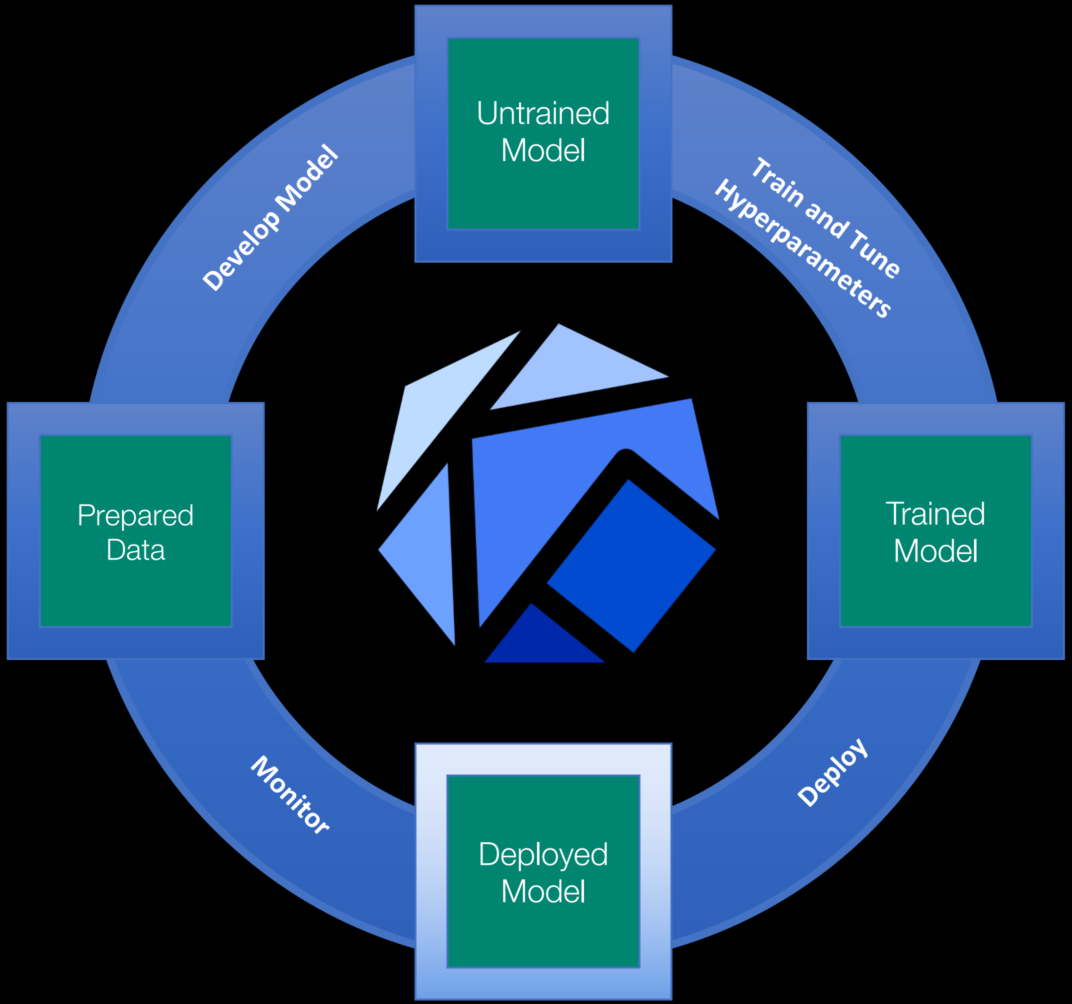
ML Lifecycle



In an MIT & BCG survey of more than 3,000 executives, managers, and analysts across industries...

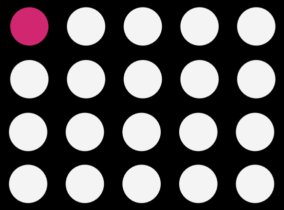
39%

Of all companies have an AI strategy in place

1/5

Has incorporated AI in *some* offerings or products

1/20



Has *extensively* incorporated AI in offerings or process

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

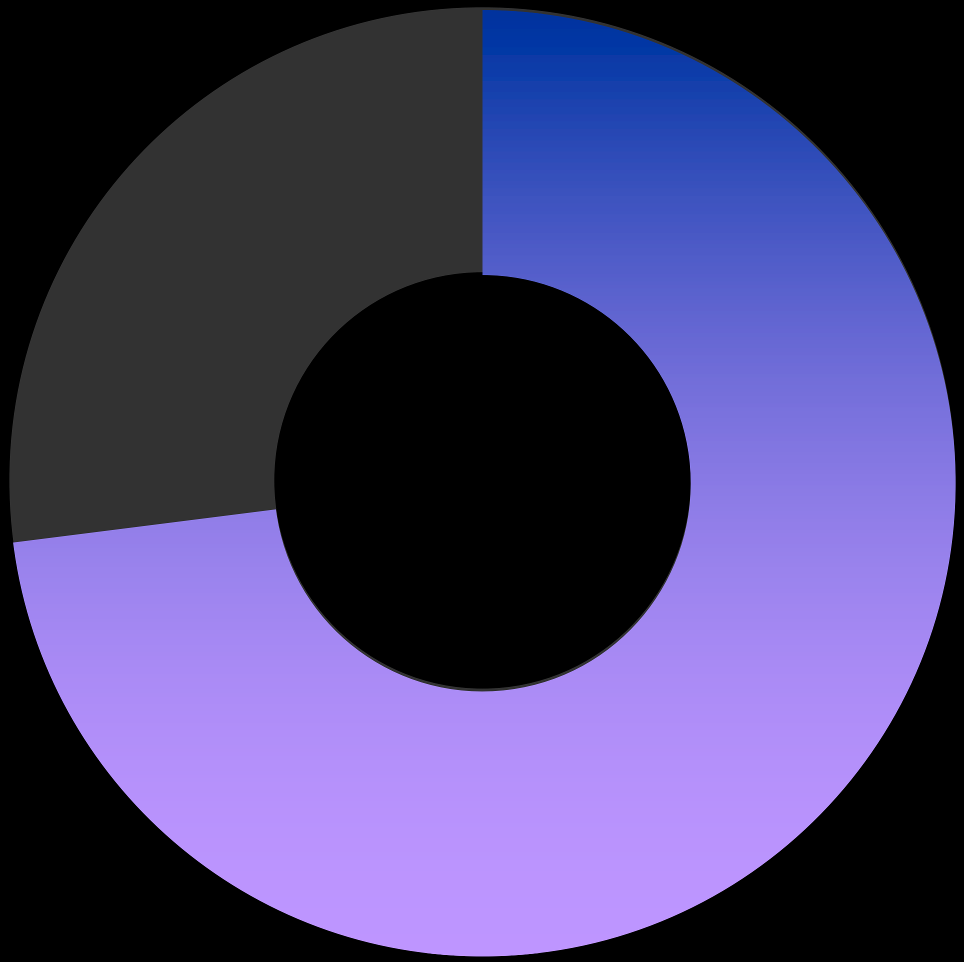
Without expensive Data Science resources handholding multiple AI models in a production application:

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**

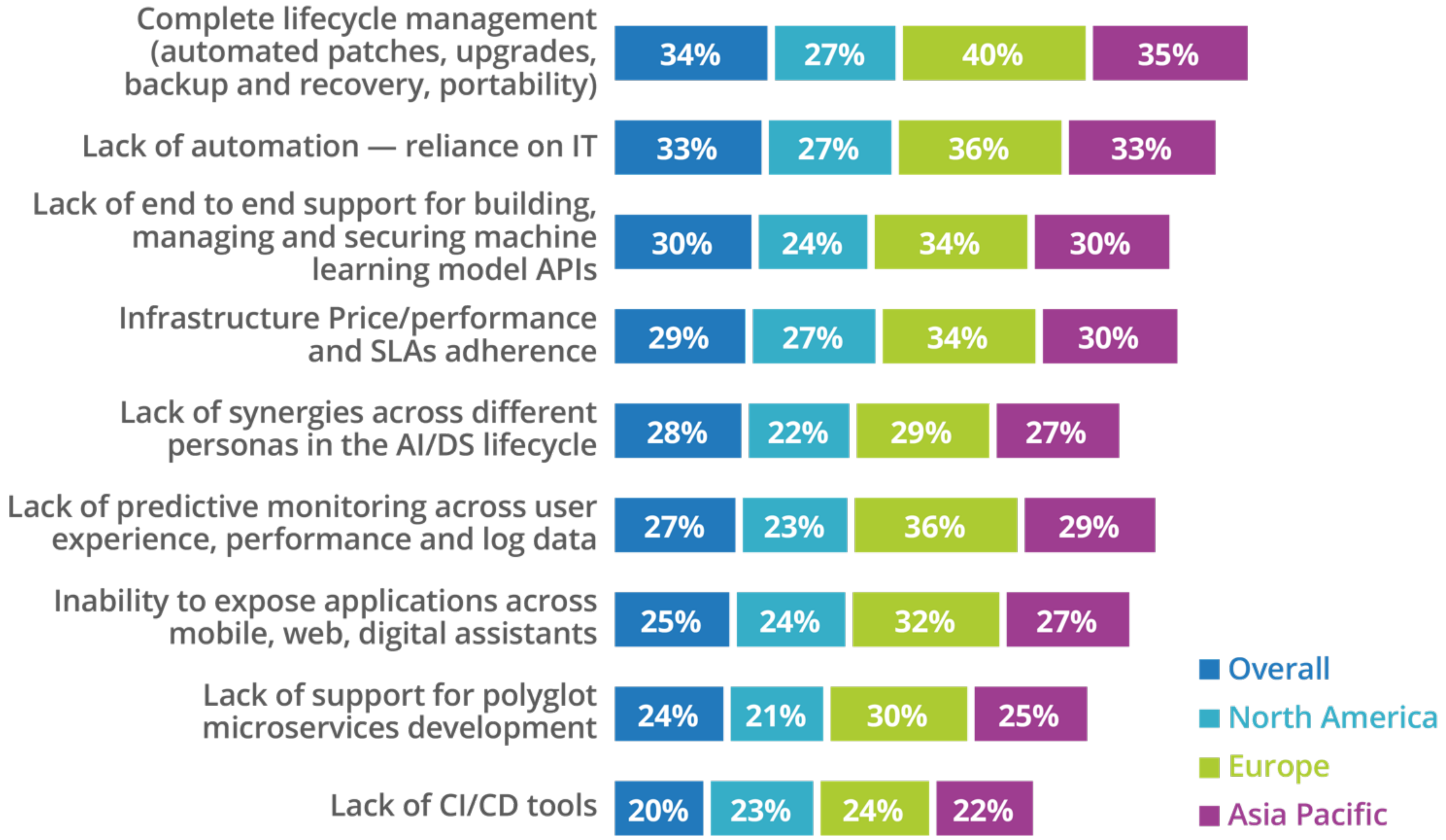
Without an AI and hybrid cloud strategy…

73%

of businesses will fail
at digital transformation



Q. What are your top challenges in end to end AI/ML lifecycle management?



IDC AI StrategiesView

MLOps is critical for realizing AI/ML at scale

Lack of integrated development environment and management are the top AI/ML solutions lifecycle management challenges.

IDC's Point of View

AI requires a radical shift from a deterministic to a probabilistic mindset, with changes required across all departments. Machine Learning Operations (MLOps) is the compound of machine learning, development and operations. It is the practice of collaboration between data scientists, business analysts, data architects and operations professionals to help manage production ML/DL lifecycle. The MLOps toolchain needs to provide visibility, managed access control, and collaboration for all such parties. MLOps should integrate with existing DevOps while also delivering the additional unique capabilities required to manage ML.

Enterprise Machine Learning

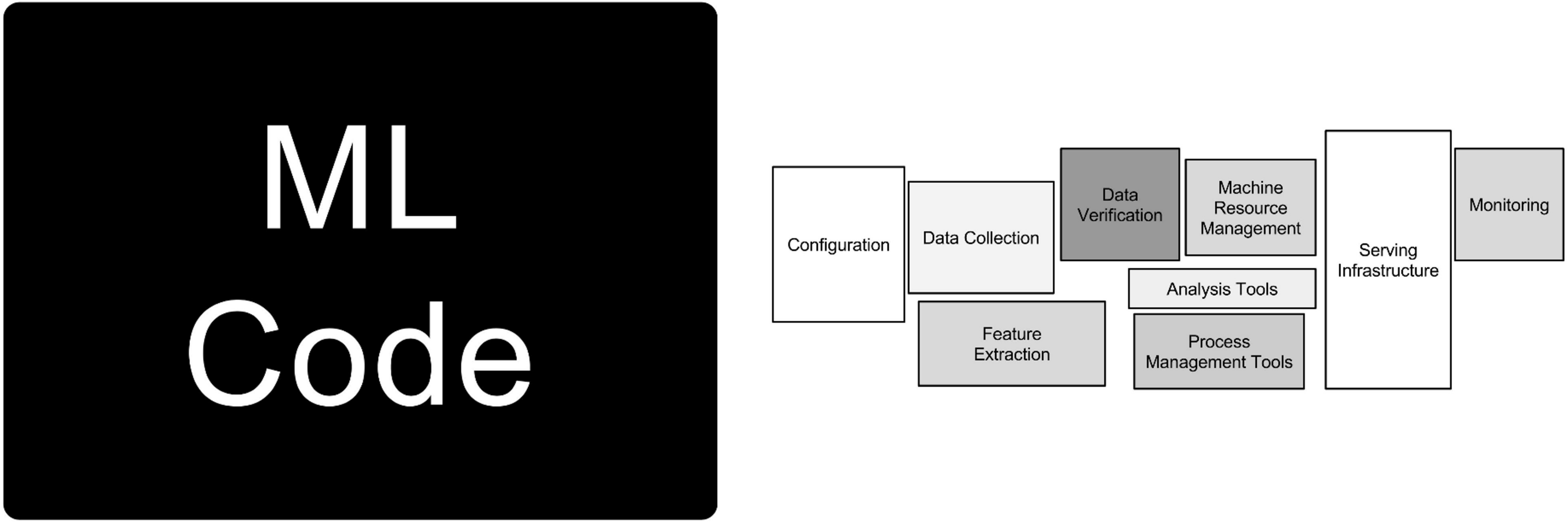
ginablaber
@ginablaber

Follow

The story of enterprise Machine Learning: “It took me 3 weeks to develop the model. It’s been >11 months, and it’s still not deployed.” @DineshNirmalIBM #StrataData #strataconf

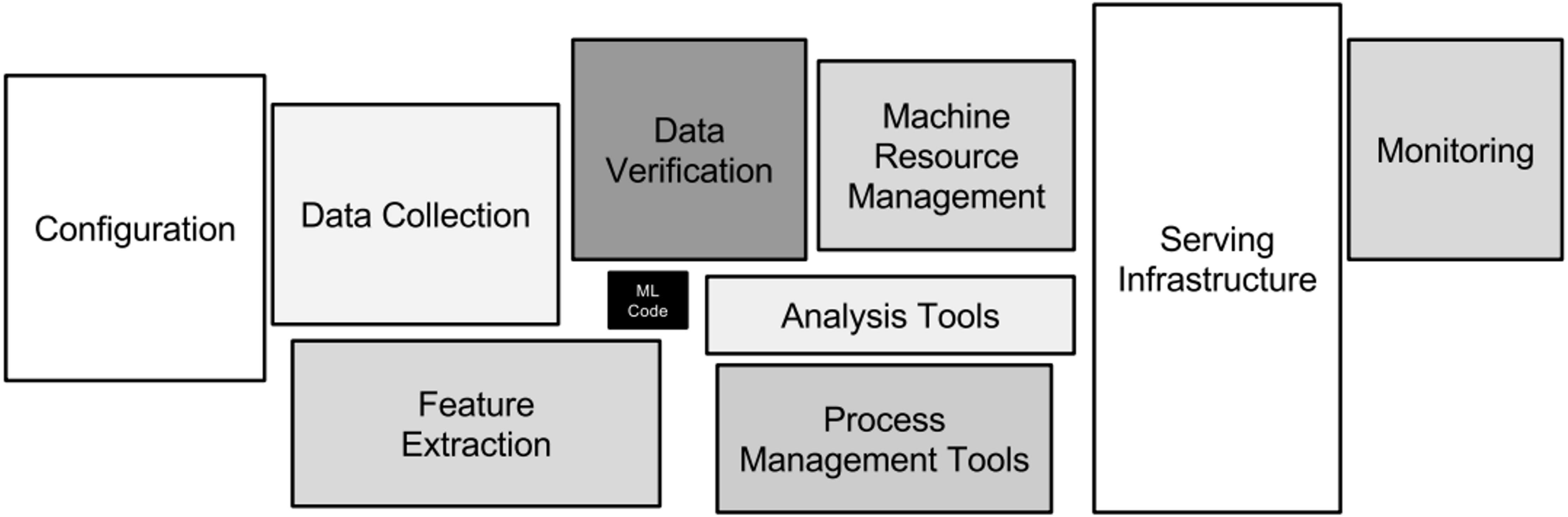
10:19 AM - 7 Mar 2018

Perception



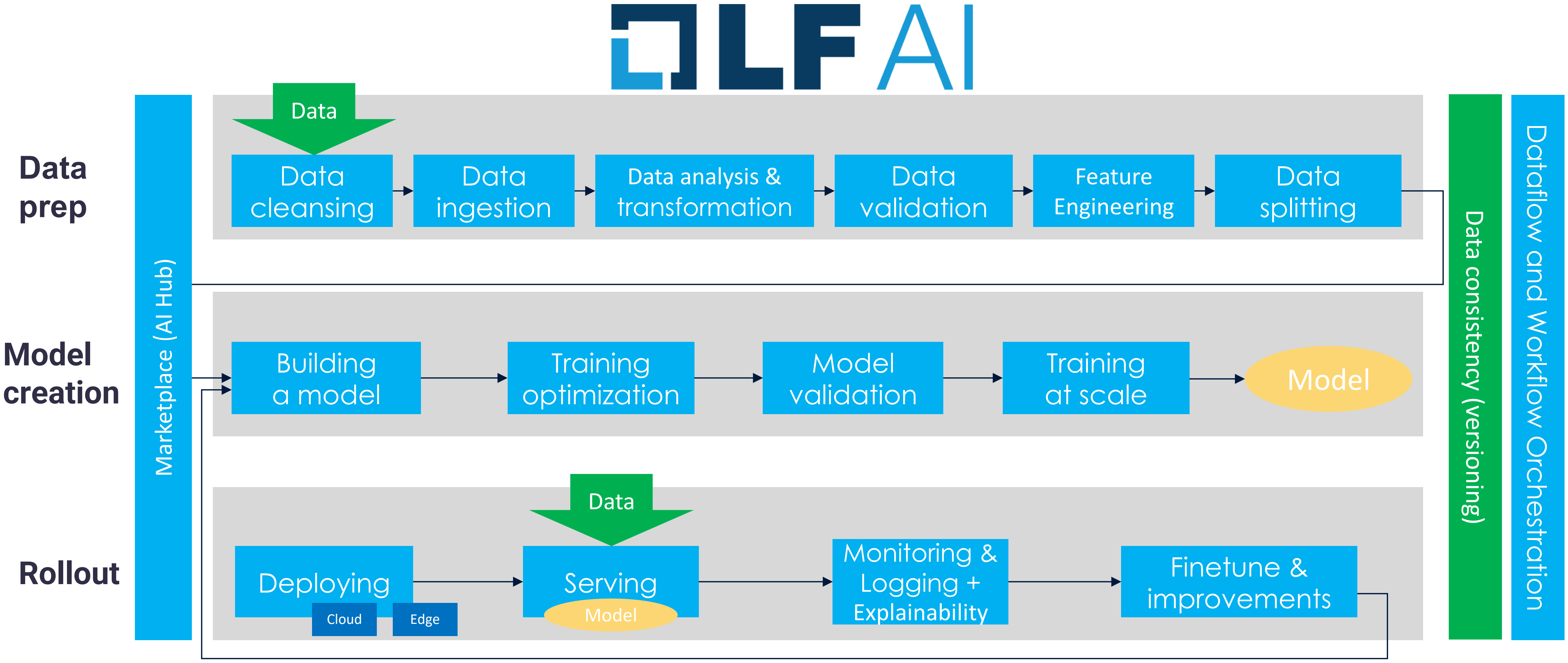
Source: Hidden Technical Debt in Machine Learning Systems

In reality…ML Code is tiny part in this overall platform

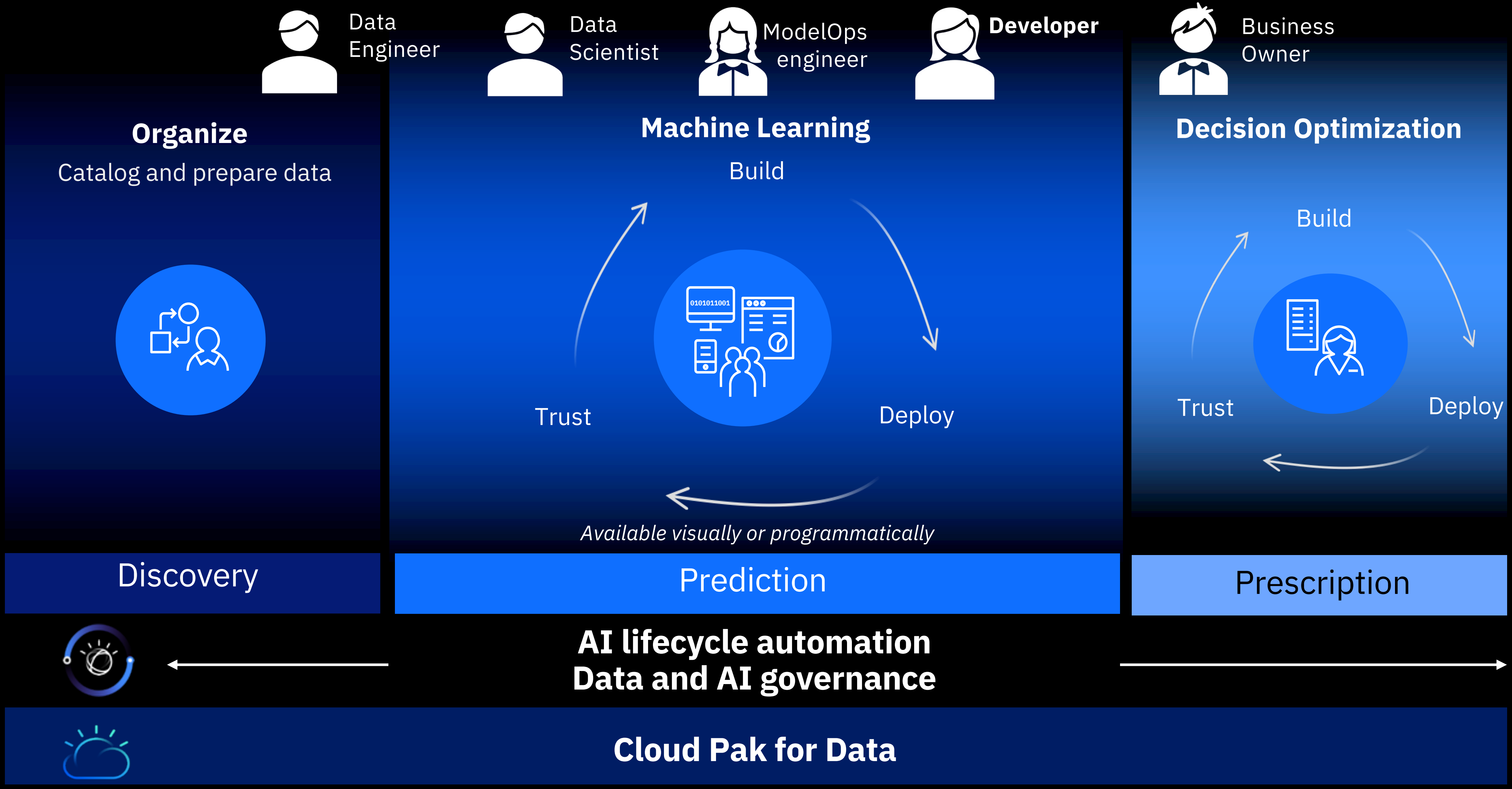


[*Source: Hidden Technical Debt in Machine Learning Systems](#)

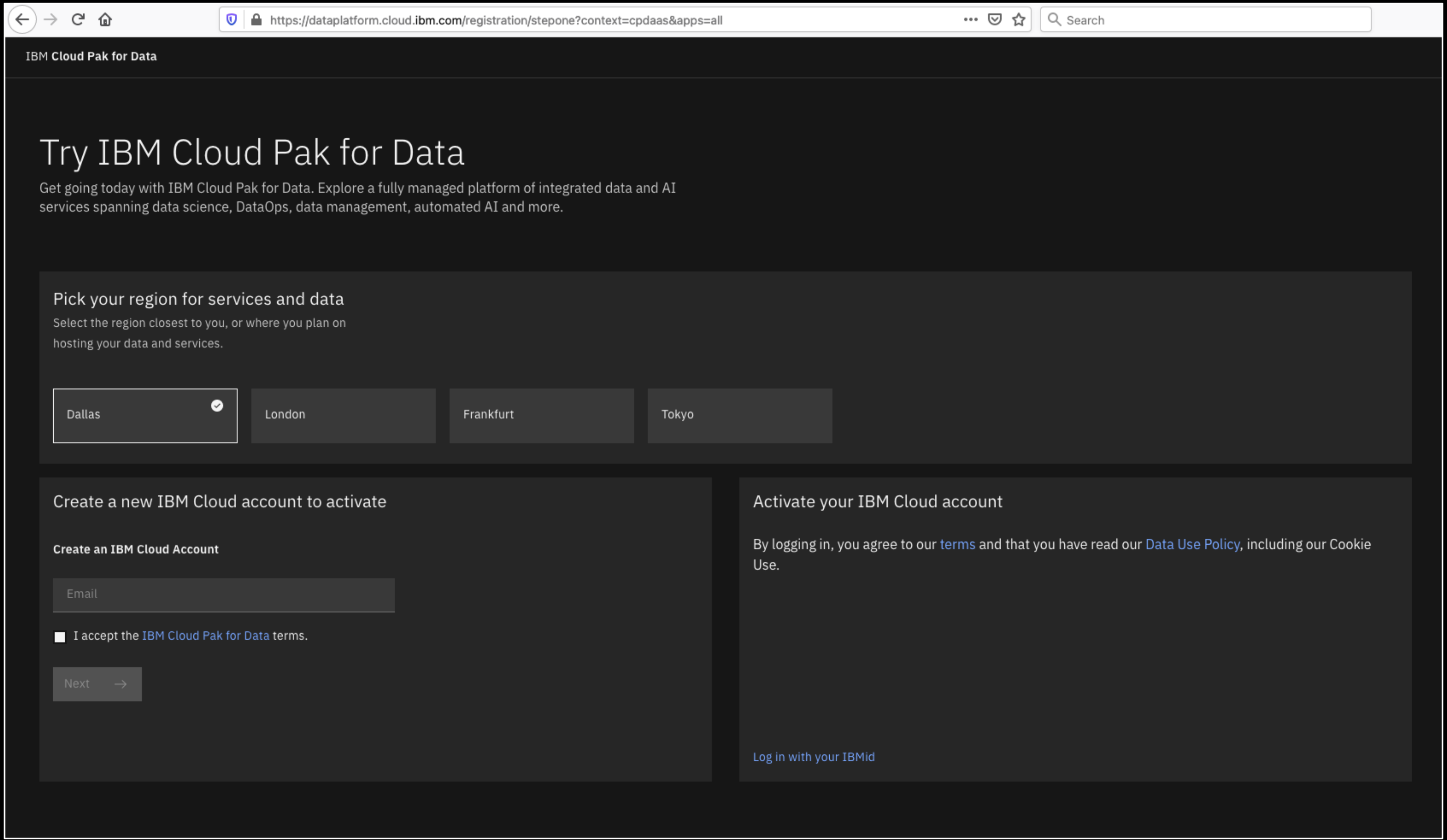
ML Lifecycle



IBM Watson Studio to build, deploy, and trust AI models



Cloud Pak for Data as a Service



dataplatform.cloud.ibm.com

Magic Quadrant for Data Science and Machine Learning Platforms

IBM is positioned as a Leader in Gartner Magic Quadrant for Data Science and Machine Learning Platforms*

Gartner Disclaimer:

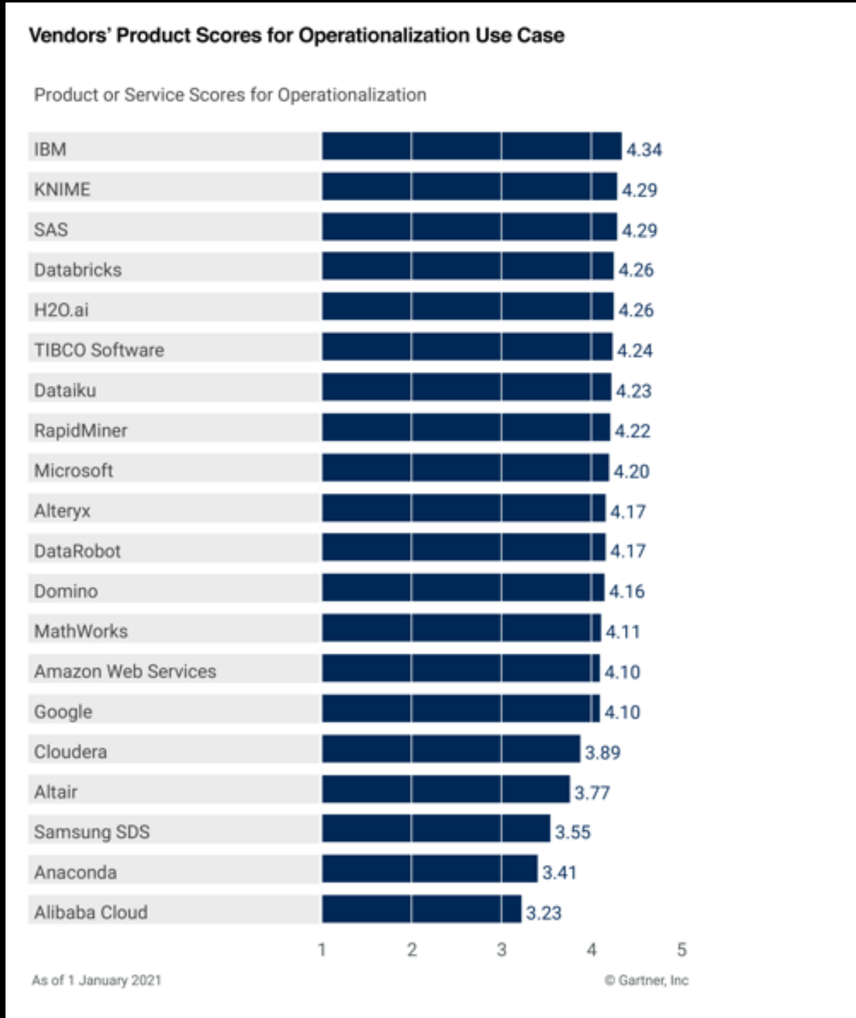
**Gartner, Magic Quadrant for Data Science and Machine Learning Platforms, Peter Krensky | Carlie Idoine | Erick Brethenoux | Pieter den Hamer | Farhan Choudhary | Afraz Jaffri | Shubhangi Vashisth, March 1, 2021*

Gartner does not endorse any vendor, product or service depicted in our research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

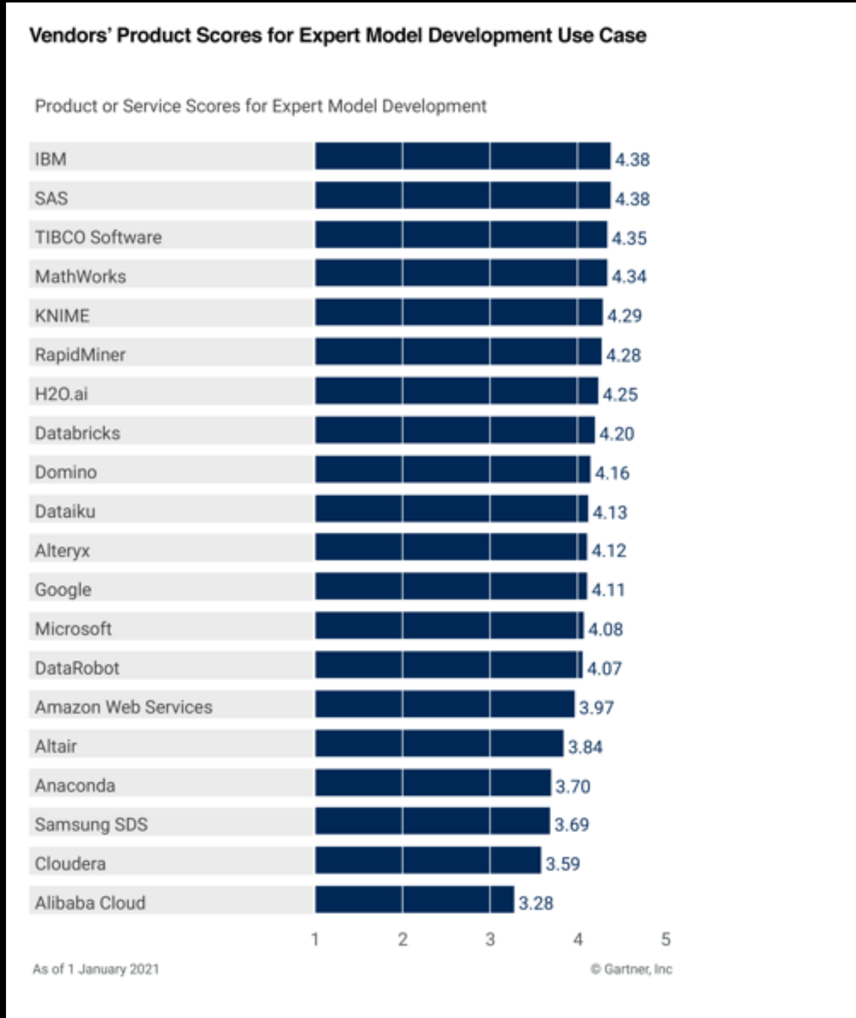
This graphic was published by Gartner, Inc. as part of a larger research document and should be evaluated in the context of the entire document. The Gartner document is available upon request from IBM.

LINK

IBM receives highest scores in two use cases from Gartner Critical Capabilities Report*



Source: Gartner (March 2021)



Source: Gartner (March 2021)

LINK

Gartner Disclaimer:

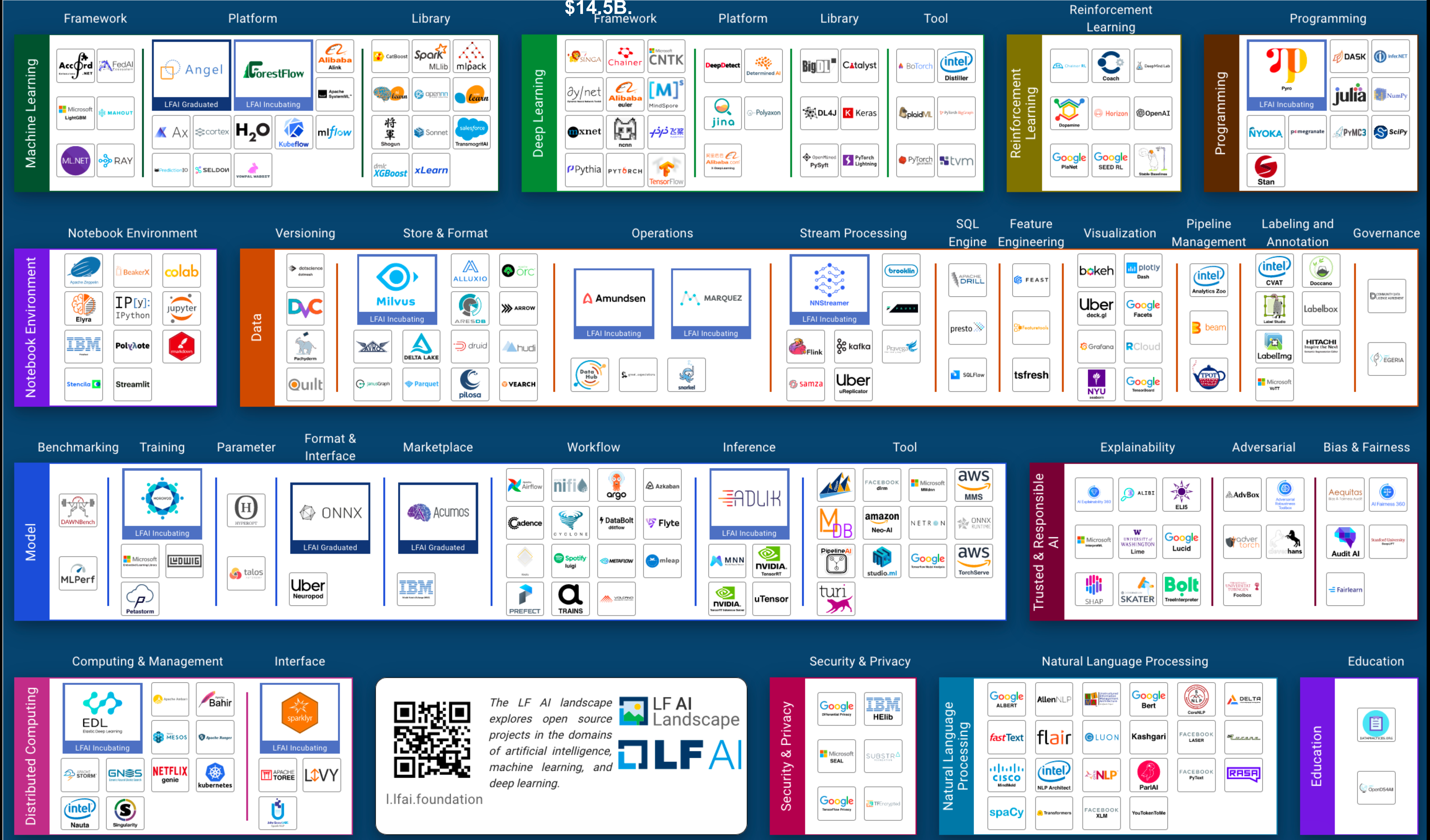
Gartner, Critical Capabilities for Data Science and Machine Learning Platforms, Pieter den Hamer, Carlie Idoine, Shubhangi Vashisth, Peter Krensky, Erick Brethenoux, Farhan Choudhary, Afraz Jaffri

The dominant use cases for AI for business

		planning to deploy within 2 years	Watson deployments
	Customer service	41%	
	Financial planning	40%	
	Employee experience	38%	
	Regulatory compliance	38%	
	IT Operations	36%	

LFAI and Data Landscape

You are viewing 316 cards with a total of 2,308,365 stars, market cap of $18T and funding of $14.5B.



Innovate with your open source tools on a multicloud data and platform

IBM's long history and dedication to open source extends to our data and AI strategies.

With IBM's creation of CODAIT (the Center for Open source, Data and AI Technology), we support multiple Open Source technologies

Languages

python™

R

Scala

Frameworks /Libraries / Runtimes

Jupyter
Apache Spark
Scikitlearn
XGBoost

TensorFlow
Keras
Caffe
PyTorch

Trusted AI

Adversarial Robustness Toolbox

AI Explainability 360

AI Fairness 360

Modern infrastructure

Kubernetes

Kubeflow

docker

Red Hat OpenShift

IBM public cloud

AWS

Microsoft Azure

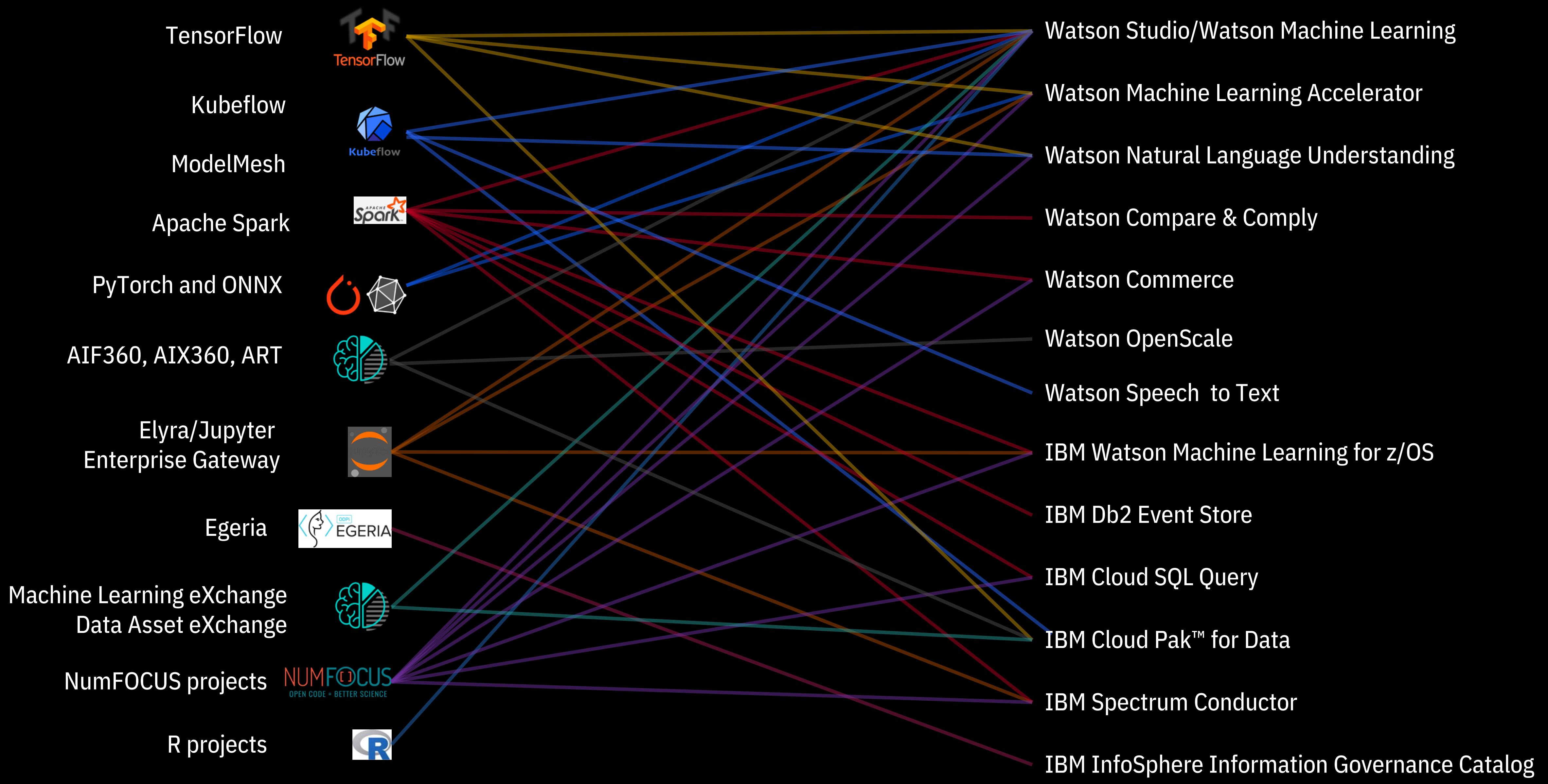
Google Cloud

Private

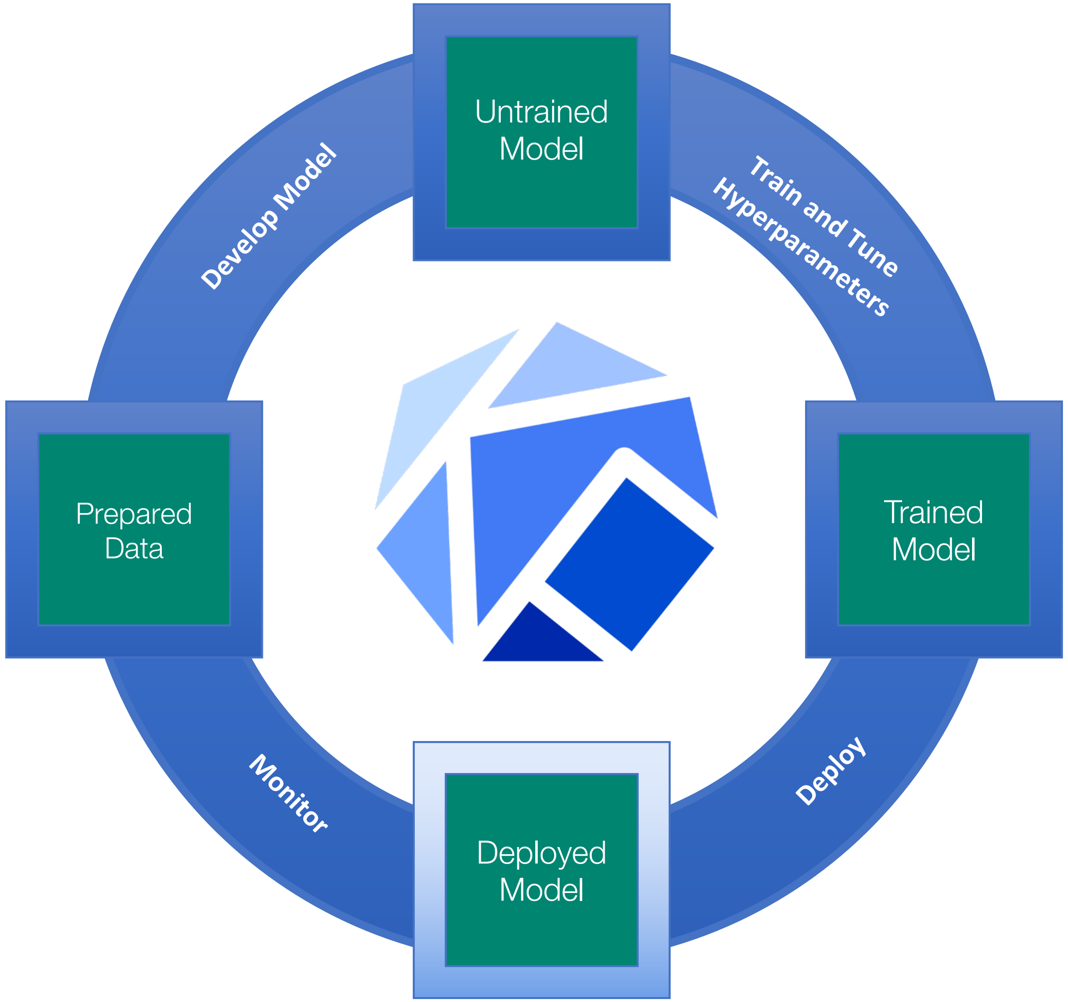
IBM **Z**
IBM **LinuxOne**
IBM **Power** Systems

End points

Open Source in Watson AI and Data



ML Lifecycle - Build: Model Development, Training and HPO

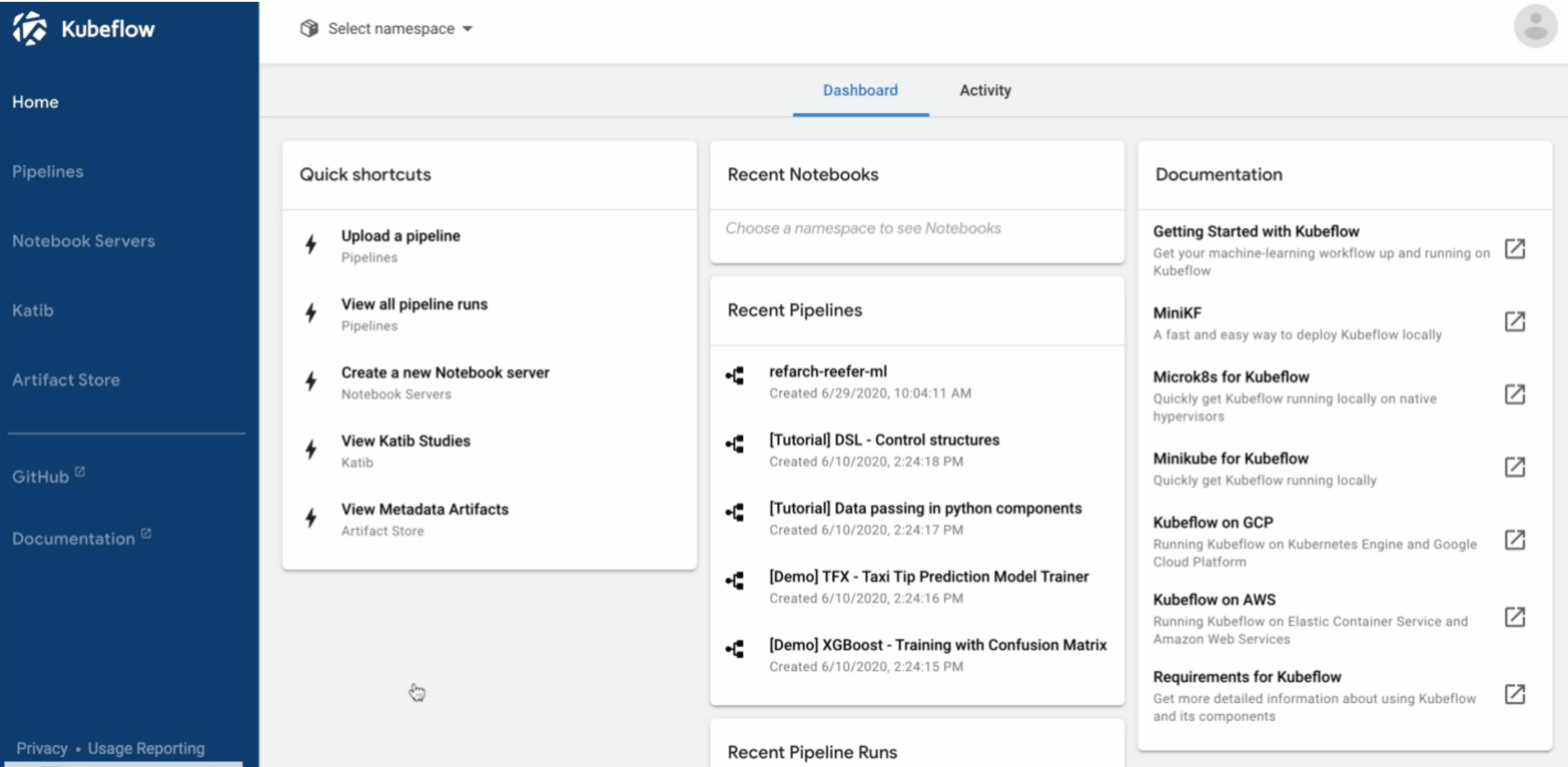


Develop (Jupyter Notebooks)

- Data Scientist
- Self-service Jupyter Notebooks provide faster model experimentation
- Simplified configuration of CPU/GPU, RAM, Persistent Volumes
- Faster model creation with training operators, TFX, magics, workflow automation (Kale, Fairing)
- Simplify access to external data sources (using stored secrets)
- Easier protection, faster restoration & sharing of "complete" notebooks

- IT Operator
- Profile Controller, Istio, Dex enable secure RBAC to notebooks, data & resources
- Smaller base container images for notebooks, fewer crashes, faster to recover

Develop (Jupyter Notebooks)



Distributed Training Operators

	TF Operator	PyTorch Operator	MPI Operator
Framework Support	TensorFlow	PyTorch	HOROVOD TensorFlow/Keras Apache MXNet/PyTorch/OpenMPI
Distribution Strategy & Backend	tf.distribute MPI/NCCL/PS/TPU	torch.distributed Gloo/MPI/NCCL	horovod DistributedOptimizer Gloo/MPI/NCCL

Distributed Training Operators

tf-operator
Tools for ML/Tensorflow on Kubernetes.
Jsonnet · Apache-2.0 · 323

pytorch-operator
PyTorch on Kubernetes
Jsonnet · Apache-2.0 · 87

mpi-operator
Kubernetes Operator for Allreduce-style
kubernetes · tensorflow · mpi · dist
horovod · kubeflow
Go · Apache-2.0 · 83 · 125

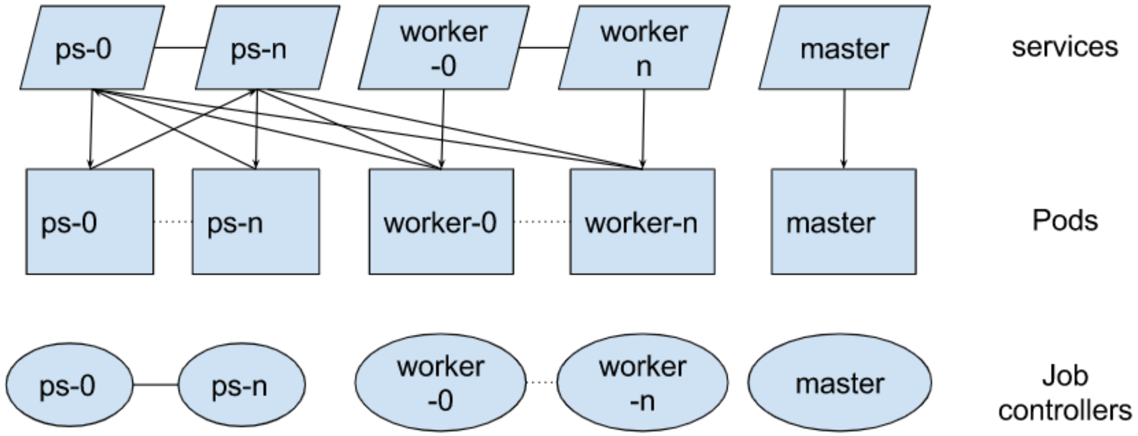
xgboost-operator
Incubating project for xgboost operator
Go · Apache-2.0 · 23 · 41

mxnet-operator
A Kubernetes operator for mxnet jobs
Go · Apache-2.0 · 20 · 50

chainer-operator
Repository for chainer operator
Go · Apache-2.0 · 9 · 12

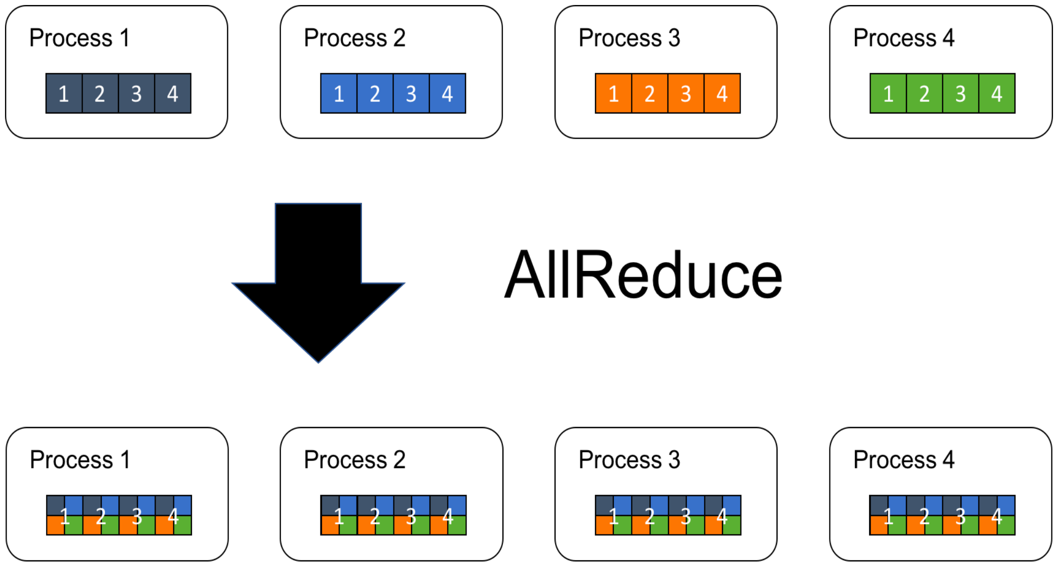
Distributed Tensorflow Operator

- A distributed Tensorflow Job is collection of the following processes
  - Chief – The chief is responsible for orchestrating training and performing tasks like checkpointing the model
  - Ps – The ps are parameters servers; the servers provide a distributed data store for the model parameters to access
  - Worker – The workers do the actual work of training the model. In some cases, worker 0 might also act as the chief
  - Evaluator - The evaluators can be used to compute evaluation metrics as the model is trained



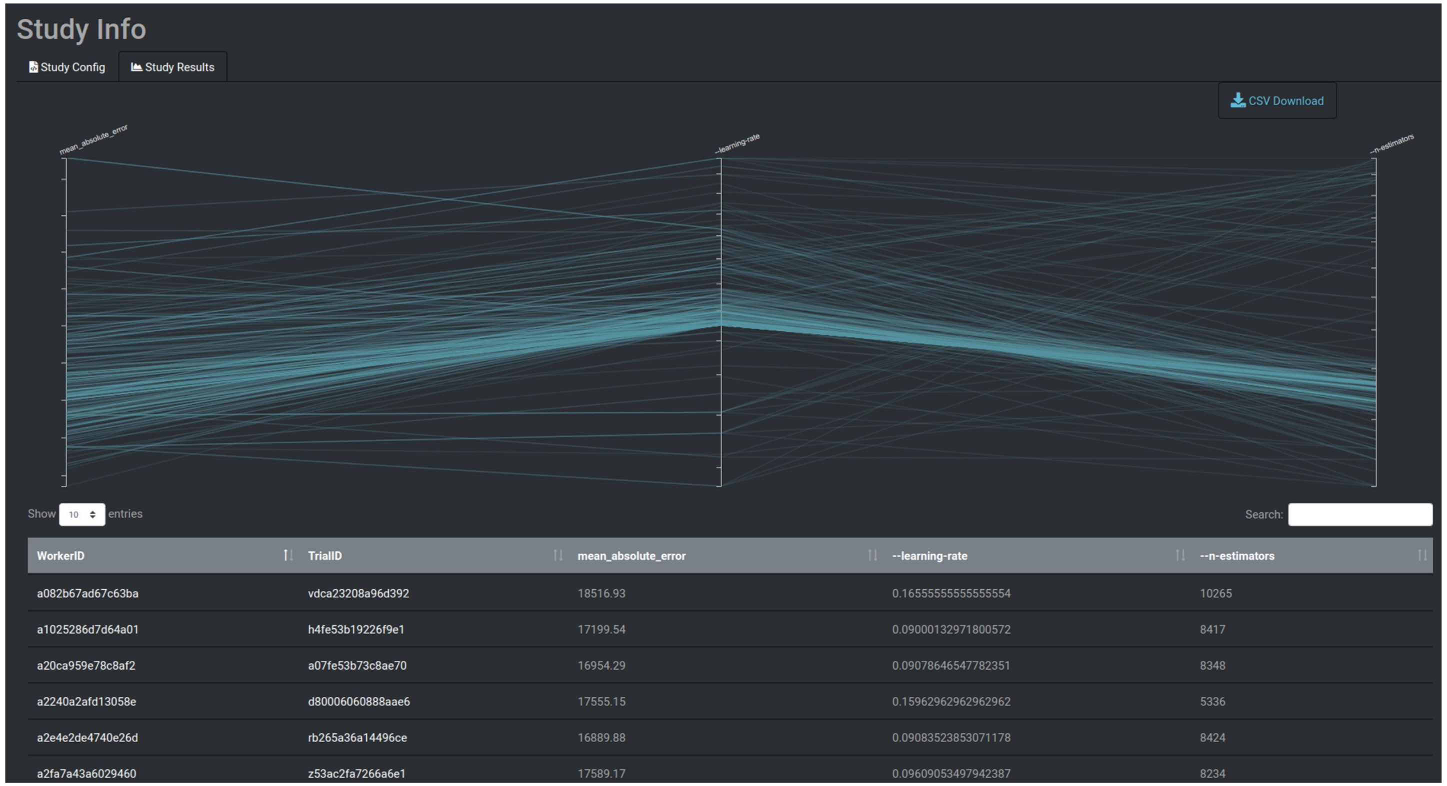
Distributed MPI Operator - AllReduce

- AllReduce is an operation that reduces many arrays spread across multiple processes into a single array which can be returned to all the processes
- This ensures consistency between distributed processes while allowing all of them to take on different workloads
- The operation used to reduce the multiple arrays back into a single array can vary and that is what makes the different options for AllReduce



Hyper Parameter Optimization and Neural Architecture Search

- Hyperparameter Tuning
  - ❑ Random Search
  - ❑ Tree of Parzen Estimators (TPE)
  - ❑ Grid Search
  - ❑ Hyperband
  - ❑ Bayesian Optimization
  - ❑ CMA Evolution Strategy
- Neural Architecture Search
  - ❑ Efficient Neural Architecture Search (ENAS)
  - ❑ Differentiable Architecture Search (DARTS)



WorkerID	TrialID	mean_absolute_error	--learning-rate	--n-estimators
a082b67ad67c63ba	vdca23208a96d392	18516.93	0.16555555555555554	10265
a1025286d7d64a01	h4fe53b19226f9e1	17199.54	0.09000132971800572	8417
a20ca959e78c8af2	a07fe53b73c8ae70	16954.29	0.09078646547782351	8348
a2240a2afd13058e	d80006060888aae6	17555.15	0.15962962962962962	5336
a2e4e2de4740e26d	rb265a36a14496ce	16889.88	0.09083523853071178	8424
a2fa7a43a6029460	z53ac2fa7266a6e1	17589.17	0.09609053497942387	8234

TensorFlow PYTORCH HOROVOD

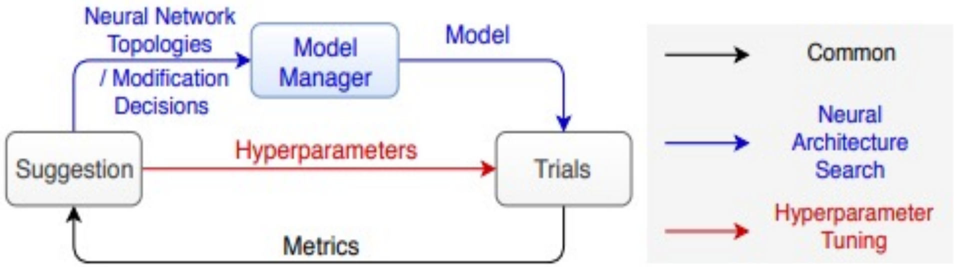
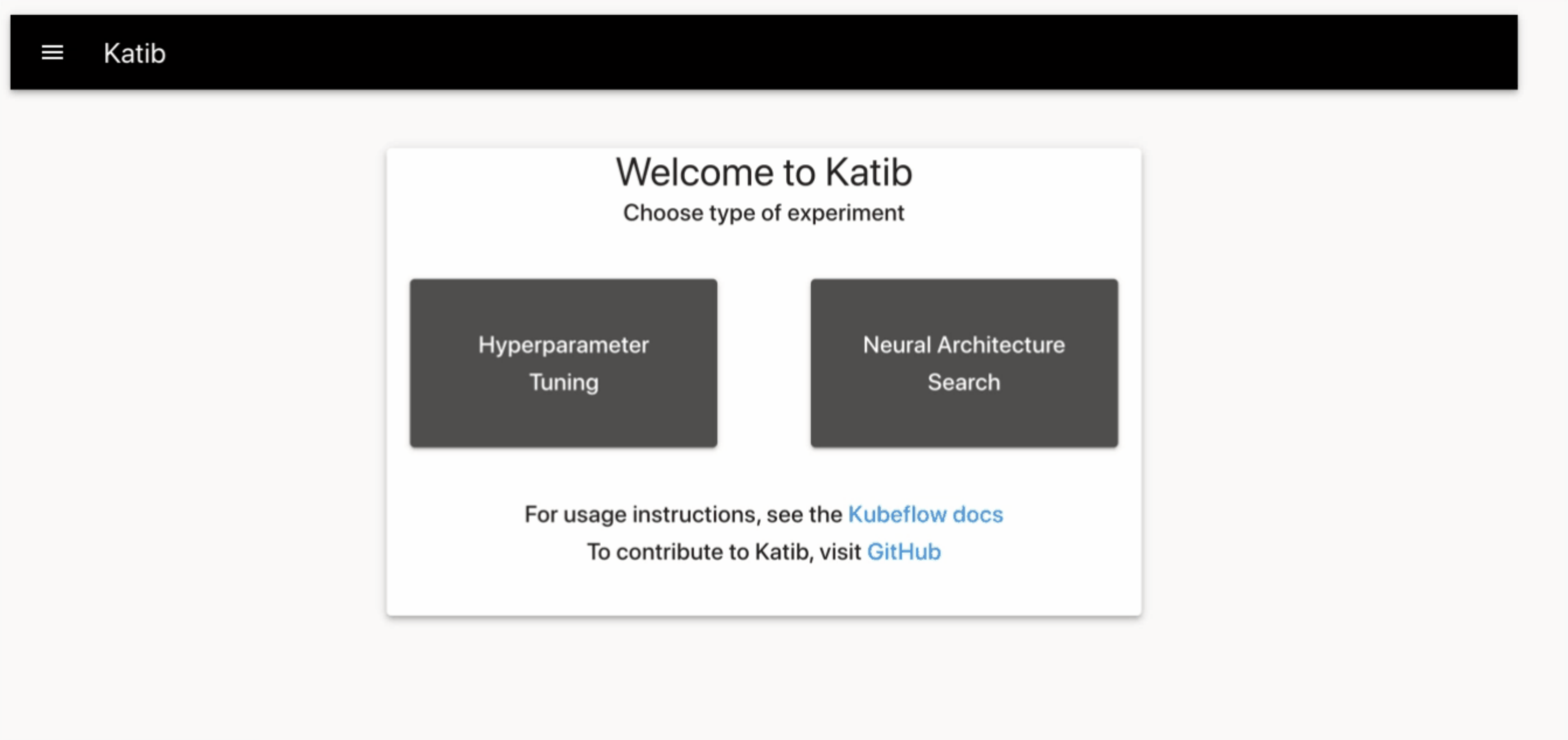


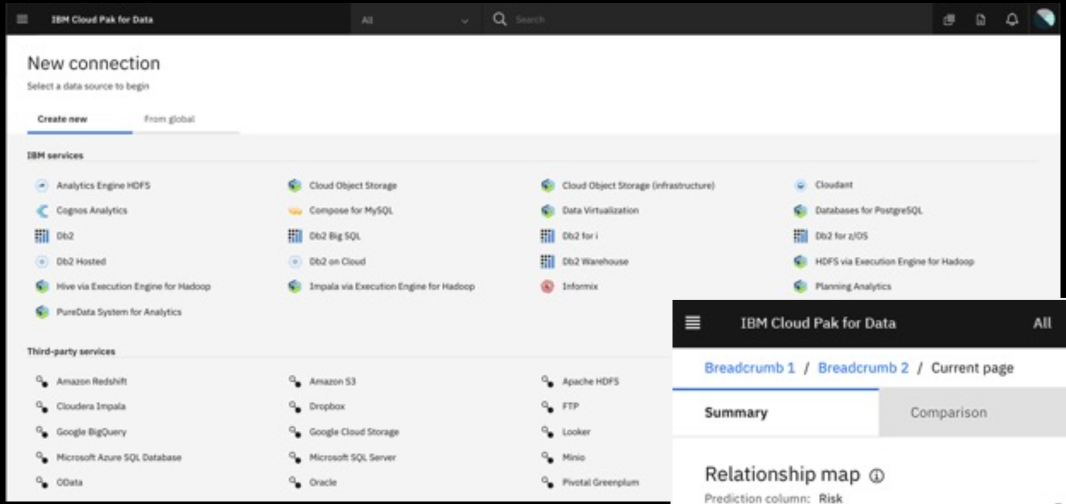
Figure 1: Summary of AutoML workflows

Hyper Parameter Optimization

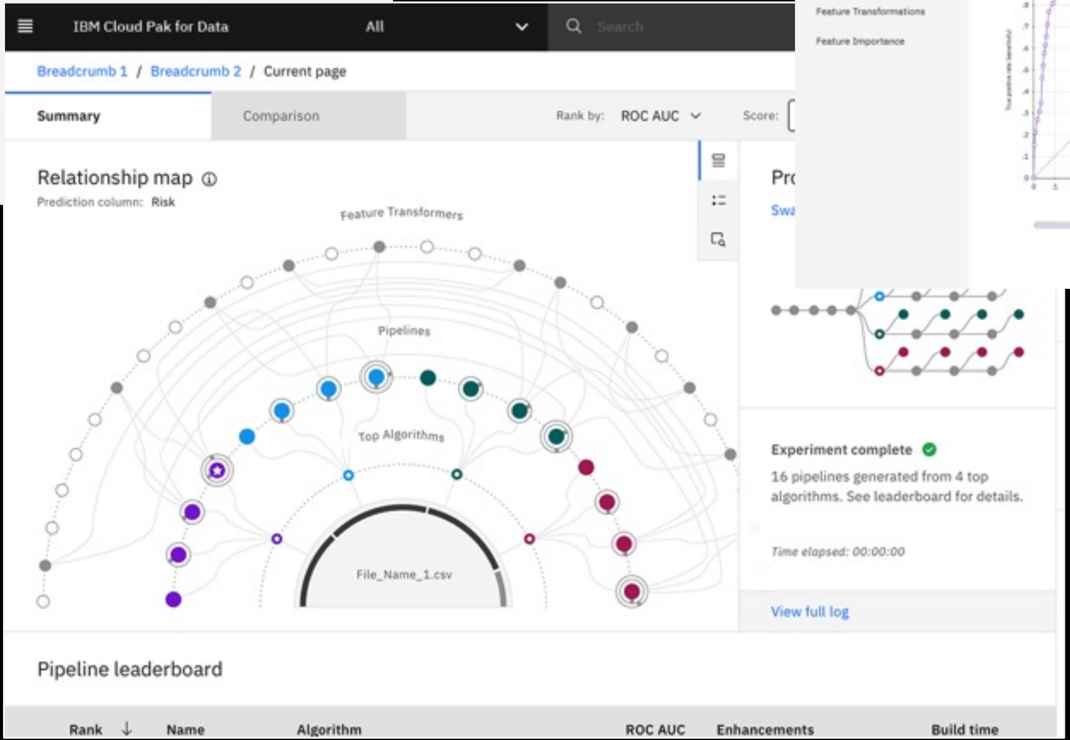


Automate and simplify AI lifecycles

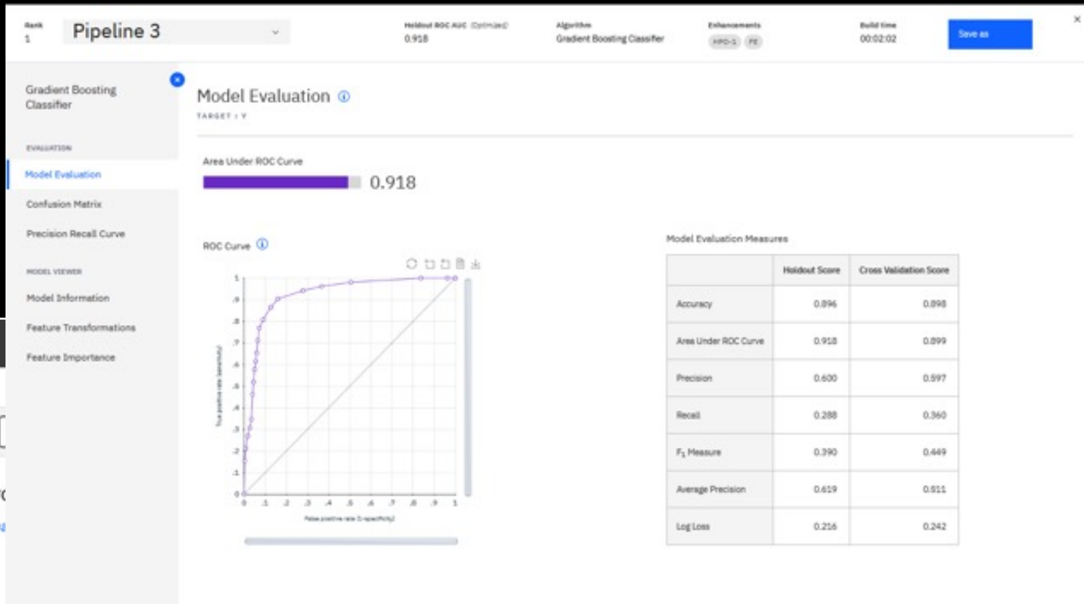
Build ModelOps with all teams



Collect: Connect and access data to create data catalogs



AutoAI: Automate data preparation, featuring engineering and hyper parameter optimization

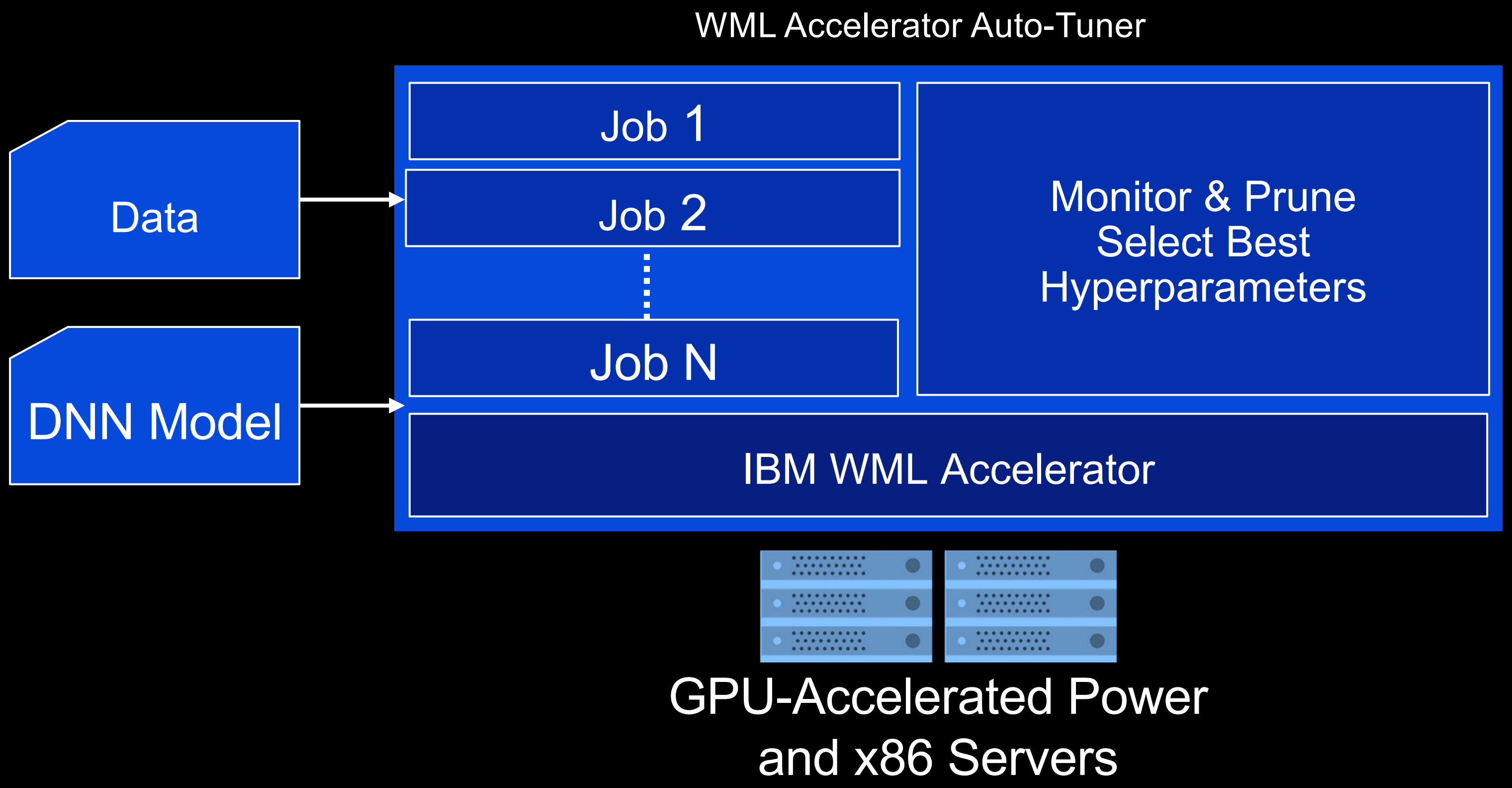


AutoAI; Evaluate models and understand relationships

Auto Hyperparameter Optimization – WML Accelerator

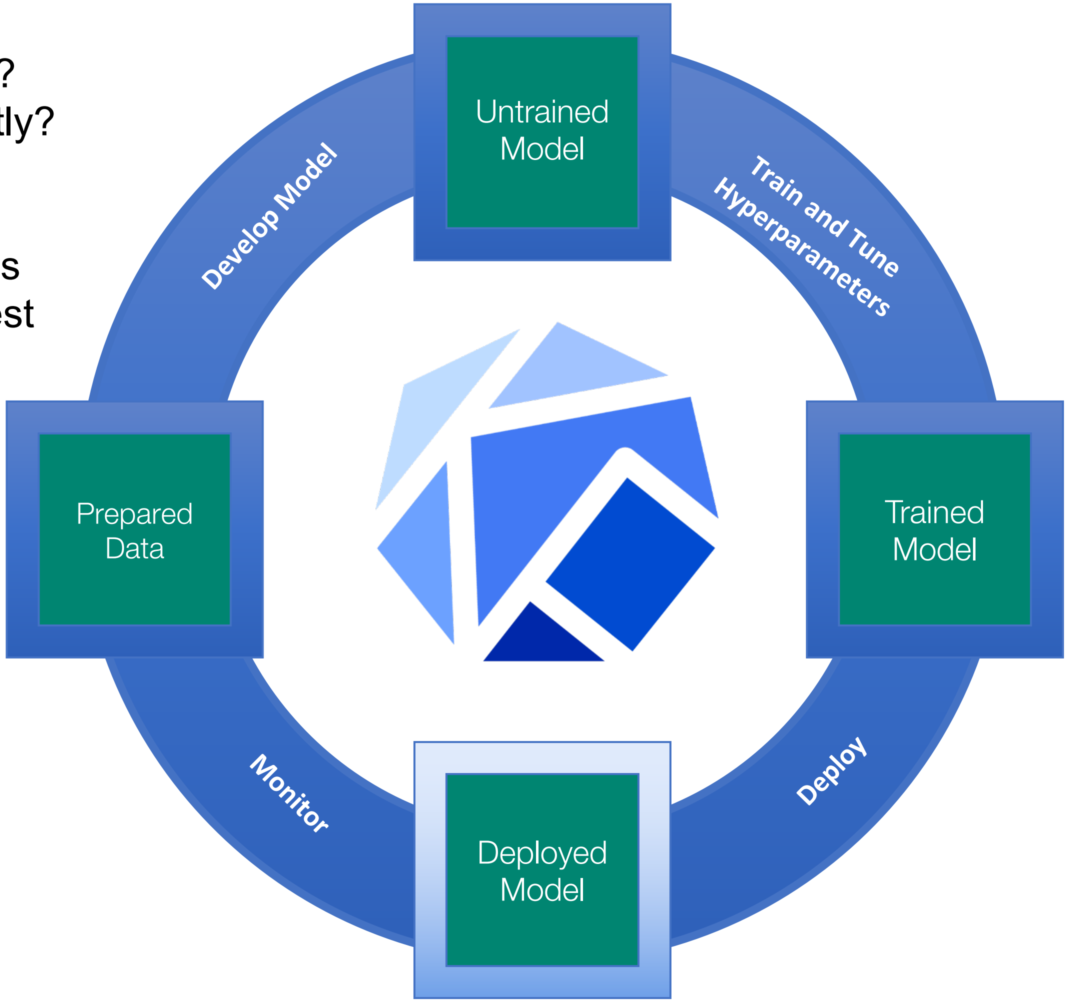
- Data scientists run 100s of jobs with different hyperparameters
  - Learning rate, Decay rate, Batch size, Optimizers (GradientDescent, Adadelta, Momentum, RMSProp, …)

- Auto-Tuner searches for good hyperparameters by launching 10s of jobs in parallel with 10,000s of iterations and selecting the best ones
  - 4 search algorithms: Random, Tree-based Parzen Estimator (TPE), Bayesian, Hyperband
  - Bring Your Own Algorithm



ML Lifecycle: Production Model Serving

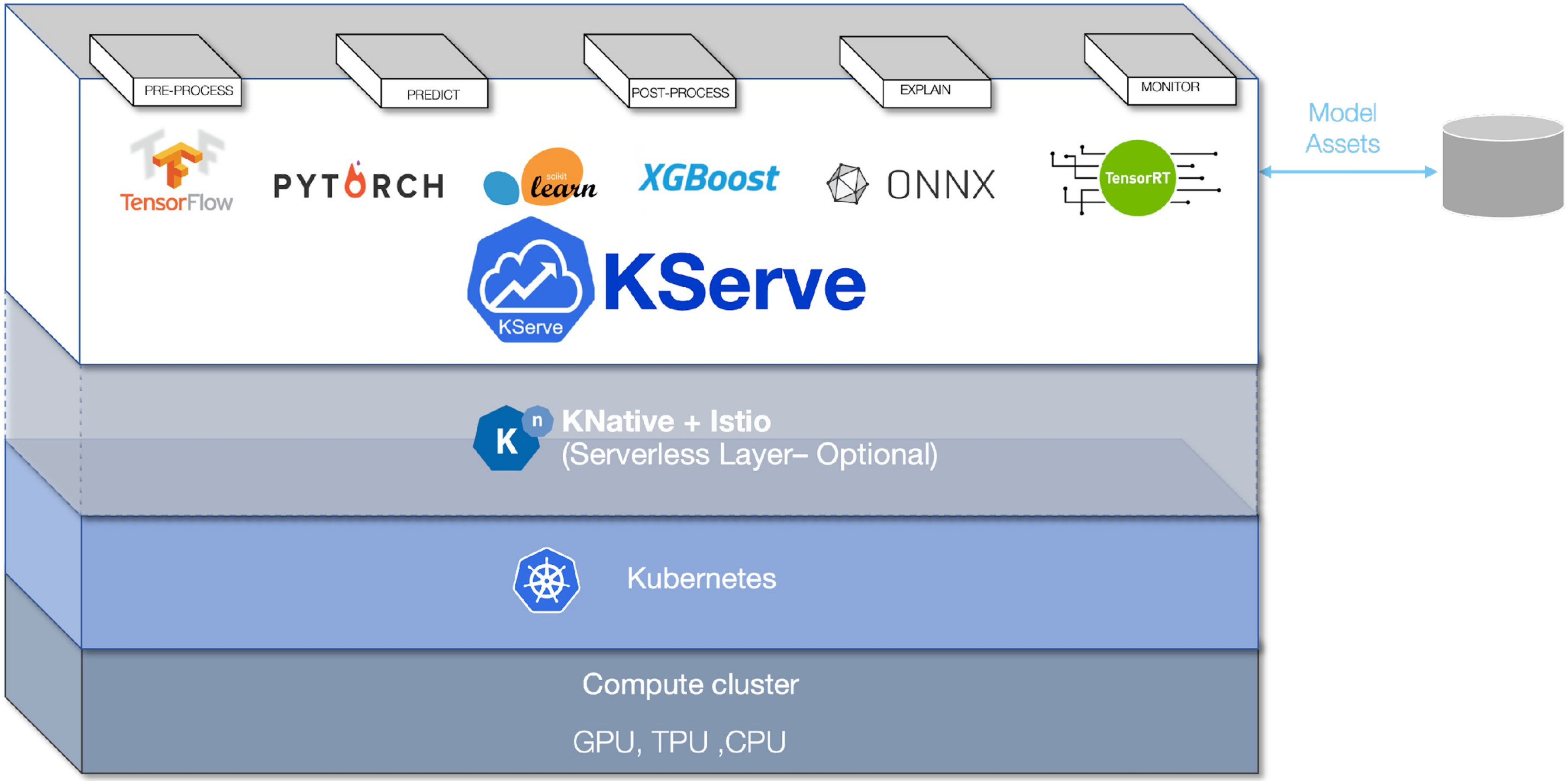
- ❑ Cost:
  Is the model over or under scaled?
  Are resources being used efficiently?
- ❑ Monitoring:
  Are the endpoints healthy? What is the performance profile and request trace?
- ❑ Rollouts:
  Is this rollout safe? How do I roll back? Can I test a change without swapping traffic?
- ❑ Protocol Standards:
  How do I make a prediction?
  GRPC? HTTP? Kafka?



- ❑ How do I handle batch predictions?
- ❑ How do I leverage standardized Data Plane protocol so that I can move my model across MLServing platforms?
- ❑ Frameworks:
  How do I serve on Tensorflow? XGBoost? Scikit Learn? Pytorch? Custom Code?
- ❑ Features:
  How do I explain the predictions? What about detecting outliers and skew? Bias detection? Adversarial Detection?
- ❑ How do I wire up custom pre and post processing

Here comes KServe!

Highly scalable and standards based Model Inferencing Platform on Kubernetes for Trusted AI



KServe overview

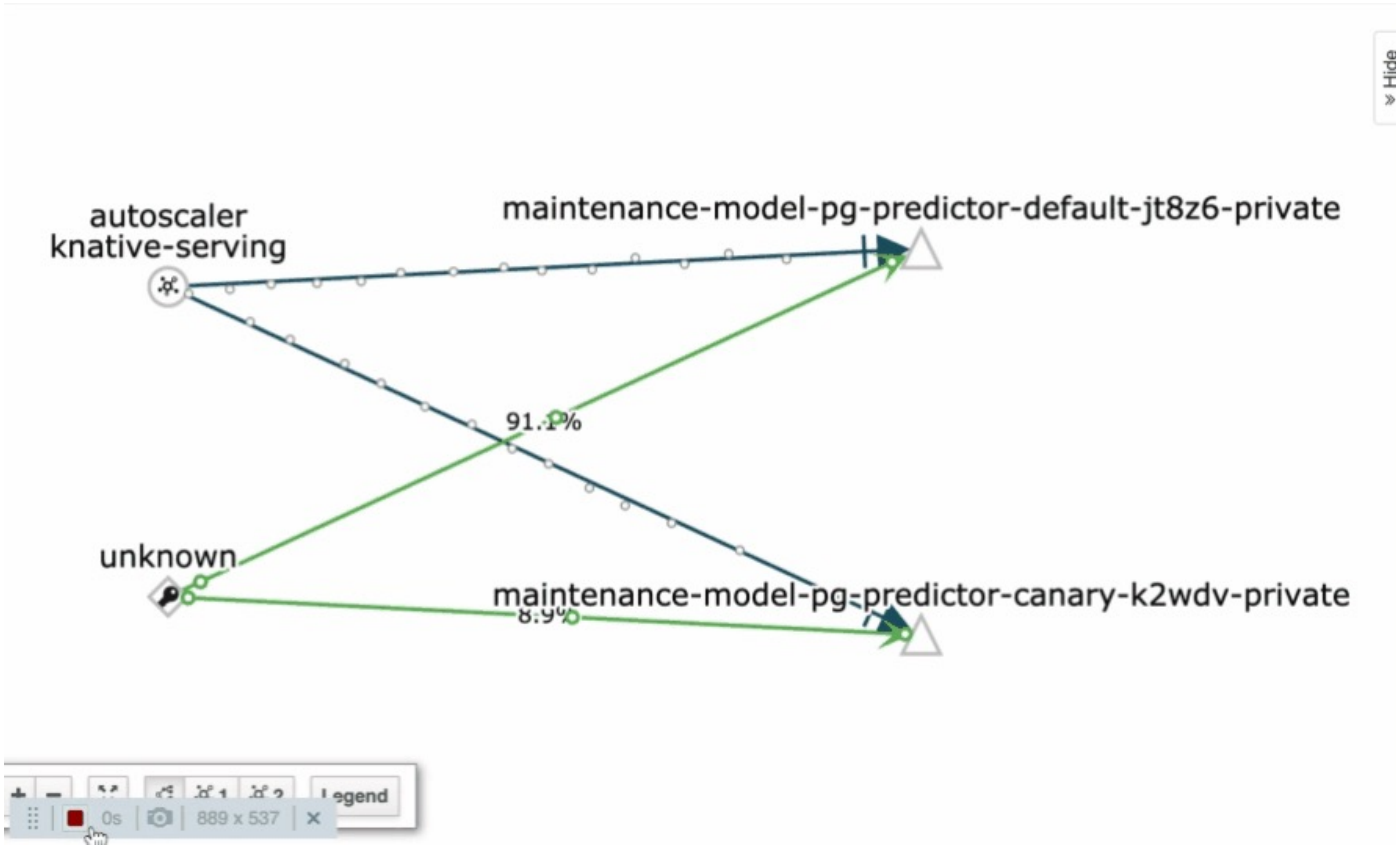
- KServe is a **Model Inferencing Platform** on **Kubernetes.** Run anywhere **Kubernetes** runs, never worry about **vendor lock-in**.

- Provides **performant, standardized inference protocol** across **ML frameworks**.

- Support modern **serverless inference workload** with **Autoscaling** including **Scale to Zero on GPU**.

- **Simple** and **Pluggable production serving** for **production ML serving** including **prediction, pre/post processing, monitoring and explainability**..

KServe: Default, Canary and Autoscaler



ModelMesh for Multi-Model Serving

- ModelMesh, model serving management layer for IBM Watson products.

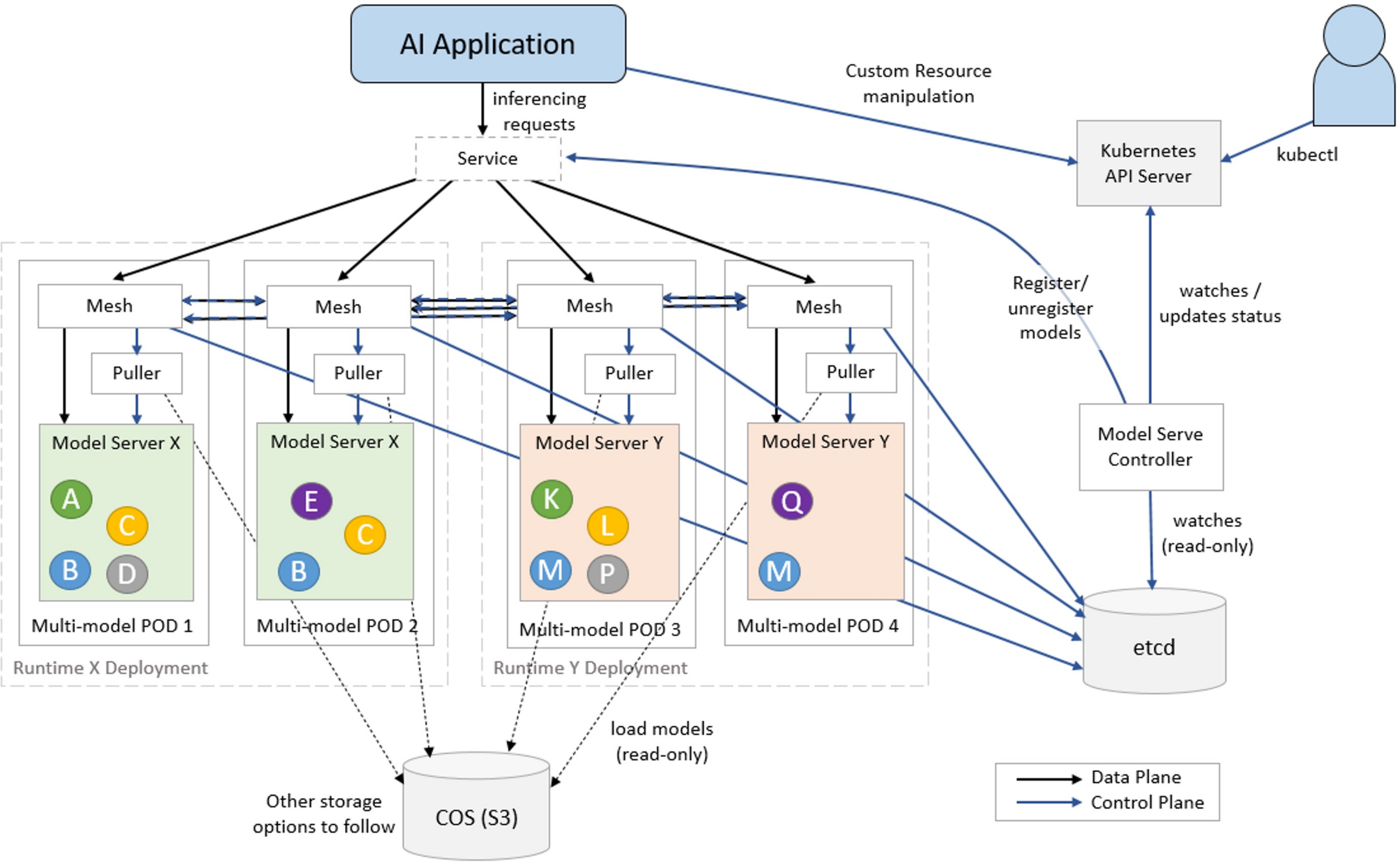
- Running successfully in production for several years, ModelMesh underpins most of the Watson cloud services, including Watson Assistant, Watson Natural Language Understanding, and Watson Discovery.

- Designed for high-scale, high-density, and frequently-changing model use cases.

- ModelMesh intelligently loads and unloads AI models to and from memory to strike an intelligent trade-off between responsiveness to users and their computational footprint.

ModelMesh Architecture

Framework for high-scale, high-density and frequently-changing model use cases.



ModelMesh Components

Core Components

- github.com/kserve/modelmesh-serving - Model serving controller.
- github.com/kserve/modelmesh - ModelMesh containers used for orchestrating model placement and routing.

Runtime Adapters

- github.com/kserve/modelmesh-runtime-adapter - the containers which run in each model serving pod and act as an intermediary between ModelMesh and third-party model-server containers. Incorporates the "puller" logic which is responsible for retrieving models from storage.

Serving Runtimes

Out-of-the-box integration with the following model servers:

[Triton Inference Server]
NVIDIA's server for frameworks like TensorFlow, PyTorch, TensorRT, or ONNX.

[MLServer]
Seldon's Python-based server for frameworks like SKLearn, XGBoost, or LightGBM.

NVIDIA
TRITON INFERENCE SERVER

SELDON®
ML SERVER

ServingRuntime custom resources can be used to add support for other existing or custom-built model servers.

ModelMesh and KServe: Better Together!

KubeCon CloudNativeCon North America 2021

Announcing

ModelMesh is being contributed to Open Source, and joining KServe!

ModelMesh
from
IBM Watson

KServe

Model

Model Model

Model Model Model

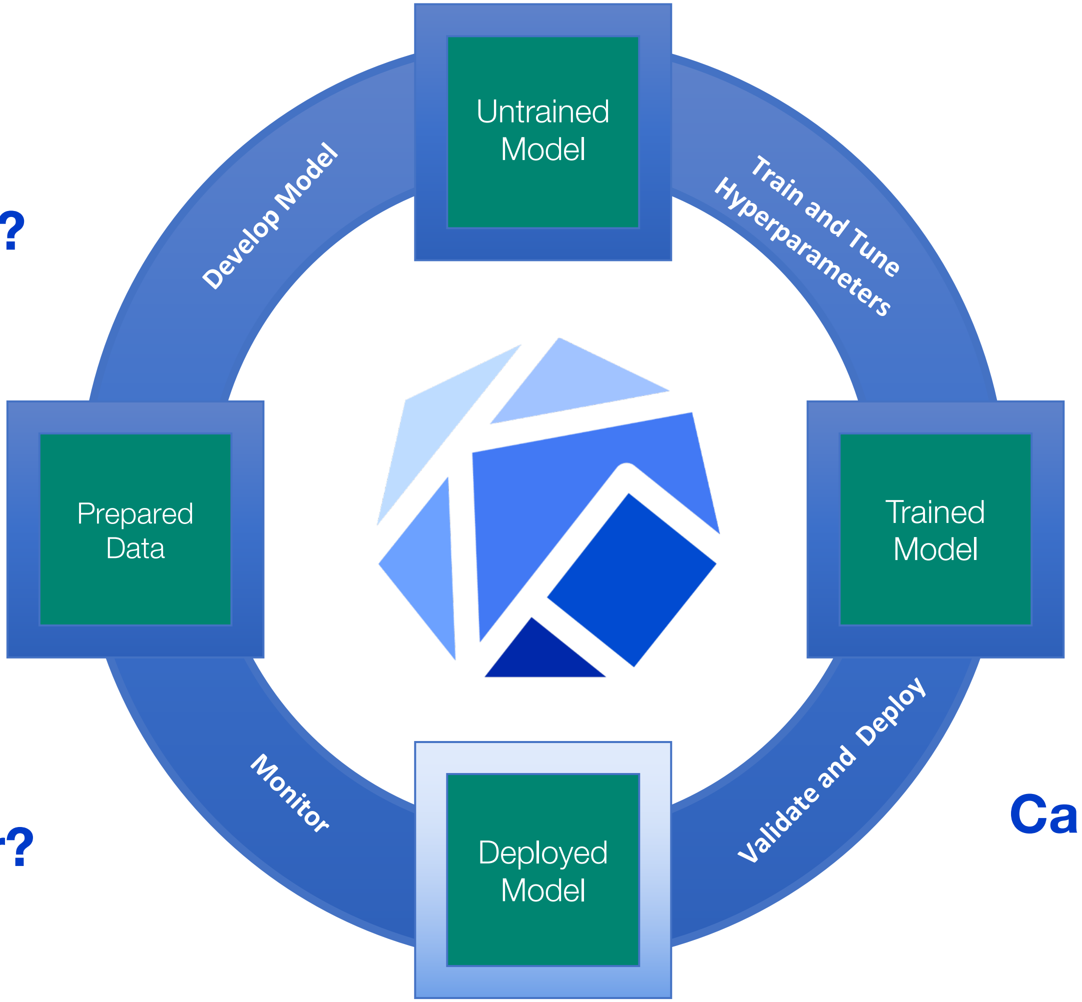
ModelMesh

General / Modelmesh monitoring dashboard
Last 5 minutes
5s
81.8
63.5
modelmesh-serving-triton-2.x...
modelmesh-serving-triton-2.x...
modelmesh-serving-triton-2.x-54b747f6b7-qglwz
modelmesh-serving-triton-2.x-54b747f6b7-5245b
Number of processed request per second
All ModelMesh Model Status
Managed
18030
Loaded
15746
Serving Runtime Cache Usage
Value
Total Cache Capacity
Cache Miss Rate
Number of missed cache models
Models on Serving Runtime
modelmesh-serving-
modelmesh-serving-
Serving Runtime Model Loading Rate
Model Unloading Rate

Model Serving is accomplished. Can the predictions be trusted?

Are there concept drifts?

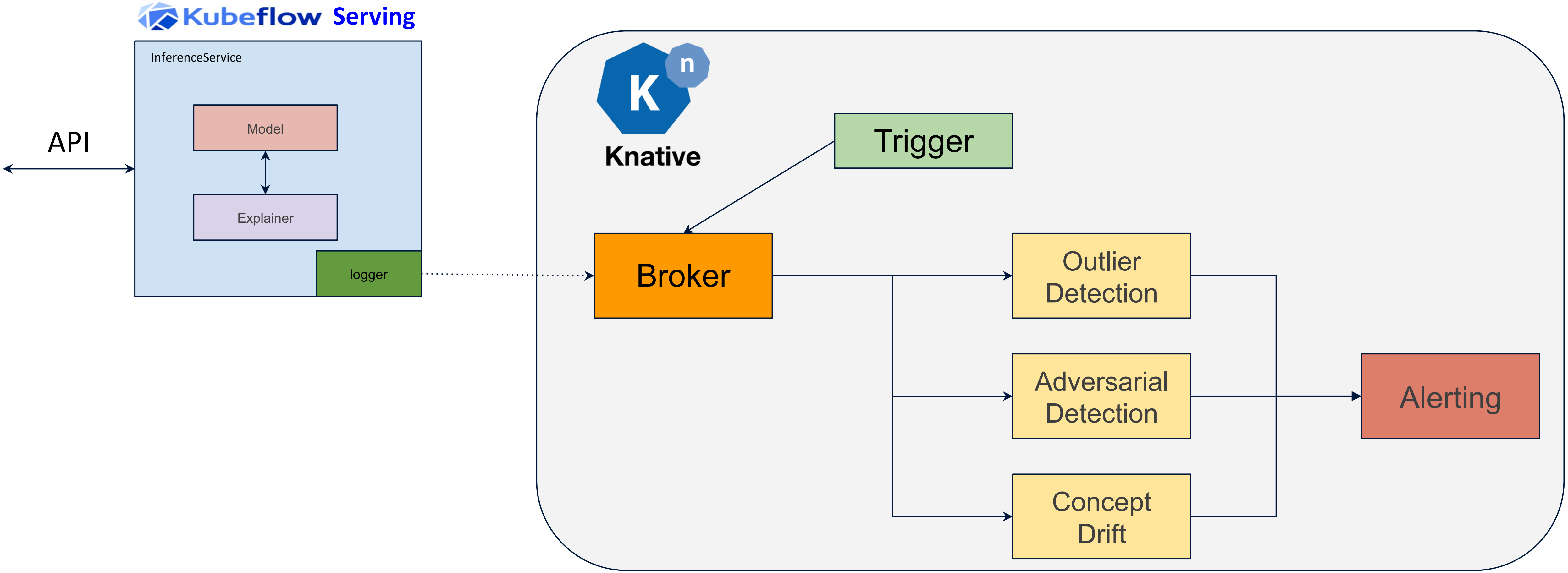
Is the model vulnerable to adversarial attacks?



Is there an outlier?

Can the model explain its predictions?

Production Model Monitoring and Mangement



ML Inference Analysis

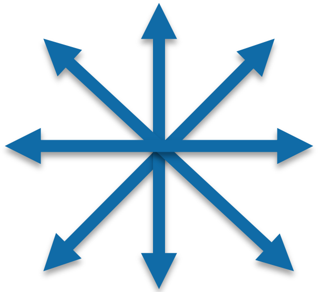
Don’t trust predictions on instances outside of training distribution!

- Outlier Detection
- Adversarial Detection
- Concept Drift

Trusted AI Lifecycle through Open Source

Pillars of trust, woven into the lifecycle of an AI application

Did anyone tamper with it?



ROBUSTNESS

Is it fair?

FAIRNESS

Is it easy to understand?

EXPLAINABILITY

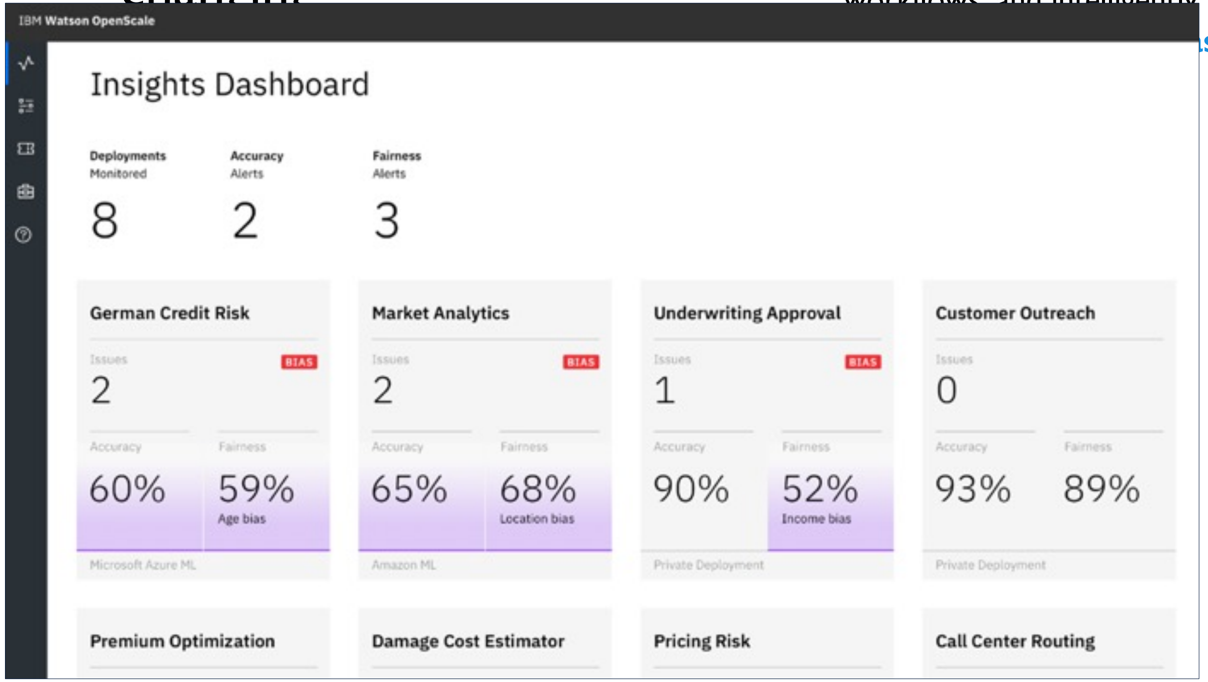
Adversarial Robustness 360 ↳ (ART)	AI Fairness 360 ↳ (AIF360)	AI Explainability 360 ↳ (AIX360)
github.com/IBM/adversarial-robustness-toolbox art-demo.mybluemix.net	github.com/IBM/AIF360 aif360.mybluemix.net	github.com/IBM/AIX360 aix360.mybluemix.net

We are making these capabilities around Trusted AI available to businesses through

Watson Studio

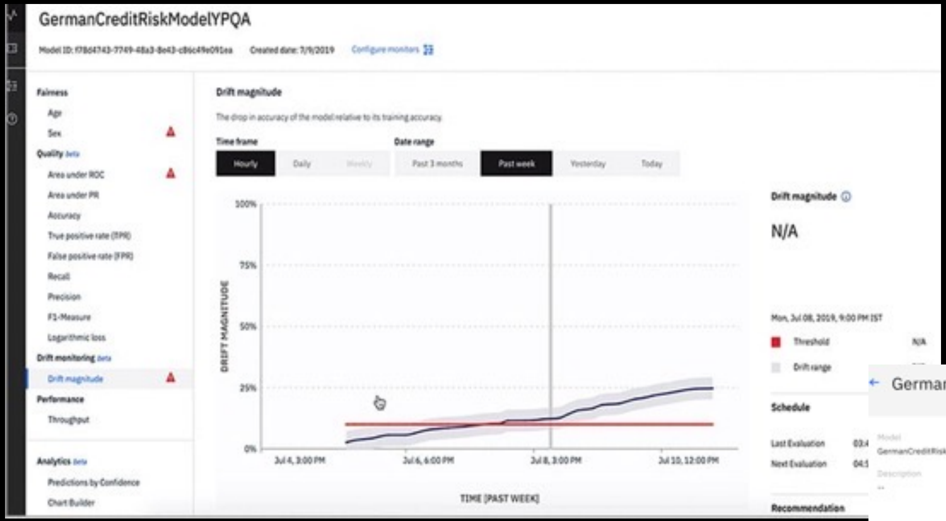
- Watson Studio **tracks and measures trusted AI outcomes across its lifecycle**, and adapts and governs AI to changing

- **Measure and track AI outcomes**
- Track performance of production AI and its impact on business goals, with actionable metrics in a single console.
- **Govern, detect bias and explain AI**
- Maintain **regulatory compliance** by **tracing and explaining AI decisions** across workflows, and intelligently

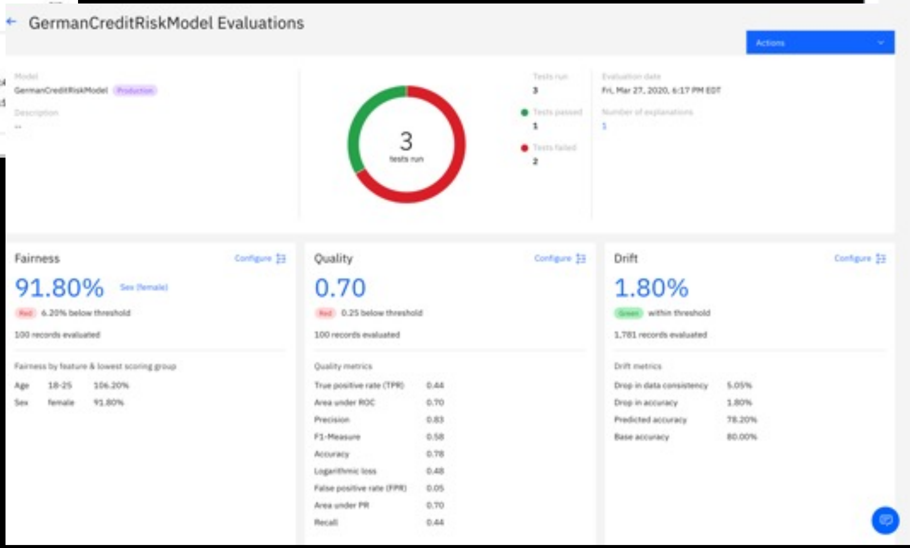


Implement responsible, explainable AI

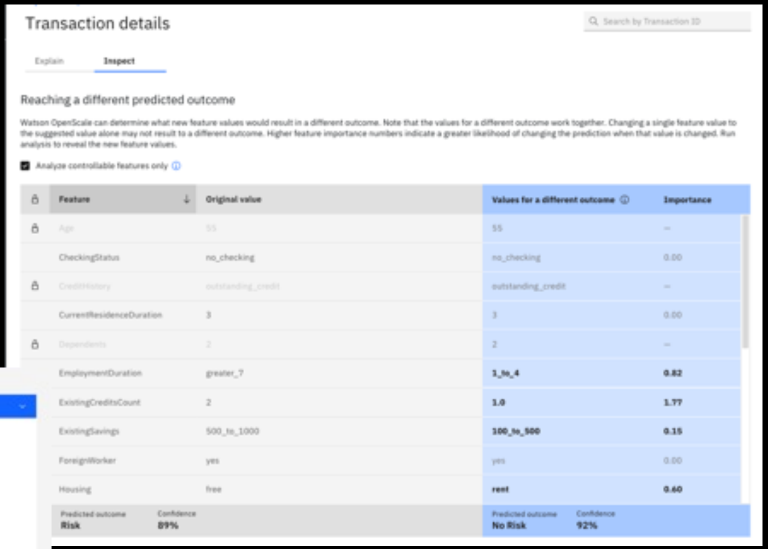
Mitigate drift, bias, and model risk



Drift: monitor model drift by hourly, daily or weekly



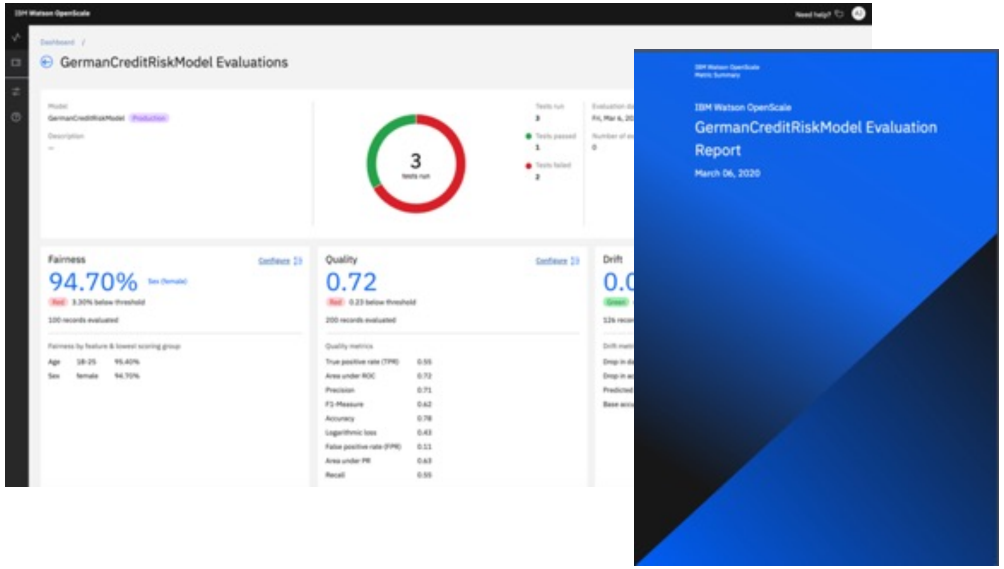
Model risk evaluation: fairness, quality and drift metrics to share model insights

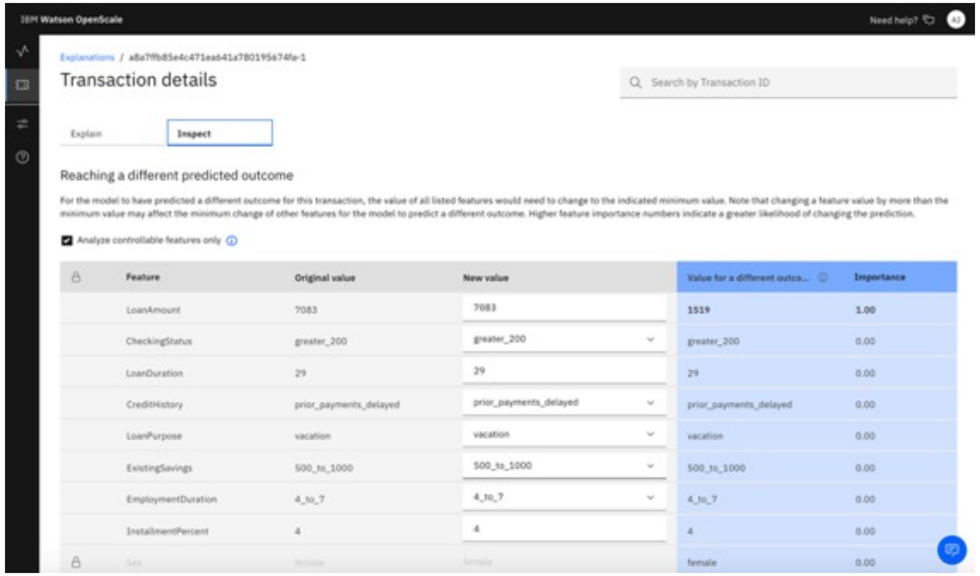


Explain transactions: Determine what features reach different outcomes

What’s are new trust capabilities in Watson Studio

- Model Validation /Model Risk Management
  - Monitor and validate pre-production models
  - Enable validators to run tests & set acceptance thresholds.
  - Compare candidate models
  - Synchronize results with OpenPages GRC solution
- Explainability Enhancements
  - Graduated Explanations
  - What If interactivity
  - Designated controllable/uncontrollable features
  - Improved performance
- Indirect Bias detection
- Cloud Pak for Data 3.0.1
  - OpenShift 4.3
  - IBM Power platform support





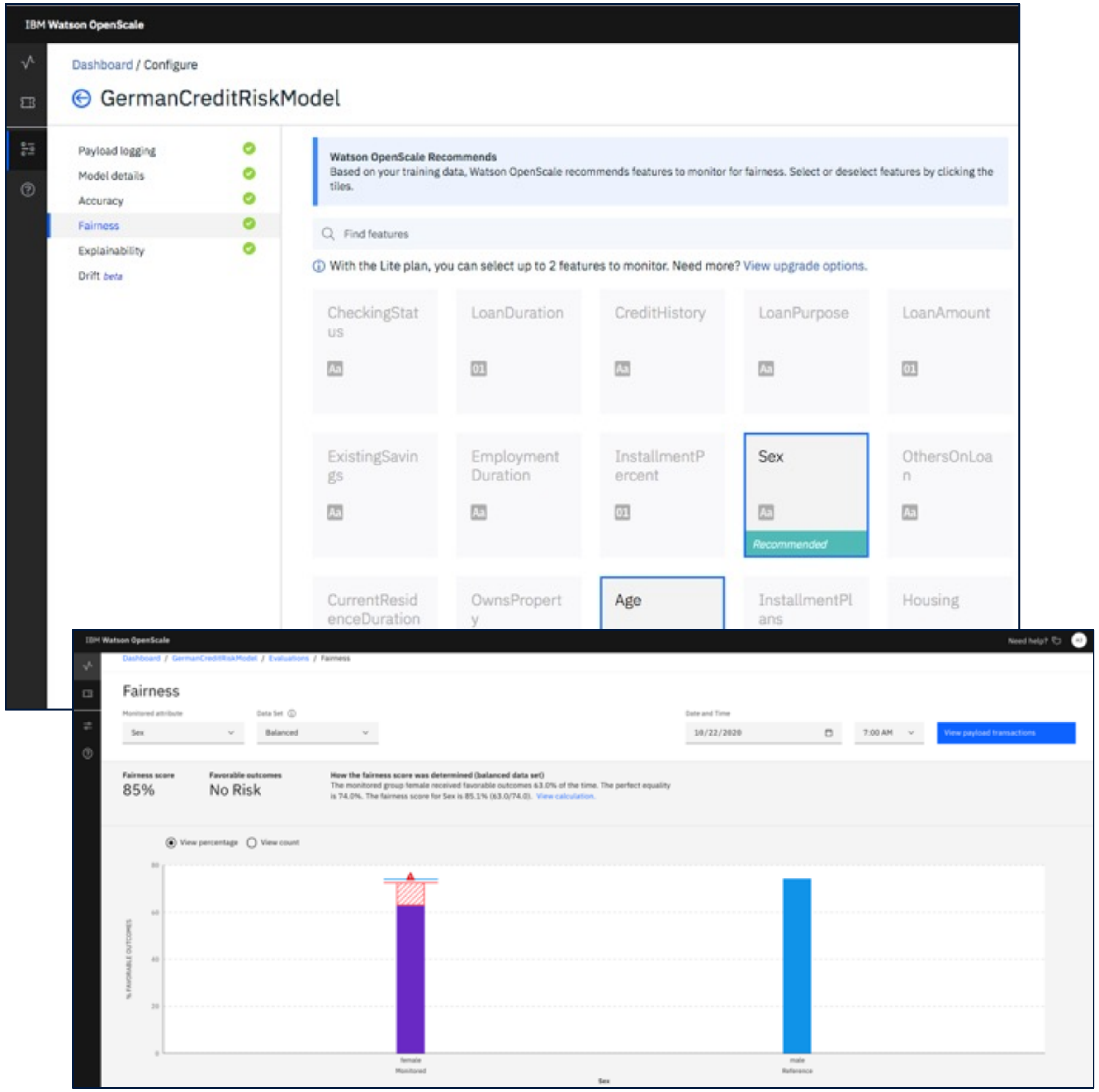
Bias Detection

- Watson Studio enables enterprises to enforce fairness in their model's outcome by analyzing transactions in production and finding biased behavior by the model

- It pinpoints the source of bias and actively mitigates the biases found in production environment

- **Value:**

- Automatically recommend common protected attributes to monitor during production
- Detect biases in runtime in order to catch impacts on business applications and compliance requirements without time consuming, manual data analysis
- Metrics and data to help data scientists further troubleshoot issues in data sets or models
- Mitigate biases in runtime in order to enforce regulatory or enterprise fairness guardrails in real time



Explainability

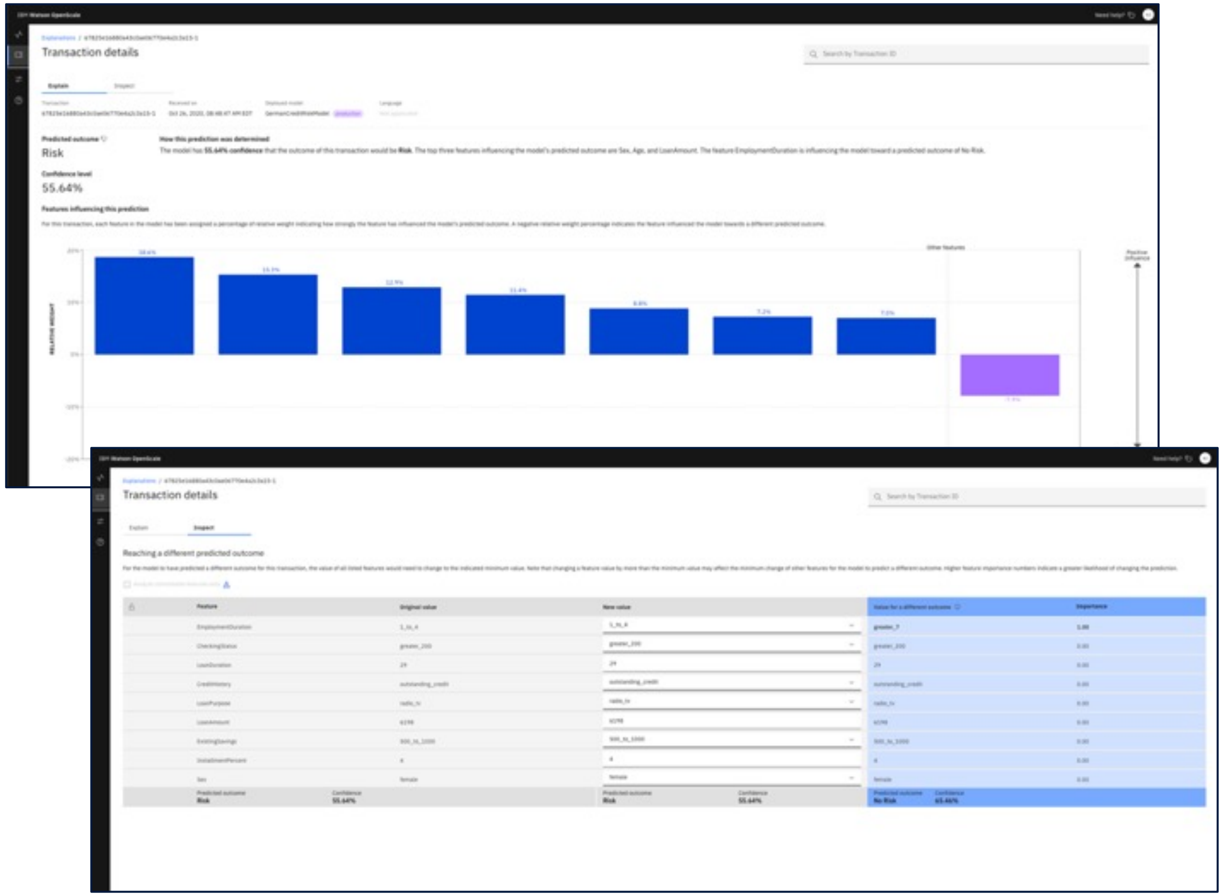
- Watson Studio records every individual transaction and drills down into its working to explain how the model makes decisions

- It provides a simple explanation that is user friendly and interactive

- **Value:**

• Explain individual transaction level decisions made by the model in run time, including details about most important attributes and their values in order to assist in compliance and customer care situations

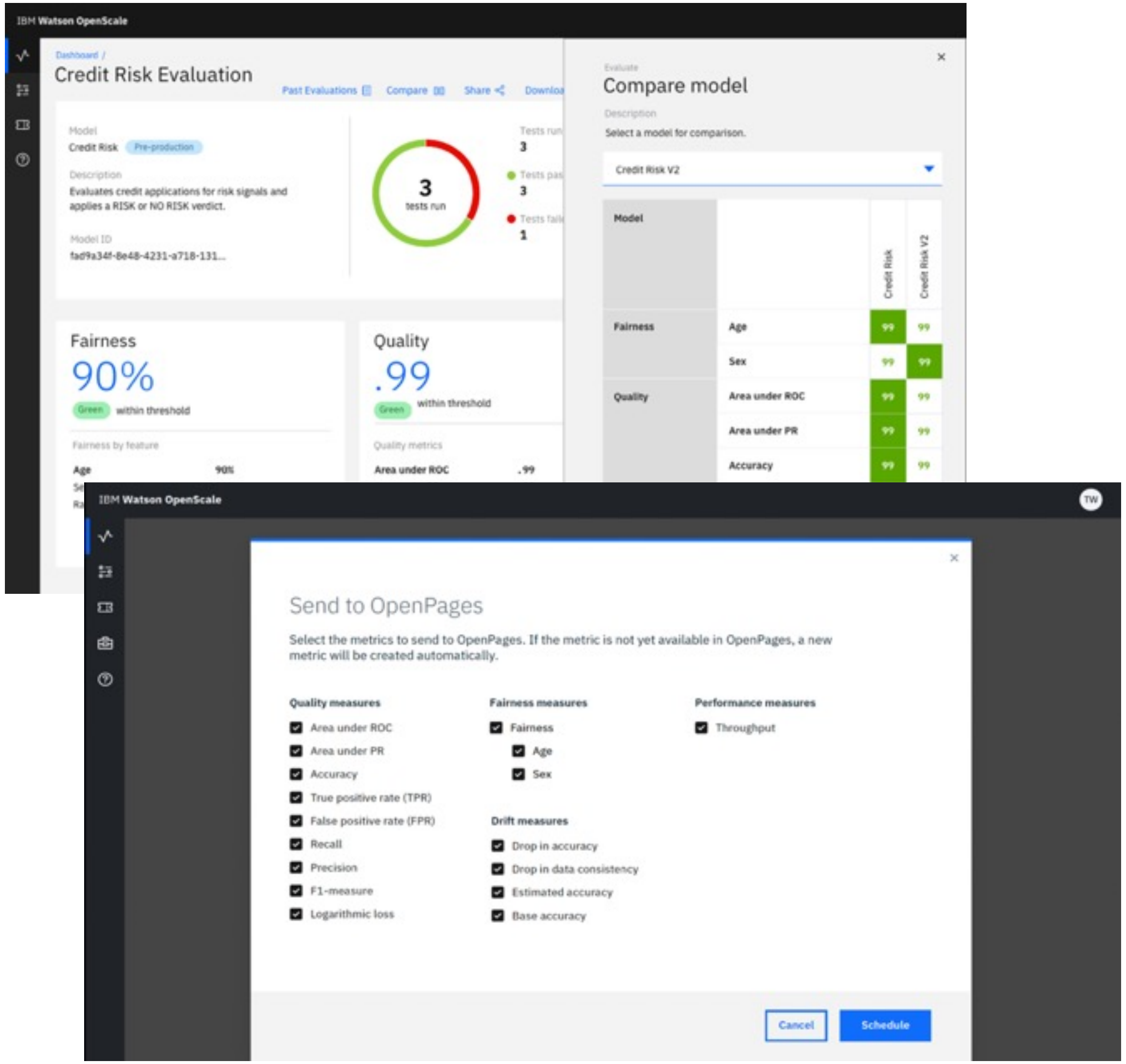
• Analyze individual transactions in a what-if manner in order to understand how model behavior will change in different business situations



Model Validation/Model Risk Management

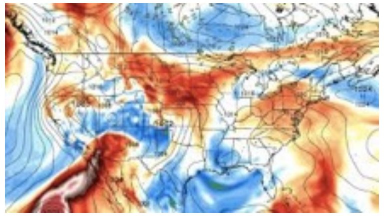
› Watson Studio enables enterprises to validate pre-production models before putting them into production to ensure they can be trusted to perform as intended.

- Validate pre-production models and generate reports of outcomes
- Enable customizable tests relevant to AI models
- Compare performance of models
- Automatically configure monitoring of production models to match pre-production settings.
- Synchronize results with Governance, Risk and Compliance (GRC) solutions (Initially with OpenPages Model Risk Governance)



Business environments are dynamic leading to “drift” in data and cause inaccuracies in model prediction

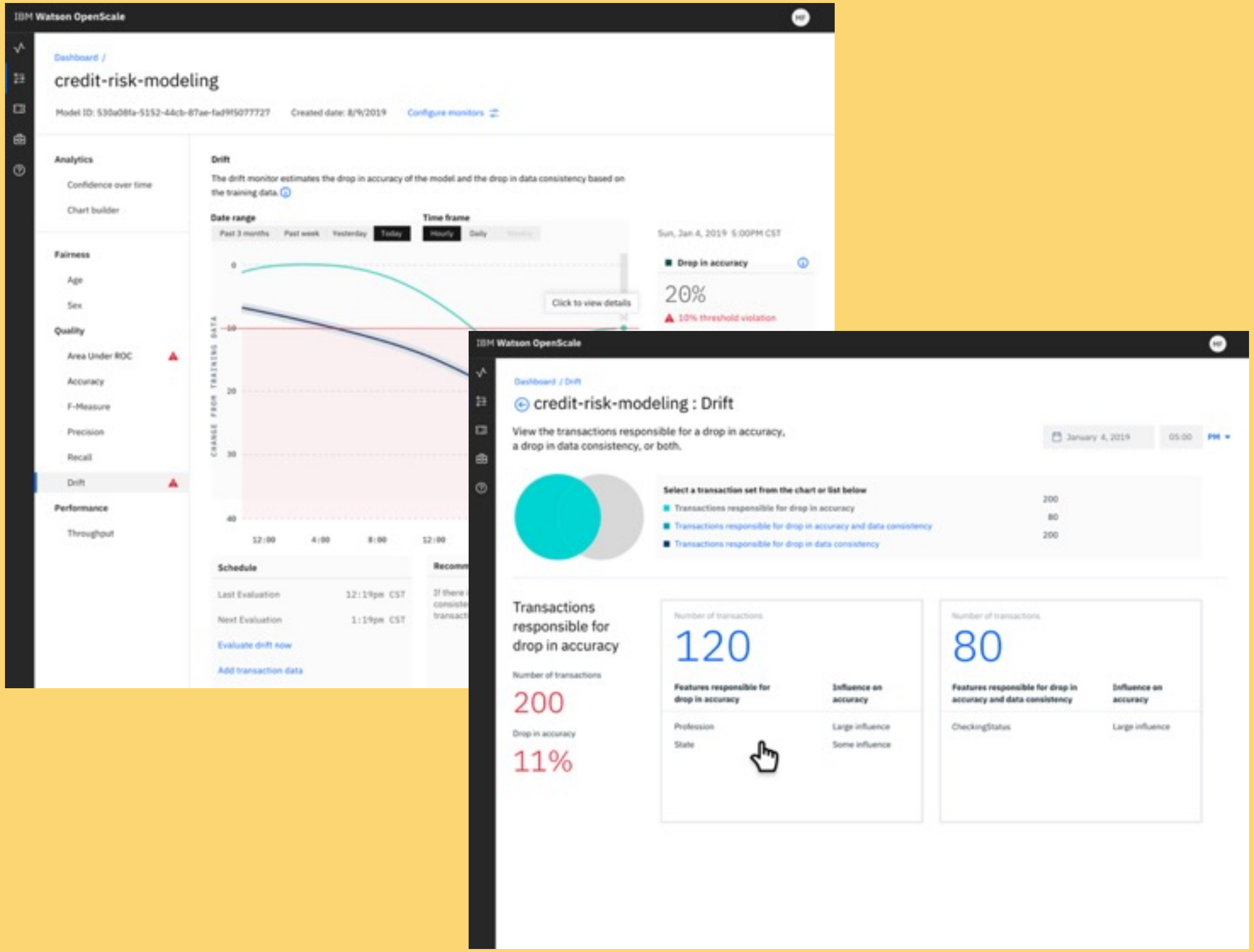
Weather data changes in short term can affect long term climate models



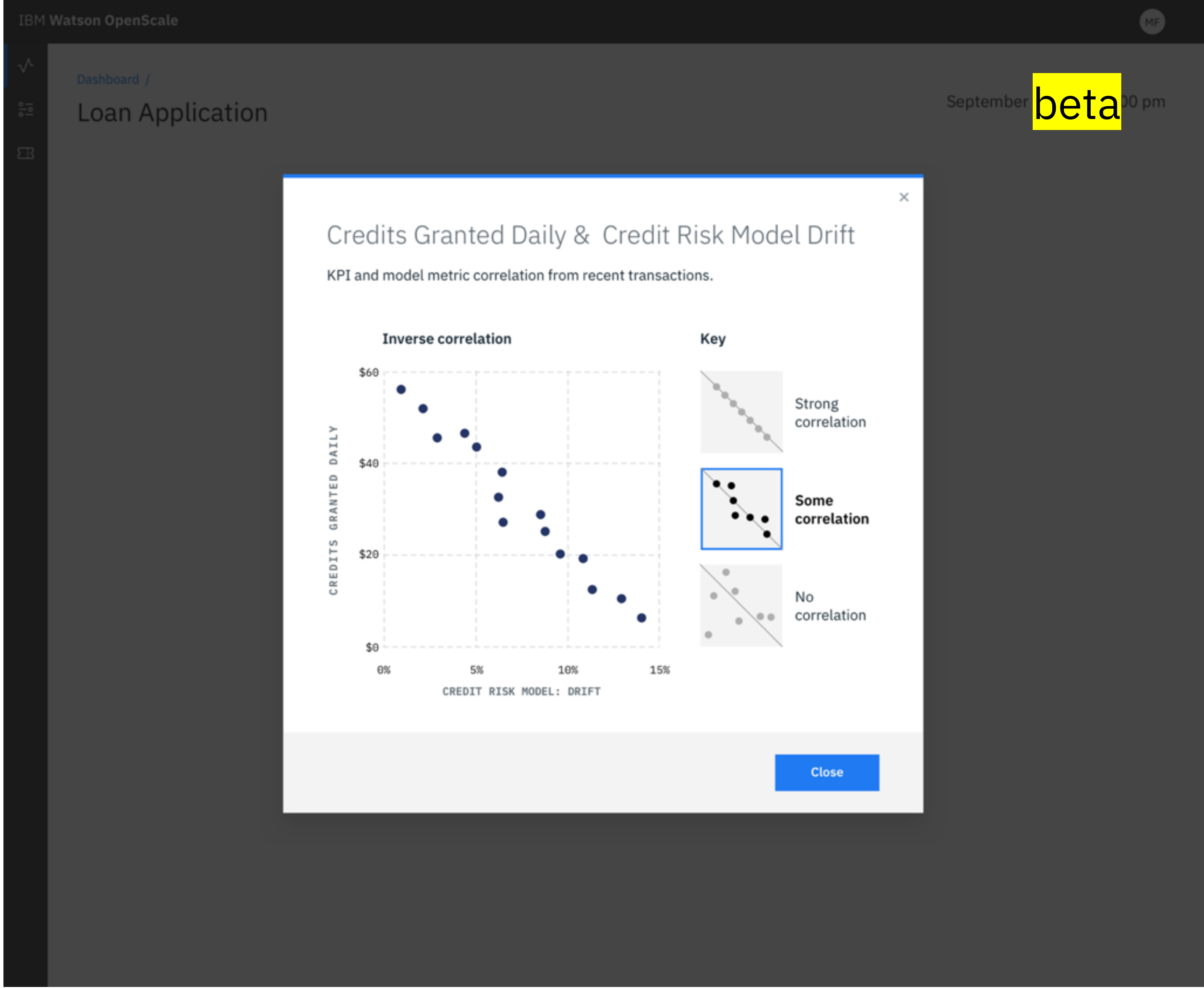
Online shopping behavior

Rise in income levels in specific geos can throw off global models

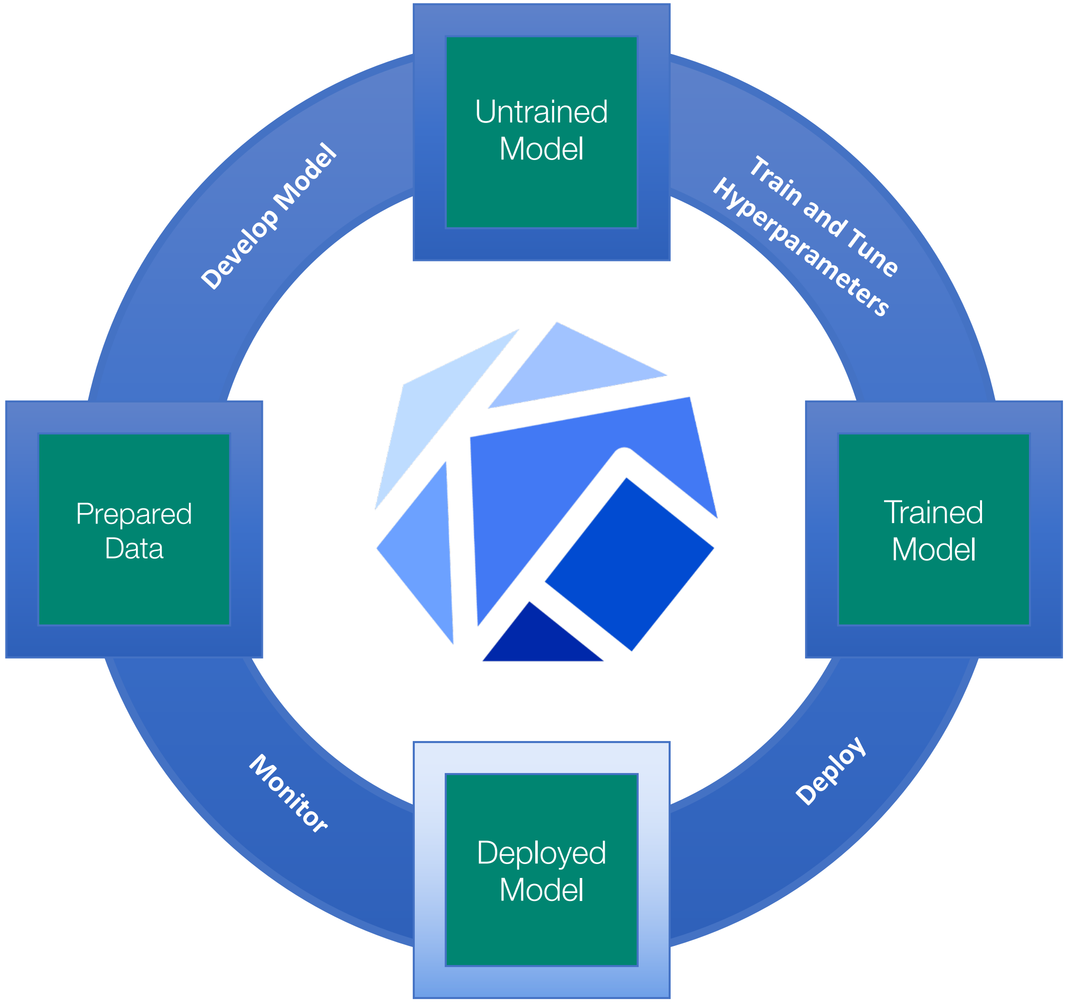
Watson Studio will **automatically detect drifted transactions and pinpoint datapoints that contribute to drift**



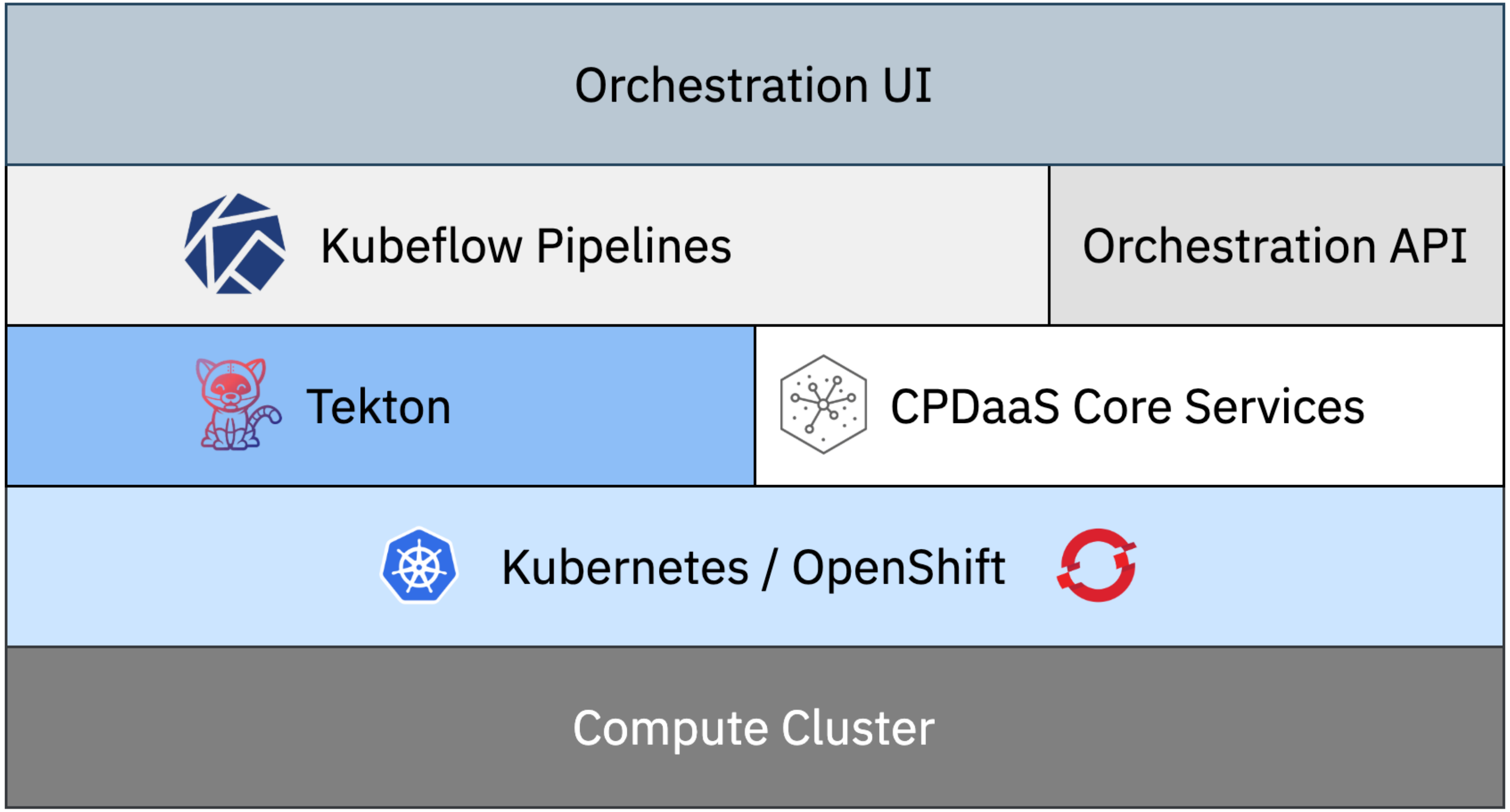
Watson Studio will **automatically correlate** model metrics in runtime with business event data to inform of **impact on business KPI** and help **easier problem resolution**



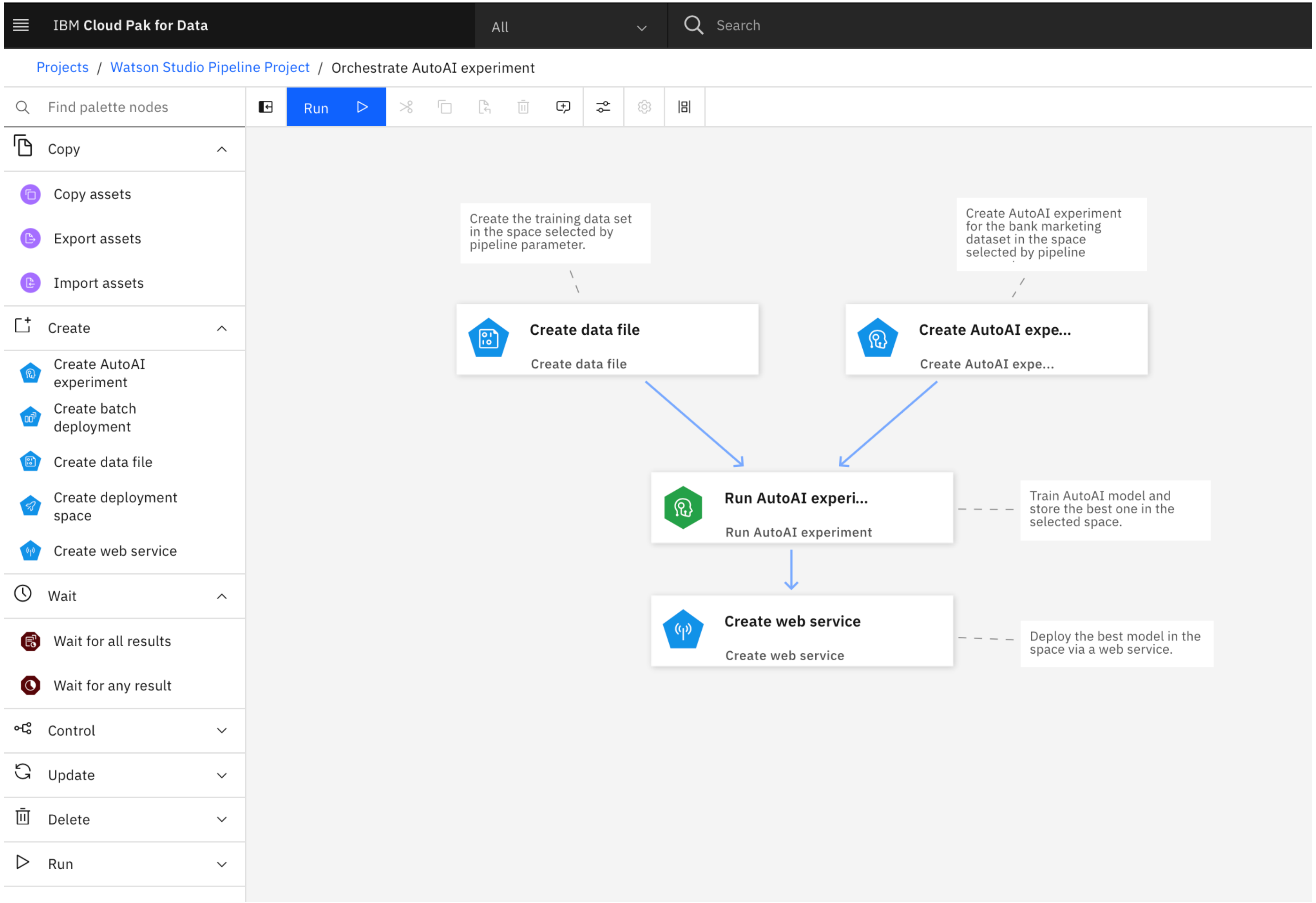
ML Lifecycle: Orchestrate Build, Train, Validate and Deploy



Watson Studio Pipelines- Architecture Overview

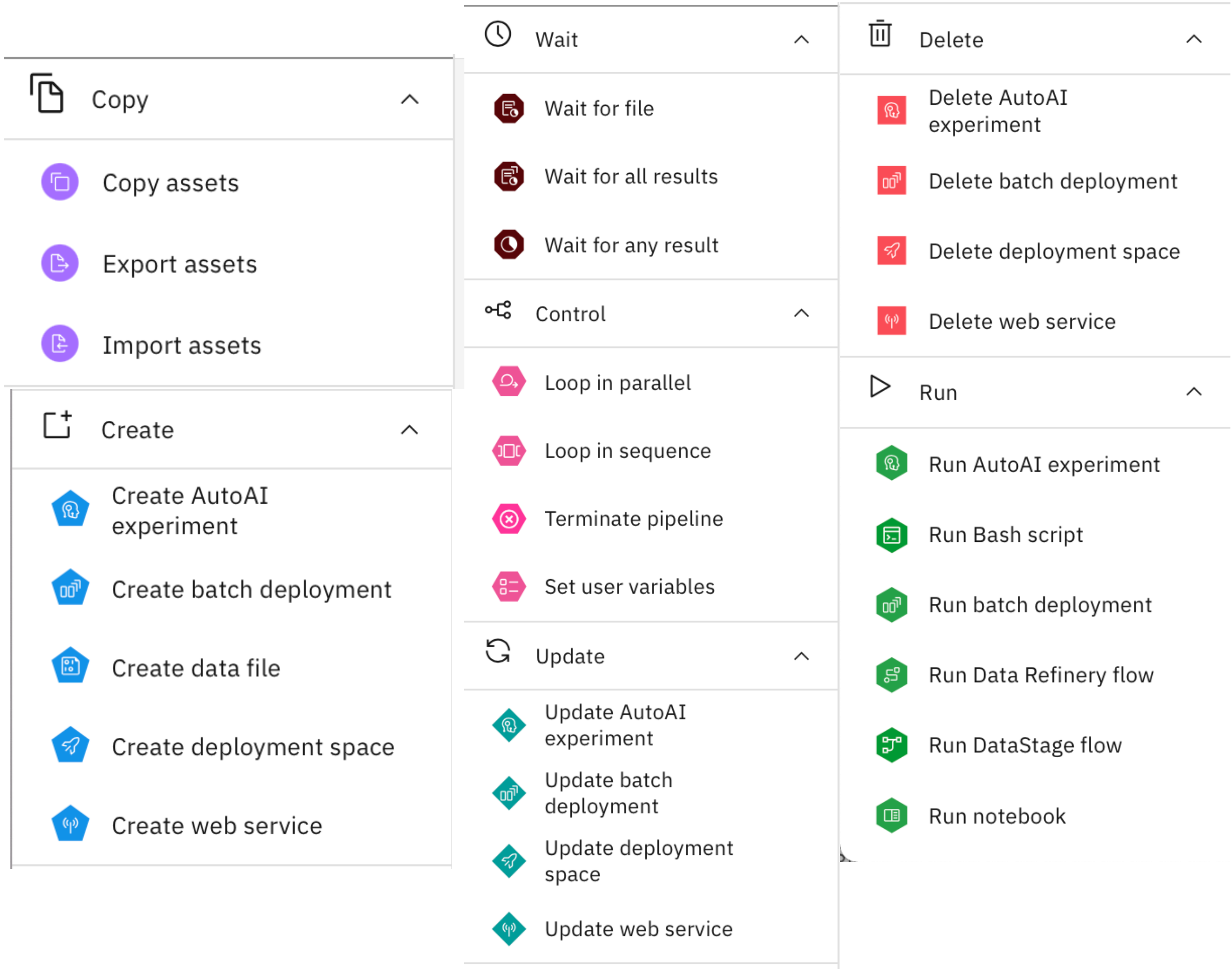


Watson Studio Pipeline



Component Catalog

- 30+ components in Beta
- Platform services integrated with Pipelines
  - AutoAI
  - Watson Machine Learning deployments
  - Data Refinery
  - Jupyter Notebook
  - DataStage
- Platform features integrated with Pipeline
- Project
- Spaces
- Jobs
- Connections



Built On: Kubeflow Pipelines

Containerized implementations of ML Tasks

- Pre-built components: Just provide params or code snippets (e.g. training code)
- Create your own components from code or libraries
- Use any runtime, framework, data types
- Attach k8s objects - volumes, secrets

Specification of the sequence of steps

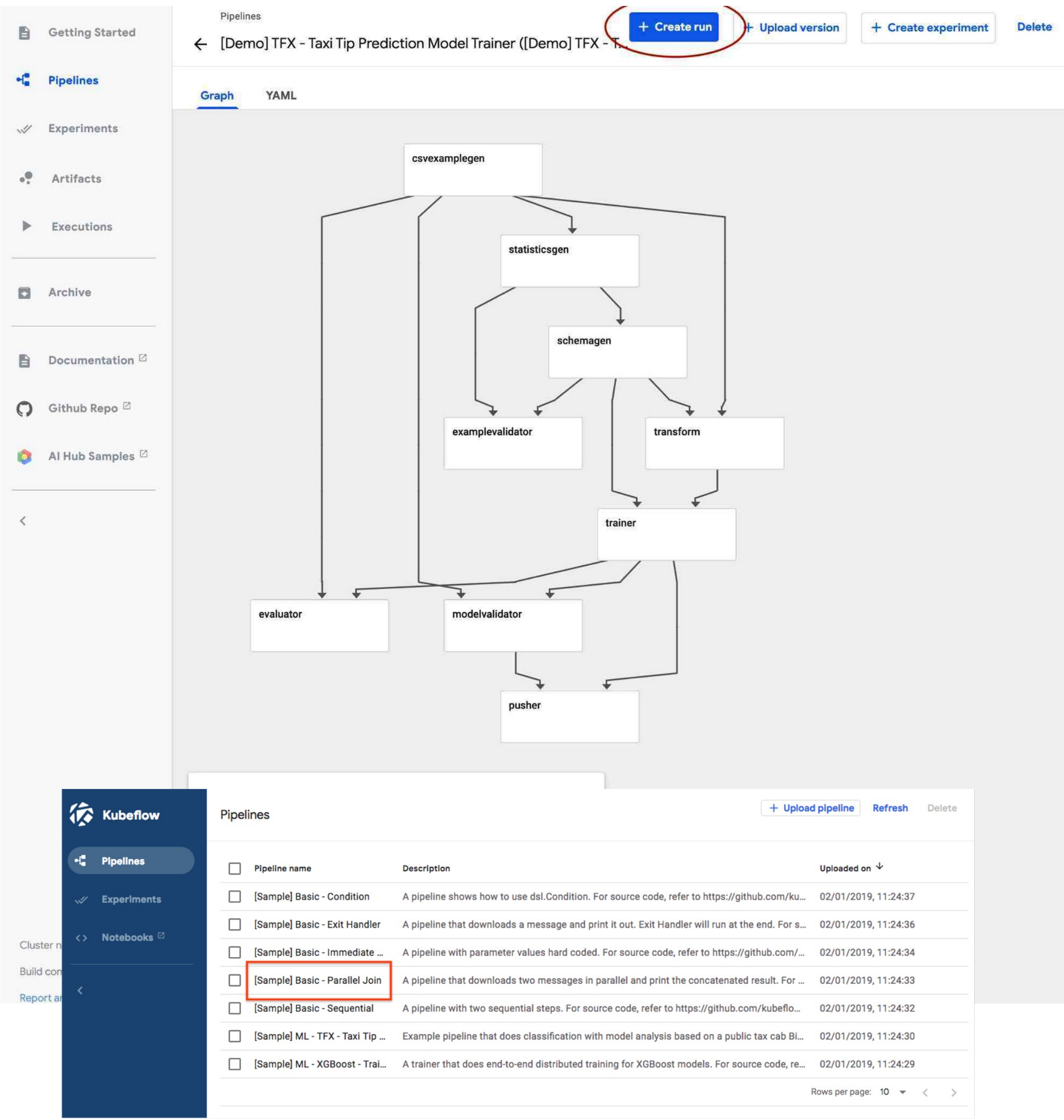
- Specified via Python DSL
- Inferred from data dependencies on input/output

Input Parameters

- A “Run” = Pipeline invoked w/ specific parameters
- Can be cloned with different parameters

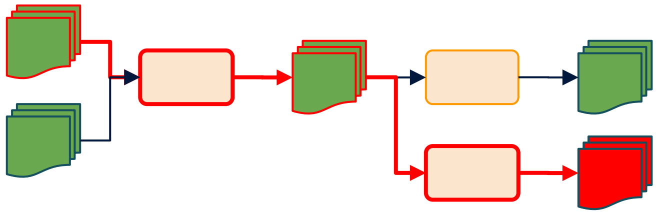
Schedules

- Invoke a single run or create a recurring scheduled pipeline

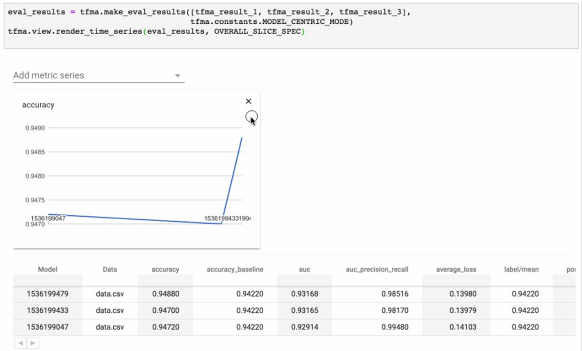


Benefits of metadata logging

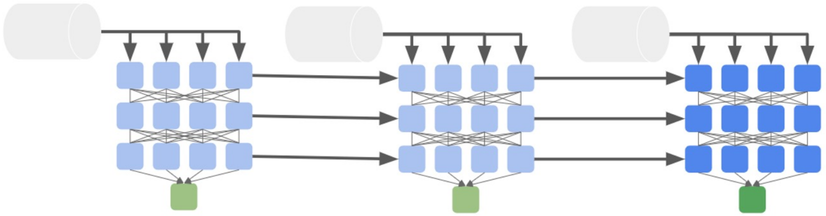
Find out which data a model was trained on



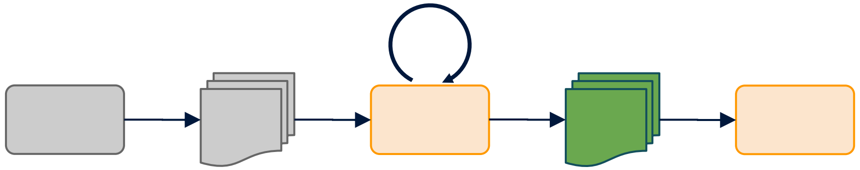
Compare previous model runs



Carry-over state from previous models

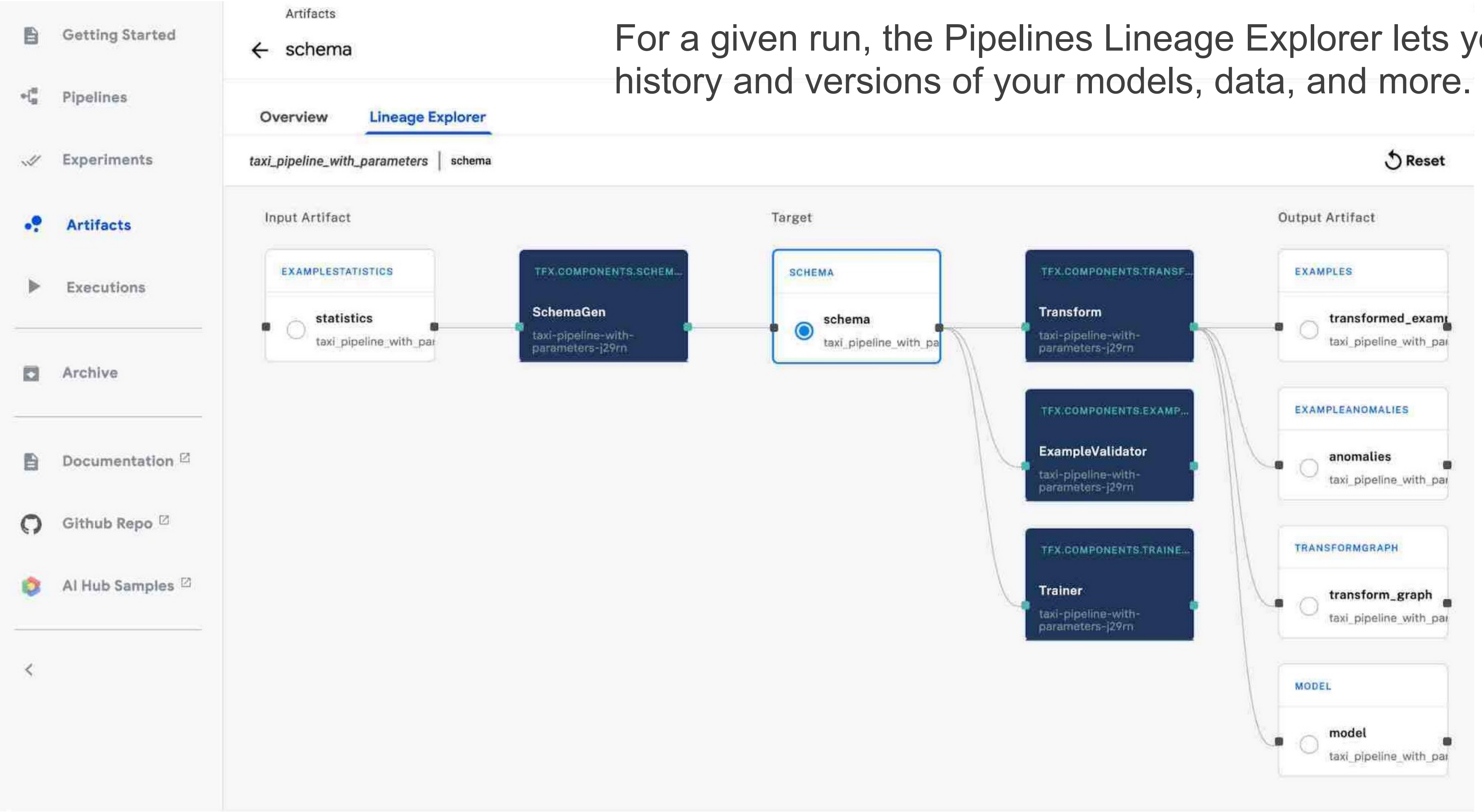


Re-use previously computed outputs

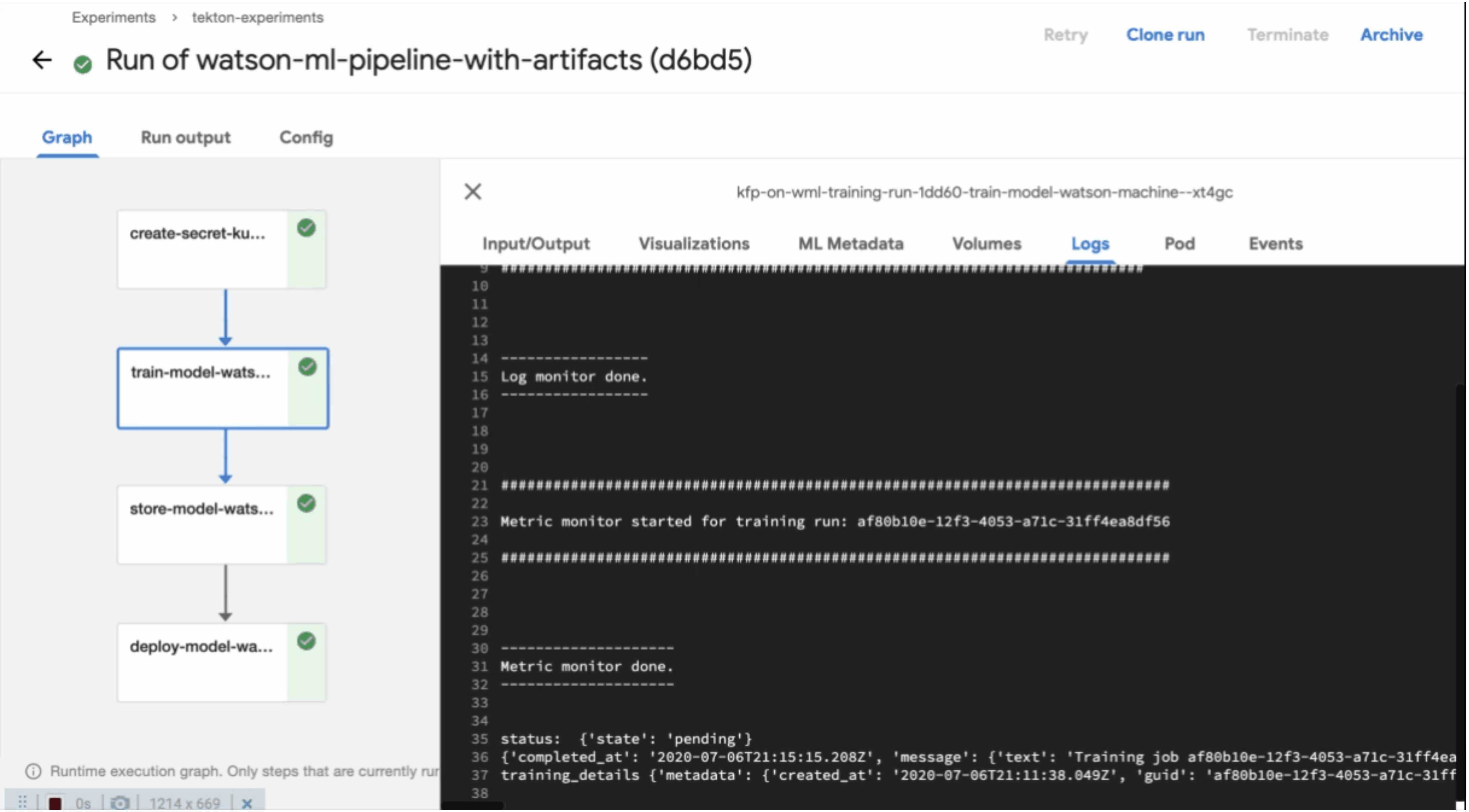


Lineage Tracking

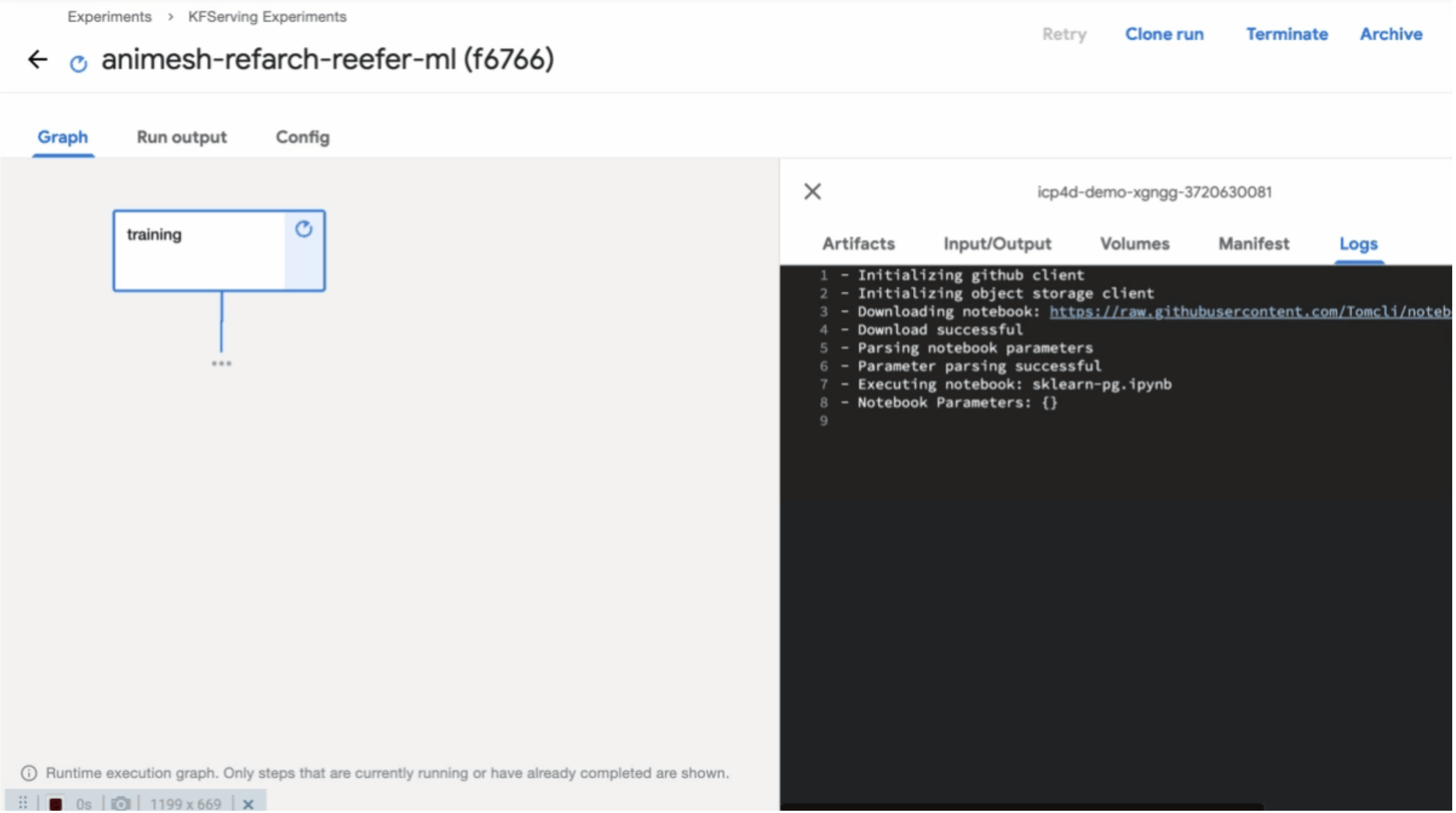
For a given run, the Pipelines Lineage Explorer lets you view the history and versions of your models, data, and more.

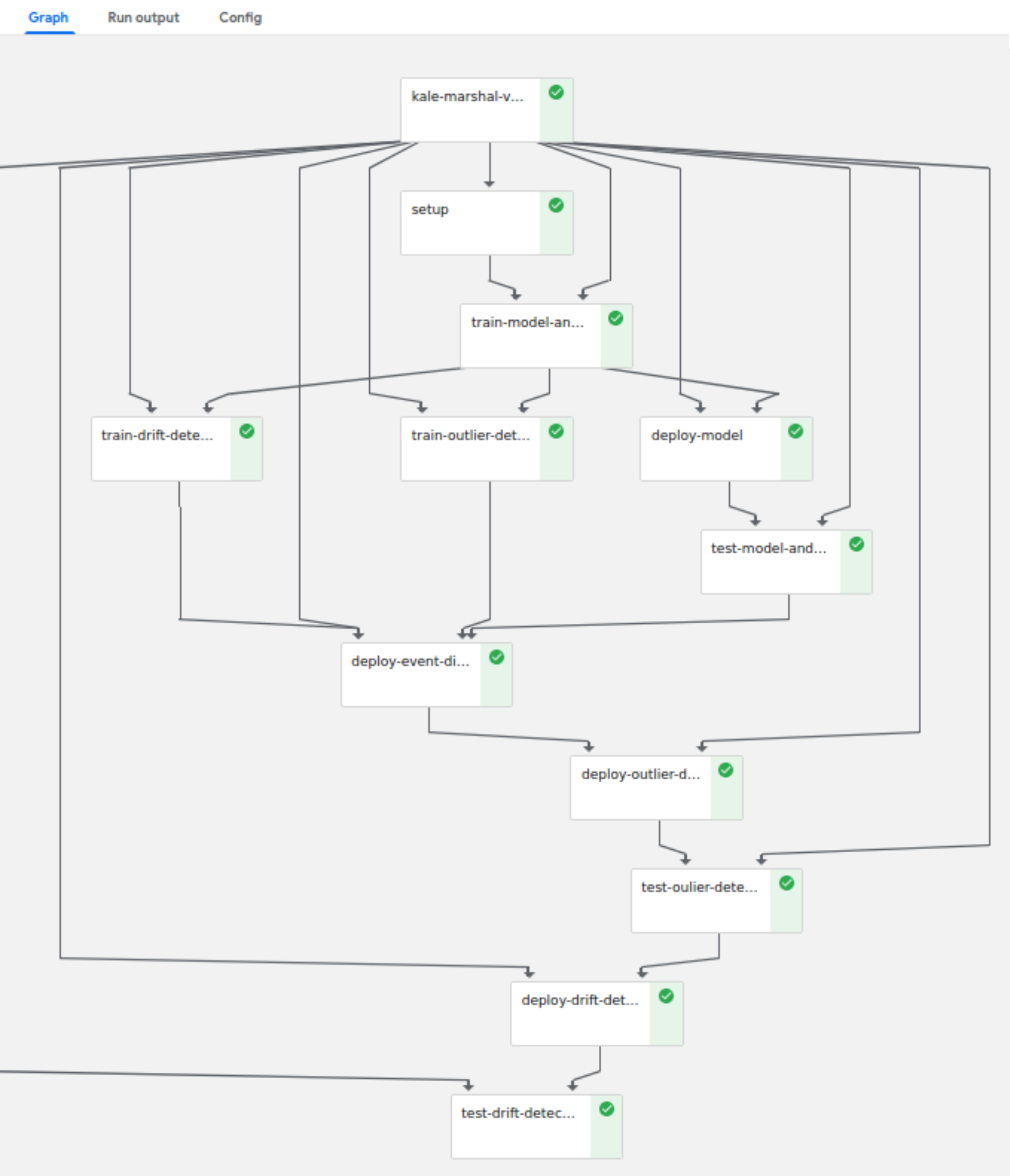


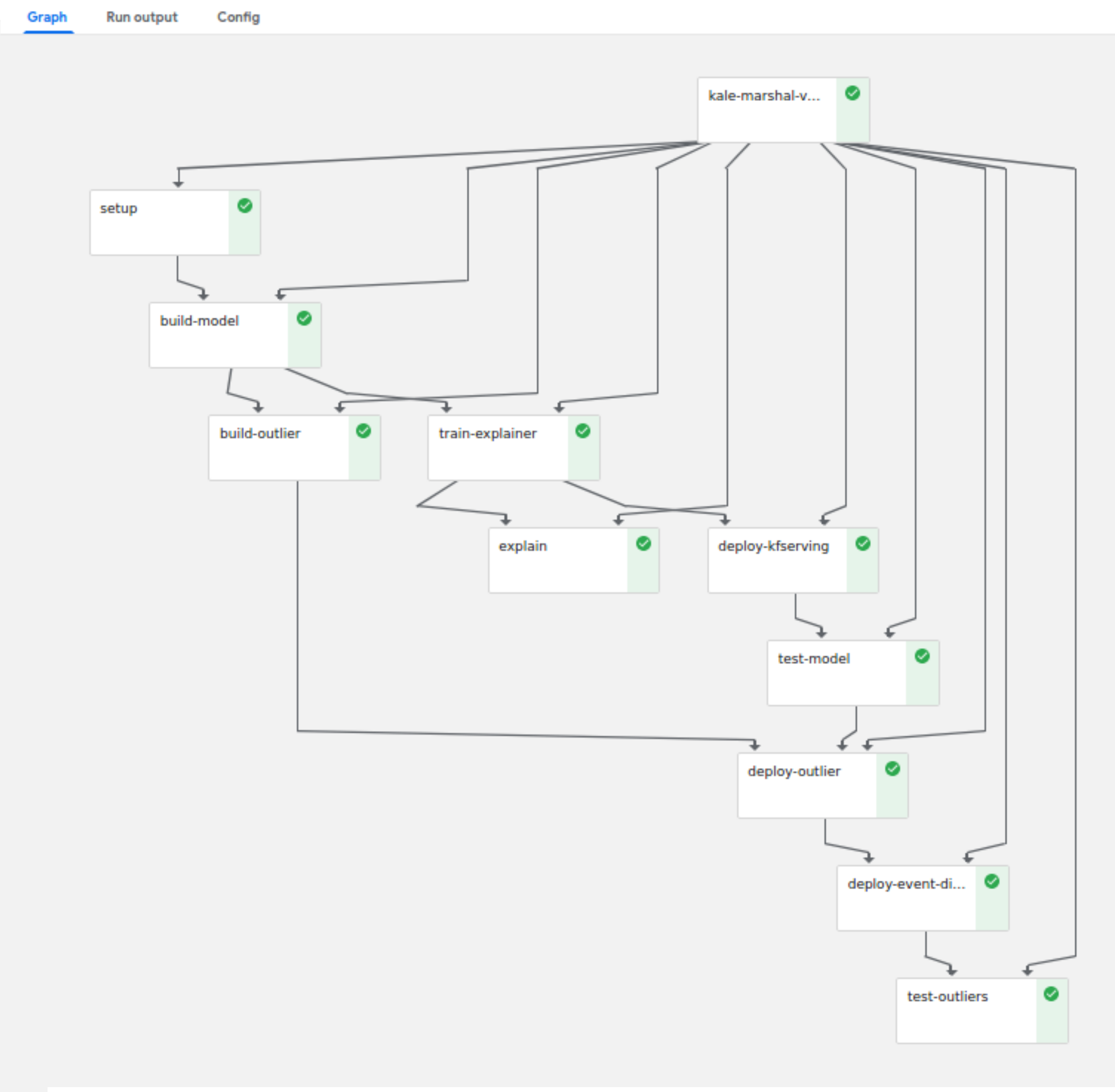
Kubeflow Pipelines on Tekton: Logs, Lineage Tracking and Artifact Tracking

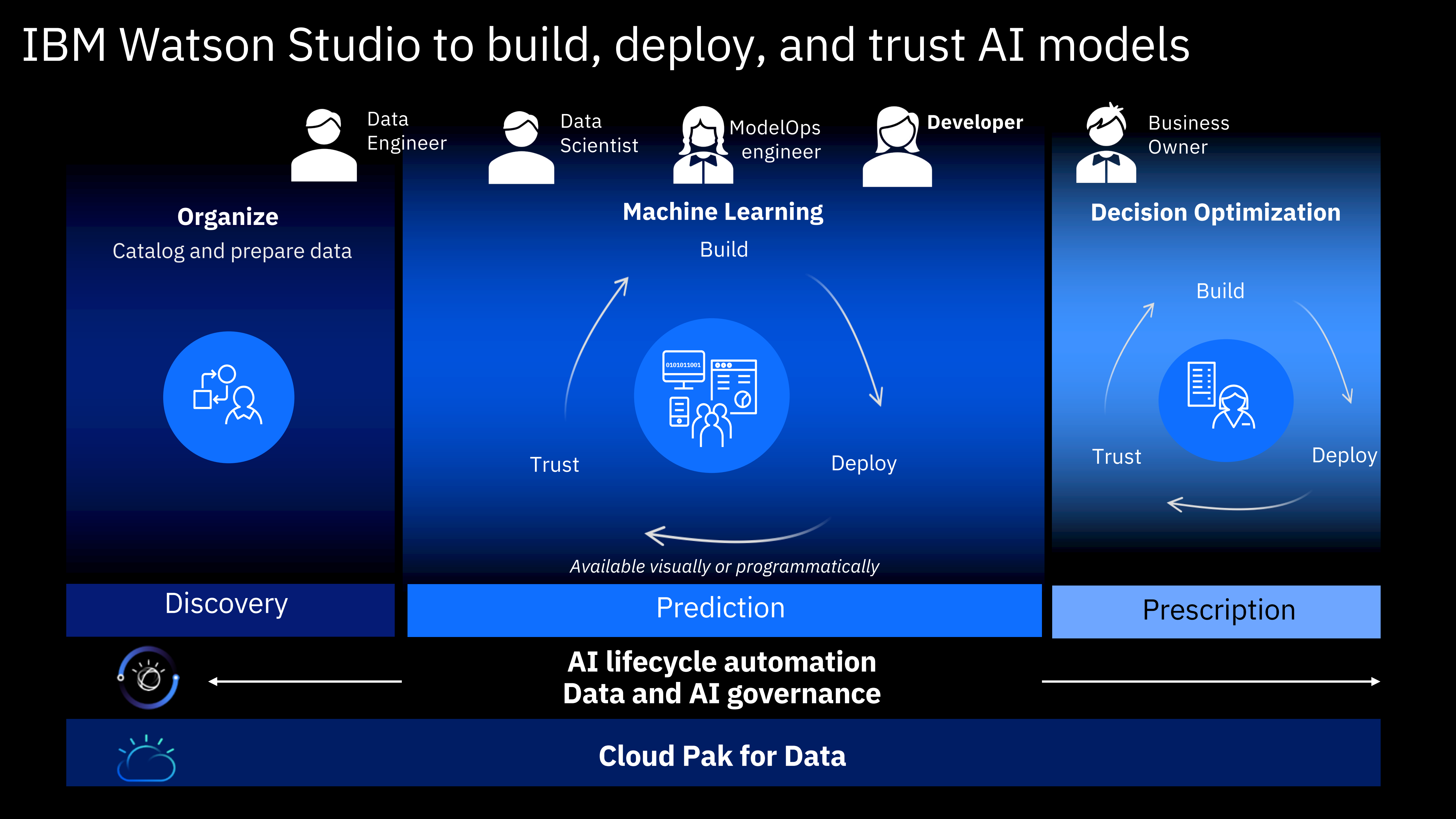


Kubeflow Pipelines can train, deploy and serve




Graph
Run output
Config
kale-marshal-v...
setup
train-model-an...
train-drift-dete...
train-outlier-det...
deploy-model
test-model-and...
deploy-event-di...
deploy-outlier-d...
test-oulier-dete...
deploy-drift-det...
test-drift-detec...


Graph
Run output
Config
kale-marshal-v...
setup
build-model
build-outlier
train-explainer
explain
deploy-kfserving
test-model
deploy-outlier
deploy-event-di...
test-outliers


IBM Watson Studio to build, deploy, and trust AI models
Data Engineer
Data Scientist
ModelOps engineer
Developer
Business Owner
Organize
Catalog and prepare data
Machine Learning
Build
Trust
Deploy
Available visually or programmatically
Decision Optimization
Build
Trust
Deploy
Discovery
Prediction
Prescription
AI lifecycle automation
Data and AI governance
Cloud Pak for Data

Open Source in Watson AI and Data

- TensorFlow
- Kubeflow
- ModelMesh
- Apache Spark
- PyTorch and ONNX
- AIF360, AIX360, ART
- Elyra/Jupyter Enterprise Gateway
- Egeria
- Machine Learning eXchange Data Asset eXchange
- NumFOCUS projects
- R projects

- Watson Studio/Watson Machine Learning
- Watson Machine Learning Accelerator
- Watson Natural Language Understanding
- Watson Compare & Comply
- Watson Commerce
- Watson OpenScale
- Watson Speech to Text
- IBM Watson Machine Learning for z/OS
- IBM Db2 Event Store
- IBM Cloud SQL Query
- IBM Cloud Pak™ for Data
- IBM Spectrum Conductor
- IBM InfoSphere Information Governance Catalog

ML Lifecycle