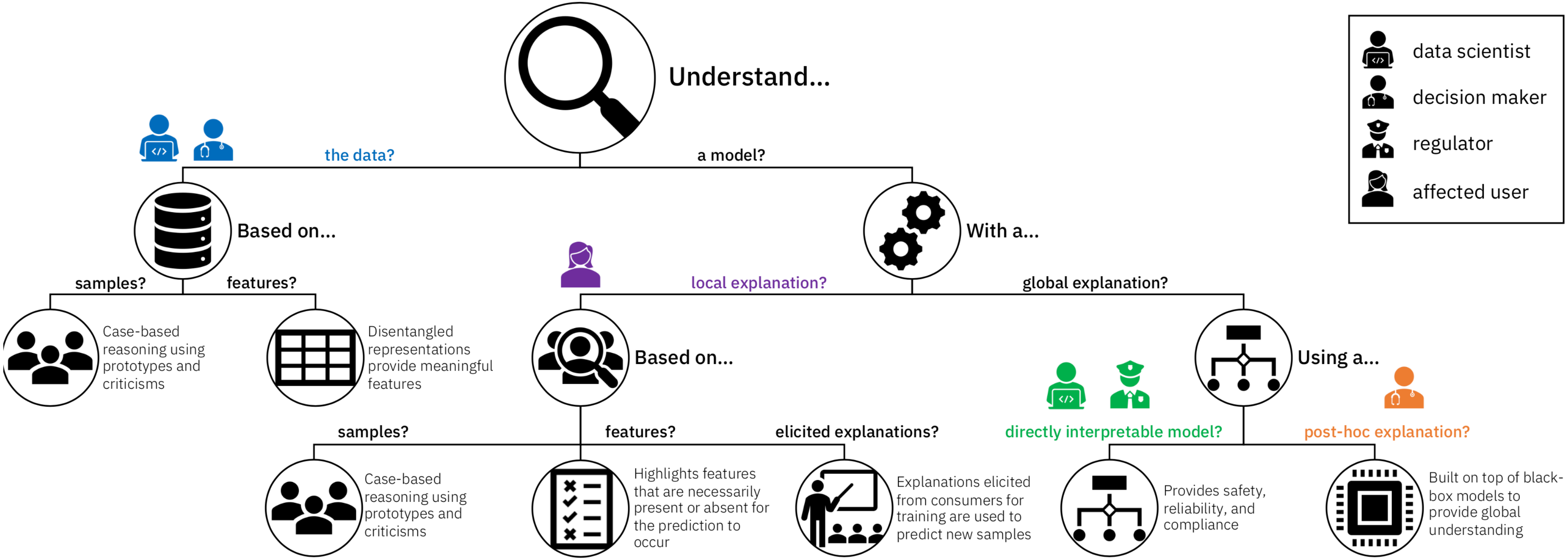
Red Hat Enterprise Linux

Infrastructure

IBM public cloud | AWS | Microsoft Azure | Google Cloud | Private | IBM **Z**
IBM **LinuxOne**
IBM **Power**
IBM **Storage** | Endpoints

How to understand or explain models?

@MargrietGr



Understand...

the data?

a model?

Based on...

samples?

features?

Case-based reasoning using prototypes and criticisms

Disentangled representations provide meaningful features

With a...

local explanation?

global explanation?

Based on...

samples?

features?

elicited explanations?

Case-based reasoning using prototypes and criticisms

Highlights features that are necessarily present or absent for the prediction to occur

Explanations elicited from consumers for training are used to predict new samples

Using a...

directly interpretable model?

post-hoc explanation?

Provides safety, reliability, and compliance

Built on top of black-box models to provide global understanding

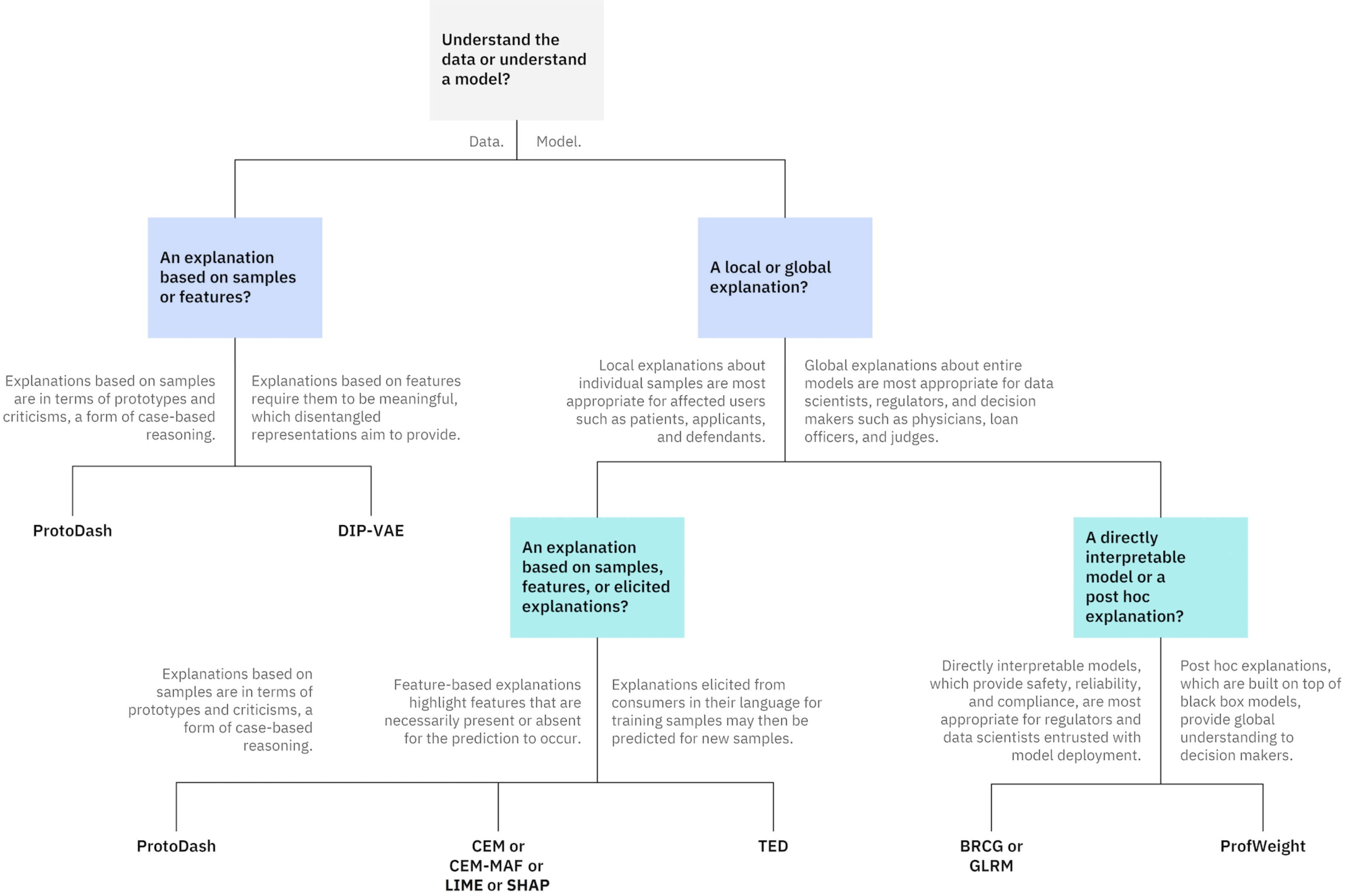
data scientist

decision maker

regulator

affected user

@MargrietGr



Understand the data or understand a model?

Data.

Model.

An explanation based on samples or features?

Explanations based on samples are in terms of prototypes and criticisms, a form of case-based reasoning.

Explanations based on features require them to be meaningful, which disentangled representations aim to provide.

ProtoDash

DIP-VAE

A local or global explanation?

Local explanations about individual samples are most appropriate for affected users such as patients, applicants, and defendants.

Global explanations about entire models are most appropriate for data scientists, regulators, and decision makers such as physicians, loan officers, and judges.

An explanation based on samples, features, or elicited explanations?

Explanations based on samples are in terms of prototypes and criticisms, a form of case-based reasoning.

Feature-based explanations highlight features that are necessarily present or absent for the prediction to occur.

Explanations elicited from consumers in their language for training samples may then be predicted for new samples.

ProtoDash

CEM or CEM-MAF or LIME or SHAP

TED

A directly interpretable model or a post hoc explanation?

Directly interpretable models, which provide safety, reliability, and compliance, are most appropriate for regulators and data scientists entrusted with model deployment.

Post hoc explanations, which are built on top of black box models, provide global understanding to decision makers.

BRCG or GLRM

ProfWeight

@MargrietGr

source: IBM Research AI Explainability 360

AIX360: Different Ways to explain

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?

AIX360: Different Ways to explain

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.

AIX360: Different Ways to explain

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”)

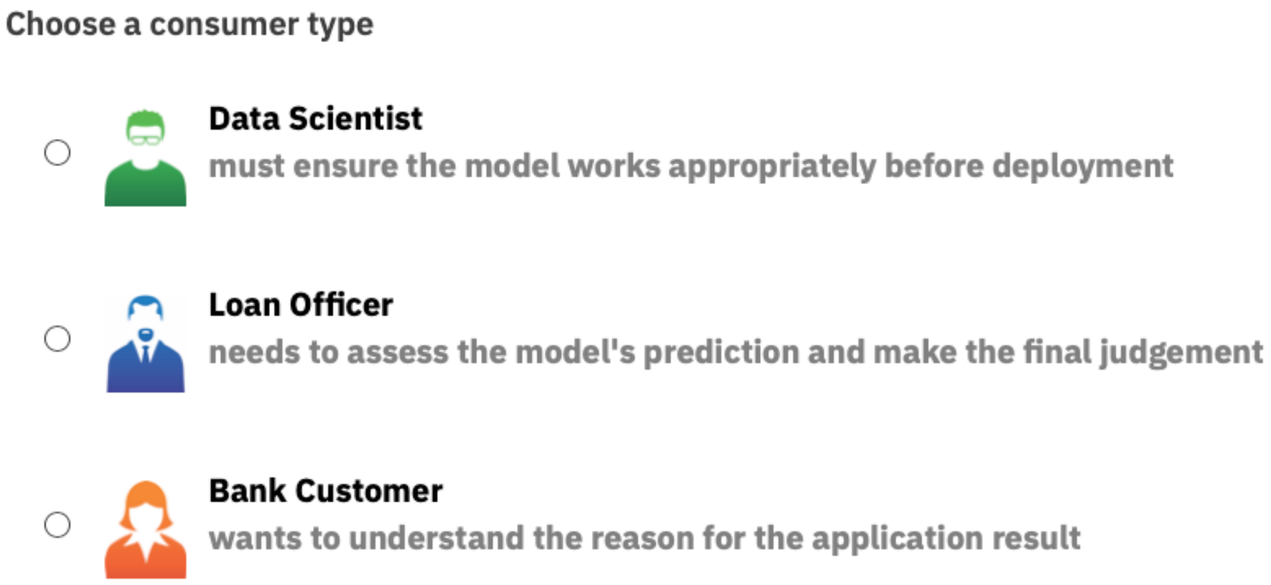
Is our system performing well?

How can we improve it?

FICO Explainable Machine Learning Challenge dataset

http://aix360.mybluemix.net/

Use the information about the applicant in their credit report to predict whether they will make timely payments over a two-year period



Choose a consumer type

○ **Data Scientist**
must ensure the model works appropriately before deployment

○ **Loan Officer**
needs to assess the model's prediction and make the final judgement

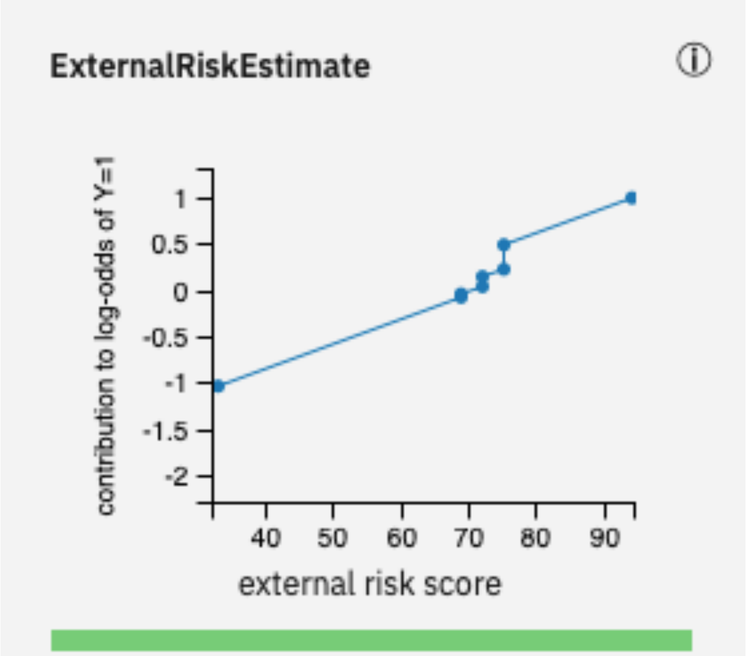
○ **Bank Customer**
wants to understand the reason for the application result

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?

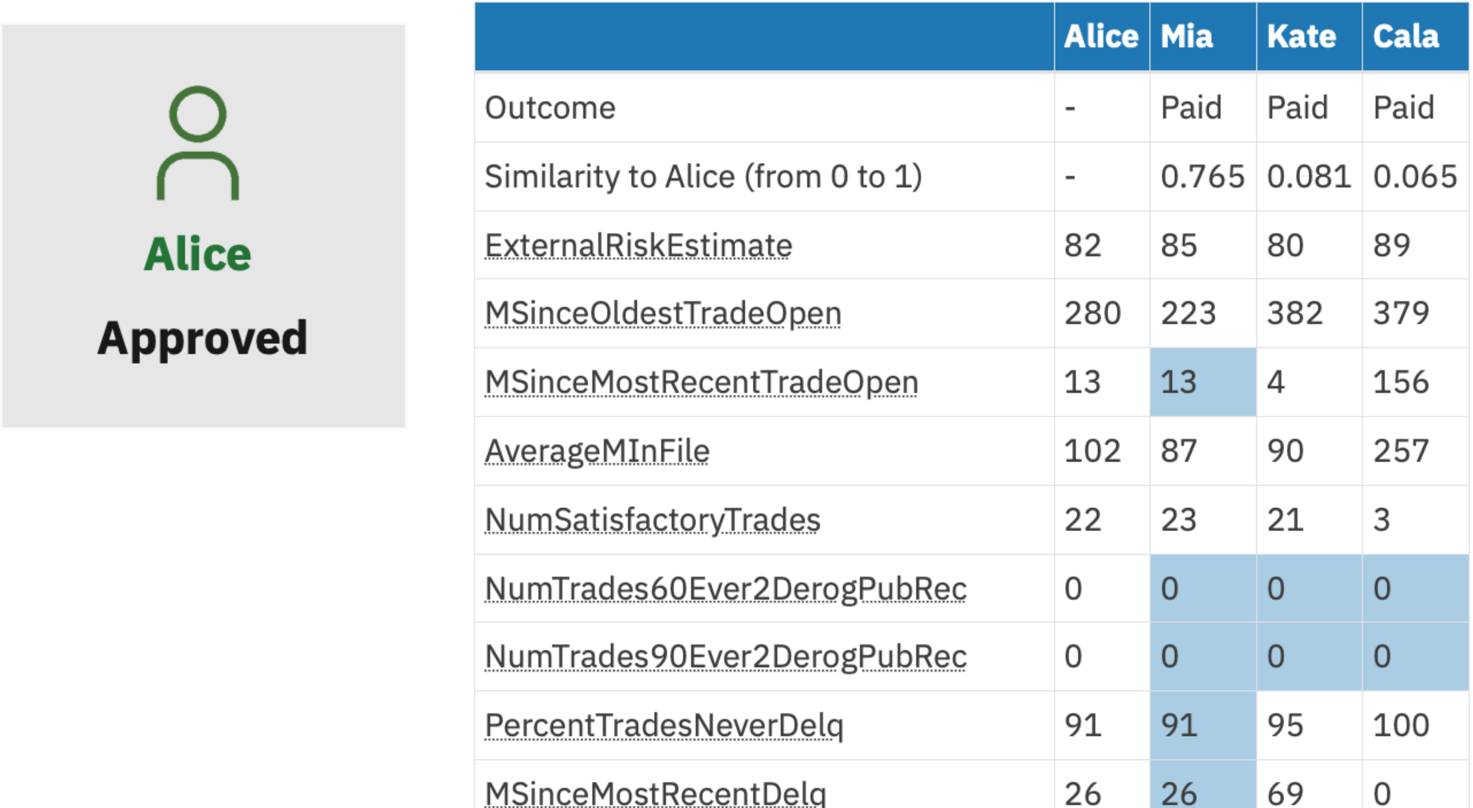
ExternalRiskEstimate is an important feature **positively correlated with good credit risk**.

The jumps in the plot indicate that applicants with above average ExternalRiskEstimate (the mean is 72) get an additional boost.



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?



Alice

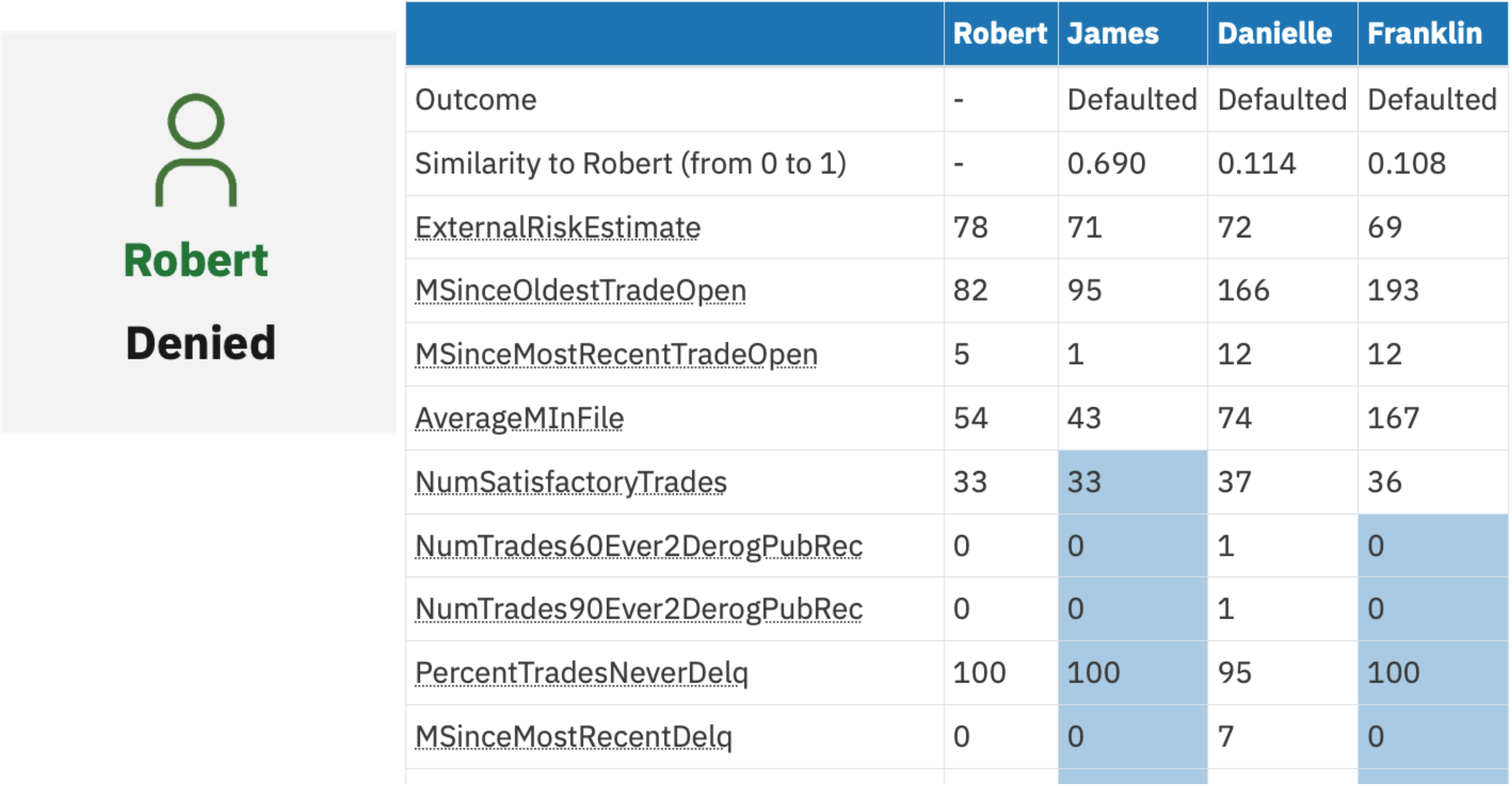
Approved

	Alice	Mia	Kate	Cala
Outcome	-	Paid	Paid	Paid
Similarity to Alice (from 0 to 1)	-	0.765	0.081	0.065
ExternalRiskEstimate	82	85	80	89
MSinceOldestTradeOpen	280	223	382	379
MSinceMostRecentTradeOpen	13	13	4	156
AverageMInFile	102	87	90	257
NumSatisfactoryTrades	22	23	21	3
NumTrades60Ever2DerogPubRec	0	0	0	0
NumTrades90Ever2DerogPubRec	0	0	0	0
PercentTradesNeverDelq	91	91	95	100
MSinceMostRecentDelq	26	26	69	0

@MargrietGr

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?



Robert

Denied

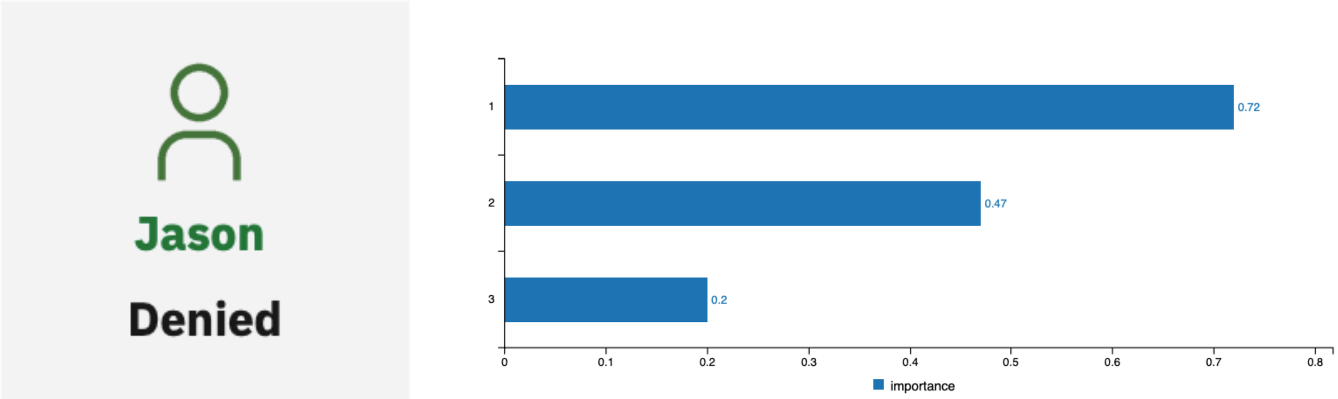
	Robert	James	Danielle	Franklin
Outcome	-	Defaulted	Defaulted	Defaulted
Similarity to Robert (from 0 to 1)	-	0.690	0.114	0.108
ExternalRiskEstimate	78	71	72	69
MSinceOldestTradeOpen	82	95	166	193
MSinceMostRecentTradeOpen	5	1	12	12
AverageMInFile	54	43	74	167
NumSatisfactoryTrades	33	33	37	36
NumTrades60Ever2DerogPubRec	0	0	1	0
NumTrades90Ever2DerogPubRec	0	0	1	0
PercentTradesNeverDelq	100	100	95	100
MSinceMostRecentDelq	0	0	7	0

@MargrietGr

A Bank Customer wants to understand:

Why was my application rejected?

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



Jason

Denied

1. The value of **Consolidated risk markers** is **65**. It needs to be around **72** for the application to be approved.
2. The value of **Average age of accounts in months** is **52**. It needs to be around **68** for the application to be approved.
3. The value of **Months since most recent credit inquiry not within the last 7 days** is **2**. It needs to be around **3** for the application to be approved.

@MargrietGr

Data explanation

ProtoDash

One way to understand a dataset is through prototypes (samples that relay the essence of a dataset) and criticisms (samples that are outliers). The **ProtoDash** algorithm will extract such prototypes and criticisms to **help a consumer** understand a dataset's properties.

DIP-VAE

Sometimes the features in dataset are meaningful to consumers, but other times they are entangled, i.e. multiple meaningful attributes are combined in a single feature. **The Disentangled Inferred Prior Variational Autoencoder (DIP-VAE)** algorithm is an unsupervised representation learning algorithm that will take the given features and learn a new representation that is disentangled in such a way that the resulting features are understandable.

Model explanation

Directly interpretable models are model formats such as decision trees, Boolean rule sets, and generalized additive models, that are easily understood by people and learned straight from the training data.

Post hoc explanation methods first train a black box model and then build another explanation model on top of the black box model.

Global explanations are for entire models whereas **local explanations** are for single sample points.

Local models are the most useful for affected user personas such as patients, defendants, and applicants who need to understand the decision on a single sample (theirs).

Global model explanation

Global **directly interpretable models** are important for **personas that need to understand the entire decision-making process** and ensure its safety, reliability, or compliance. Such personas include regulators and data scientists responsible for the deployment of systems.

Two global directly interpretable model learning algorithms: Boolean Decision Rules via Column Generation (Light Edition) and Generalized Linear Rule Models. Both are applicable for classification problems whereas Generalized Linear Rule Models also applies to regression problems. Both have logical conjunctions, i.e. ‘and’-rules of features as their starting point. Boolean Decision Rules combines ‘and’-rules with a logical ‘or’ whereas Generalized Linear Rule Models combines them with weights. For classification problems, Boolean Decision Rules tends to return simple models that can be quickly understood, whereas Generalized Linear Rule Models can achieve higher accuracy while retaining the interpretability of a linear model.

Global **post hoc explanations** are useful for **decision maker personas** that are being supported by the machine learning model. Physicians, judges, and loan officers develop an overall understanding of how the model works, but there is necessarily a gap between the black box model and the explanation. Therefore, a global post hoc explanation may hide some safety issues, but its antecedent black box model may have favourable accuracy.

One algorithm for producing a global post hoc explanation specifically from a neural network as the base black box model. ProfWeight probes into the neural network and produces instance weights that are then applied to training data to learn a directly interpretable model.

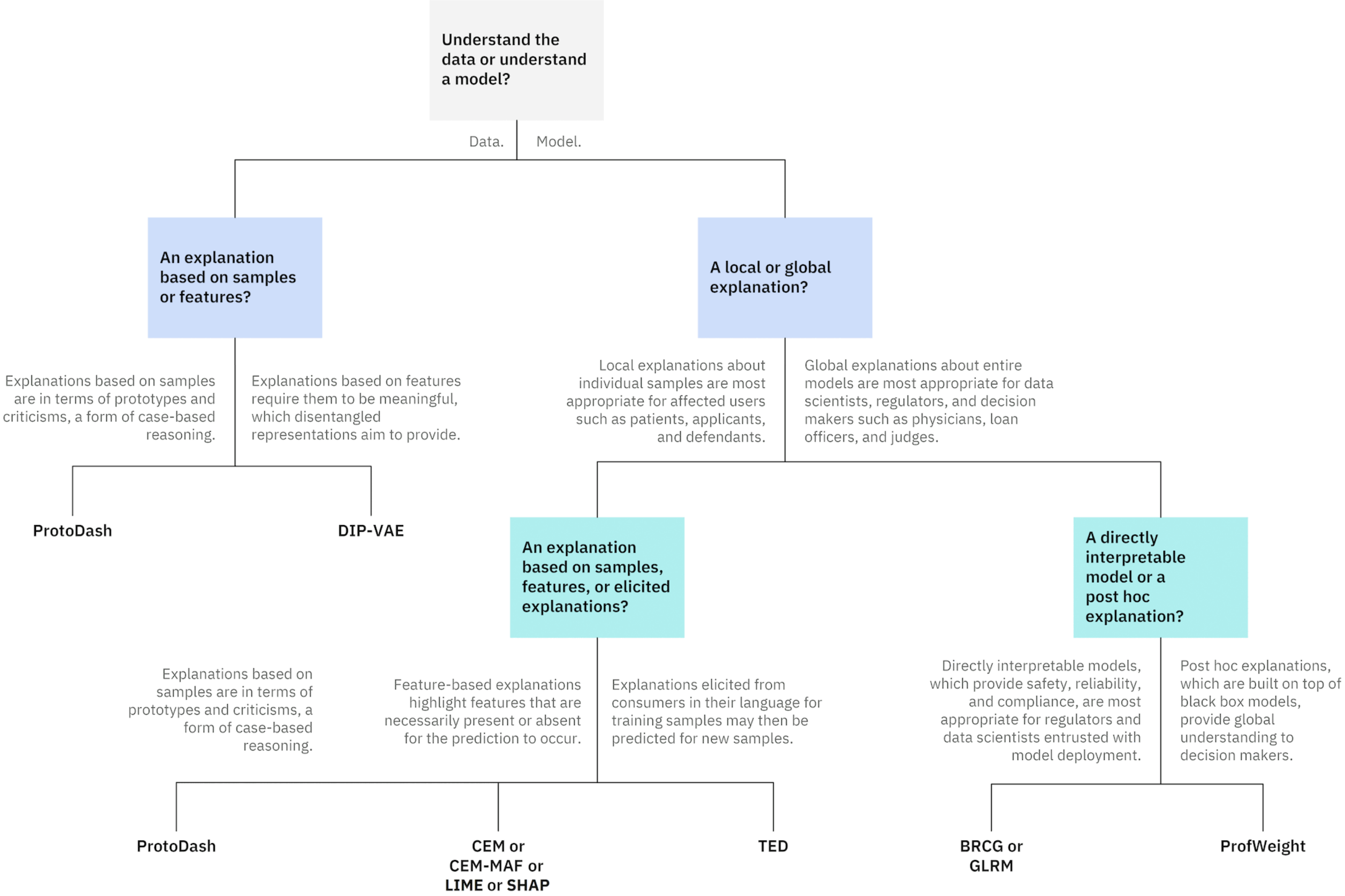
Local model explanation

Local directly interpretable models

The initial release of AIX360 contains one method, Teaching AI to Explain Its Decisions (TED), that directly learns a model to provide explanations at the sample level. This algorithm is unique in that it requires a training set to have not only features and labels, but also training explanations for each sample collected in the language of the consumer. It then predicts an explanation along with a label from the features of new unseen samples.

Local post hoc explanations

Among local post hoc explanation methods, the initial release of AIX360 contains two variants of the Contrastive Explanations Method. The first variant of the Contrastive Explanations Method is the basic version for classification with numerical features and presents minimally sufficient features as well as minimally and critically absent features for a prediction. The second variant, Contrastive Explanations Method with Monotonic Attribute Functions, is specific for image data, with a particular focus on colored images and images with rich structure. ProtoDash, discussed earlier in data explanation, can also be used for local post hoc model explanation via prototypes.



Understand the data or understand a model?

Data.

Model.

An explanation based on samples or features?

A local or global explanation?

Explanations based on samples are in terms of prototypes and criticisms, a form of case-based reasoning.

Explanations based on features require them to be meaningful, which disentangled representations aim to provide.

Local explanations about individual samples are most appropriate for affected users such as patients, applicants, and defendants.

Global explanations about entire models are most appropriate for data scientists, regulators, and decision makers such as physicians, loan officers, and judges.

ProtoDash

DIP-VAE

An explanation based on samples, features, or elicited explanations?

A directly interpretable model or a post hoc explanation?

Explanations based on samples are in terms of prototypes and criticisms, a form of case-based reasoning.

Feature-based explanations highlight features that are necessarily present or absent for the prediction to occur.

Explanations elicited from consumers in their language for training samples may then be predicted for new samples.

Directly interpretable models, which provide safety, reliability, and compliance, are most appropriate for regulators and data scientists entrusted with model deployment.

Post hoc explanations, which are built on top of black box models, provide global understanding to decision makers.

ProtoDash

CEM or CEM-MAF or LIME or SHAP

TED

BRCG or GLRM

ProfWeight

@MargrietGr

source: IBM Research AI Explainability 360

Resources

AIX360

https://aix360.mybluemix.net

https://github.com/Trusted-AI/AIX360

https://developer.ibm.com/open/projects/ai-explainability/

IBM Cloud and Cloud Pak for Data aaS

http://ibm.biz/explainai_mg

https://dataplatform.cloud.ibm.com/ OR

https://eu-gb.dataplatform.cloud.ibm.com/

https://developer.ibm.com/

Interpretable Machine Learning

https://christophm.github.io/interpretable-ml-book/

AIX360 workshop

https://github.com/IBMDeveloperUK/AIX360-workshop

AIF360 workshop

https://github.com/IBMDeveloperUK/AIF360-workshop