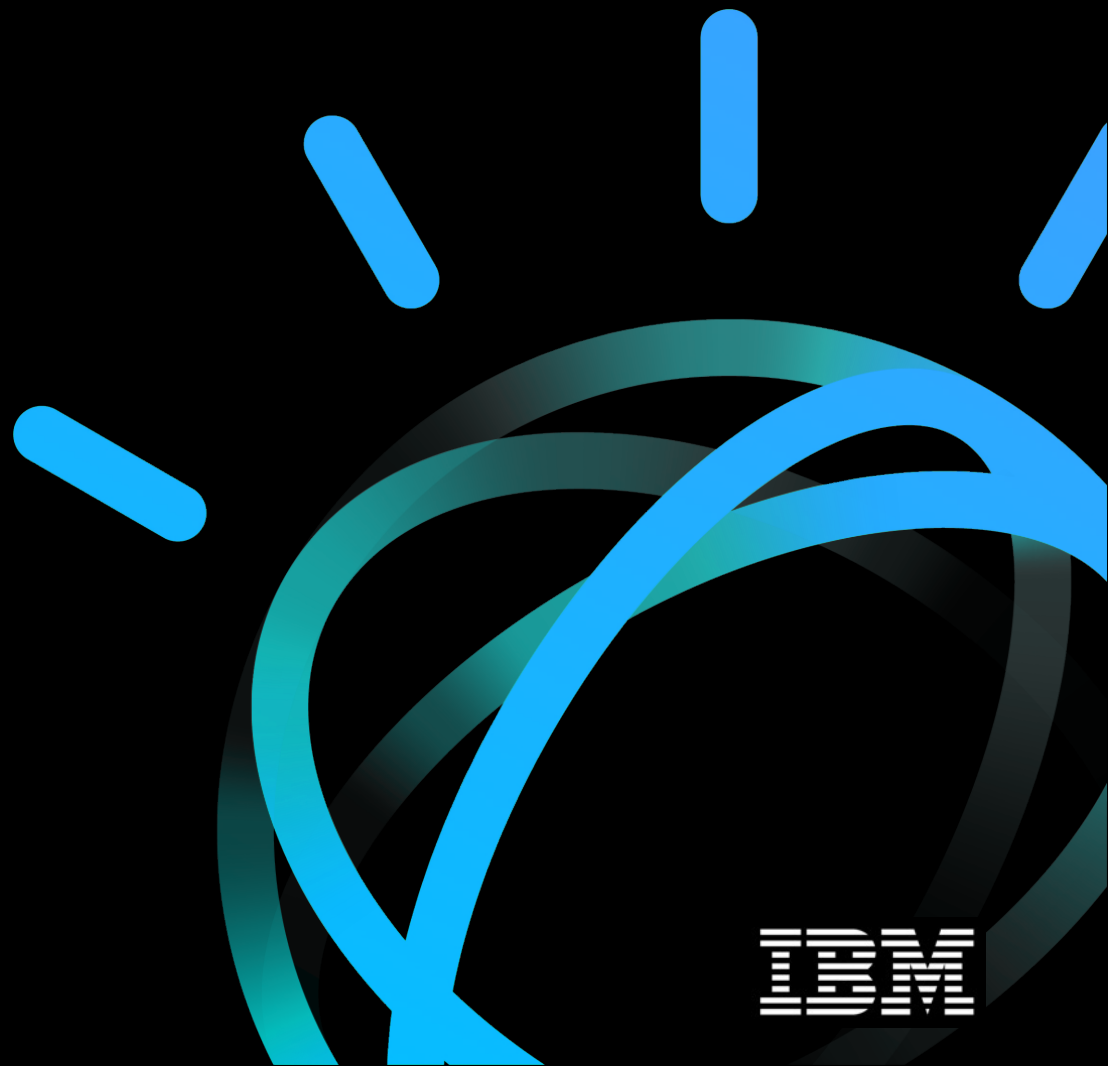
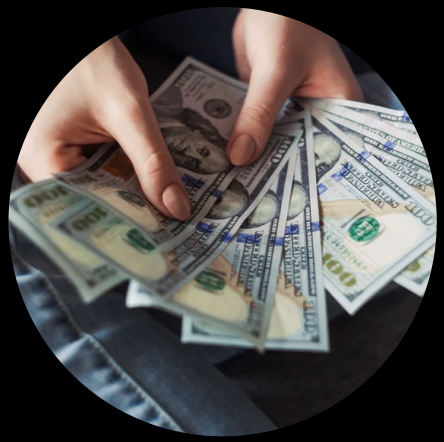
Trustworthy AI

John J Thomas, IBM

@johnjaithomas



AI is powering critical workflows and trust is essential



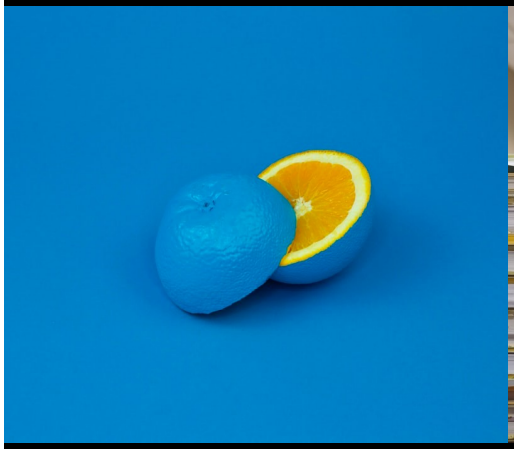
credit

employment

customer management

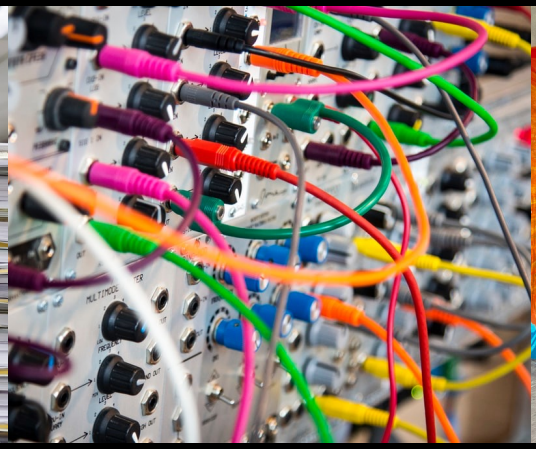
healthcare

Multiple factors are placing trust in AI as a top priority



brand reputation

increased regulation



complexity of AI deployments

focus on social justice

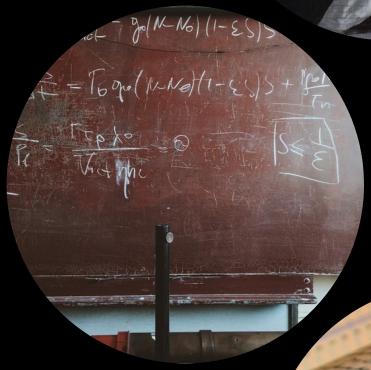
What does it take to trust a decision made by a machine?

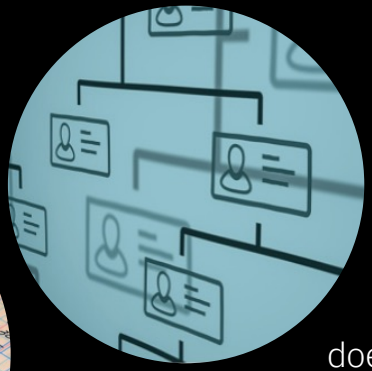
is it accurate?

is it fair?

did anyone tamper with it?

is it easy to understand?





does it handle privacy?

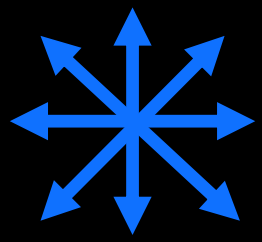
is it transparent?

Aspects of Trustworthy AI

Fair

Impartial and addressing bias

Are privileged groups at a systematic advantage compared to other groups?



Robust

Handle exceptional conditions effectively

Can we evaluate and defend against a variety of threats?

Privacy

High integrity data and business compliance

How do we ensure owners retain control of data and insights?

Explainable

Easy to understand outcomes/decisions

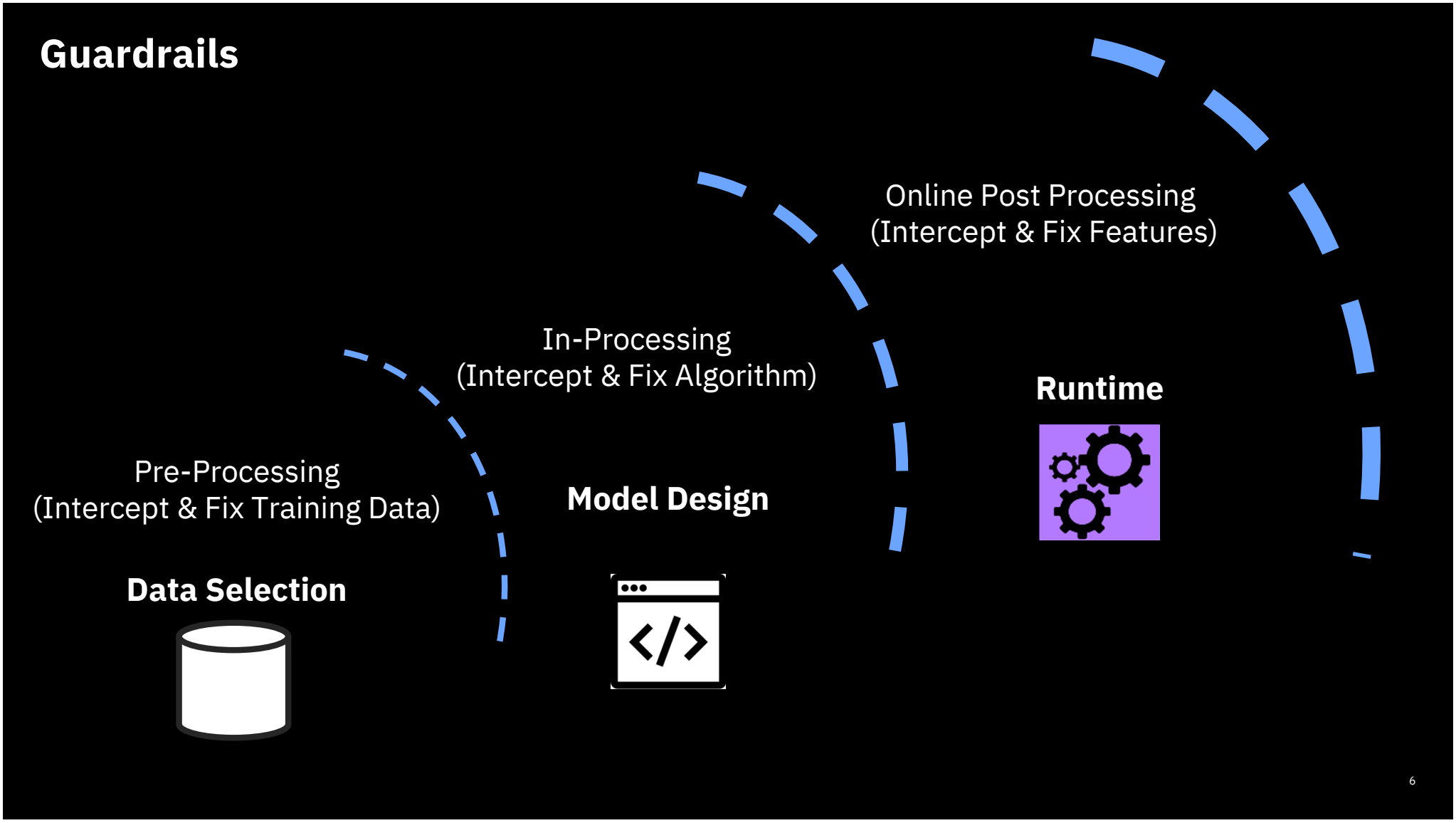
Why did the AI arrive at an outcome? When would it have been different?

Transparent

Open to inspecting facts and details

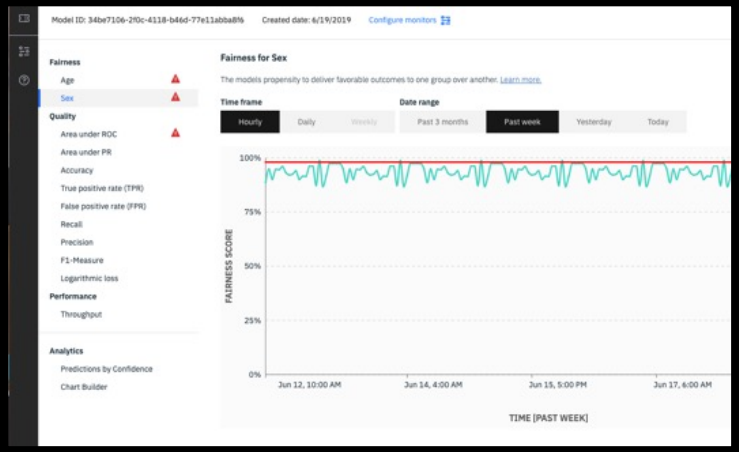
Can we increase understanding of why and how AI was created?

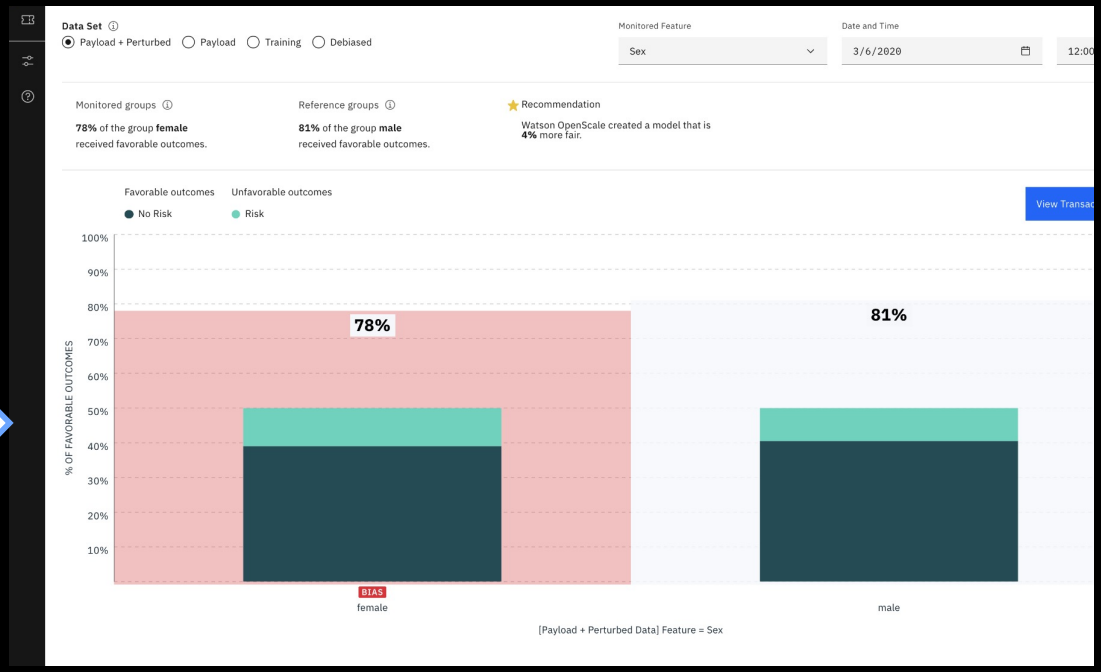
Guardrails


Online Post Processing
(Intercept & Fix Features)
In-Processing
(Intercept & Fix Algorithm)
Runtime
Pre-Processing
(Intercept & Fix Training Data)
Model Design
Data Selection

Example of a runtime guardrail

Bias Monitoring and Mitigation with Watson OpenScale





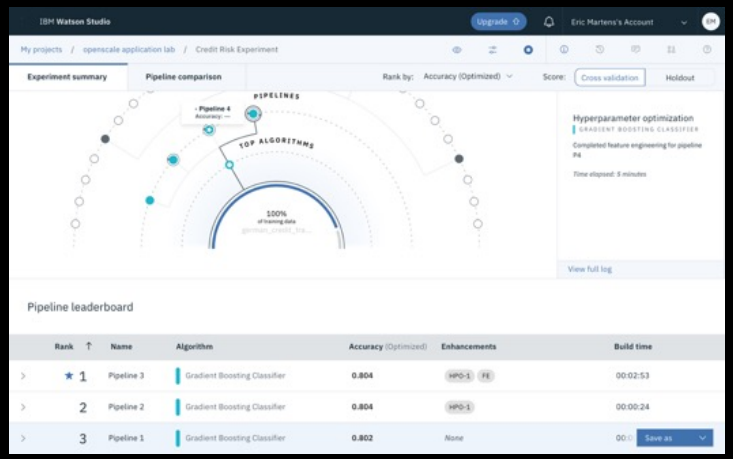
Setup ongoing monitoring of deployed model
Define monitored and reference groups

Calculate Disparate impact Value

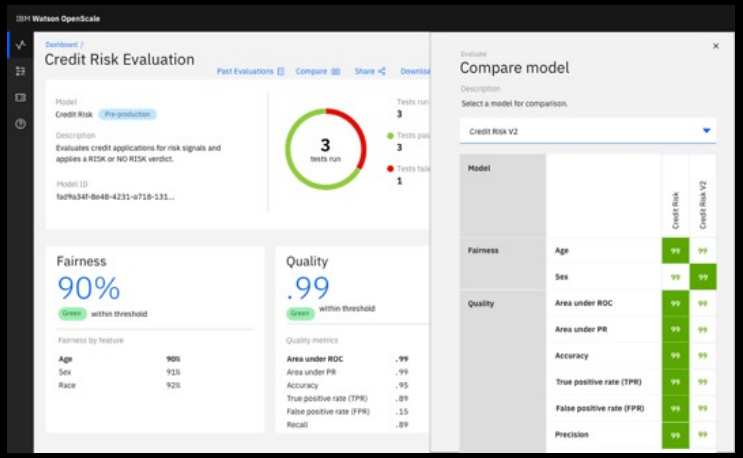
78% of the monitored group (female) have a favorable output
81% of the reference group (male) get a favorable output

Disparate impact Value: 96%
Mitigation based on policy

Guardrails across the AI lifecycle

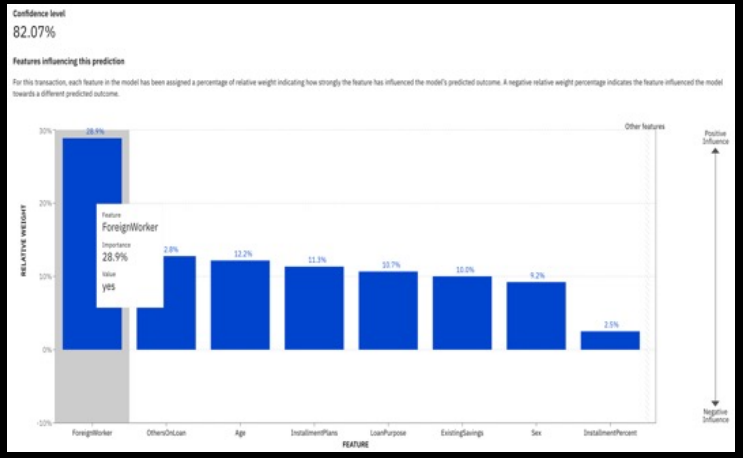


Challenger models



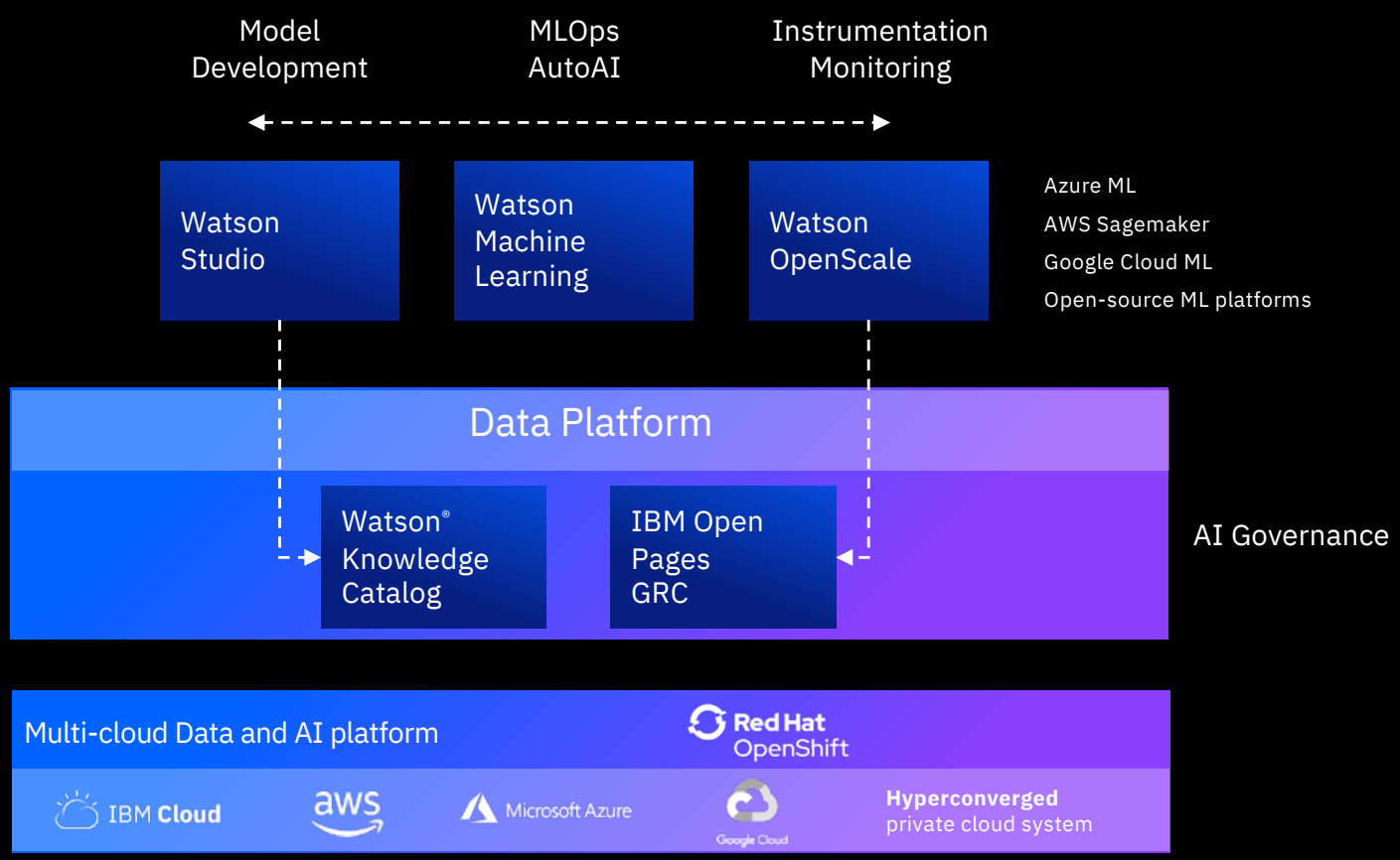
Validation, Model Risk Management

Drift in data consistency, Drift in accuracy



Local and contrastive explanations

IBM Cloud Pak for Data enables a Governed, Automated AI Lifecycle



Customer patterns for Trustworthy AI

Assess, Audit & Mitigate Risk

Guidance and tooling to help customers assess, audit and mitigate risk in existing AI solutions

Full AI lifecycle

Partnering with customers to plan, build, deploy and manage new AI solutions while ensuring trustworthiness

AI Governance frameworks

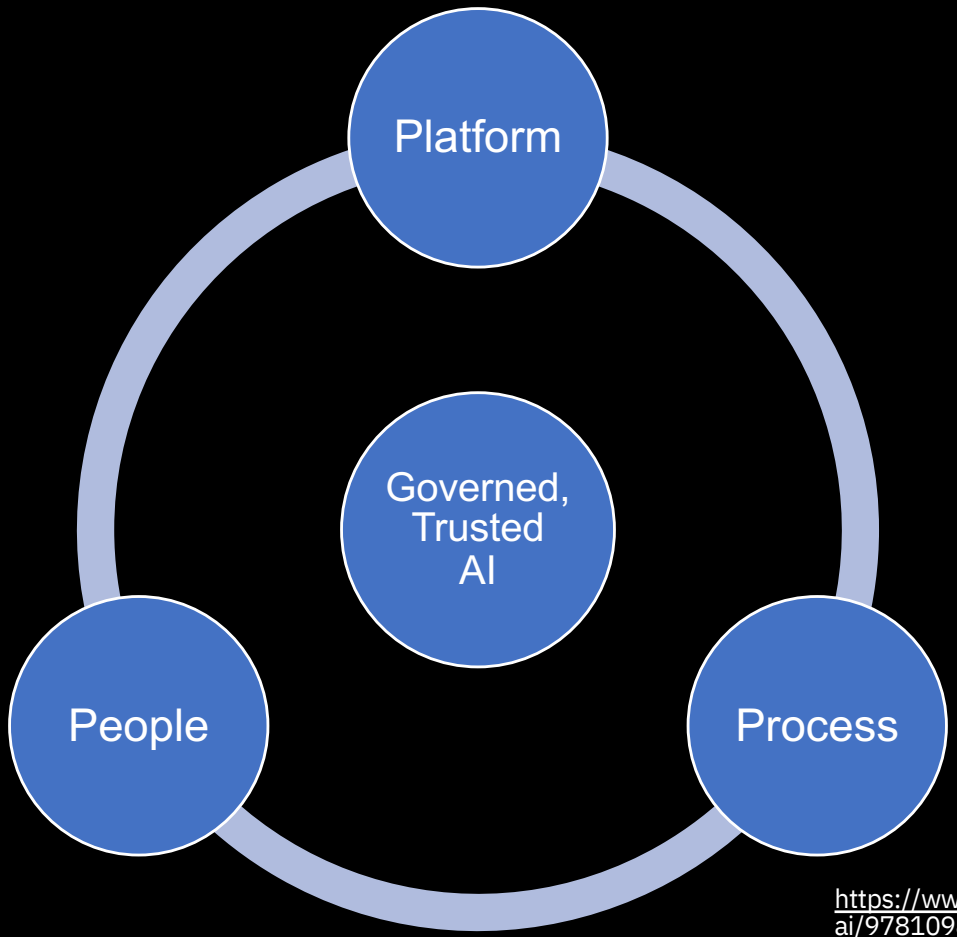
Partnering with customers to set up an AI Governance framework and implement enterprise-scale AI Governance solutions

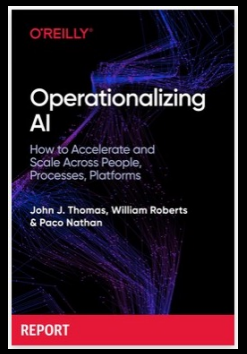
Education and Guidance

Best practices for building trustworthy AI solutions with education and guidance for data scientists, developers, and decision-makers

Standalone courses & certifications or integrated with above solutions

An AI Factory methodology operationalizes Trustworthy AI





https://www.oreilly.com/library/view/operationalizing-ai/9781098101329/

Get Started!

https://ibm.biz/trustworthyAI