Trusted AI

IBM Watson OpenScale

Dr. Rudolf Pailer
Artificial Intelligence Practice Austria
IBM Services

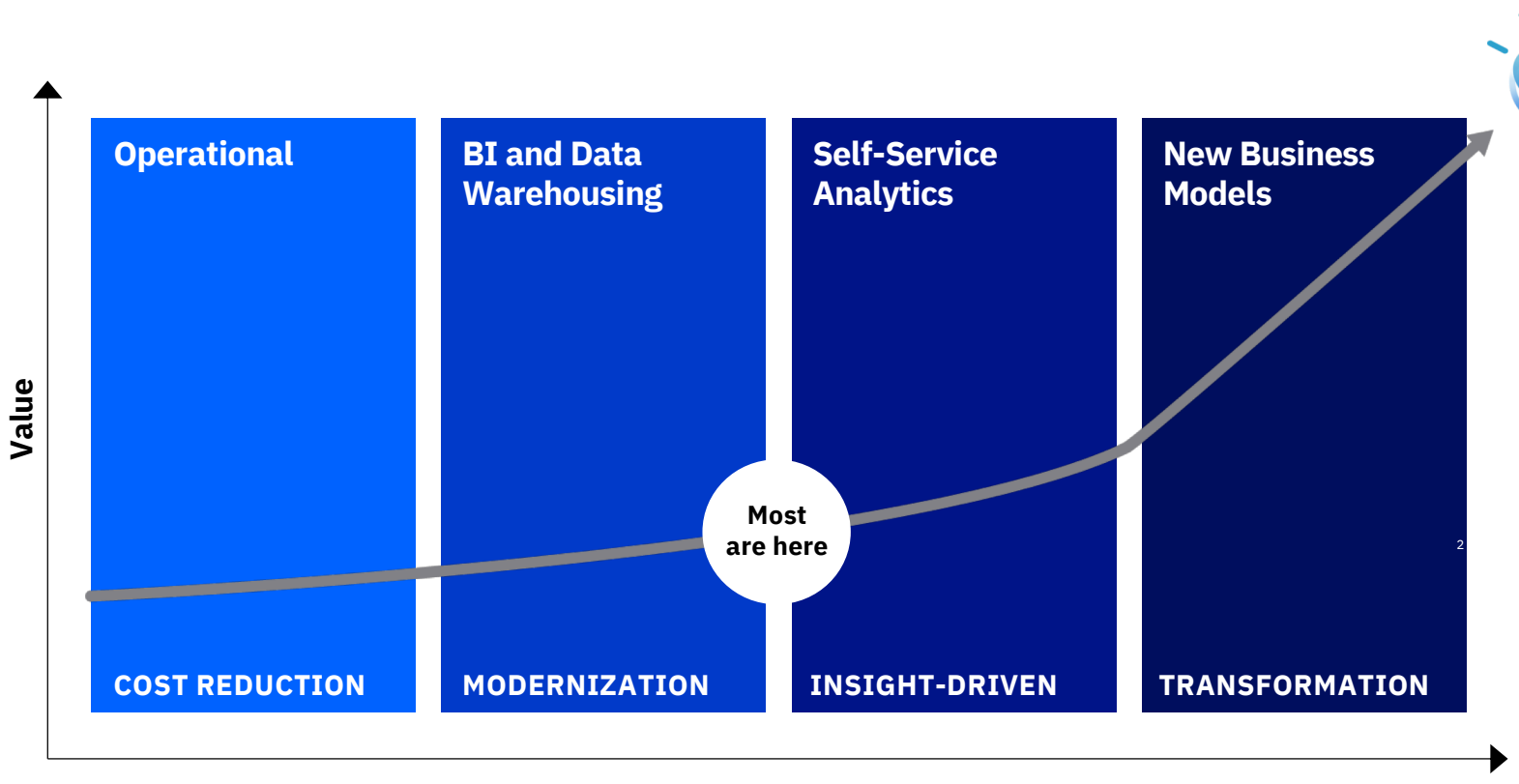
3. Oktober 2019

linkedin.com/in/rudolfpailer

papers 4you by CONECT

CONECT INFORMUNITY

I want AI !



85%

view AI as a strategic opportunity

MITSloan

BUT..., business stakeholders do not trust AI.

60%

of companies see **regulatory constraints** as a barrier to implementing AI.

– *IBM IBV AI 2018*

63%

cite availability of **technical skills** as a challenge to implementation.

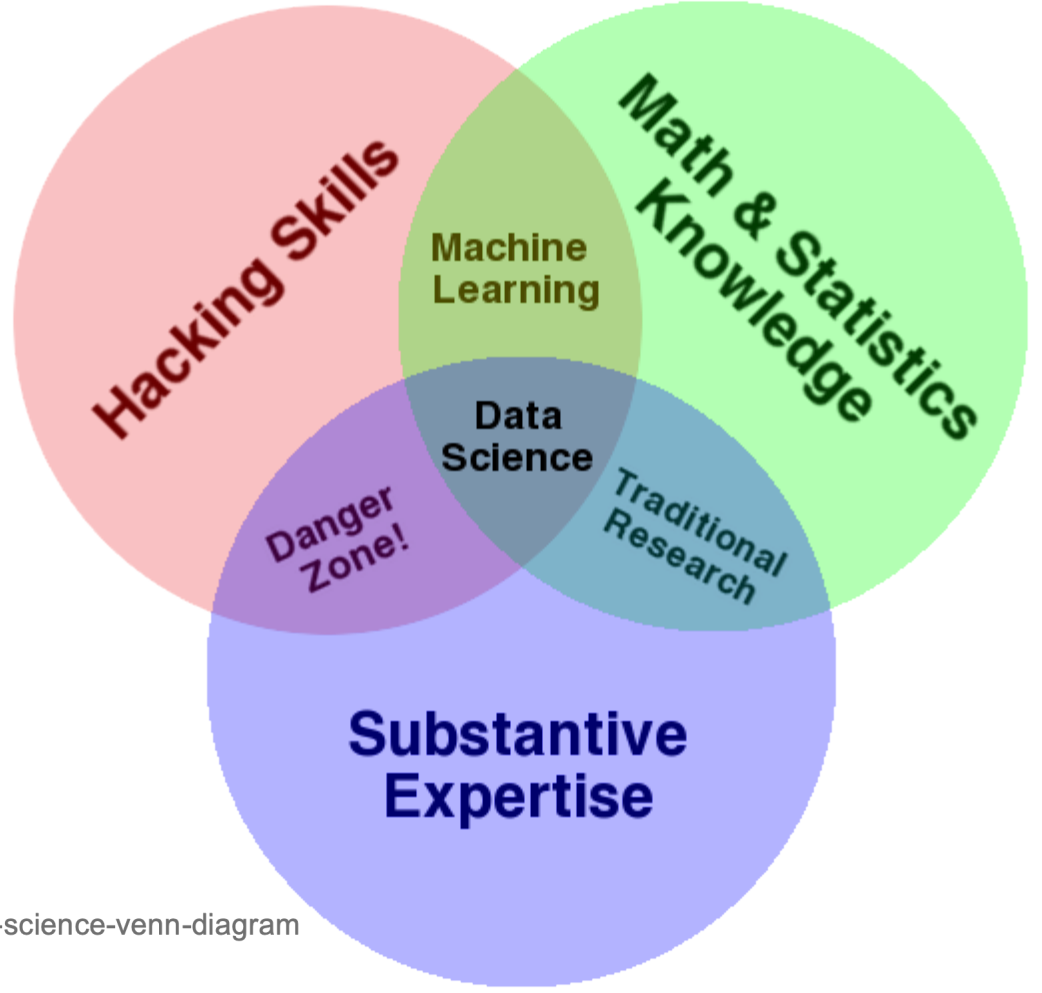
- IBM IBV AI 2018

Without expensive Data Science resources handholding multiple AI models in a production application:

1. No way to **validate** if AI models are **compliant with regulations** and will achieve expected business outcomes before deploying
2. Difficult to **track and measure** indicators of business success in production
3. Resource intensive and unreliable processes for **ongoing business monitoring and compliance**
4. Impossible for business users to **feedback** subtle domain knowledge into model lifecycle

IBM

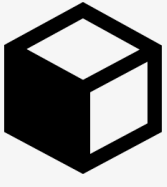
Skill Requirements in Data Science & AI Projects



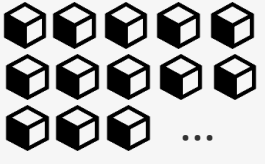
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

IBM

Model management



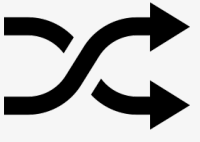
Modeling & Evaluation



Model Versioning

Model Deployment

Model Monitoring

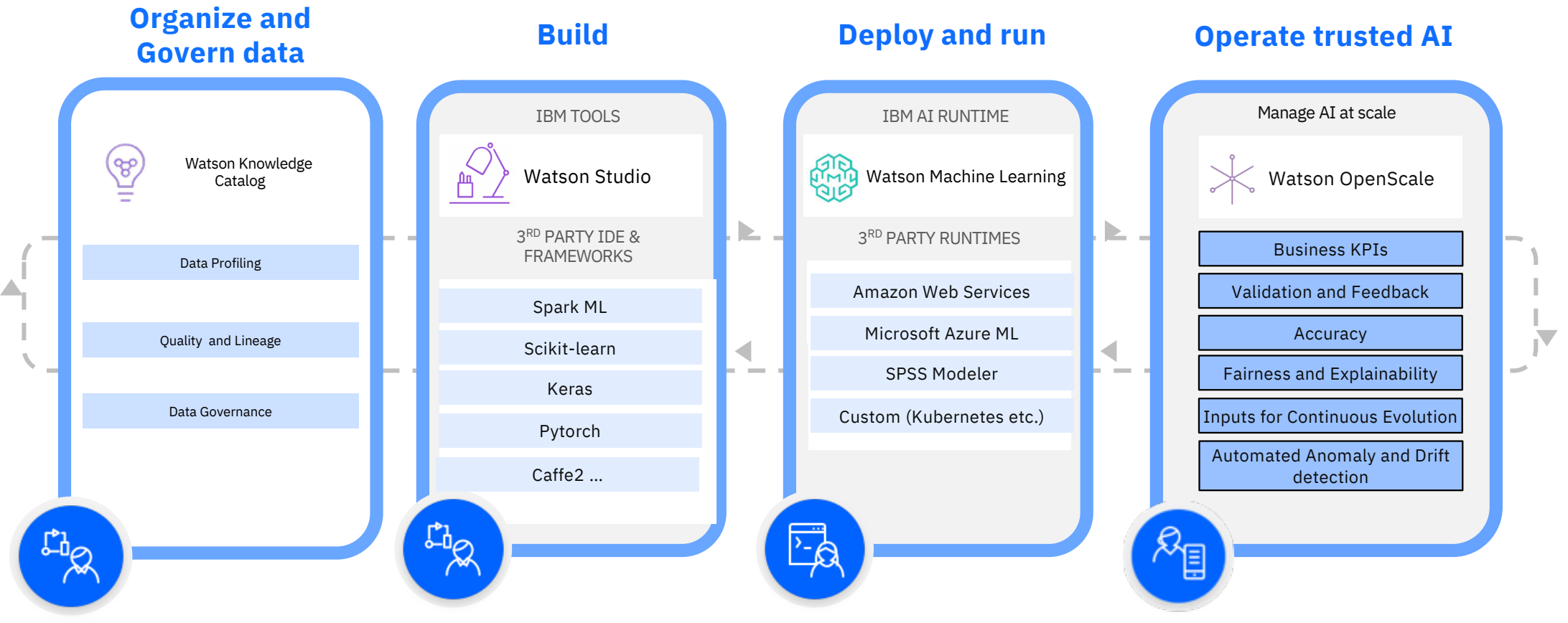


Dynamic Model Selection & Retraining

Data Science Solutions are **not** static by definition!

IBM

Watson OpenScale along with Watson Studio and WML enables enterprises to operationalize AI across the enterprise



Watson OpenScale will help validate and monitor AI models, deployed anywhere, to help comply with regulations and mitigate business risk

Production monitoring for compliance and safeguards
Detect and mitigate model biases
Audit and Explain model decisions
Model Validation and acceptance

Required in regulated industries and use cases – FSS, HR etc. in short term; others longer term*

Ensure that models are resilient to changing situations
Detect drift in data and anomaly in model behavior
Specific inputs and triggers to model lifecycle

Required to meet transformational goals

Align model performance with business outcomes
Correlate model metrics and business KPIs to measure business impact
Actionable metrics and alerts

Foundational to all AI implementations

Current capability

Upcoming capability

https://cloud.ibm.com/docs/services/ai-openscale-icp?topic=ai-openscale-icp-gs-get-started

IBM

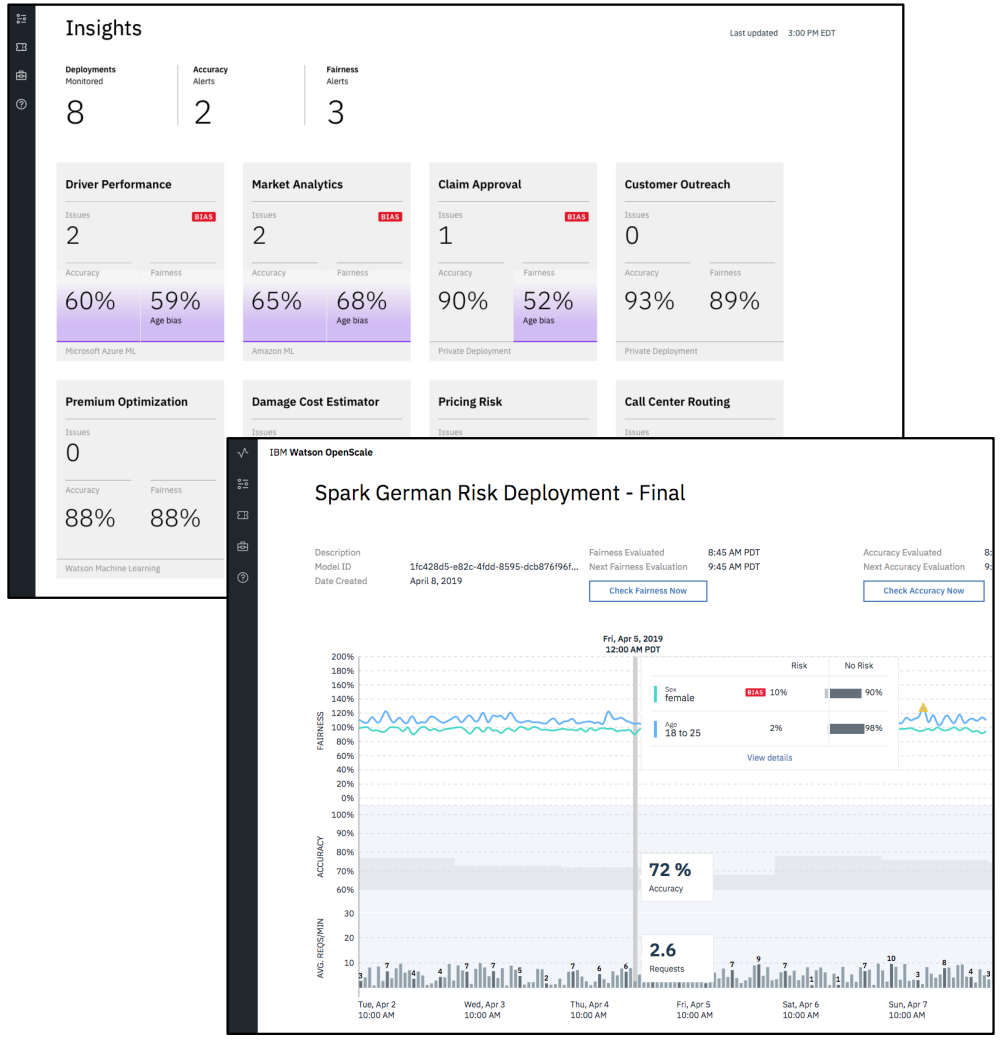
Operations Dashboard

Description:

Monitor deployed models in a single dashboard that can be filtered by deployment making it easy to manage AI in apps

Value:

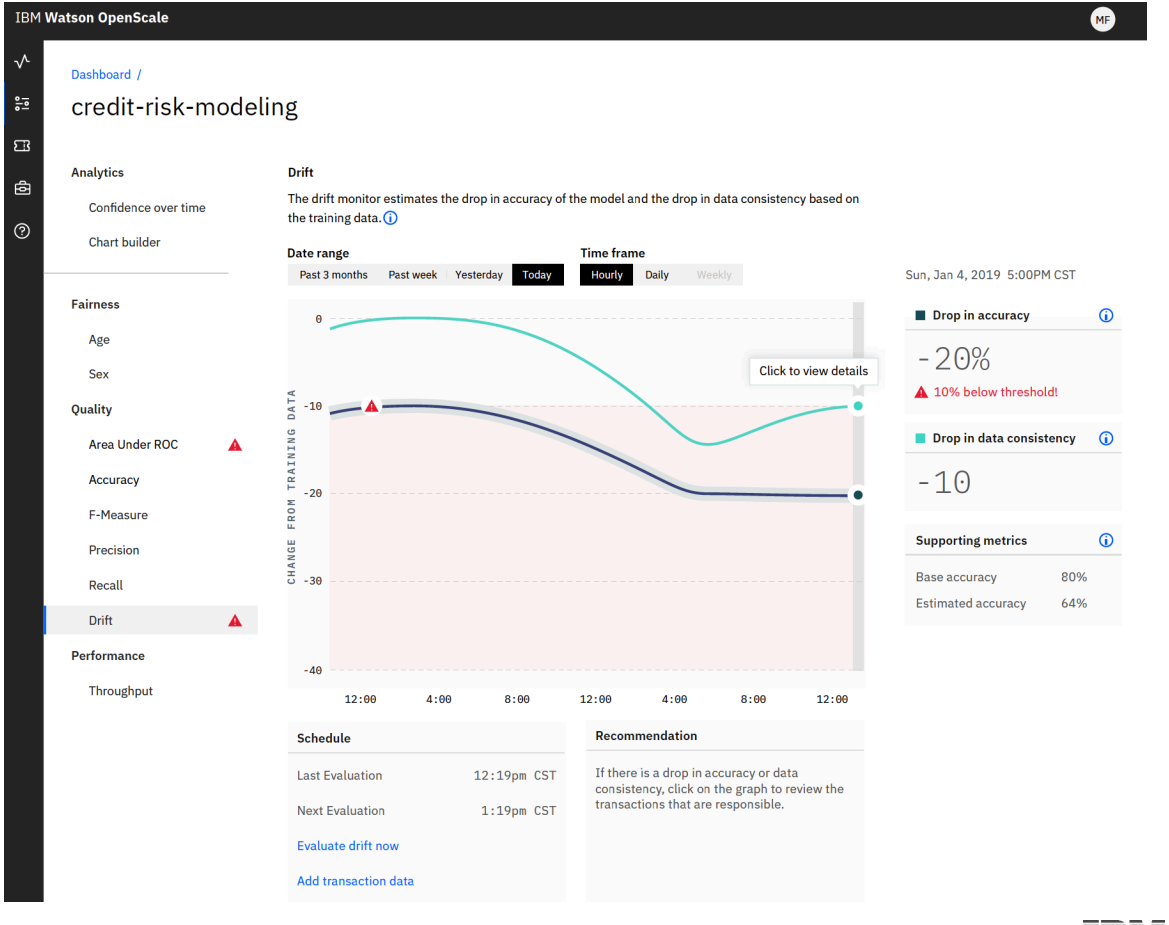
- Configure alerts or actions to be triggered when KPIs exceed threshold, ensuring model quality for improve business outcomes
- Measure model accuracy as it pertains to it’s ability to deliver outcomes more accurate than knowledge workers



Drift Detection in OpenScale

Drift Monitor in OpenScale measures two types of drifts:

- **Drop in accuracy**: It estimates the drop in accuracy of the model at runtime. The model accuracy could drop if there is an increase in transactions similar to those which the model was unable to evaluate correctly in the training data.
- **Drop in data consistency**: It estimates the drop in consistency of the data at runtime as compared to the characteristics of the data at training time.



Bias Mitigation – Guardrails against AI Bias

Watson OpenScale

Online Post Processing
(Intercept & Fix Features)

AI Fairness 360

In-Processing
(Intercept & Fix Algorithm)

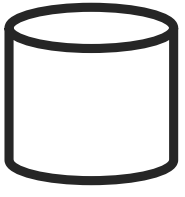
Runtime

AI Fairness 360

Pre-Processing
(Intercept & Fix Training Data)

Algorithm Design

Data Selection

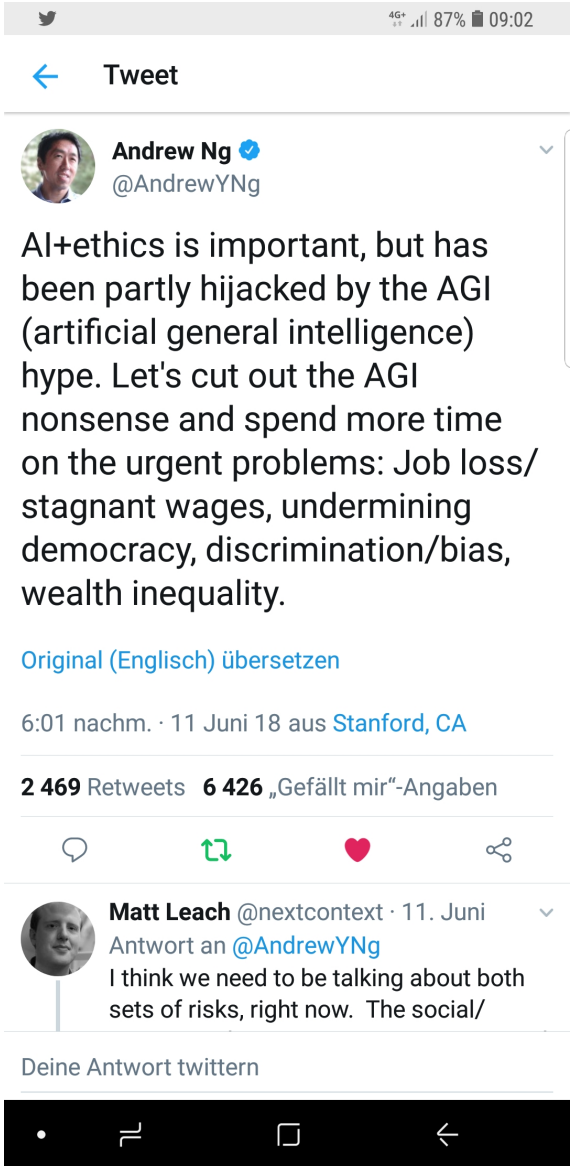


https://aif360.mybluemix.net/

IBM

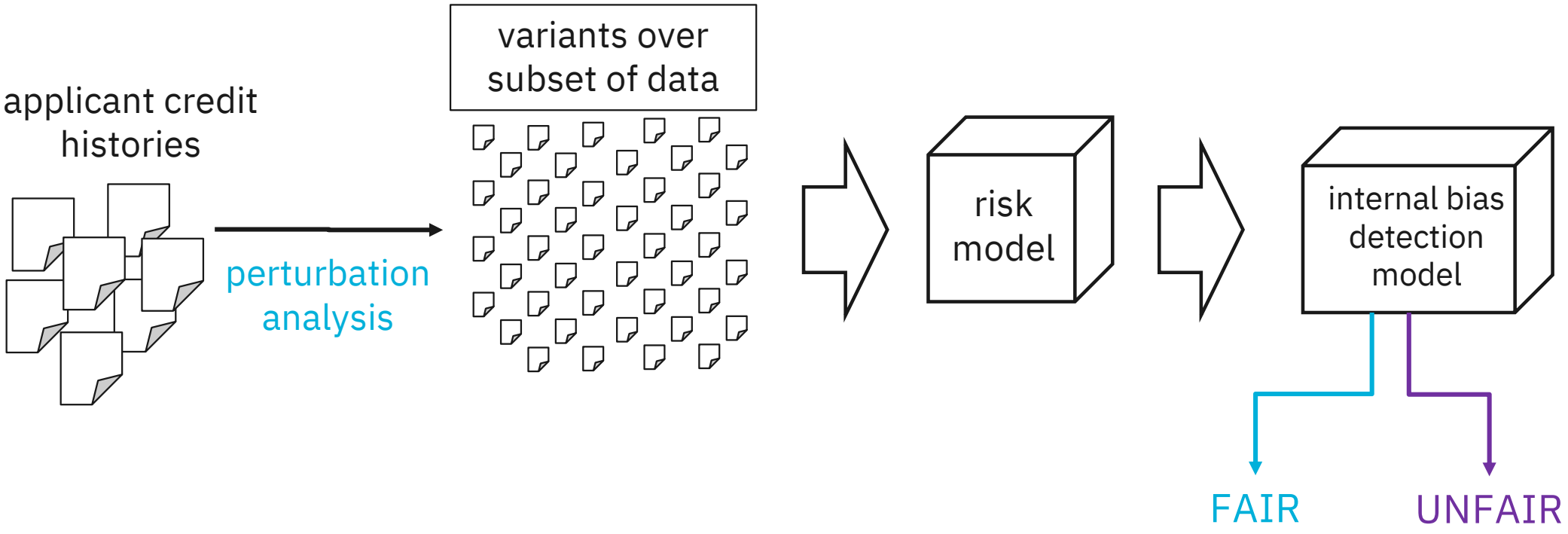
Ethics & Bias, COMPAS Recidivism

- Machine Bias: Pro Publica
  "There's software used across the country to predict future criminals. And it's biased against blacks."
  https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5777393/
  "…We further show that a simple linear predictor provided with only two features is nearly equivalent to COMPAS with its 137 features."



How does Watson OpenScale mitigate fairness issues?

Calculated hourly over a sliding window



BIAS MITIGATION POST-PROCESSING FOR INDIVIDUAL AND GROUP FAIRNESS, Lohia et al.
https://arxiv.org/abs/1812.06135

Model Fairness

Description:

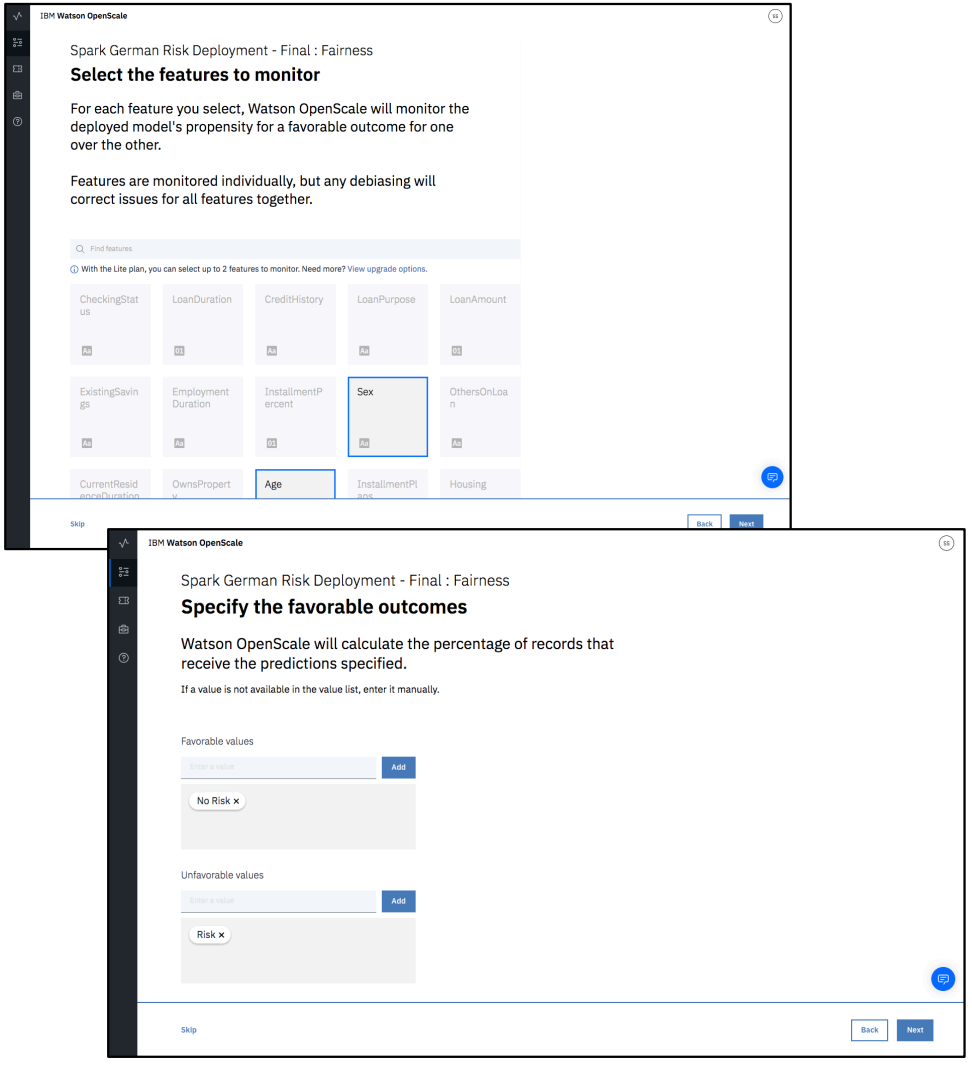
Models in production need to make fair decisions and can not be biased in their recommendations

How Does it Work?

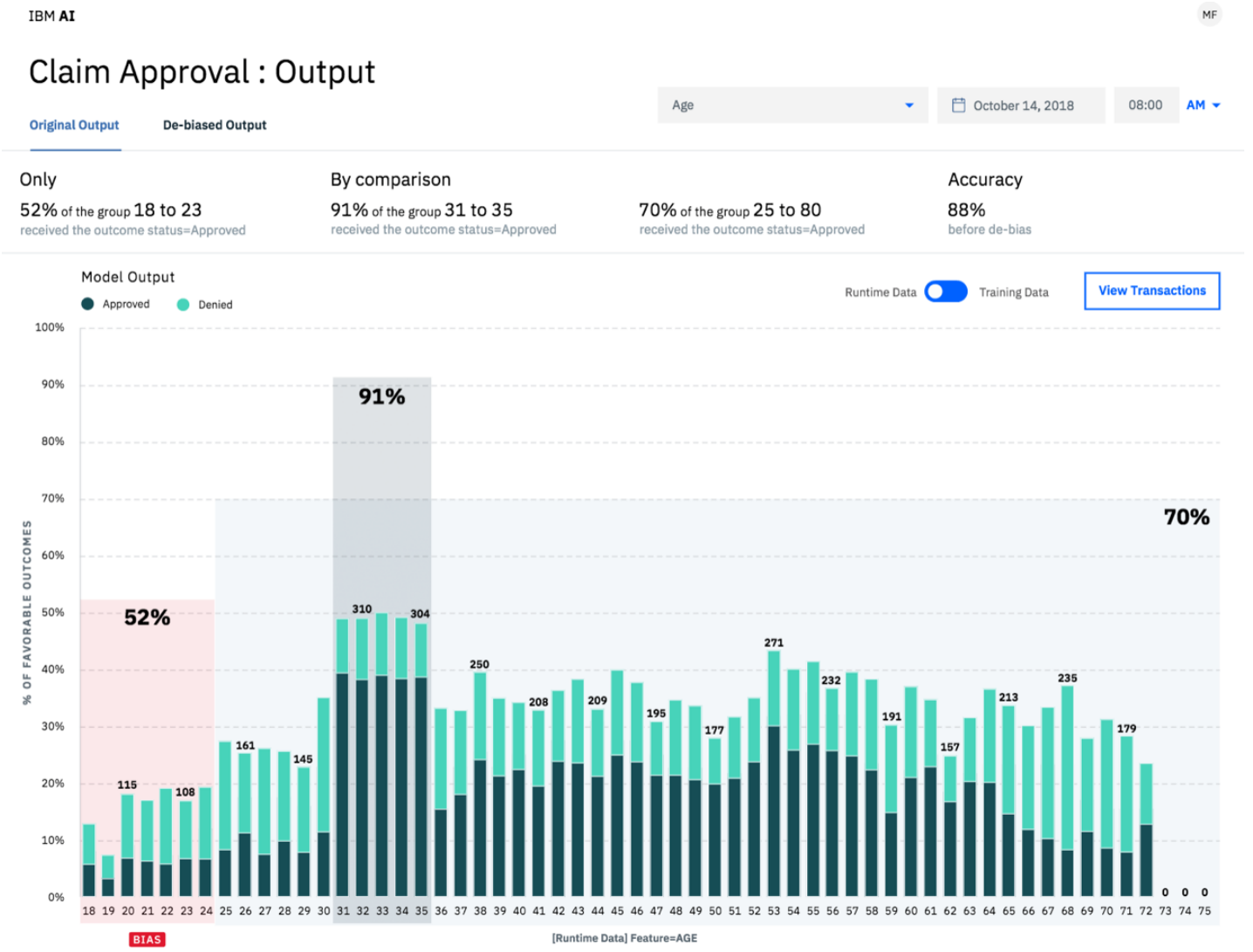
- Outcomes are selected as “favorable or unfavorable”
- “Favored Populations” and “protected populations” are selected where majority and minority groups are found
- A score is calculated based on the probability of favorable outcome for minority vs. probability of favorable outcome for majority

Value Add

- Configure fairness metrics for continuous evolution – this provides quality checks so no biased data or models can drive unfair decisions in business applications
- Visualize the fairness of deployed models on the OpenScale dashboard, so model builders can act swiftly to mitigate bias in production deployments

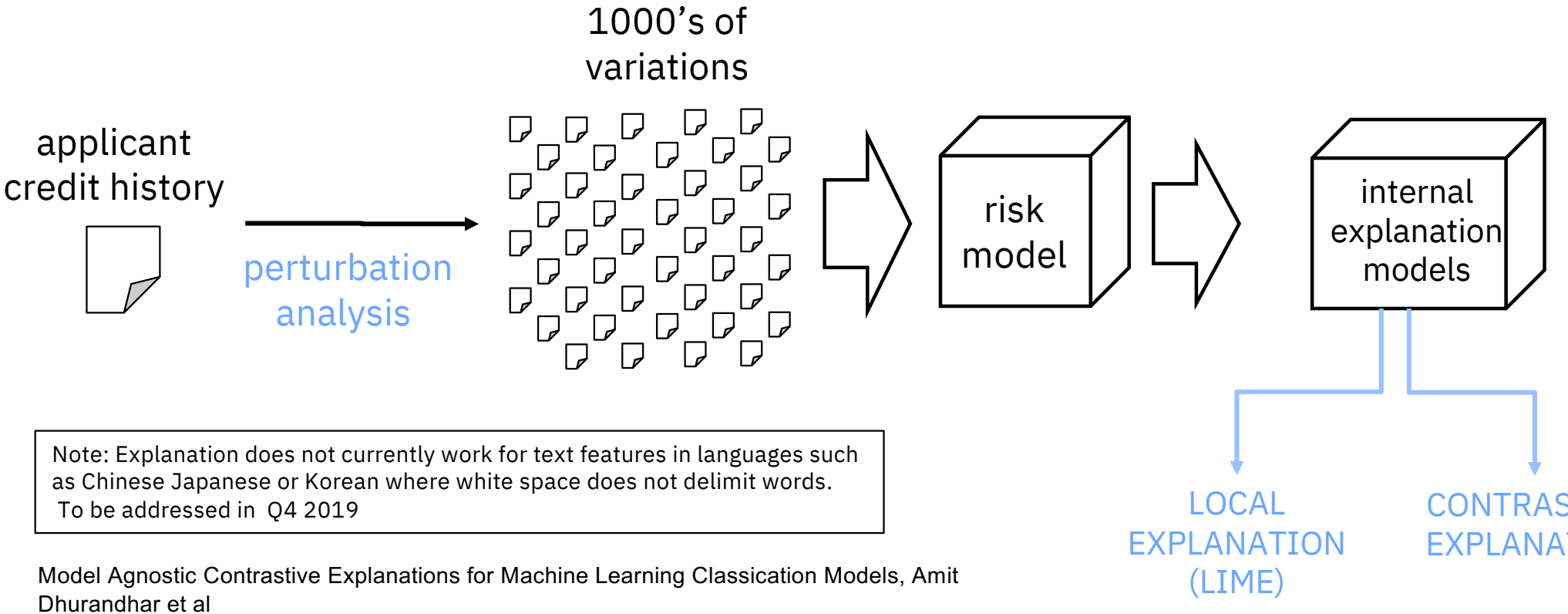


Bias Mitigation – Original Output



How does Watson OpenScale explain a prediction?

Calculated upon request

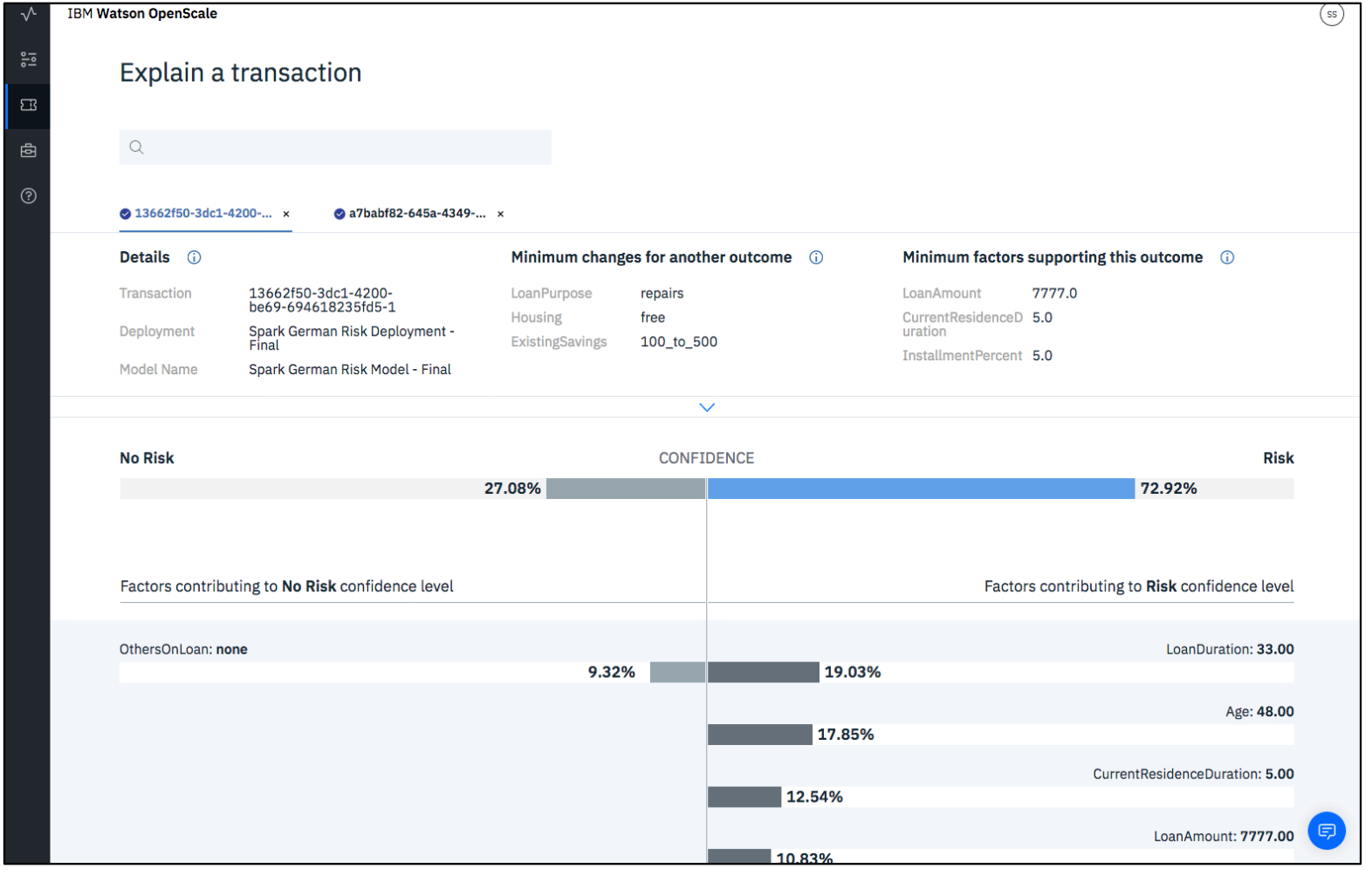


Note: Explanation does not currently work for text features in languages such as Chinese Japanese or Korean where white space does not delimit words.
To be addressed in Q4 2019

Model Agnostic Contrastive Explanations for Machine Learning Classication Models, Amit Dhurandhar et al
https://www.ibm.com/downloads/cas/0ZRZNR8E

IBM

Explainability



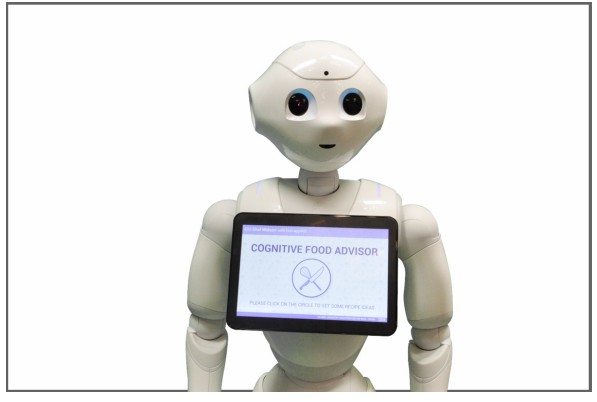
Links and related Material

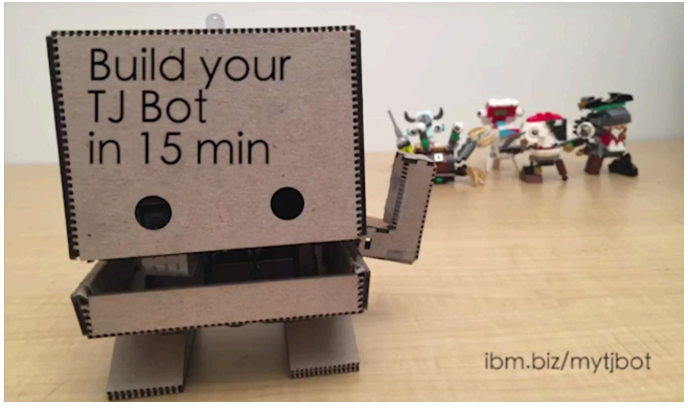
- IBM Watson OpenScale, Getting Started
  https://cloud.ibm.com/docs/services/ai-openscale-icp?topic=ai-openscale-icp-gs-get-started

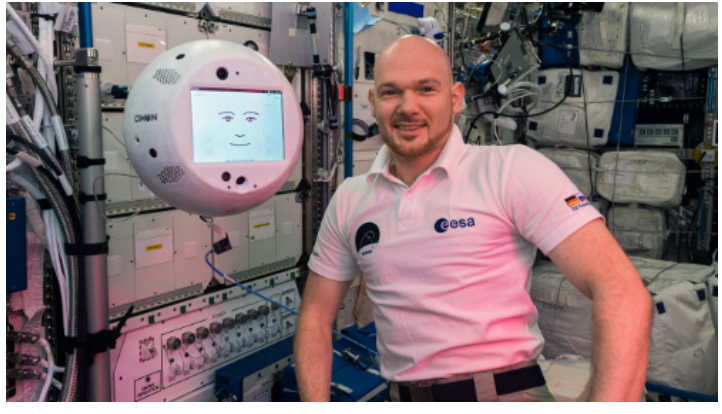
- IBM AI Fairness 360 Framework (open source)
  https://aif360.mybluemix.net/

- IBM Watson Studio
  https://cloud.ibm.com/catalog/services/watson-studio

- IBM Cloud
  https://cloud.ibm.com







© 2019 IBM Corporation