Fair and Explainable AI

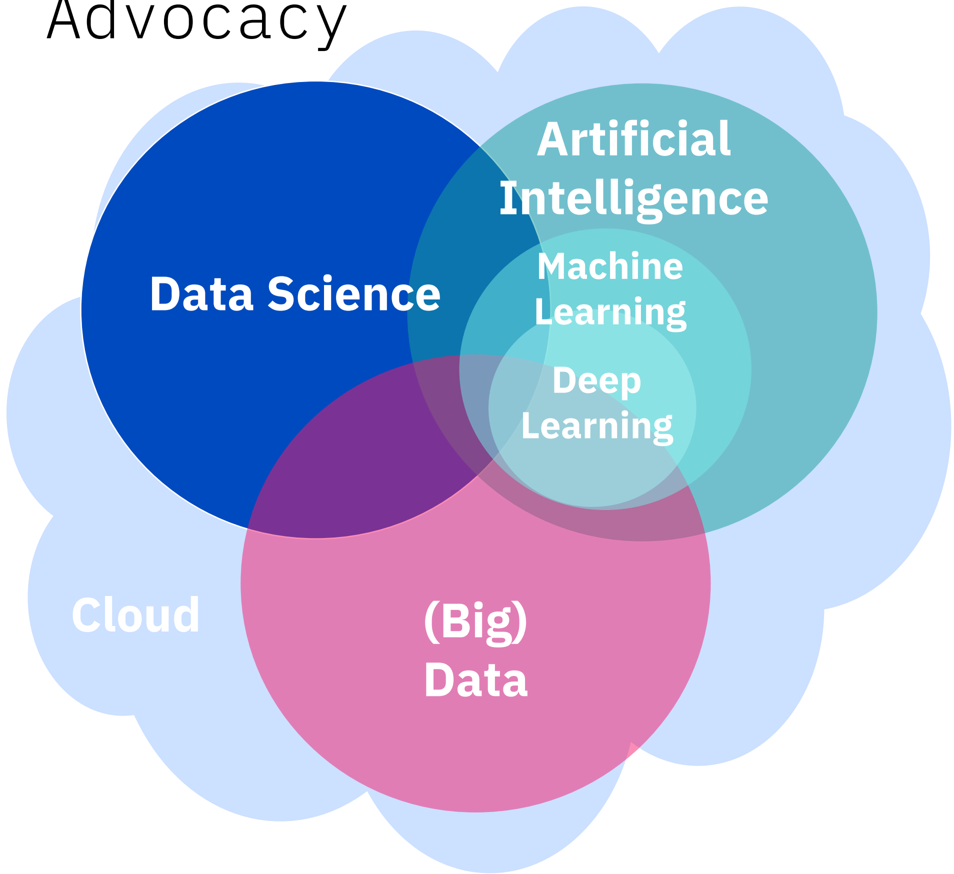
Dr. Margriet Groenendijk

Data Science & AI Developer Advocate

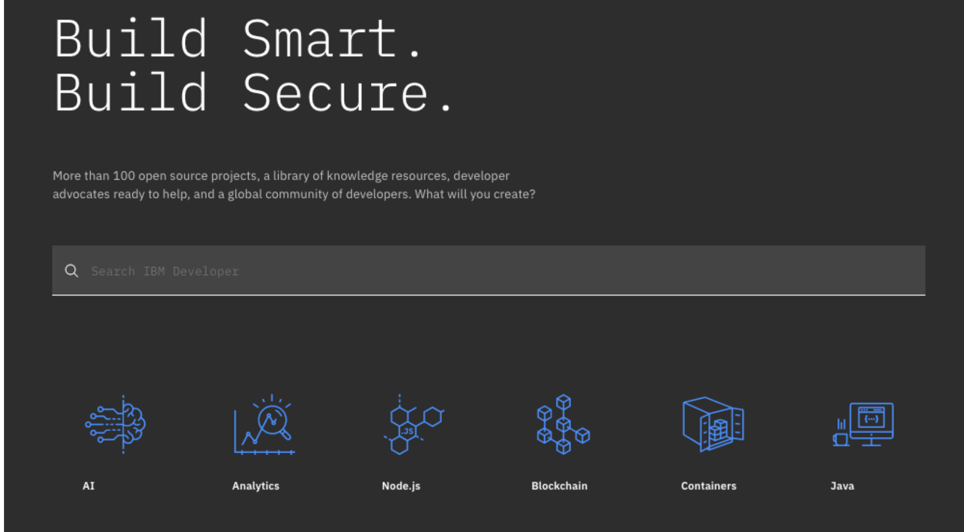
IBM

@MargrietGr

Data & AI Developer Advocacy



@MargrietGr



developer.ibm.com

Code patterns
Tutorials
Blogs, articles
Models, data
Open source projects
Events, podcasts, videos

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

AI is used in many decision-making applications

Credit

Employment

Admission

Sentencing

Healthcare

Fair and explainable AI pipelines

@MargrietGr

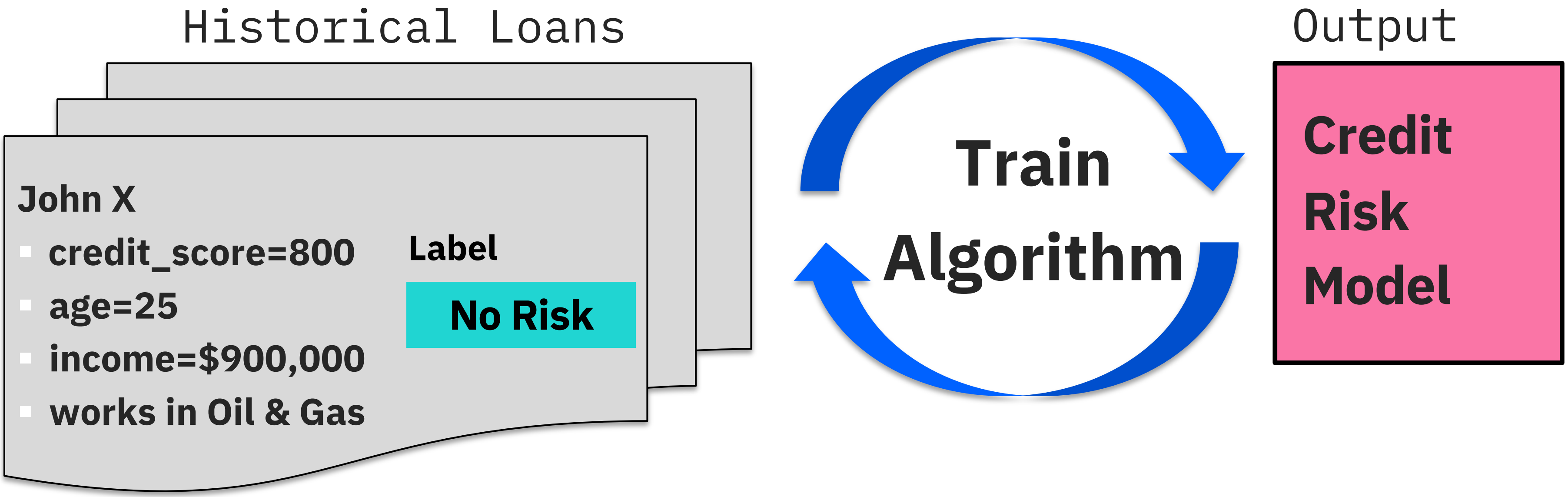
Machine learning
Algorithm selection

Deep learning
Neural network design

Natural Language Processing
Interactions between computers and human languages

Artificial intelligence
Systems architecture

Example: credit risk



@MargrietGr

Example: credit risk

New Applicant

James Y

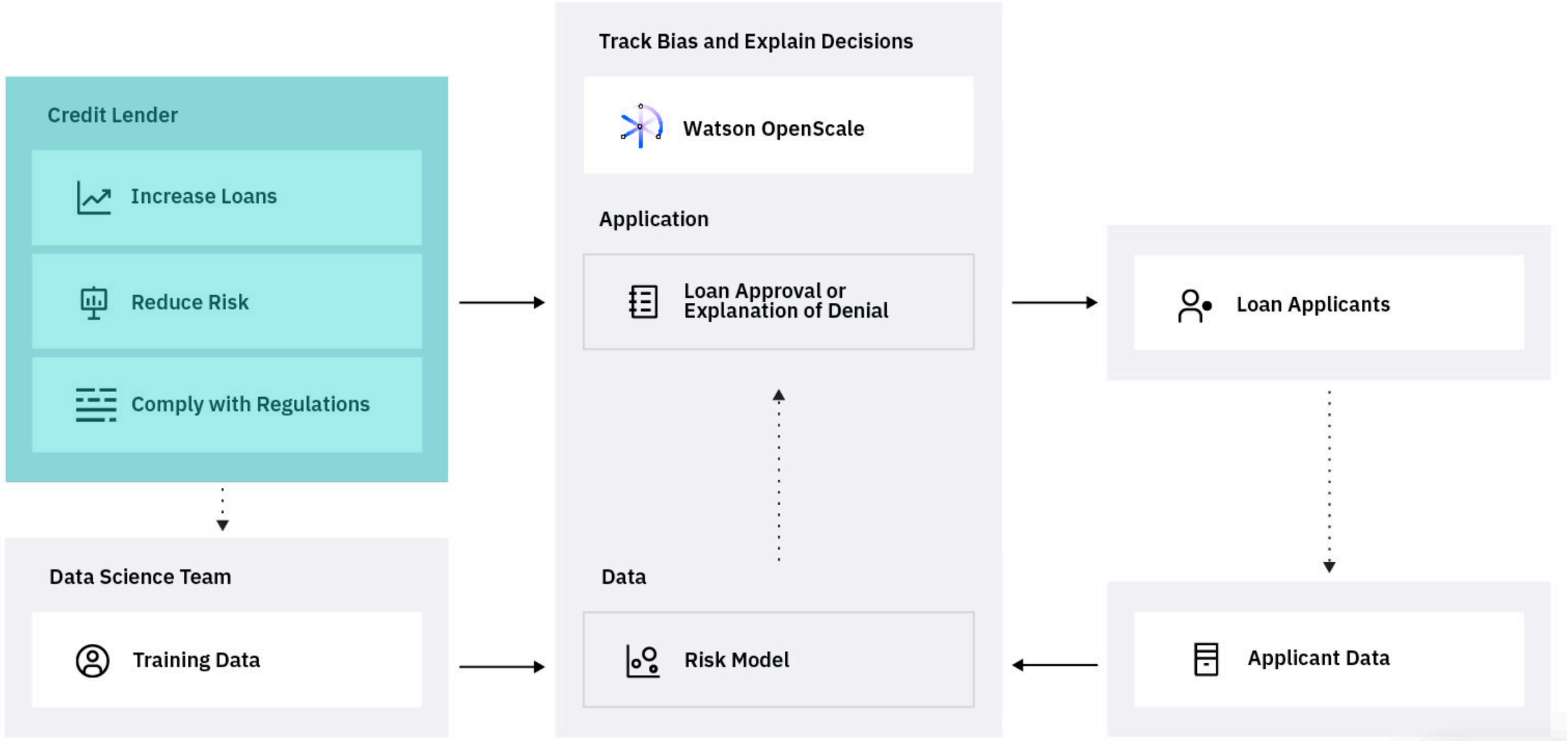
- **credit_score=900**
- **age=55**
- **income=$1,200,000**
- **works in Insurance**

Credit Risk Model

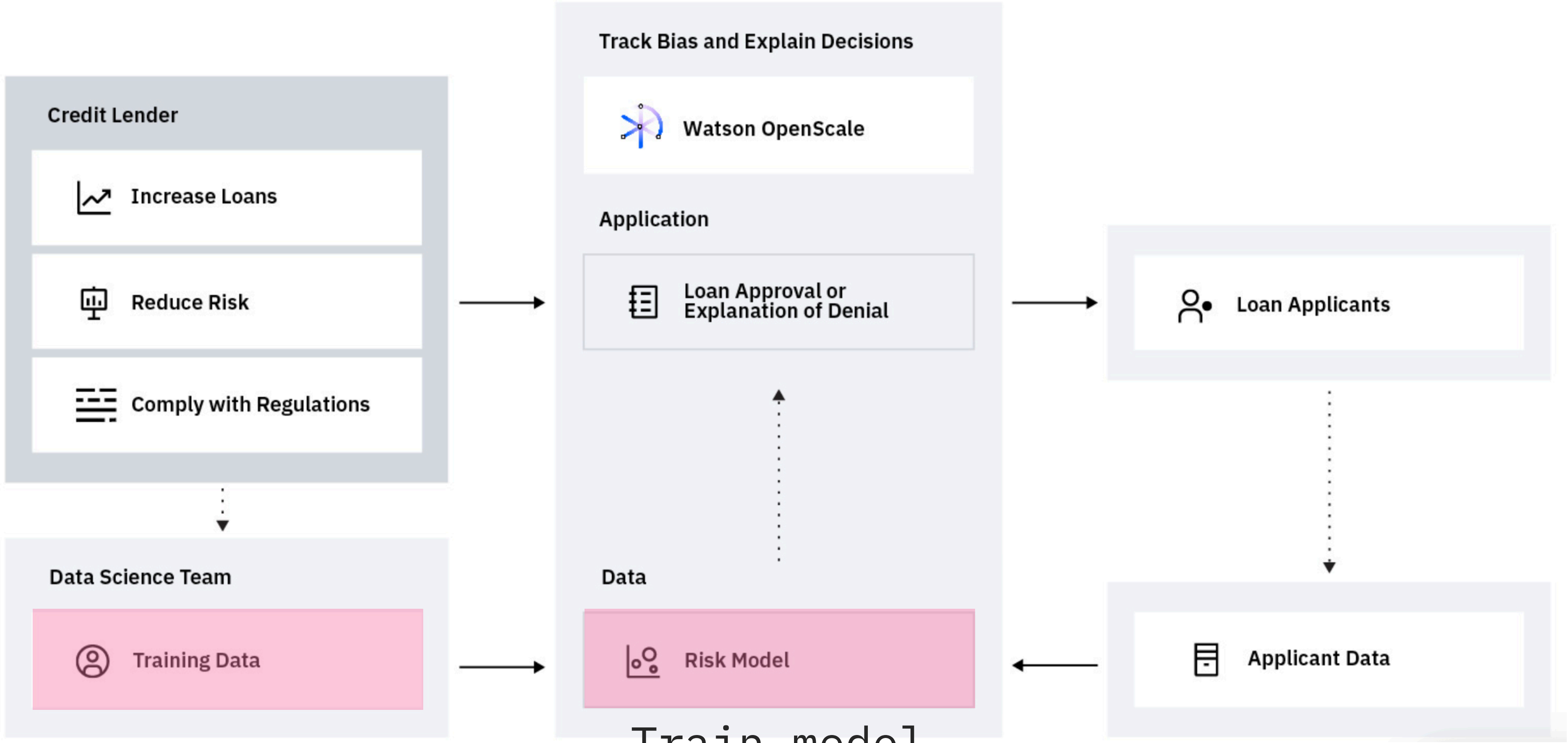
Output

No Risk

Risk

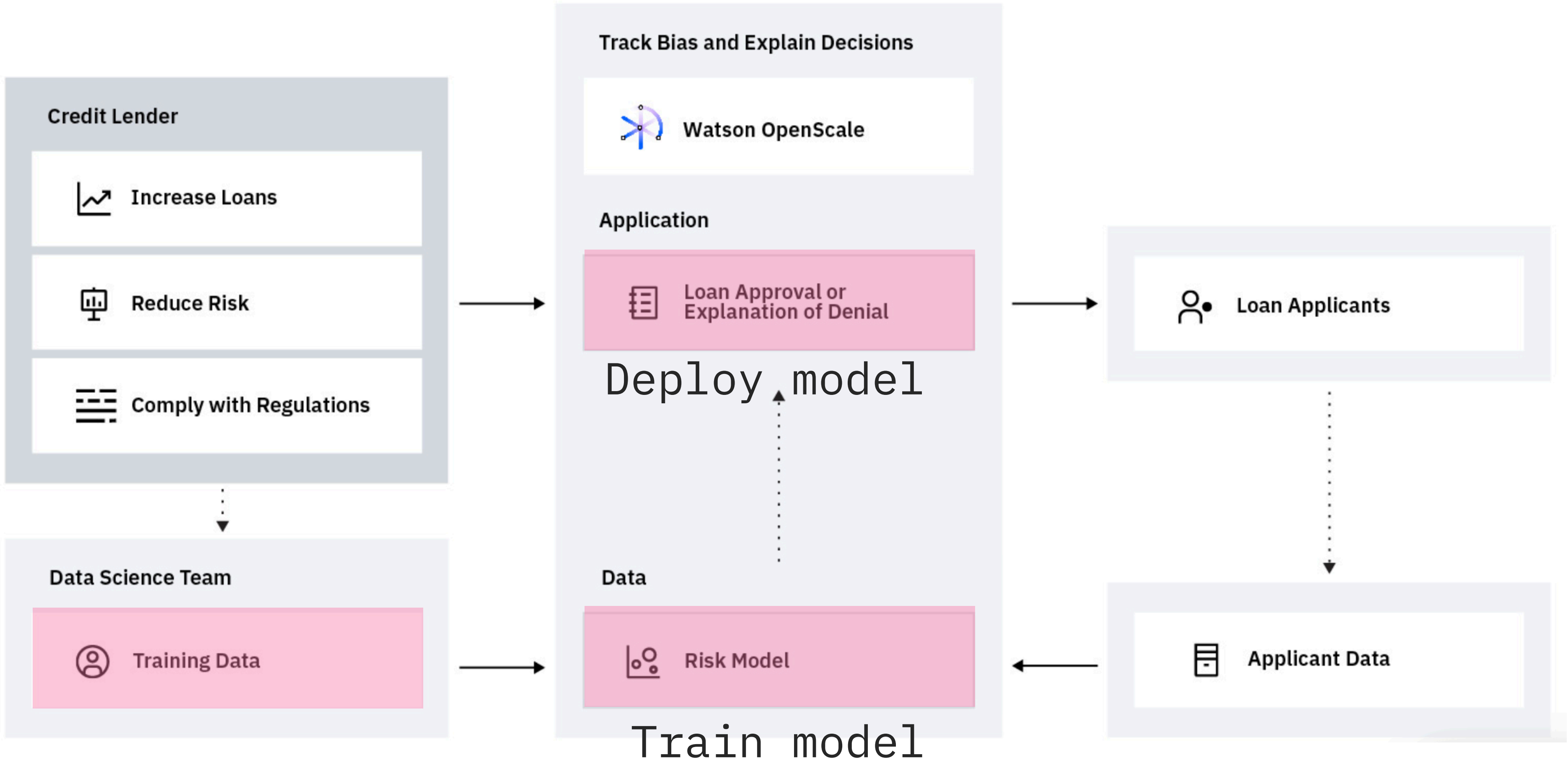

Track Bias and Explain Decisions
Watson OpenScale
Credit Lender
Increase Loans
Reduce Risk
Comply with Regulations
Application
Loan Approval or Explanation of Denial
Loan Applicants
Data Science Team
Training Data
Data
Risk Model
Applicant Data

@MargrietGr

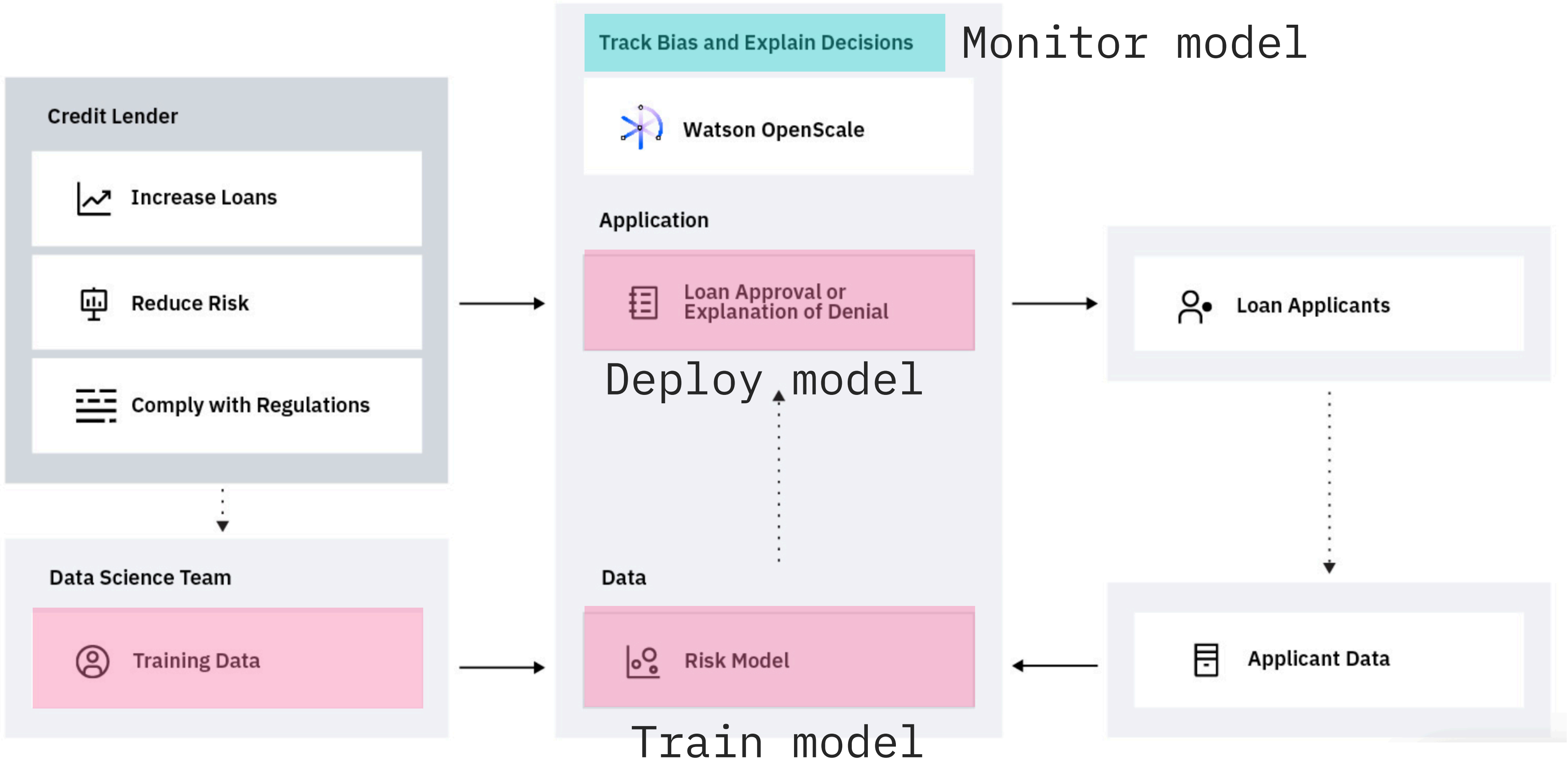


Train model

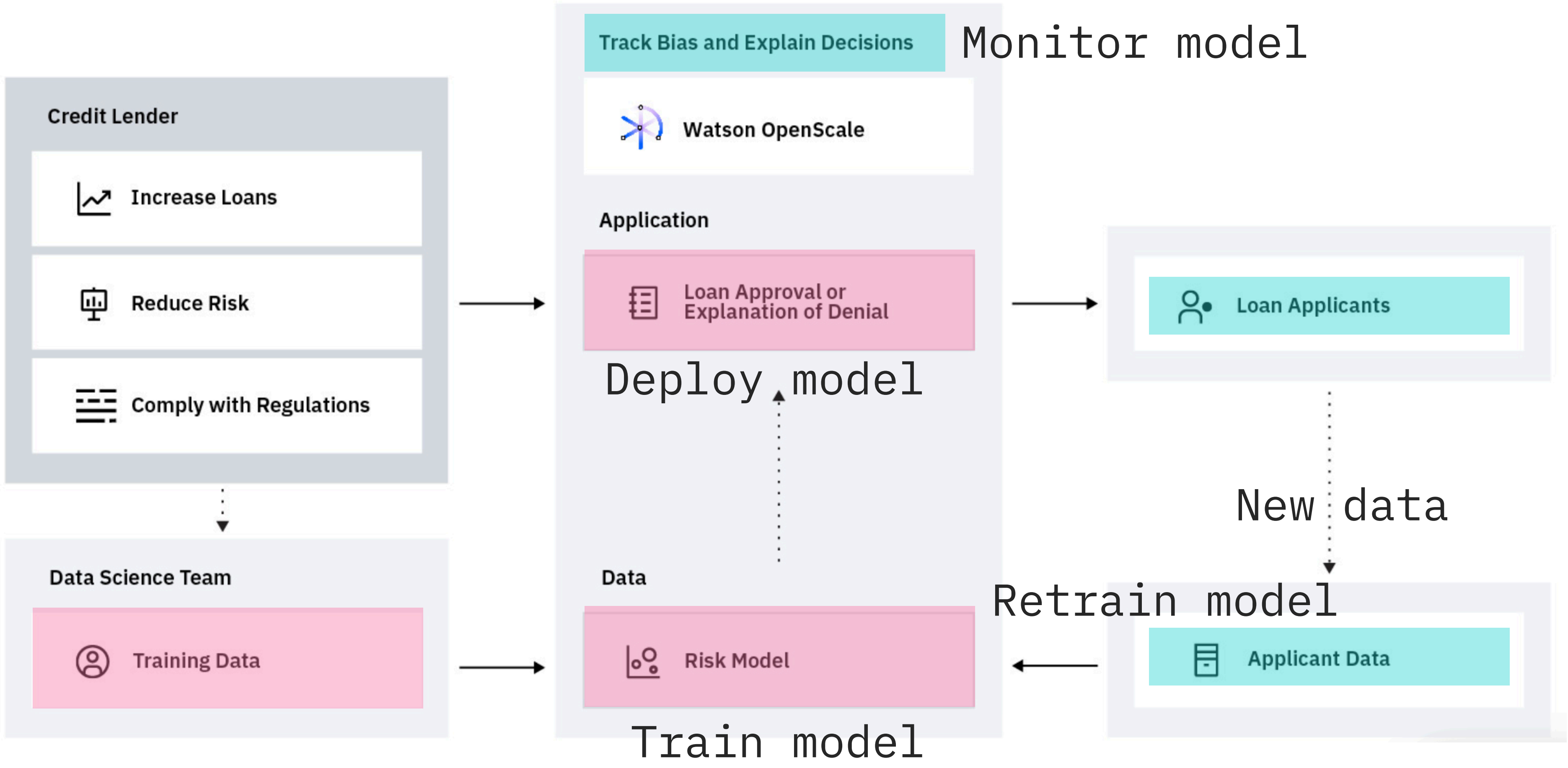
@MargrietGr


Credit Lender
Increase Loans
Reduce Risk
Comply with Regulations
Data Science Team
Training Data
Track Bias and Explain Decisions
Watson OpenScale
Application
Loan Approval or
Explanation of Denial
Deploy model
Data
Risk Model
Train model
Loan Applicants
Applicant Data

@MargrietGr

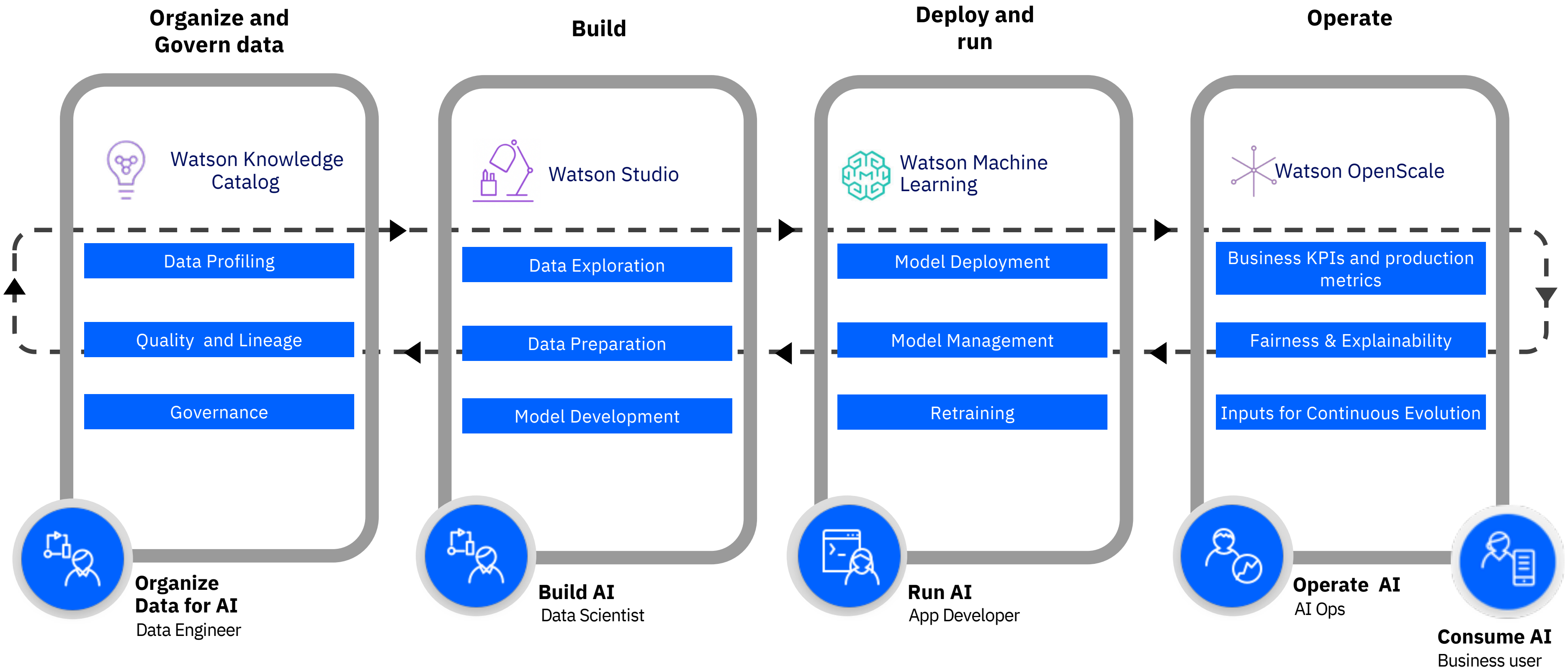

Track Bias and Explain Decisions
Monitor model
Watson OpenScale
Credit Lender
Increase Loans
Reduce Risk
Comply with Regulations
Application
Loan Approval or Explanation of Denial
Deploy model
Loan Applicants
Data Science Team
Training Data
Data
Risk Model
Applicant Data
Train model

@MargrietGr


Track Bias and Explain Decisions
Monitor model
Watson OpenScale
Credit Lender
Increase Loans
Reduce Risk
Comply with Regulations
Application
Loan Approval or Explanation of Denial
Deploy model
Loan Applicants
New data
Data Science Team
Training Data
Data
Risk Model
Retrain model
Applicant Data
Train model

@MargrietGr

AI pipeline



Organize and Govern data
Build
Deploy and run
Operate
Watson Knowledge Catalog
Watson Studio
Watson Machine Learning
Watson OpenScale
Data Profiling
Quality and Lineage
Governance
Data Exploration
Data Preparation
Model Development
Model Deployment
Model Management
Retraining
Business KPIs and production metrics
Fairness & Explainability
Inputs for Continuous Evolution
Organize Data for AI
Data Engineer
Build AI
Data Scientist
Run AI
App Developer
Operate AI
AI Ops
Consume AI
Business user

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

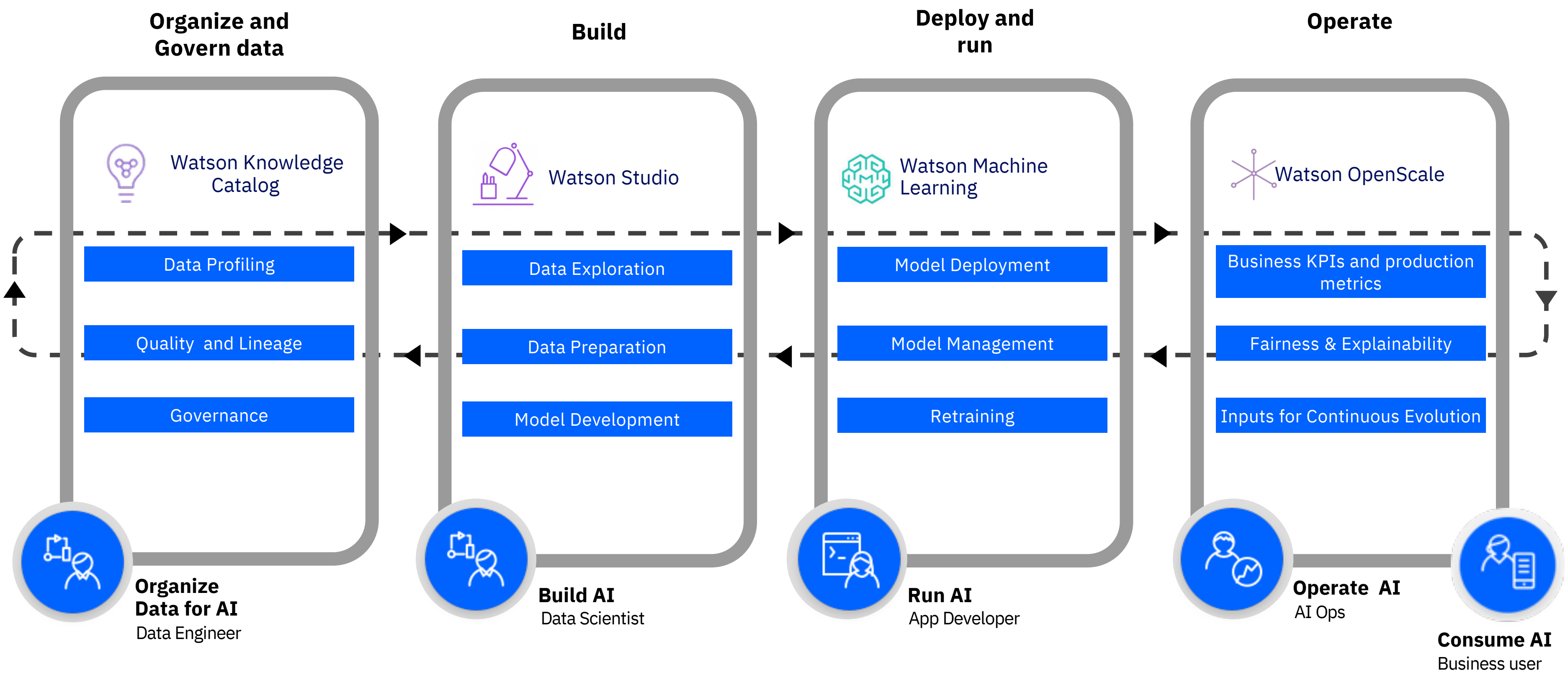
- Red Hat OpenShift
- Red Hat Enterprise Linux

Infrastructure

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

@MargrietGr

AI pipeline in Cloud Pak for Data (aaS)



Fair and explainable AI pipelines

Is your model treating different classes fairly?

WILL KNIGHT BUSINESS 11.19.2019 09:15 AM

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

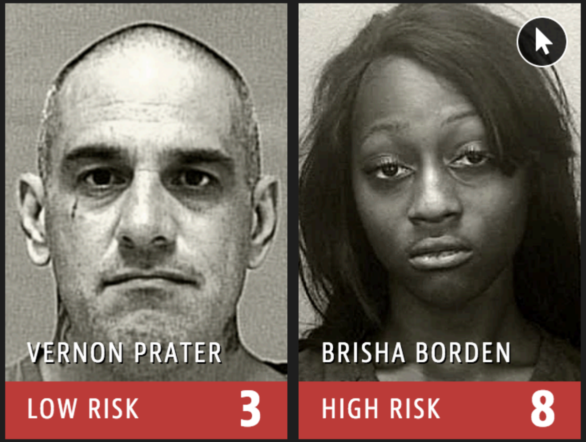
The way its algorithm determines credit lines makes the risk of bias more acute.

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

Can you explain your model results?

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

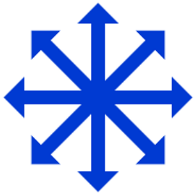
https://www.bbc.co.uk/news/education-53787203

@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

Agenda

08:45 - 09:00: Enrolment & Setup

09:00 - 09:10: Introductory remarks

09:10 - 09:30: Fair and Explainable AI

09:30 - 10:15: Remove Unfair Bias in Machine Learning

10:15 - 10:30: Break

10:30 - 11:05: Explain Machine Learning Models

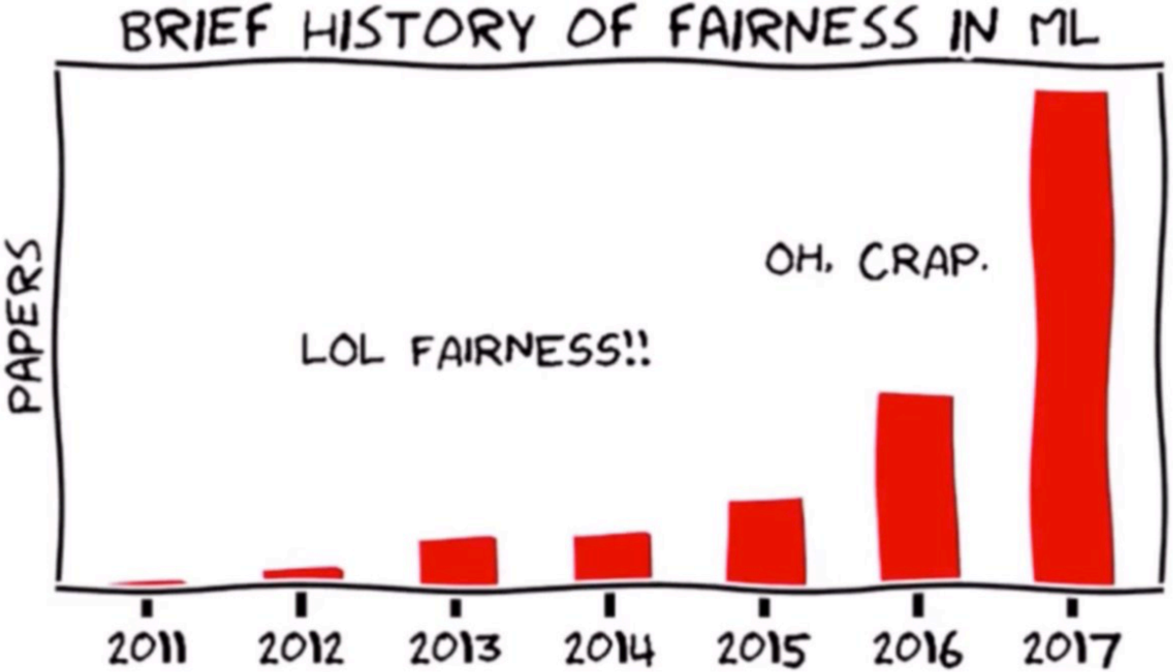
11:05 - 11:40: Build a machine learning model and monitor the performance, bias and drift

11:40 - 11:50: Summary & Next Steps including Q&A

11:50 - 12:00: Closing remarks

https://margriet-groenendijk.gitbook.io/trusted-ai-workshop

Part 1: Remove Unfair Bias in Machine Learning


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

(Hardt, 2017)

What is Fairness?

There are 21 definitions of fairness

Many of the definitions conflict

The way you define fairness impacts bias

AI Fairness 360
↳ (AIF360)

https://github.com/IBM/AIF360

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

http://aif360.mybluemix.net/

Extensible Toolkit for Detecting, Understanding, & Mitigating Unwanted **Algorithmic Bias**

Leading Fairness Metrics and Algorithms from **Industry** & **Academia**

Designed to **translate new research** from the **lab to industry practitioners** (using Scikit Learn's fit/predict paradigm)

Fairness Terms

Protected Attribute – an attribute that partitions a population into groups whose outcomes should have parity (ex. race, gender, caste, and religion)

Privileged Protected Attribute – a protected attribute value indicating a group that has historically been at systemic advantage

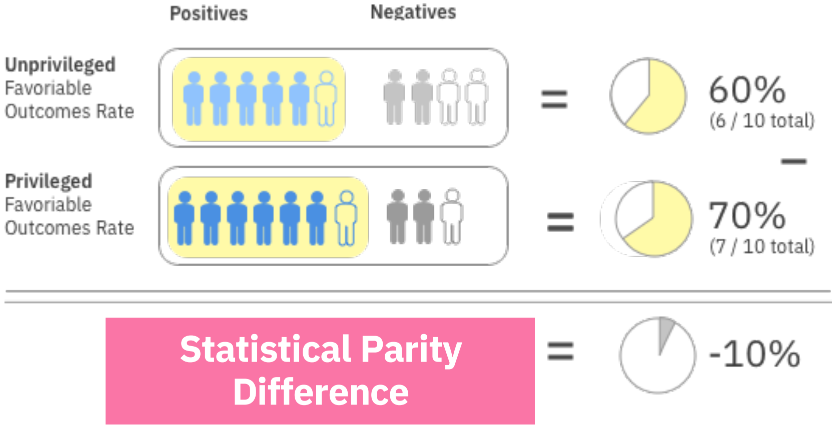
Group Fairness – Groups defined by protected attributes receiving similar treatments or outcomes

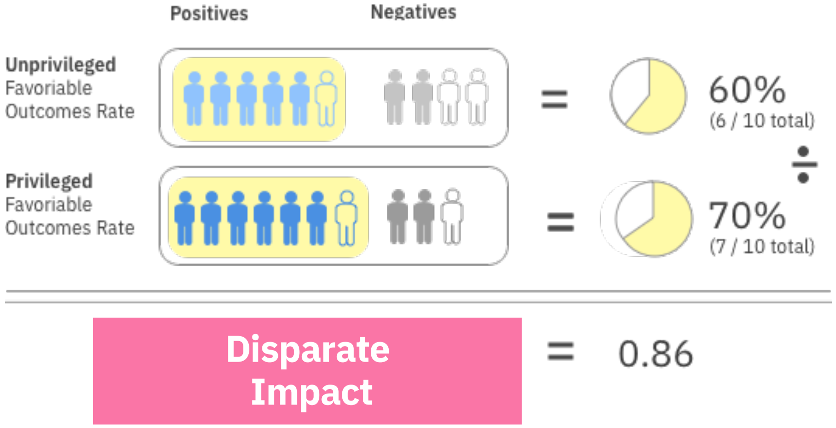
Individual Fairness – Similar individuals receiving similar treatments or outcomes

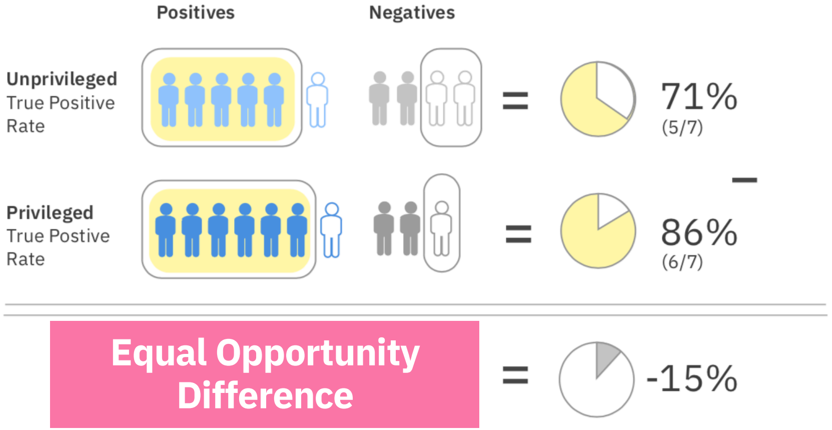
Fairness Metric – a measure of unwanted bias in training data or models

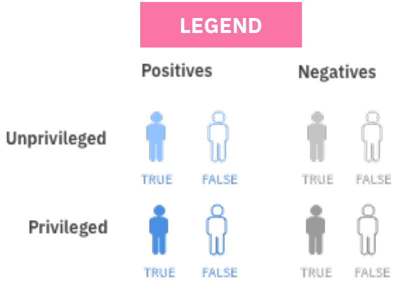
Favorable Label – a label whose value corresponds to an outcome that provides an advantage to the recipient

How To Measure Fairness – Some Group Fairness Metrics









@MargrietGr

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

Tradeoffs - Bias vs. Accuracy

1. Is your model doing good things or bad things to people?
   - If your model is sending people to jail, may be better to have more false positives than false negatives
   - If your model is handing out loans, may be better to have more False Negatives than False Positives
2. Determine your threshold for accuracy vs. fairness based upon your legal, ethical and trust guidelines

LEGAL
Doing what is legal is top priority (Penalties)

ETHICAL
What's your company's Ethics (Amazon Echo)

TRUST
Losing customer's Trust costly (Facebook)

Preventing Bias Is Hard!

Work with your stakeholders early to define fairness, protected attributes & thresholds

Apply the earliest mitigation in the pipeline that you have permission to apply

Check for bias as often as possible using any metrics that are applicable

Caveat: AIF360 should only be used with well defined data sets & well-defined use cases

@MargrietGr

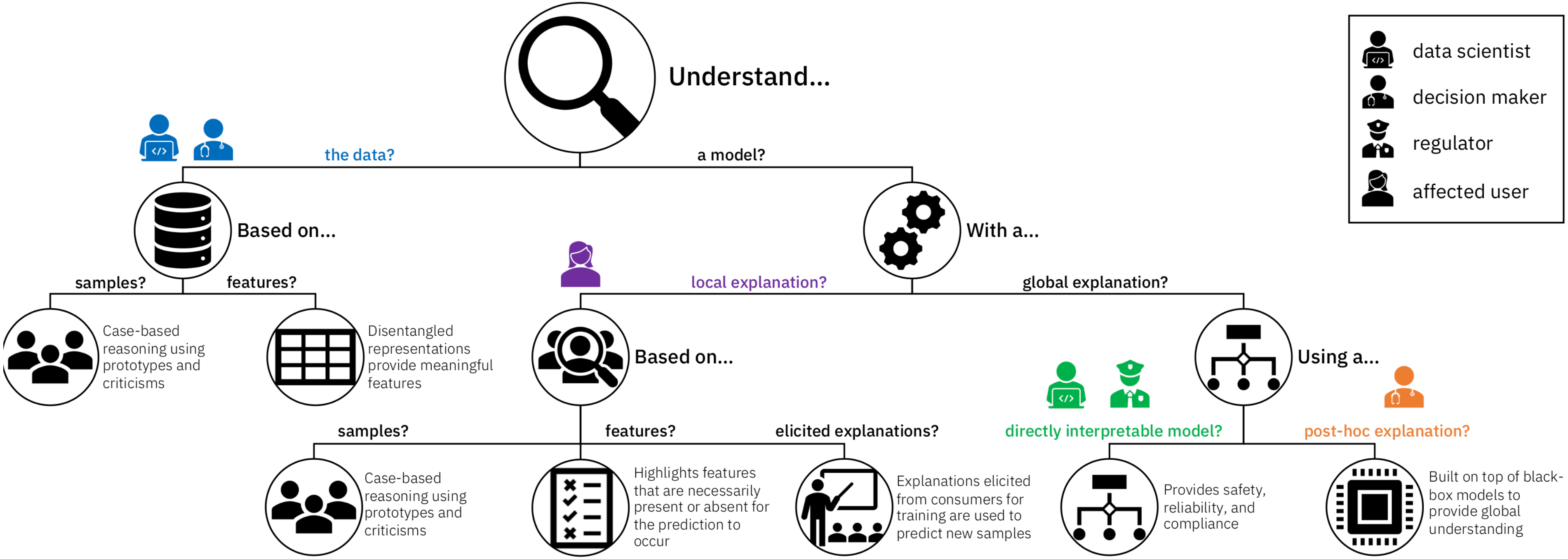
Part 1: Remove Unfair Bias in Machine Learning

Hands-on

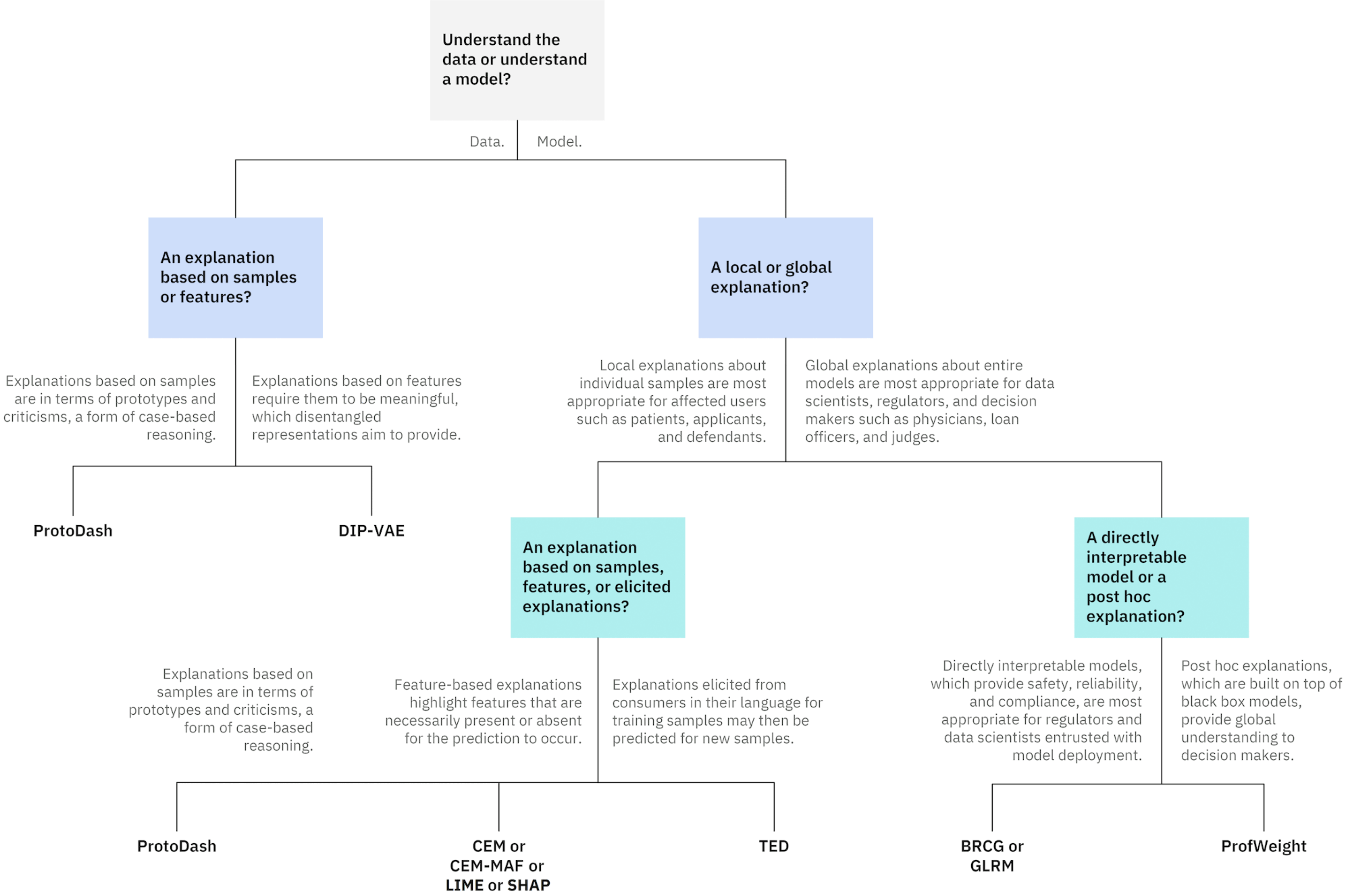
Break until 10:30

@MargrietGr

Part 2: Explain Machine Learning Models


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



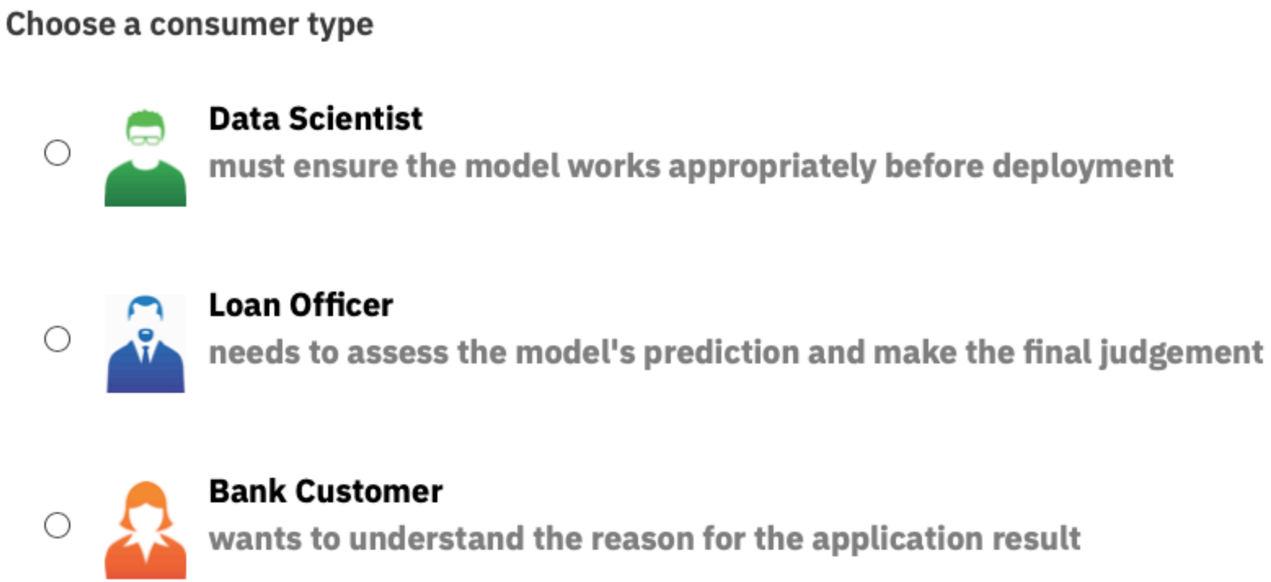
@MargrietGr

source: IBM Research AI Explainability 360

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

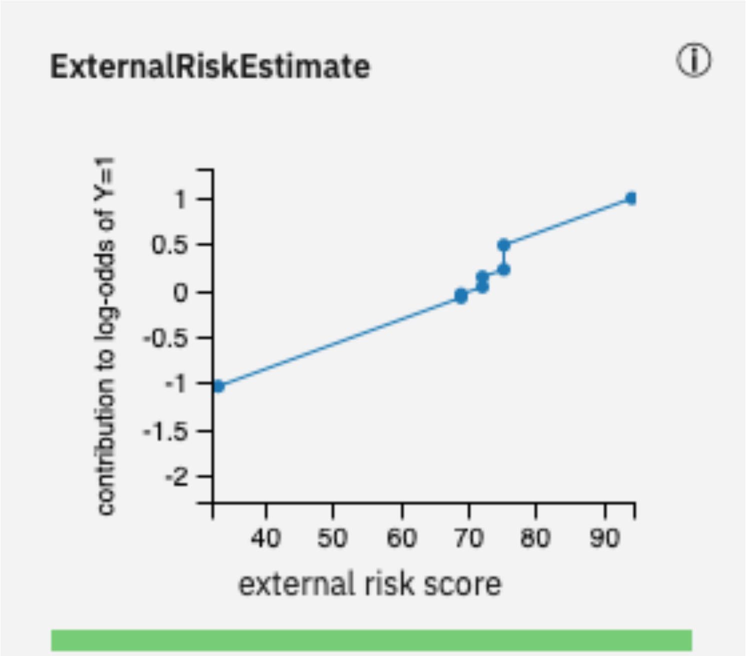


A Data Scientist wants to understand:

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.



@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?

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

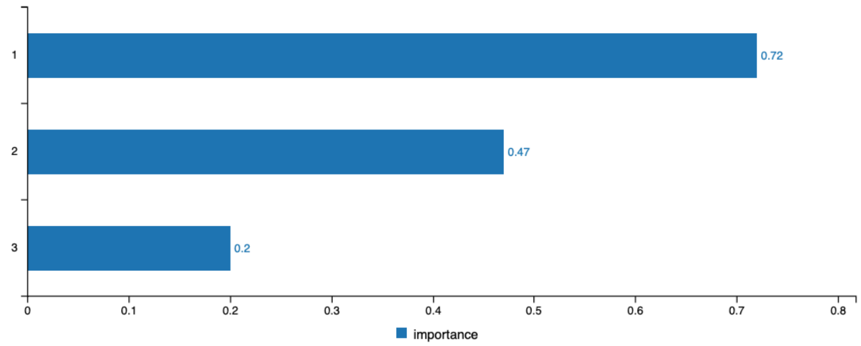
	Robert	James	Danielle	Franklin
Outcome	-	Defaulted	Defaulted	Defaulted
Similarity to Robert (from 0 to 1)	-	0.690	0.114	0.108
ExternalRiskEstimate	78	71	72	69
MSinceOldestTradeOpen	82	95	166	193
MSinceMostRecentTradeOpen	5	1	12	12
AverageMInFile	54	43	74	167
NumSatisfactoryTrades	33	33	37	36
NumTrades60Ever2DerogPubRec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



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.

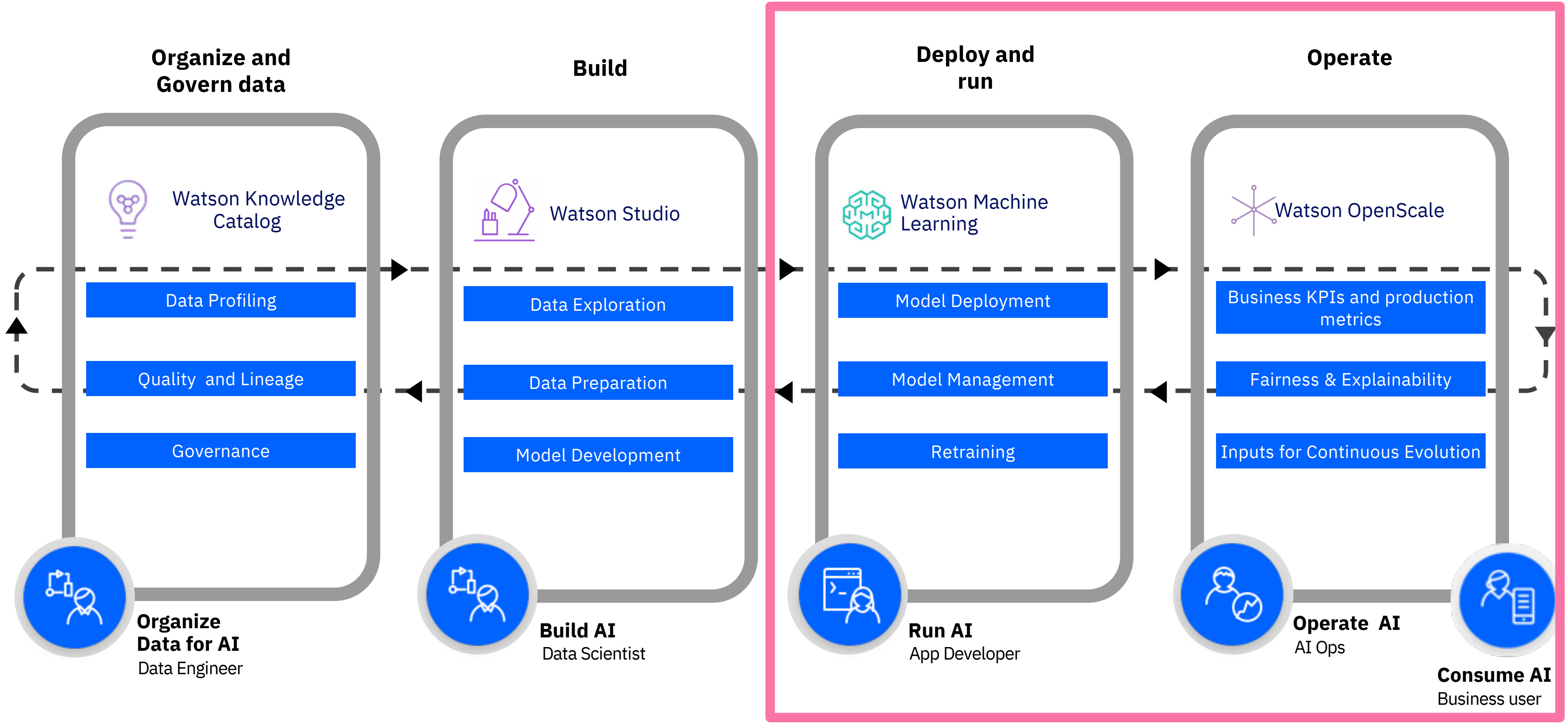
@MargrietGr

Part 2: Explain Machine Learning Models

Hands-on

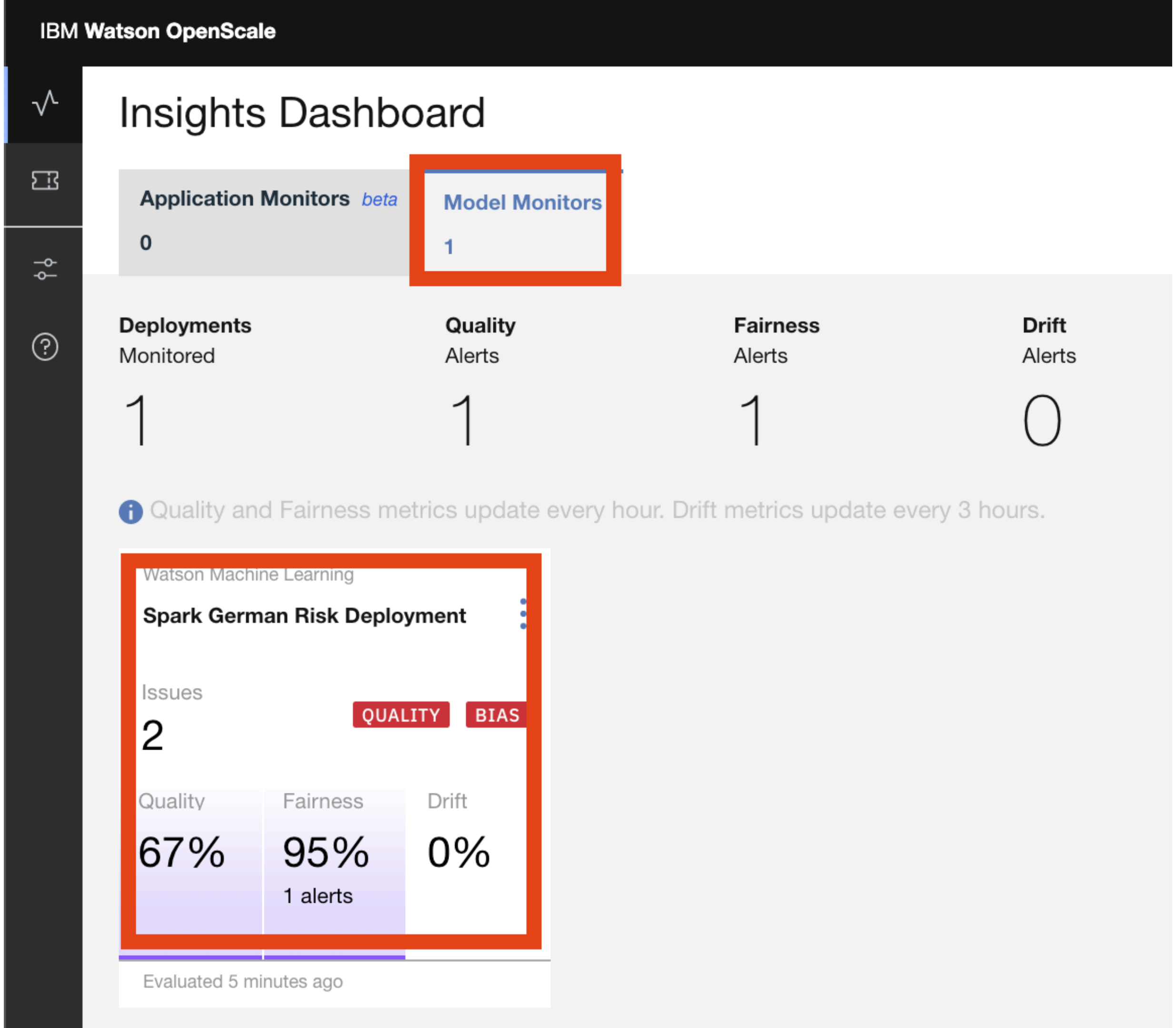
Part 3: Monitor the performance, bias and drift

AI pipeline in Cloud Pak for Data (aaS)



Part 3: Monitor the performance, bias and drift

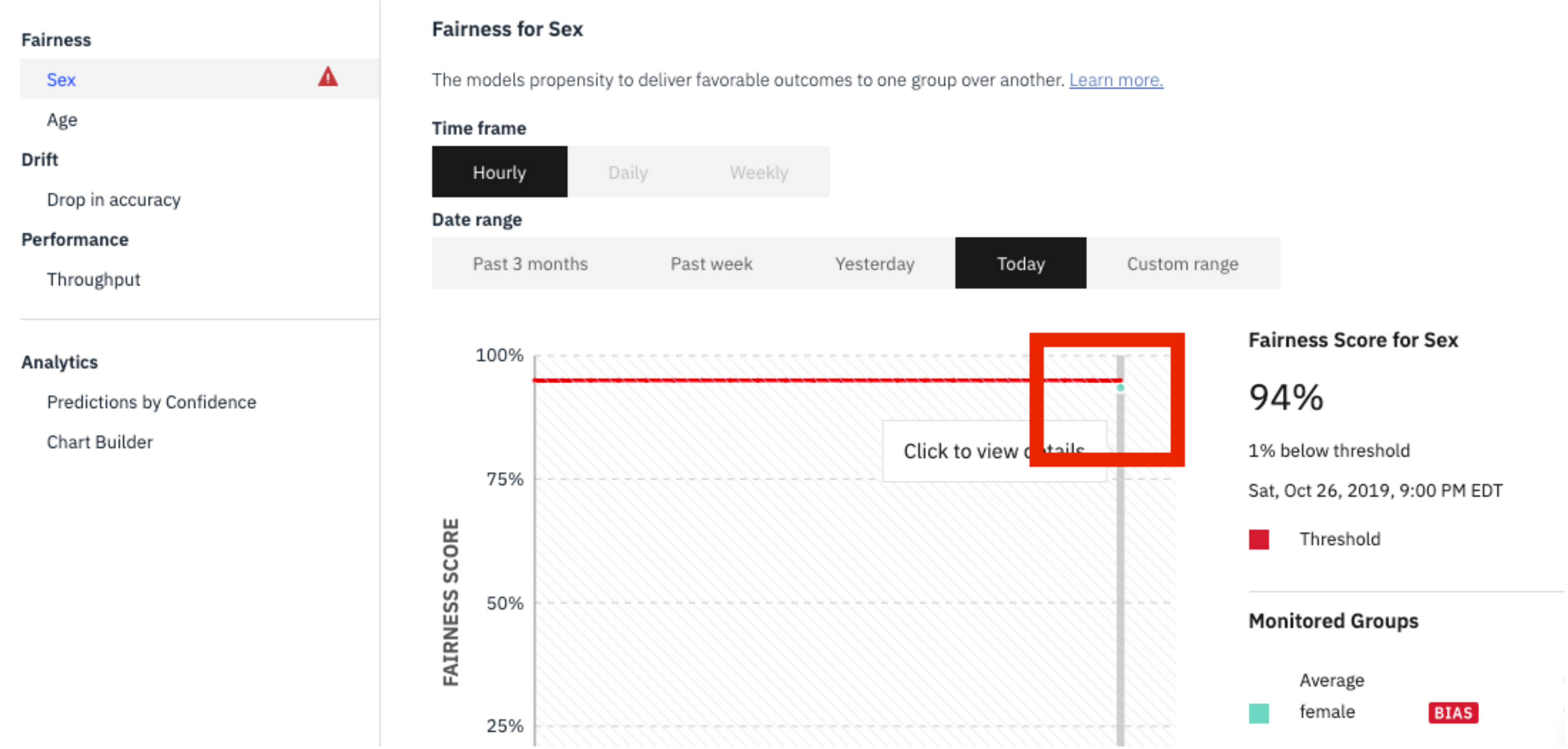
Hands-on

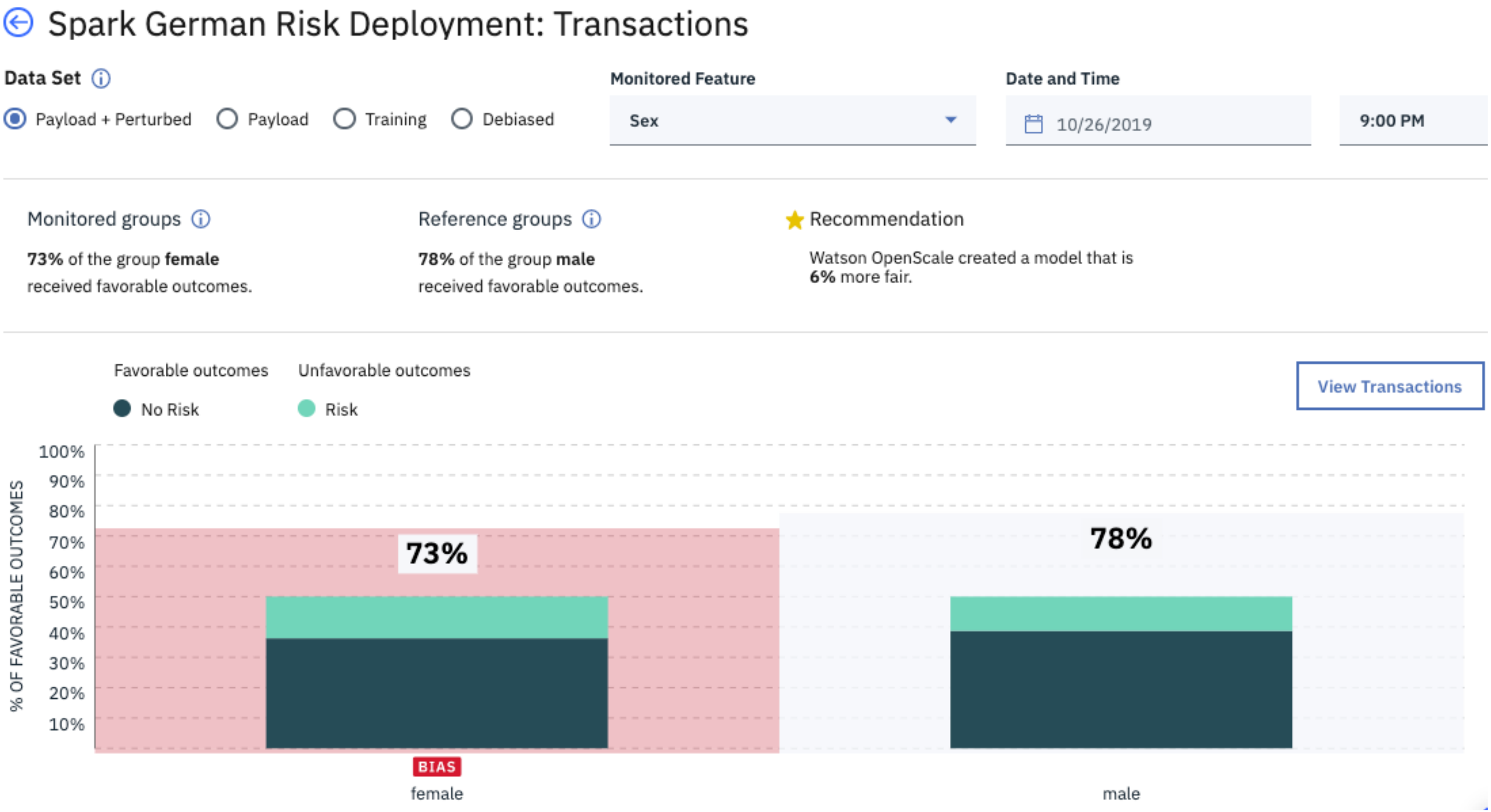

IBM Watson OpenScale
Insights Dashboard
Application Monitors beta
0
Model Monitors
1
Deployments
Monitored
1
Quality
Alerts
1
Fairness
Alerts
1
Drift
Alerts
0
Quality and Fairness metrics update every hour. Drift metrics update every 3 hours.
Watson Machine Learning
Spark German Risk Deployment
Issues
2
QUALITY
BIAS
Quality
67%
Fairness
95%
1 alerts
Drift
0%
Evaluated 5 minutes ago

@MargrietGr

Date range

Past 3 months | Past week | Yesterday | Today | Custom range


Fairness
Sex
Age
Drift
Drop in accuracy
Performance
Throughput
Analytics
Predictions by Confidence
Chart Builder
Fairness for Sex
The models propensity to deliver favorable outcomes to one group over another. Learn more.
Time frame
Hourly
Daily
Weekly
Date range
Past 3 months
Past week
Yesterday
Today
Custom range
100%
75%
50%
25%
FAIRNESS SCORE
Fairness Score for Sex
94%
1% below threshold
Sat, Oct 26, 2019, 9:00 PM EDT
Threshold
Monitored Groups
Average
female
BIAS


Spark German Risk Deployment: Transactions
Data Set
Payload + Perturbed
Payload
Training
Debiased
Monitored Feature
Sex
Date and Time
10/26/2019
9:00 PM
Monitored groups
73% of the group female received favorable outcomes.
Reference groups
78% of the group male received favorable outcomes.
Recommendation
Watson OpenScale created a model that is 6% more fair.
Favorable outcomes
No Risk
Unfavorable outcomes
Risk
View Transactions
% OF FAVORABLE OUTCOMES
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
73%
78%
BIAS
female
male

Spark German Risk Deployment: Transactions

October 31, 2019, 2:00 AM

Sex

View

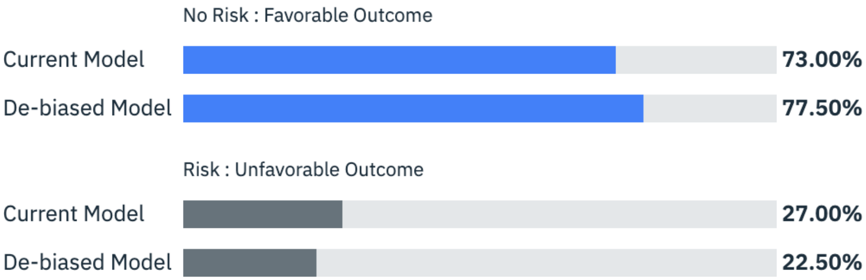
All transactions

Biased transactions

This subset of transactions received biased outcomes. Click the Explain link to determine how the monitored feature contributed to each unfavorable outcome.

Transaction ID	Type	Outcome	Action
dc552a8ff80c6d30a1a15c875f7ed6c3-168	Original	Risk	Explain
dc552a8ff80c6d30a1a15c875f7ed6c3-53	Altered	Risk	Explain
dc552a8ff80c6d30a1a15c875f7ed6c3-199	Altered	Risk	Explain
dc552a8ff80c6d30a1a15c875f7ed6c3-37	Original	No Risk	Explain
dc552a8ff80c6d30a1a15c875f7ed6c3-101	Altered	No Risk	Explain

Fairness Correction Table Manual_Labeling_430ce9f4-72c6-48fa-b98e-662a18211bdb



dc552a8ff80c6d30a1... ×

Details

Transaction	dc552a8ff80c6d30a1a15c875f7ed6c3-168
Deployment	Spark German Risk Deployment
Model Name	Spark German Risk Model
Type	Original

Minimum changes for No Risk outcome

LoanDuration	21.0
Sex	male
InstallmentPercent	3.0

Maximum changes allowed for the same outcome

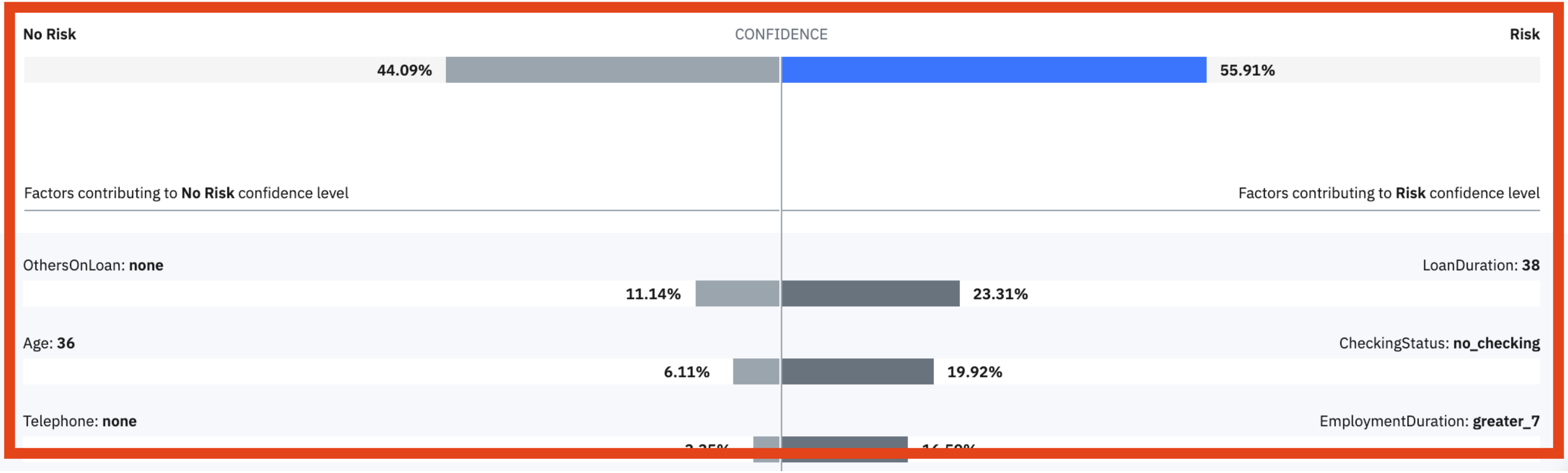
CheckingStatus	no_checking
LoanDuration	38.0
CreditHistory	credits_paid_to_date

How this prediction was determined

The **Spark German Risk Model** predicts **Risk** with 55.91% confidence. The following features were most important in determining this prediction: LoanDuration (23.31%), CheckingStatus (19.92%), and EmploymentDuration (16.59%).

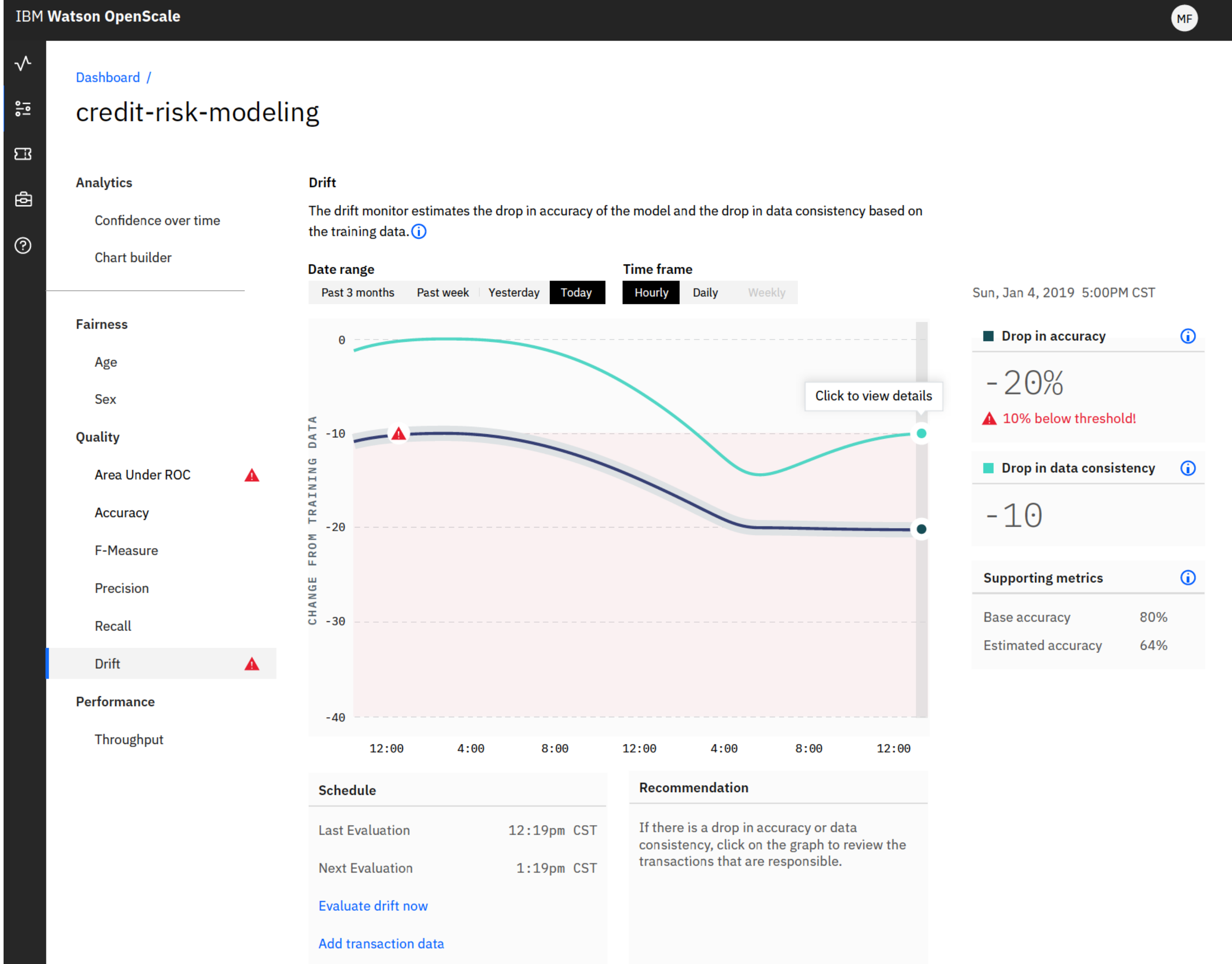
Most important factors influencing prediction

Feature	Value	Weight
LoanDuration	38	23.31%
CheckingStatus	no_checking	19.92%
EmploymentDuration	greater_7	16.59%



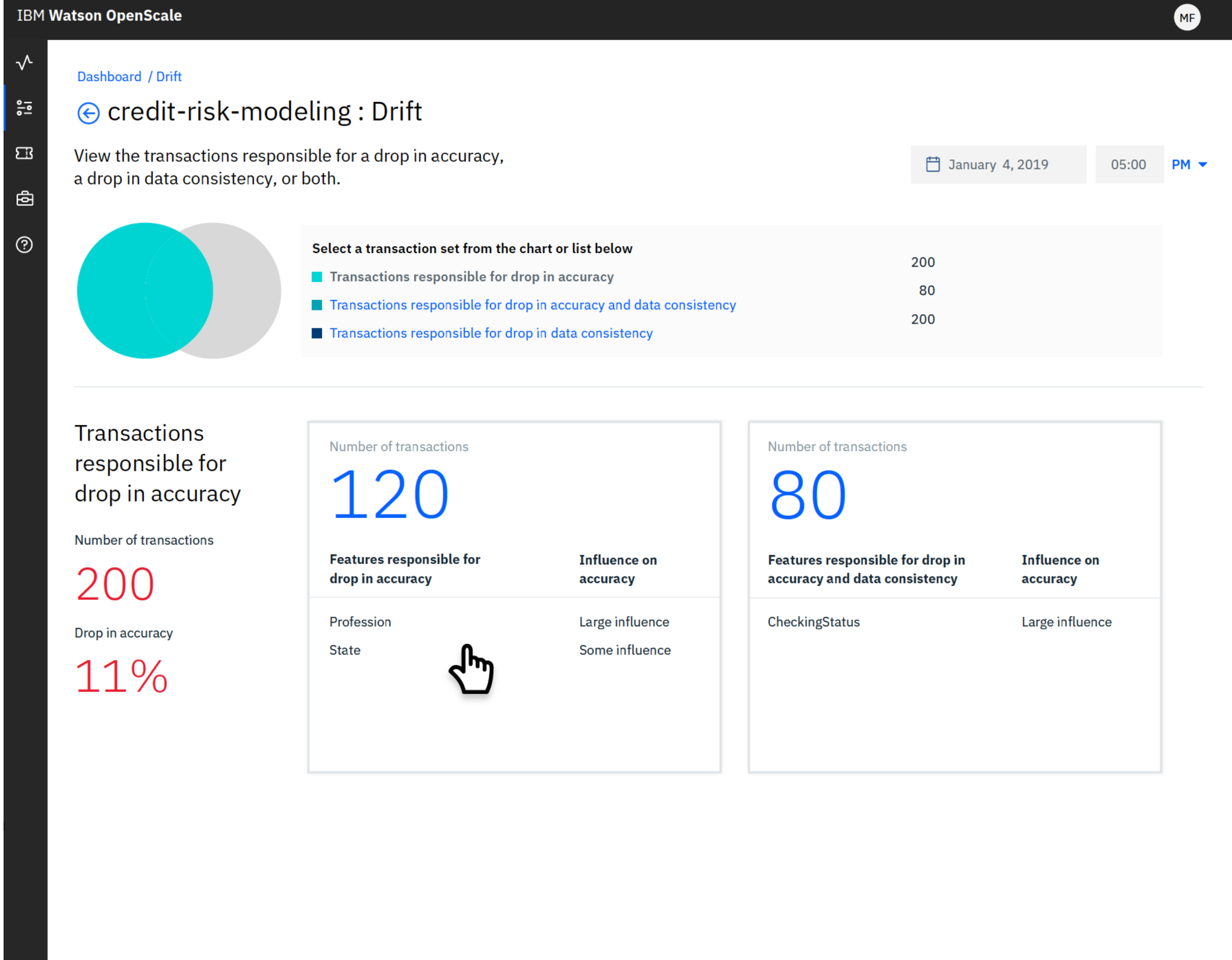
@MargrietGr

Drift



@MargrietGr

Drift



@MargrietGr

Summary

Is it fair?

FAIRNESS

AI Fairness 360

↳ (AIF360)

github.com/Trusted-AI/AIF360

aif360.mybluemix.net

Is it easy to understand?

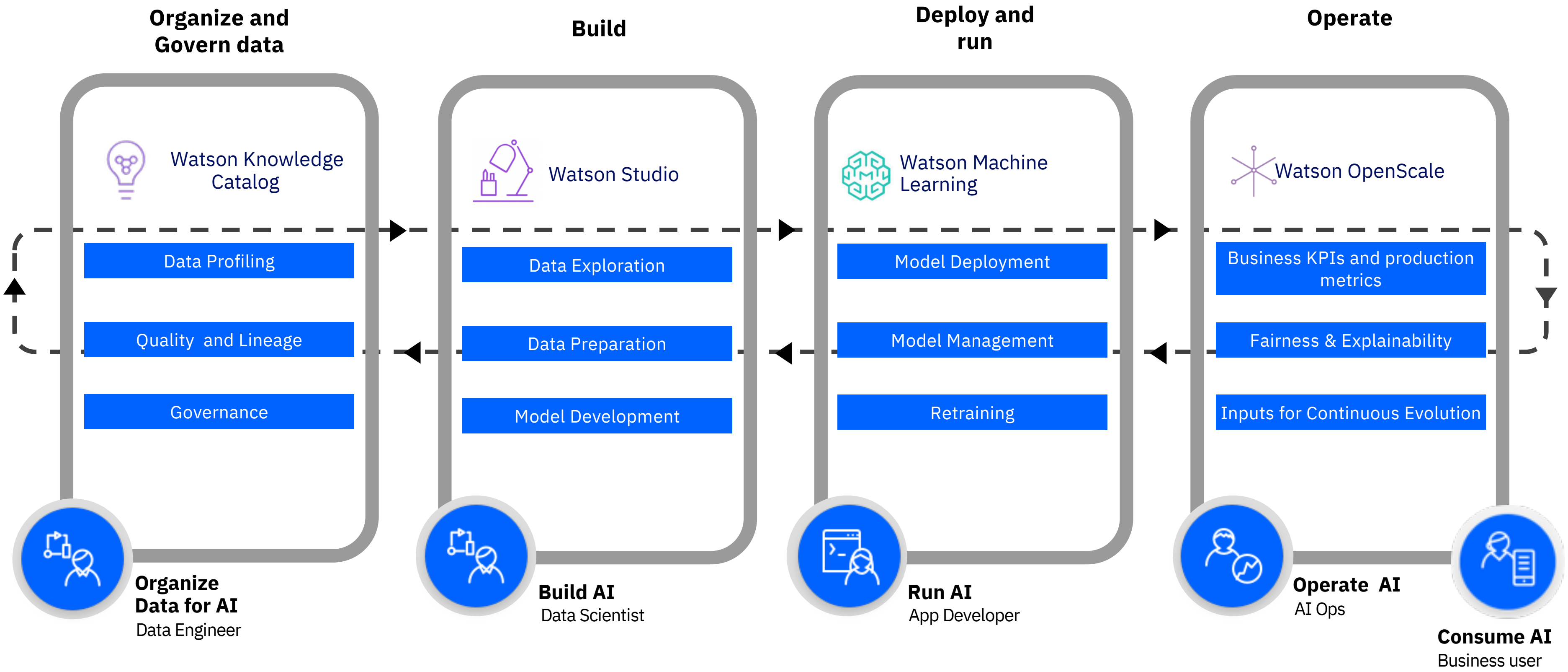
EXPLAINABILITY

AI Explainability 360

↳ (AIX360)

github.com/Trusted-AI/AIX360

aix360.mybluemix.net


Organize and Govern data
Build
Deploy and run
Operate
Watson Knowledge Catalog
Watson Studio
Watson Machine Learning
Watson OpenScale
Data Profiling
Quality and Lineage
Governance
Data Exploration
Data Preparation
Model Development
Model Deployment
Model Management
Retraining
Business KPIs and production metrics
Fairness & Explainability
Inputs for Continuous Evolution
Organize Data for AI
Data Engineer
Build AI
Data Scientist
Run AI
App Developer
Operate AI
AI Ops
Consume AI
Business user