

Árnyék a gépezetben: mit kezdünk a mesterséges intelligencia fekete dobozával?

Husztai Dániel

daniel.husztai1@ibm.com

+36 20 823 5737

IDEAS

Algorithmic bias isn't just unfair — it's bad for business

If it's not deployed wisely, artificial intelligence can turn consumers off.

By Kalinda Ukanwa Updated May 23, 2021, 3:00 a.m.

YouTube sued for using AI to racially profile content creators

They claim YouTube's algorithms discriminate against black users

Data science during COVID-19: Some reassembly required

Most likely, the assumptions behind your data science model or the patterns in your data did not survive the coronavirus pandemic. Here's how to address the challenges of model drift

2018 / 4:04 PM / UPDATED 2 YEARS AGO

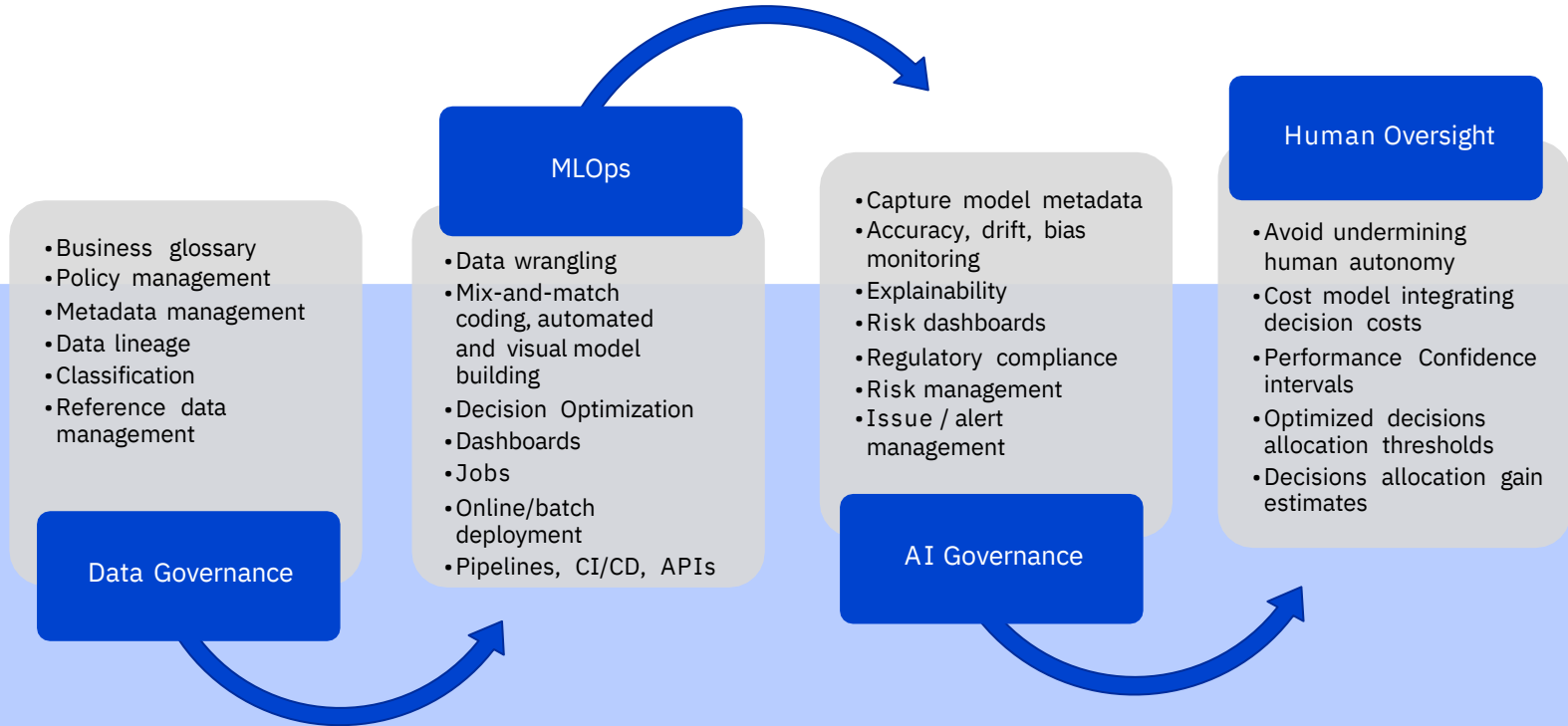
Amazon scraps secret AI recruiting tool that showed bias against women

The \$300m flip flop: how real-estate site Zillow's side hustle went badly wrong

The Washington Post
Democracy Dies in Darkness

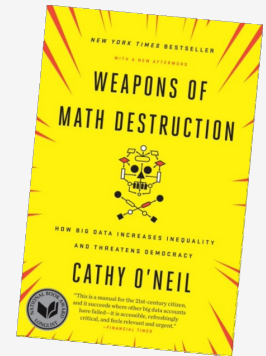
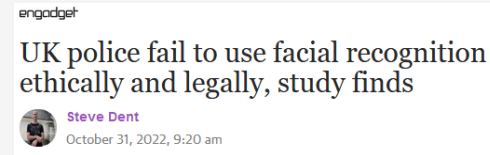
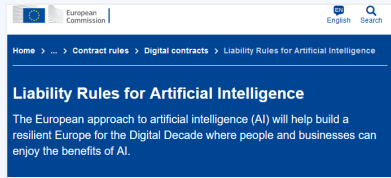
Apple Card algorithm sparks gender bias allegations against Goldman Sachs

Putting AI Governance in perspective



IBM Cloud Pak for Data

Regulation and reputation are the driving factors behind AI governance



Nearly all (97%) respondents believe that regulation will impact them to some extent and 95% believe that at least part of their business will be affected by the EU regulations specifically. 25% have yet to establish any meaningful Responsible AI capabilities.

Accenture - From AI compliance to competitive advantage, 2022

“Fewer than 20% of executives strongly agree that their organizations’ practices and actions on AI ethics match (or exceed) their stated principles and values.”

IBM and Oxford Economics – AI ethics in action, 2021

AI Governance solution

IBM solution is built on 3 pillars to meet clients on their maturity curve

Lifecycle governance

Monitor, catalog, and govern AI models from anywhere, throughout the AI lifecycle

Risk management

Manage risk and compliance to business standards, through automated facts and workflow management

Regulatory compliance

Ensure clients adhere to external AI regulations for audit and compliance

IBM's differentiation

Comprehensive: only vendor to support all three layers of AI governance

Automated: automated facts collection and lineage tracking within python notebooks

Open: AI vendor-neutral, supports governance of models built and deployed in third party tools

Unified: unified cataloging for models & data and unified notions of data quality for AI and reporting

Regulations-driven: support for AI regulations natively

Active policy enforcement: support for policies and rules to automate regulatory compliance

In practice, this means that...



Business stakeholders need to ... (C-Suite, CPO/CRO, Business Owner)

- ✓ Define new responsibilities, policies, guidelines and processes based on AI principles & enterprise values
- ✓ Establish organizational structures
- ✓ Understand regulation and translate to business & technical requirements
- ✓ Raise awareness and train leaders and practitioners
- ✓ Assess maturity, risk level and conformity
- ✓ Oversee, manage and mitigate risks
- ✓ Design architecture, pipelines and governance rules for trust and scale

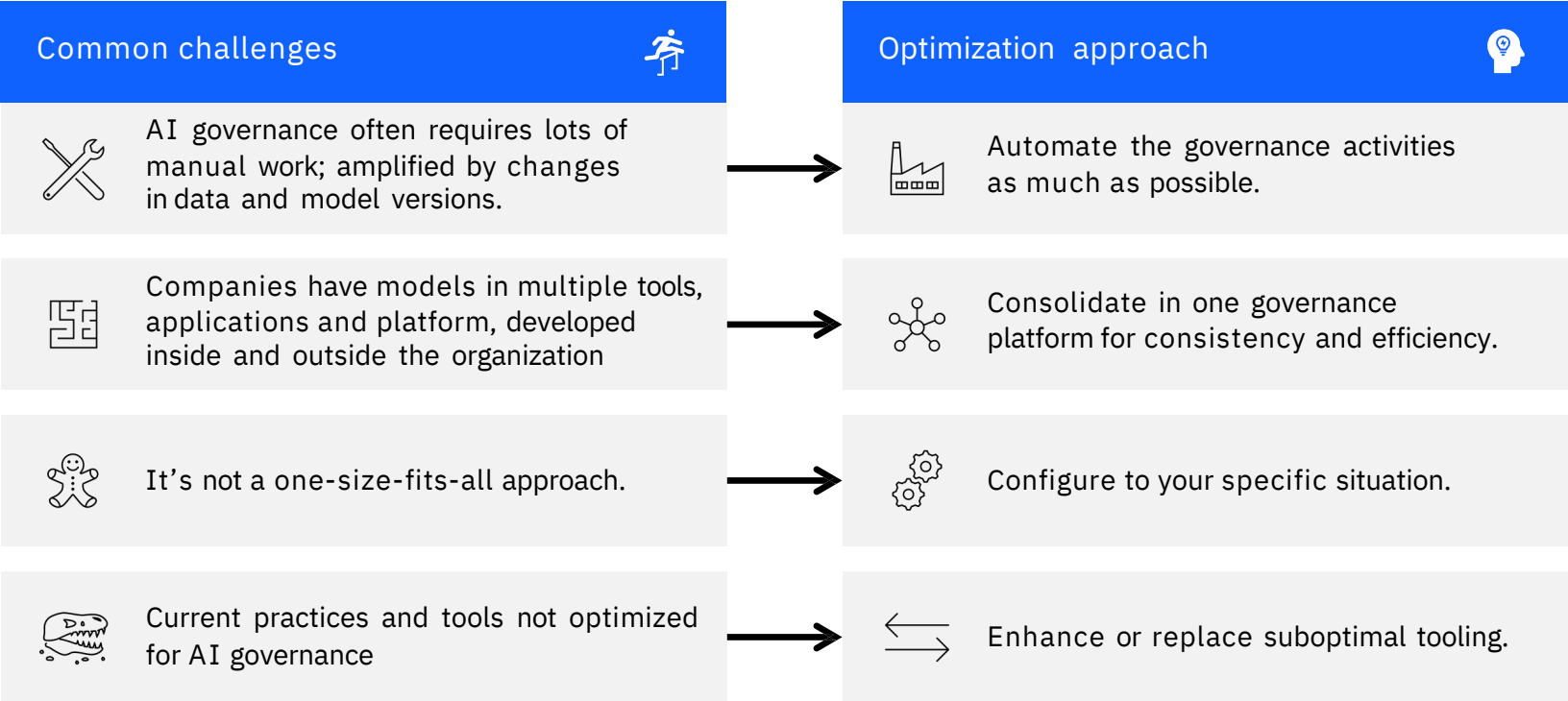


Technical stakeholders need to ... (Data Scientist, Ops Lead, Risk Manager)

- ✓ Adhere to data governance rules
- ✓ Keep track of security risks
- ✓ Document every new and updated model
- ✓ Find and fix direct and indirect bias in each data set and model
- ✓ Find and fix model accuracy drift & data consistency drift in each model
- ✓ Explain model decisions as requested
- ✓ Find and fix adversarial attacks on your data and models
- ✓ Complete all model review process activities



Which is easier said than done



Putting AI Governance to work

A typical customer context

AI Governance functions

OpenPages MRG Governance control

Build and validate models

Watson Knowledge Catalog Model metadata

Deploy models

Watson OpenScale Monitor and explain

AI Fact Sheets Centralized facts

Line of business A

Line of business B

Line of business C

AI Governance solution

Watson Studio

DataRobot

In-house solution

AI Governance solution

Watson Studio

DataRobot

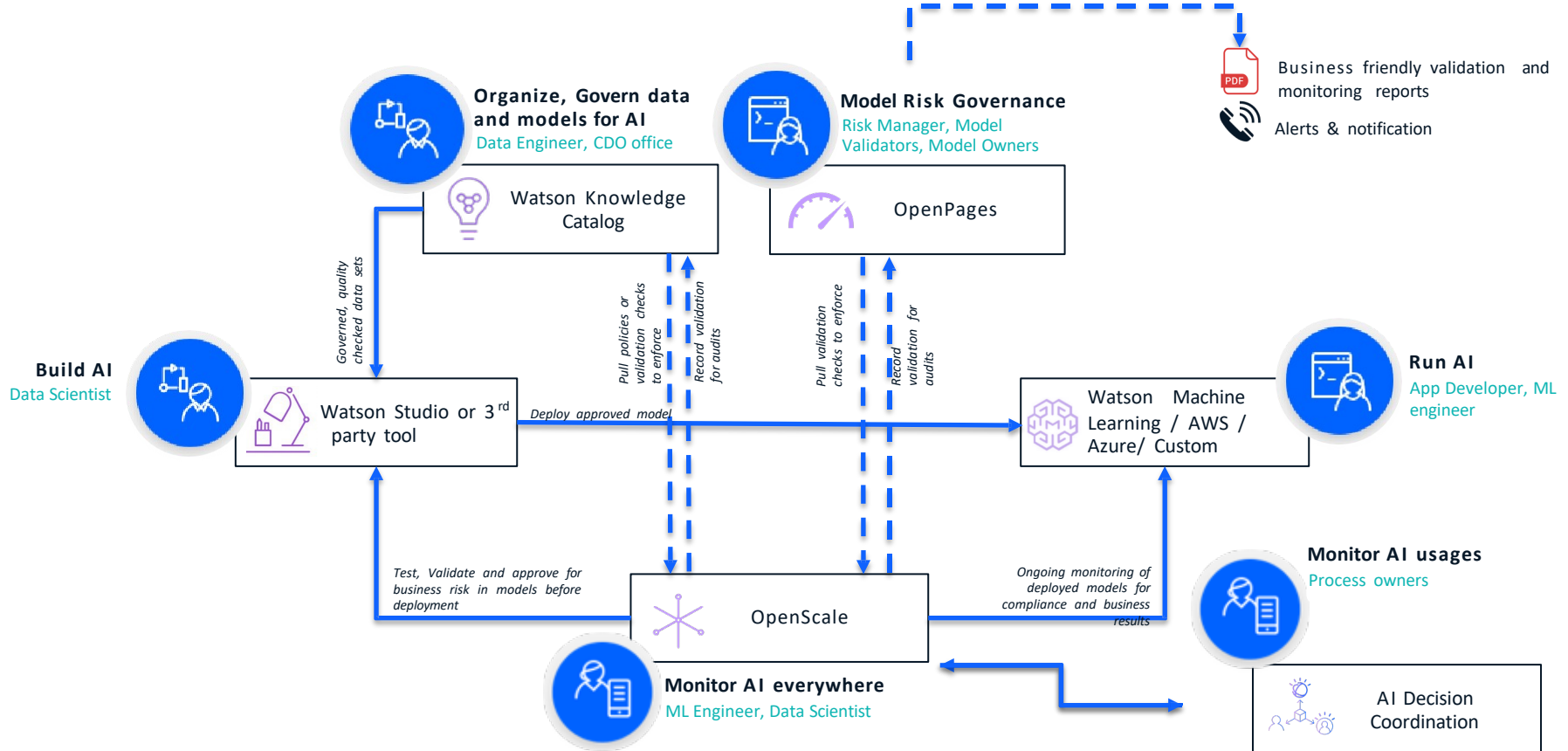
In-house solution

AI Governance solution

AI Governance solution

Technology View – AI Governance

Cloud Pak for Data components in scope for Trusted AI

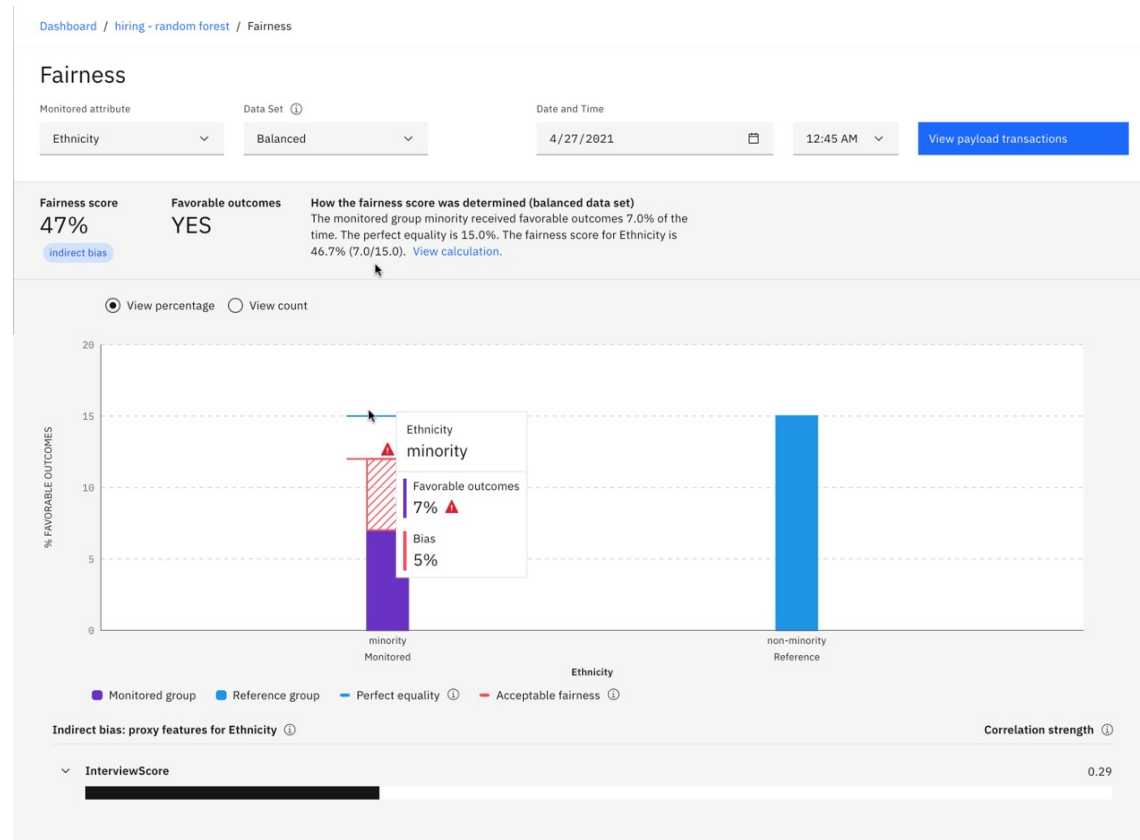


Bias detection

Continuous calculation of model fairness

- Analyze deployed model predictions for bias
- Collect and aggregate bias data for dashboards and alerts
- Find non-feature data correlations
- Use a corrected model for “de-biased” predictions

Ensuring fairness in model scoring



Model Explainability

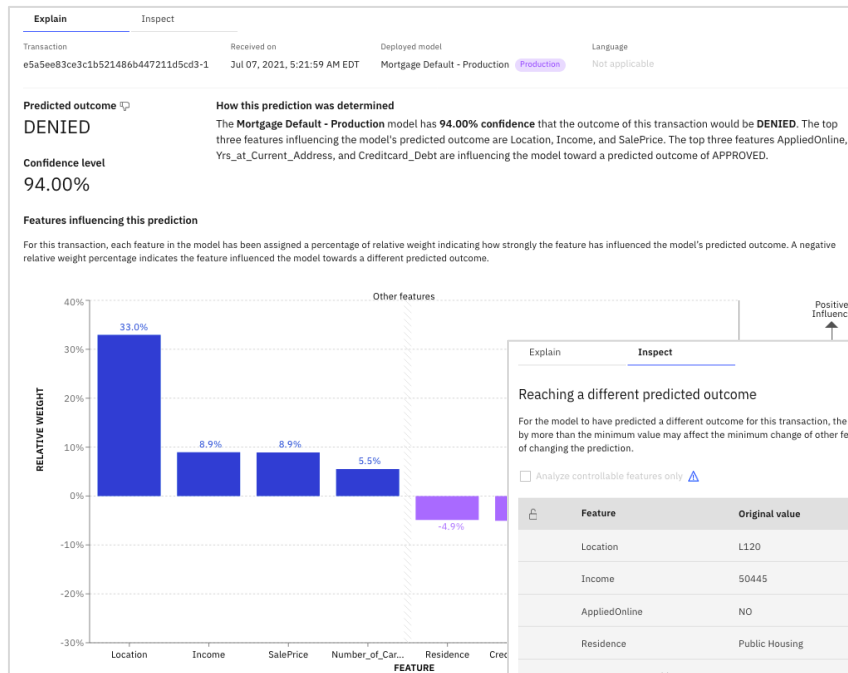
Explain model predictions

- Show the most influential features
- Explain in natural language
- Available API for prediction explanations

What-if analysis

- Experiment with values
- Assess effects of changes to features

Understand model outcomes



Explain **Inspect**

Reaching a different predicted outcome
For the model to have predicted a different outcome for this transaction, the value of all listed features would need to change to the indicated minimum value. Note that changing a feature value by more than the minimum value may affect the minimum change of other features for the model to predict a different outcome. Higher feature importance numbers indicate a greater likelihood of changing the prediction.

Analyze controllable features only

Feature	Original value	New value	Value for a different ou...	Importance	
Location	L120	L120	L100	1.00	
Income	50445	50445	50445	0.00	
AppliedOnline	NO	NO	NO	0.00	
Residence	Public Housing	Public Housing	Public Housing	0.00	
Yrs_at_Current_Address	10	10	10	0.00	
Yrs_with_Current_Employer	8	8	8	0.00	
Number_of_Cards	1	1	1	0.00	
Creditcard_Debt	237	237	237	0.00	
Loans	1	1	1	0.00	
Predicted outcome	Confidence	Predicted outcome	Confidence	Predicted outcome	Confidence
DENIED	94.00%	DENIED	94.00%	APPROVED	54.00%



Drift detection

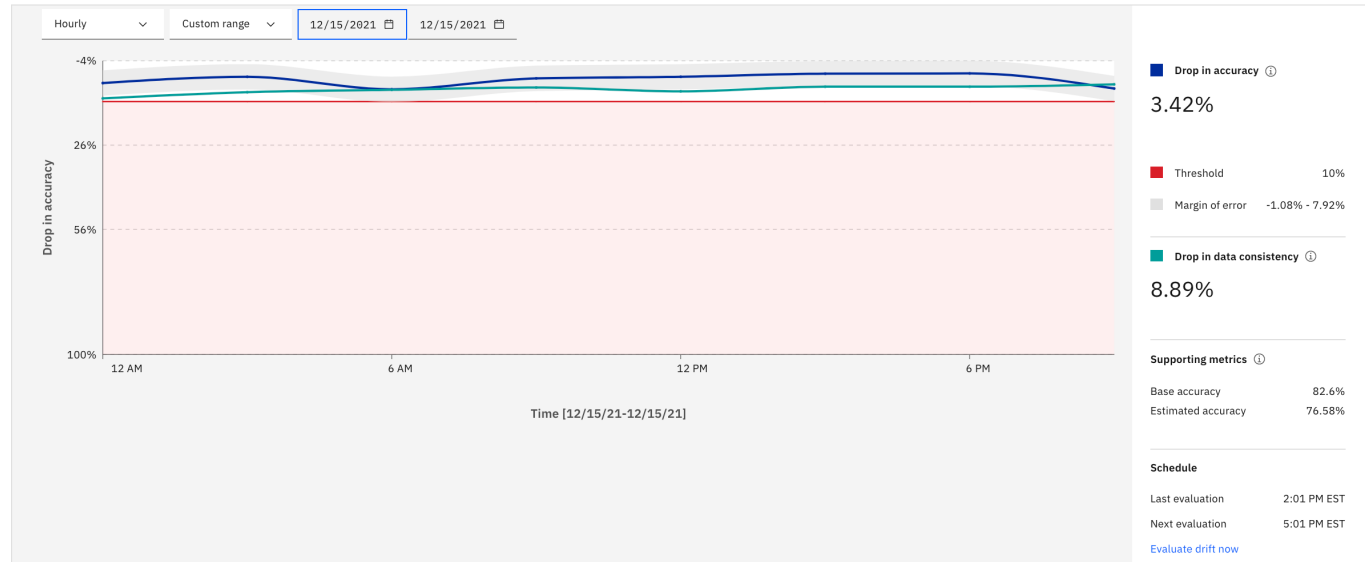
Measure the degree to which a model has moved away from reality

- Drop in accuracy – reality has changed, as shown by the scoring data
- Drop in consistency – reality is the same, the events vary

Drift monitoring and alerts

- Degradation of model performance can trigger retraining and redeployment

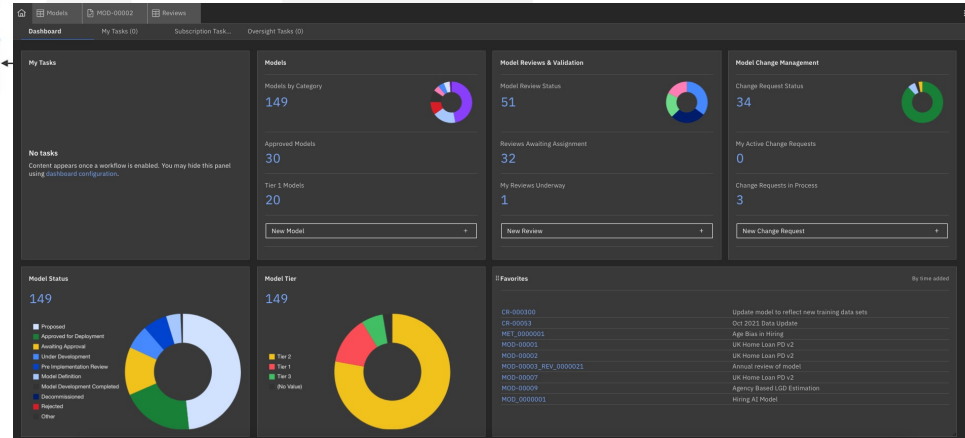
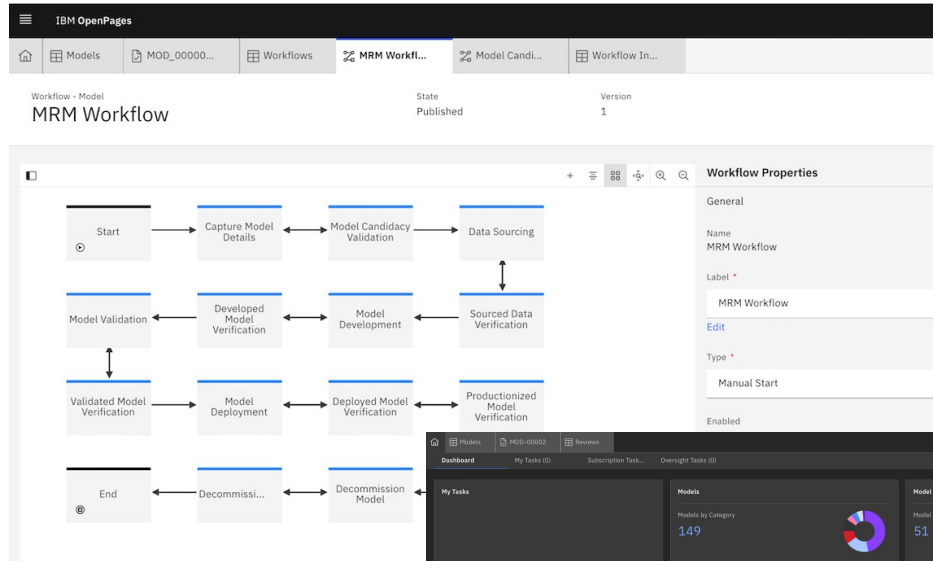
Handle changing scenarios





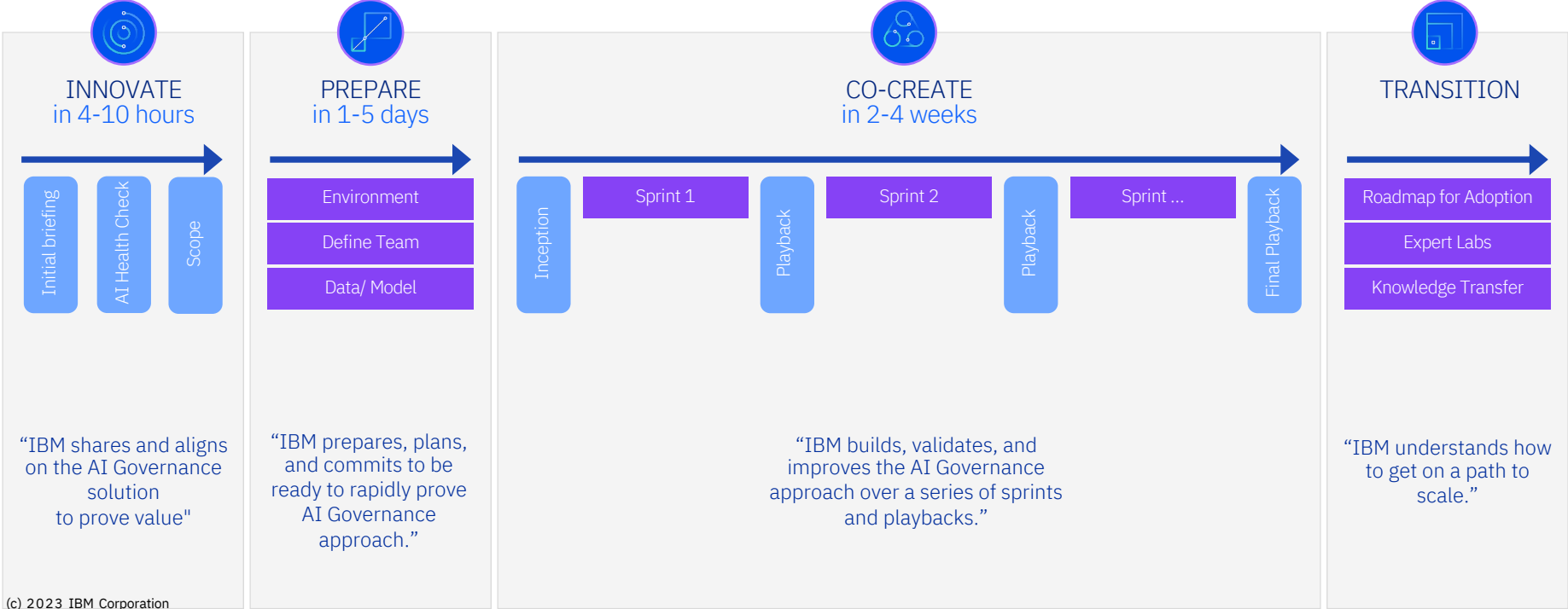
Manage risk across the enterprise – Open Pages

- Consistent holistic views of risk and compliance
- Drive GRC adoption
- Embedded self-service reporting, analytics, and dashboarding



How to get started?

Launch a pilot project on AI Governance



Recently announced at IBM Think 2023

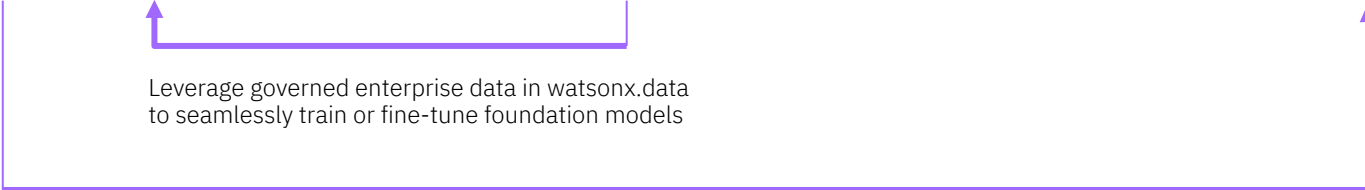
Put AI to work with **watsonx**

Scale and accelerate the impact of AI with trusted data.

Leverage foundation models to automate data search, discovery, and linking in watsonx.data



Leverage governed enterprise data in watsonx.data to seamlessly train or fine-tune foundation models



Enable fine-tuned models to be managed through market leading governance and lifecycle management capabilities

IBM