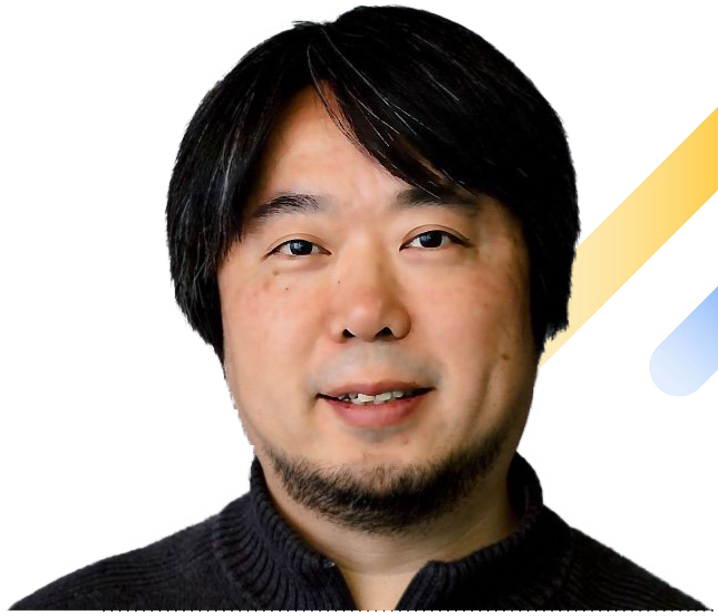
Google Cloud

Model Monitoring

Google Cloud Applied ML Summit

Solving for the future.

06/10/21



Kaz Sato
Developer Advocate
Google Cloud

Marc Cohen
Developer Advocate
Google Cloud

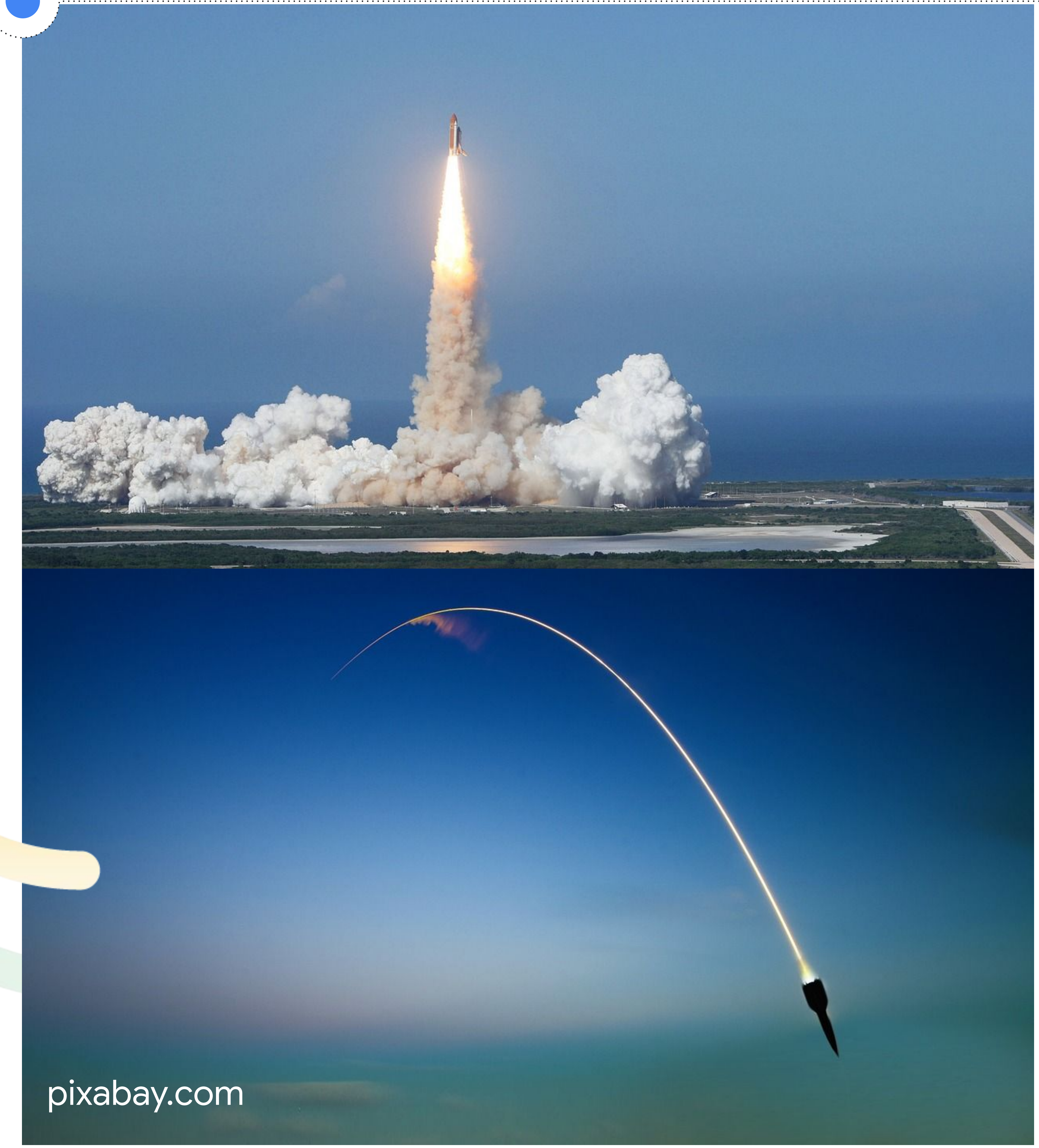
Table of Contents

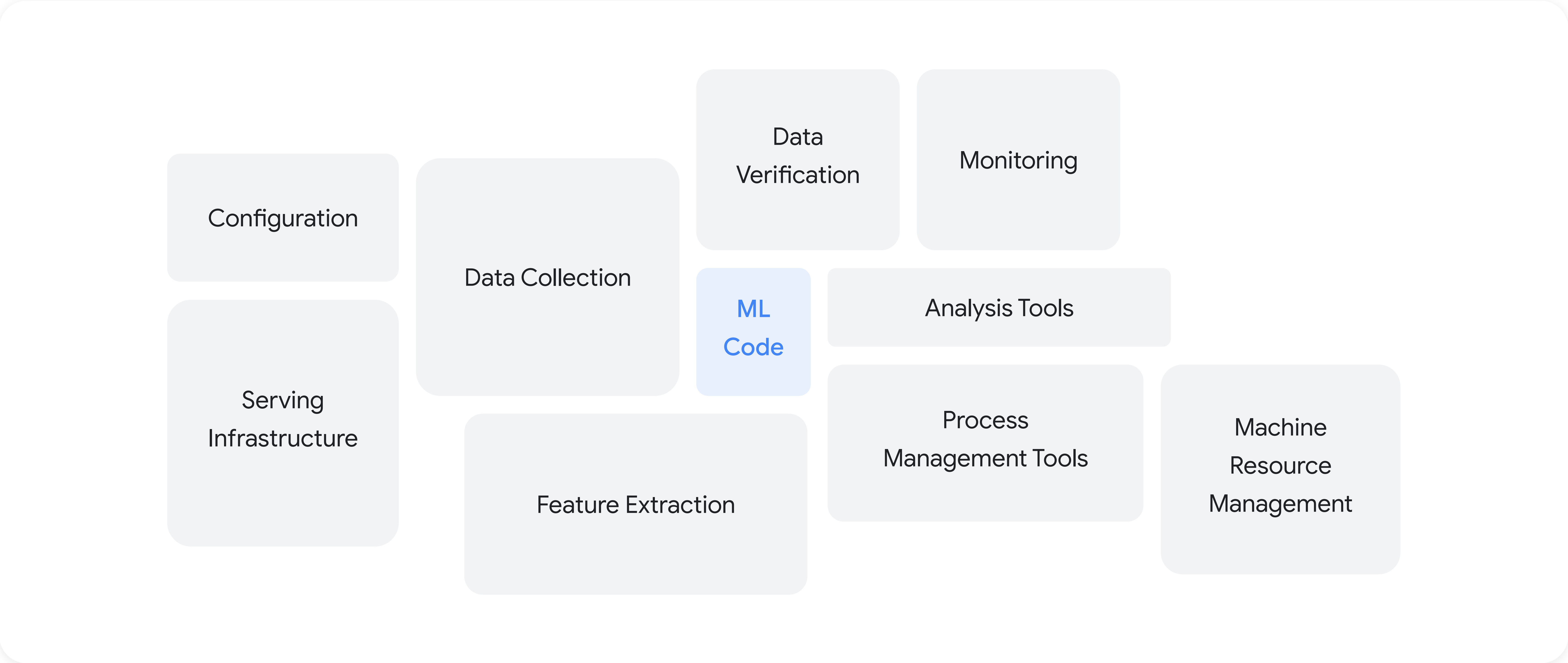
01

MLOps at Google

Launching is easy, Operating is hard.

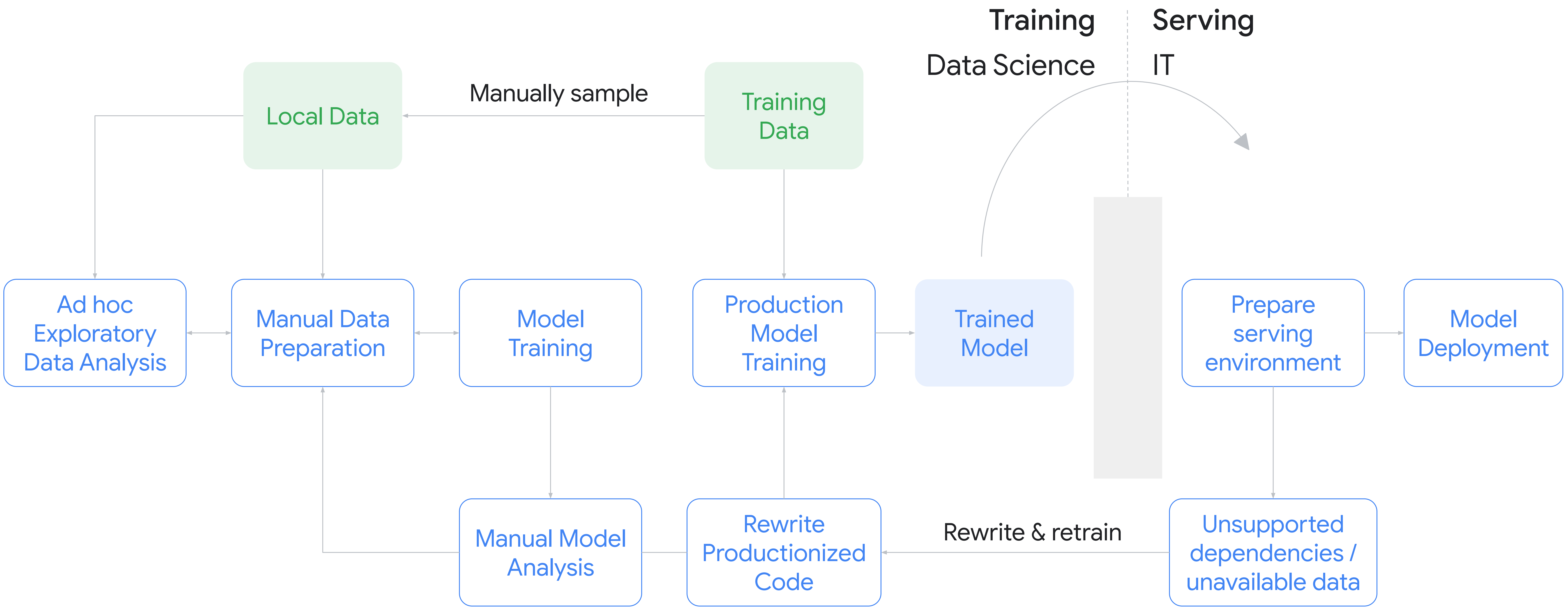
> “The real problems with a ML system will be found while you are continuously operating it for the long term”





Hidden Technical Debt in Machine Learning Systems, NeurIPS 2015

What's happening today:
Data Science and IT (Ops) are isolated



MLOps

An ML engineering culture and practice that aims at **unifying** ML system development (Dev) and ML system operation (Ops)

The Challenges on Productionizing ML

ML challenges

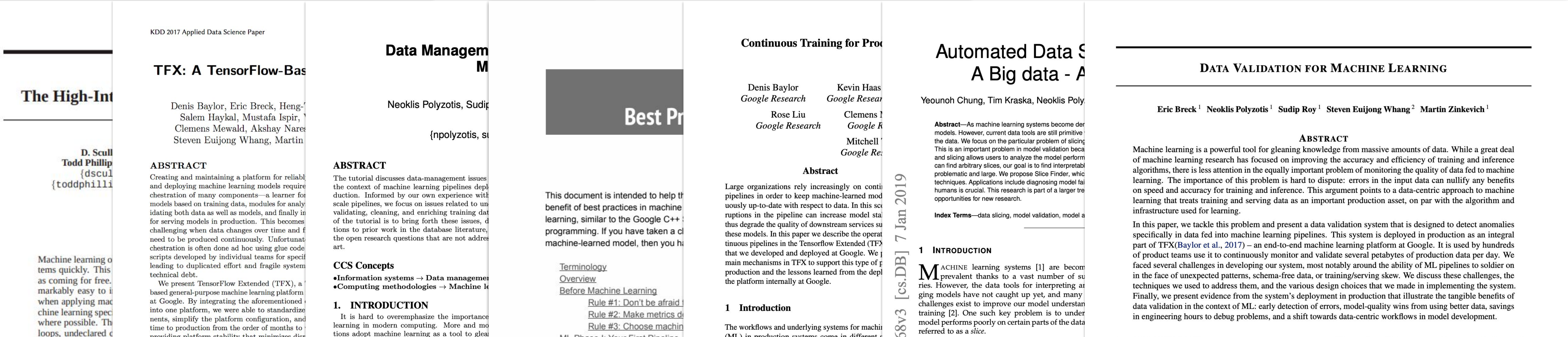
- Governance of data, features, models, pipelines and experiments
- Continuous training and deployment
- Training-serving skew
- Data validation
- Model analysis
- Fairness and Explainability

Production system challenges

- Scalability
- Availability
- Portability
- Reproducibility
- Modularity
- Monitoring and Alerting
- Security
- Hosted or Serverless

The History of Productionizing ML at Google

Continuous Training for Production ML in the TFX Platform. OpML (2019).
Slice Finder: Automated Data Slicing for Model Validation. ICDE (2019).
Data Validation for Machine Learning. SysML (2019).
TFX: A TensorFlow-Based Production-Scale Machine Learning Platform. KDD (2017).
Data Management Challenges in Production Machine Learning. SIGMOD (2017).
Rules of Machine Learning: Best Practices for ML Engineering. Google AI Web (2017).
Machine Learning: The High Interest Credit Card of Technical Debt. NeurIPS (2015).
Hidden Technical Debt in Machine Learning Systems. NIPS (2015).

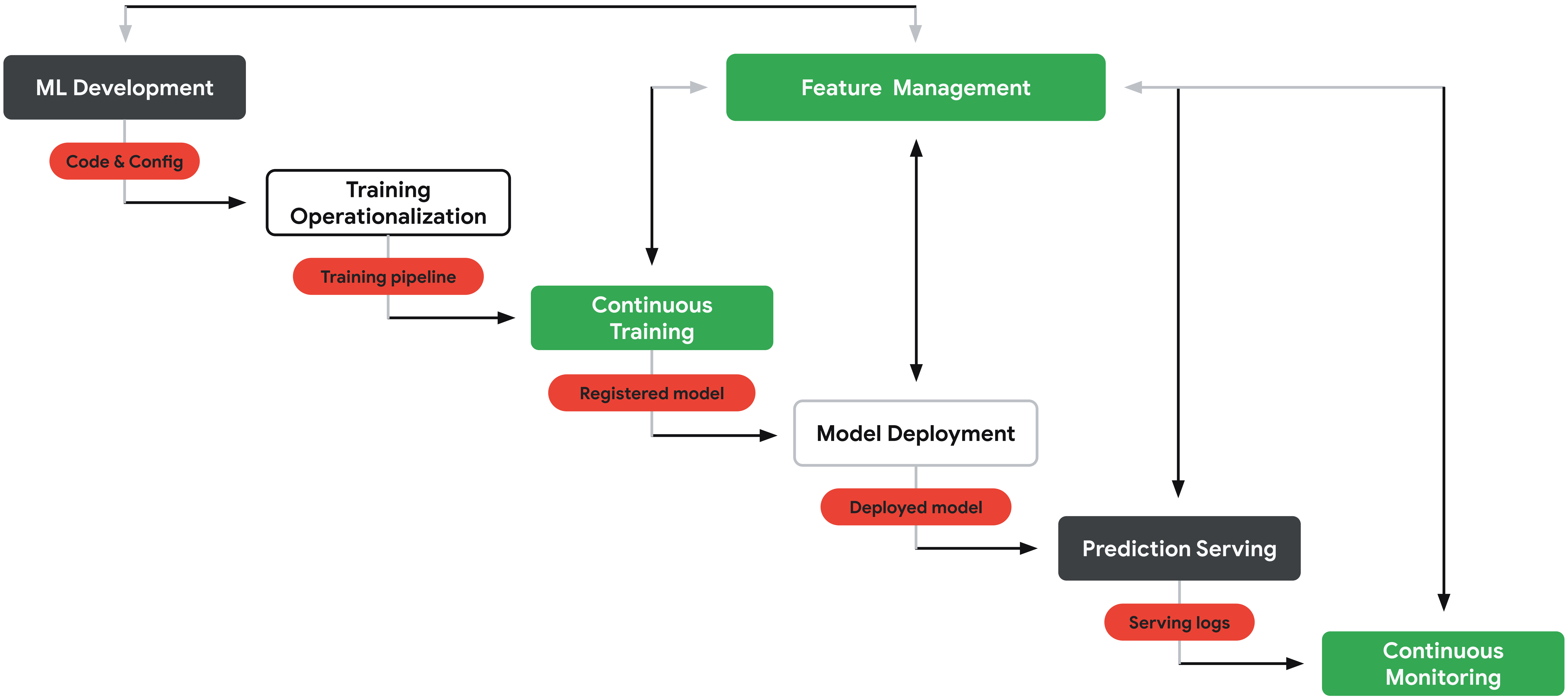


02

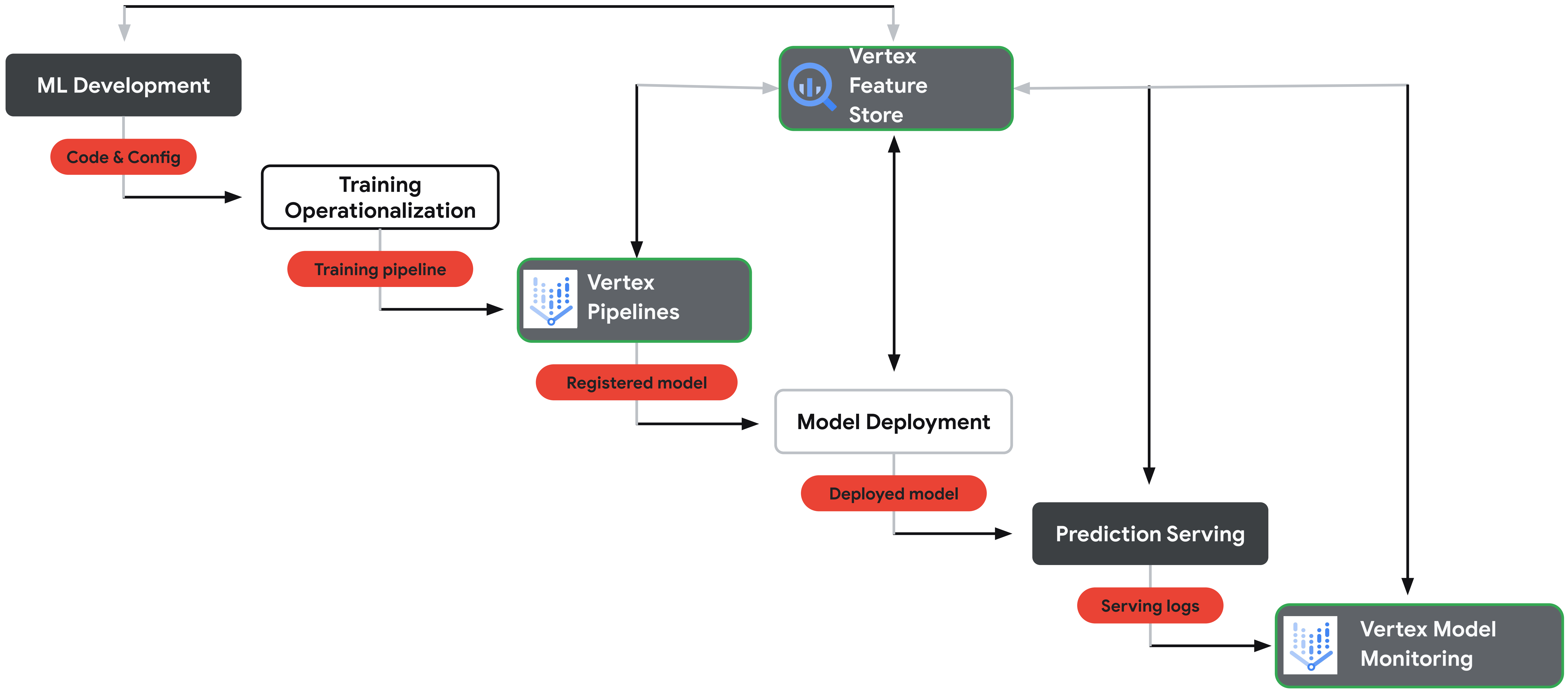
MLOps with Vertex AI

Vertex AI is a **managed ML platform** for practitioners to accelerate **experiments** and **deploy** AI models.

MLOps challenges in ML lifecycle



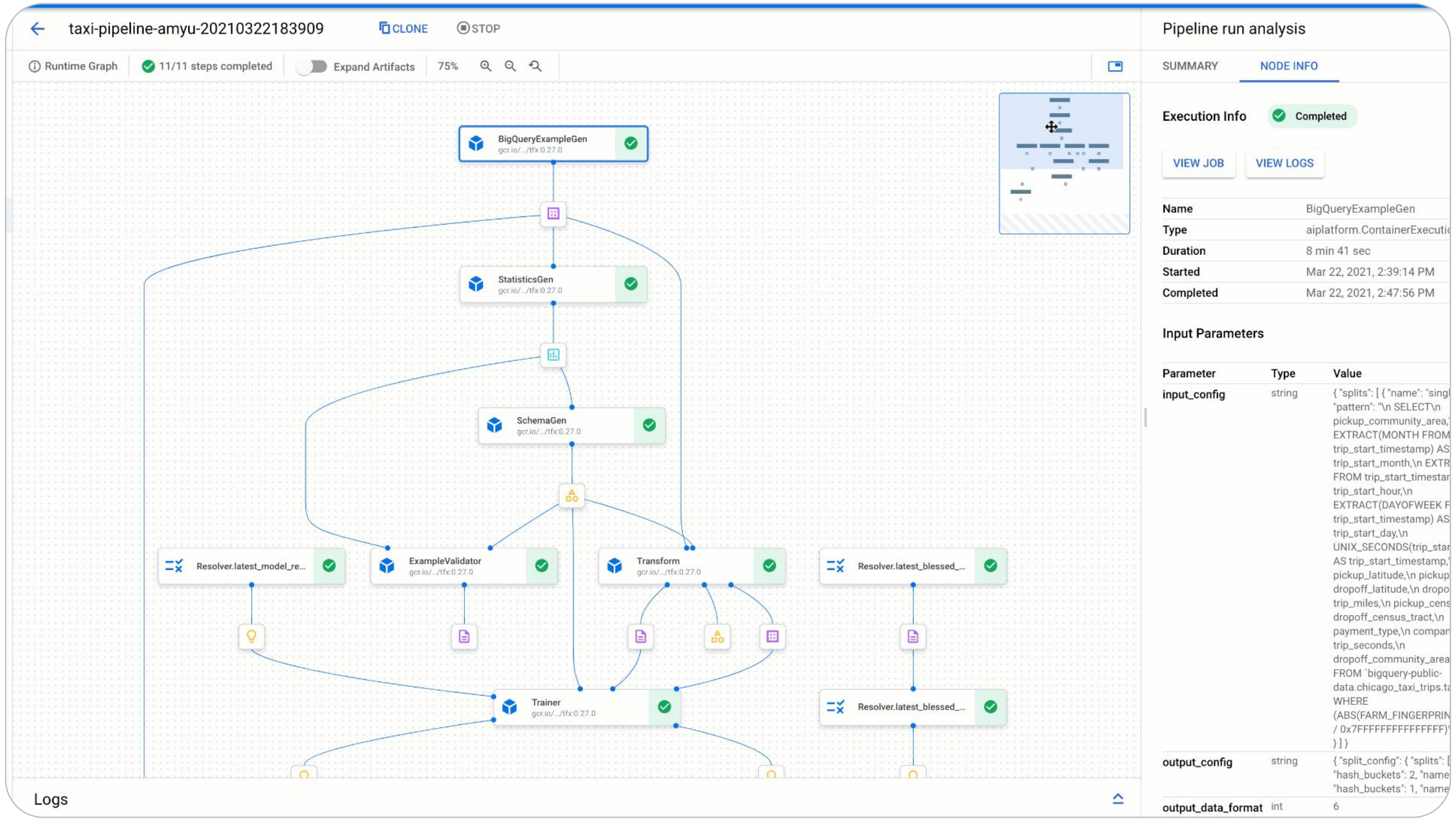
MLOps challenges with Vertex AI



Vertex Pipelines

Build pipelines with familiar open-source **Kubeflow Pipelines**

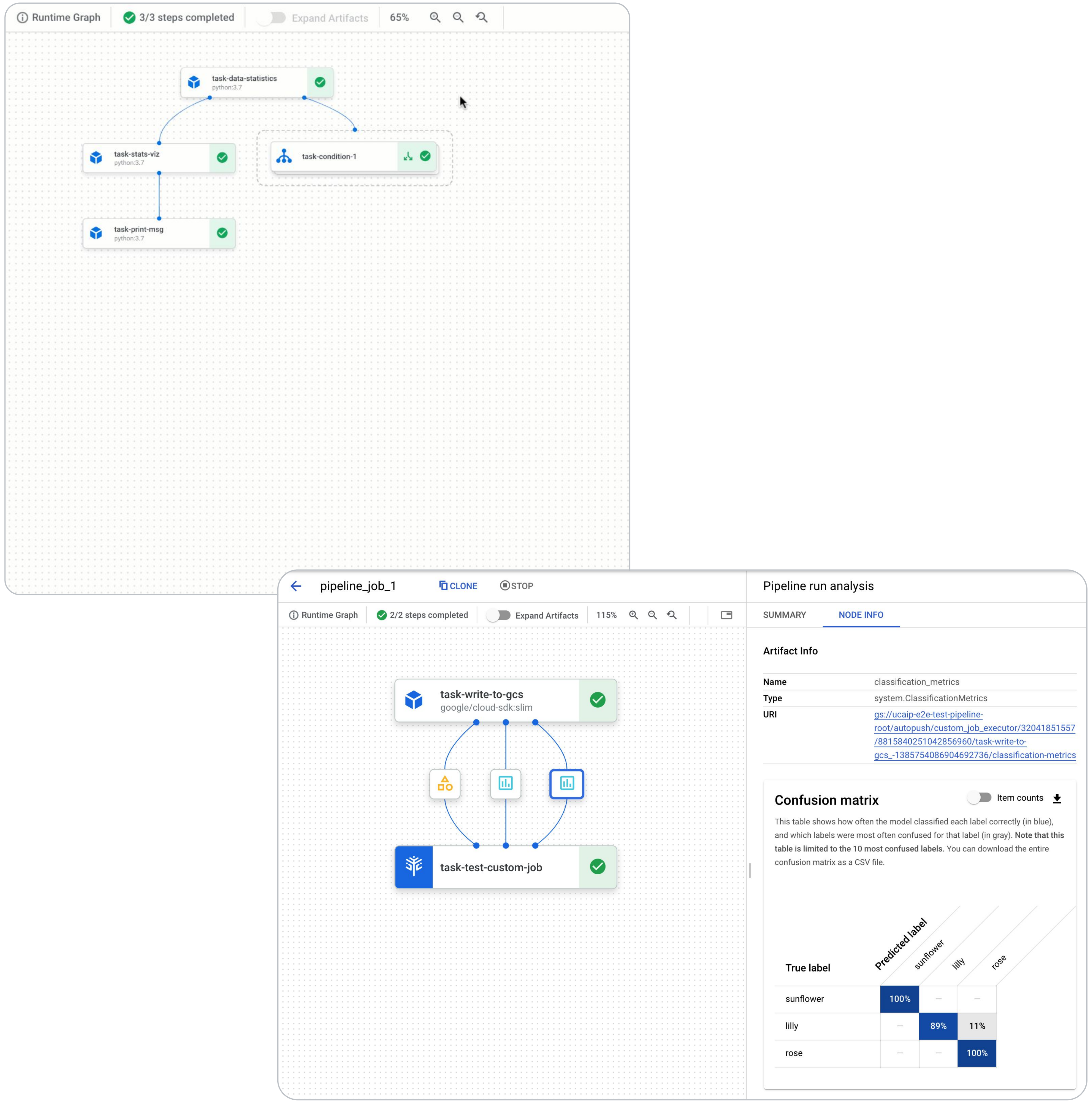
Serverless, automated, scalable: pay only for what you use



Vertex Pipelines

Define **pipelines** with your **components** for data ingestion, preprocessing, training, validation and deployment

Track **artifacts** and **lineage** of data, feature, model and experiment metrics across your ML workflow with metadata store

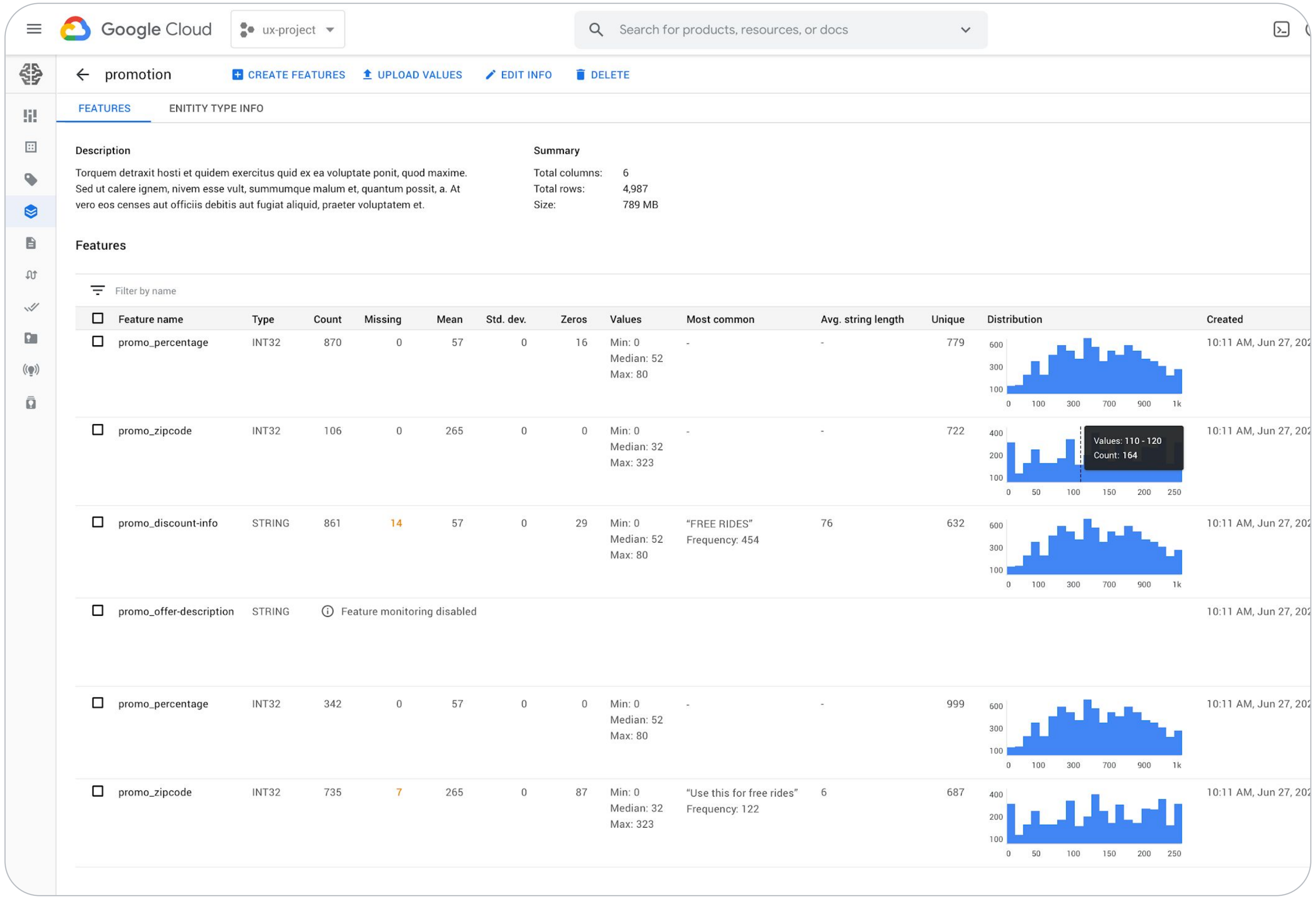


Vertex Feature Store

Share and reuse ML **features** across use cases

Serve the features at scale with low latency

Alleviate training serving skew



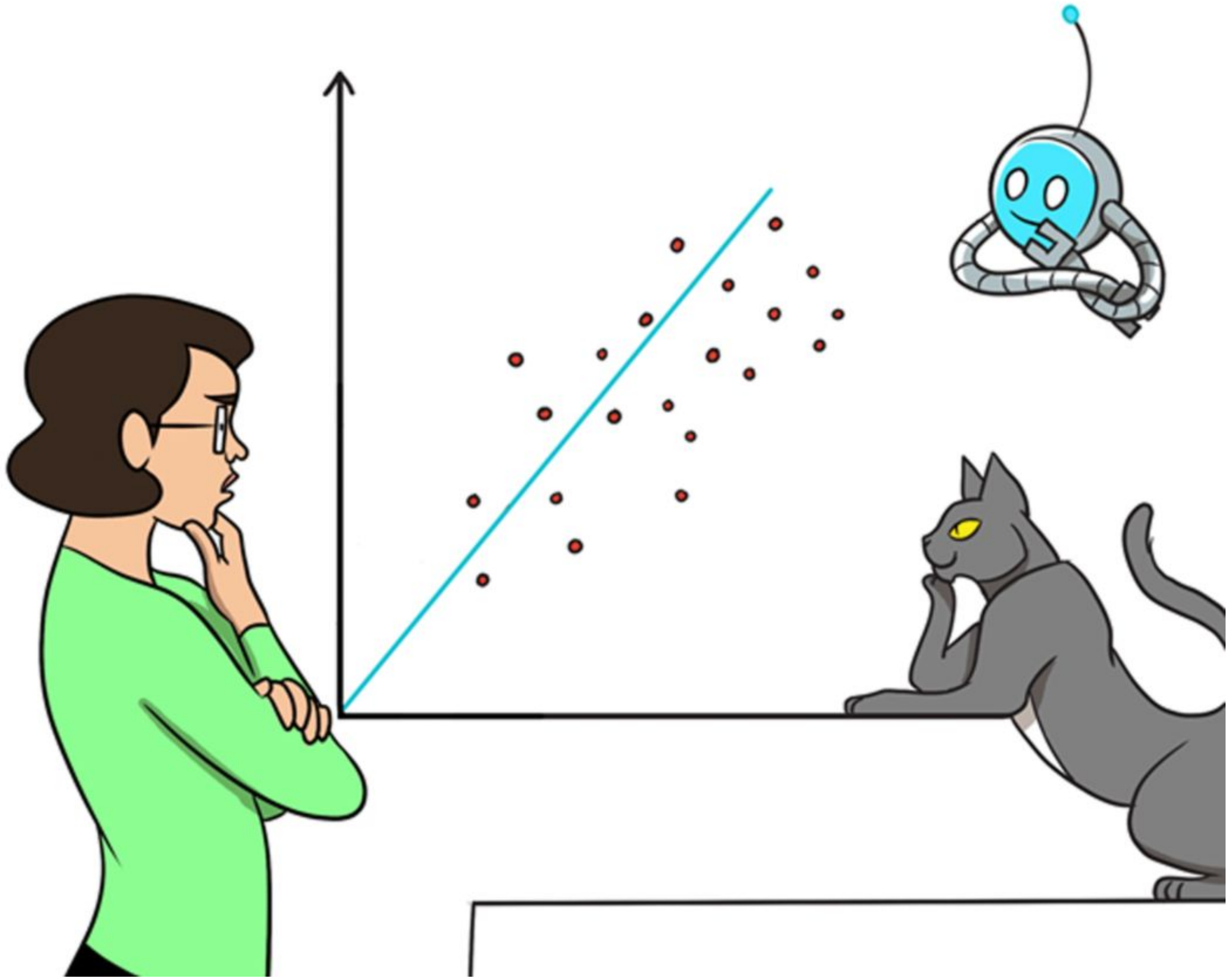
03

Model Monitoring

Why my model doesn't work well?

Any models **get stale**, as they are trained with data in the past

After the deployment, you would see **degradation** of performance and UX



Google

The Silent Killer: Skew and Drift

The important lesson learned at Google:
Skew and **Drift** are the silent killers of your ML models

Training-Serving Skew:
Feature at training: green banana
Feature at serving: yellow banana

Prediction Drift:
Feature at serving: changing from green to yellow

“We want the user to treat data errors with the same rigor and care that they deal with bugs in code”

Google Play app install rate improved **2%** after introducing data validation, finding stale table

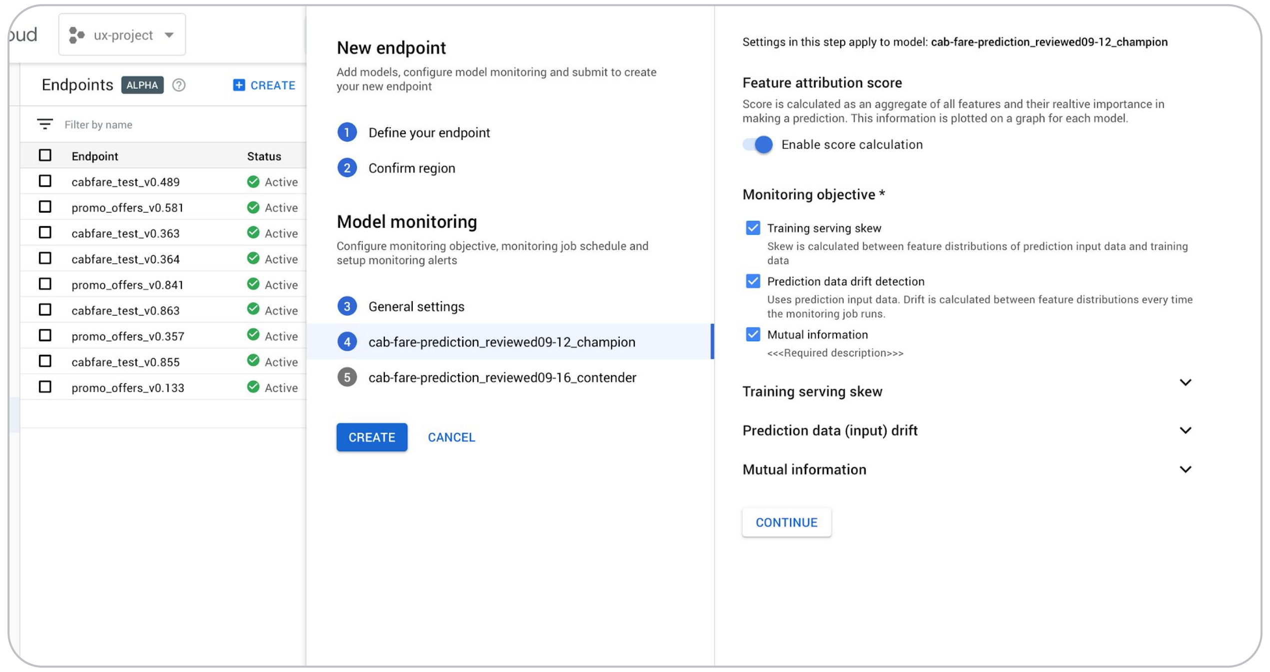
- [TFX paper]

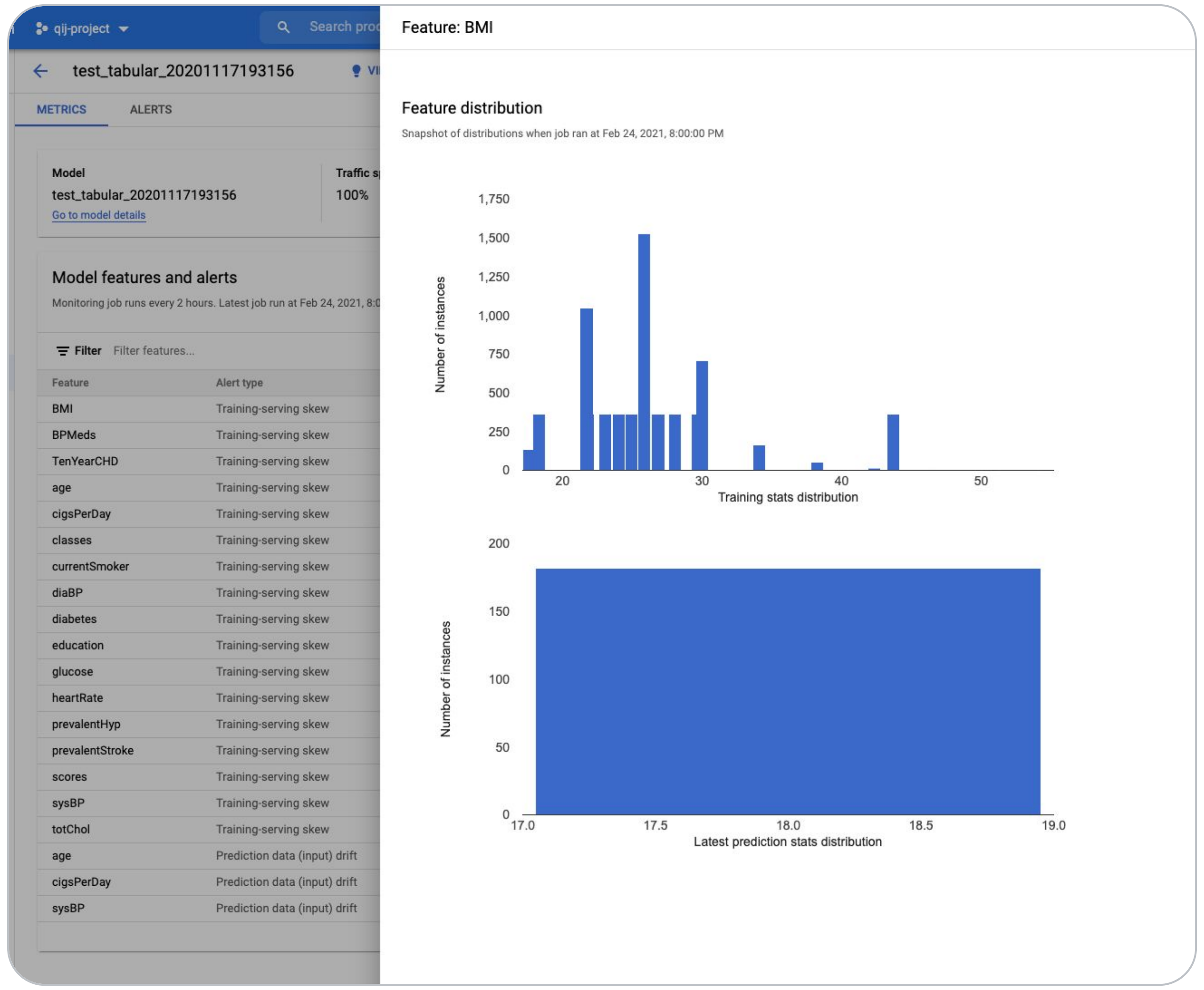
Vertex Model Monitoring

Automatically alert your data scientists and ML engineers when model performance changes

Detect **drift** and **training-serving skew**

Provides confidence in model **reliability**





Easy and proactive monitoring of model performance

Monitor and alert

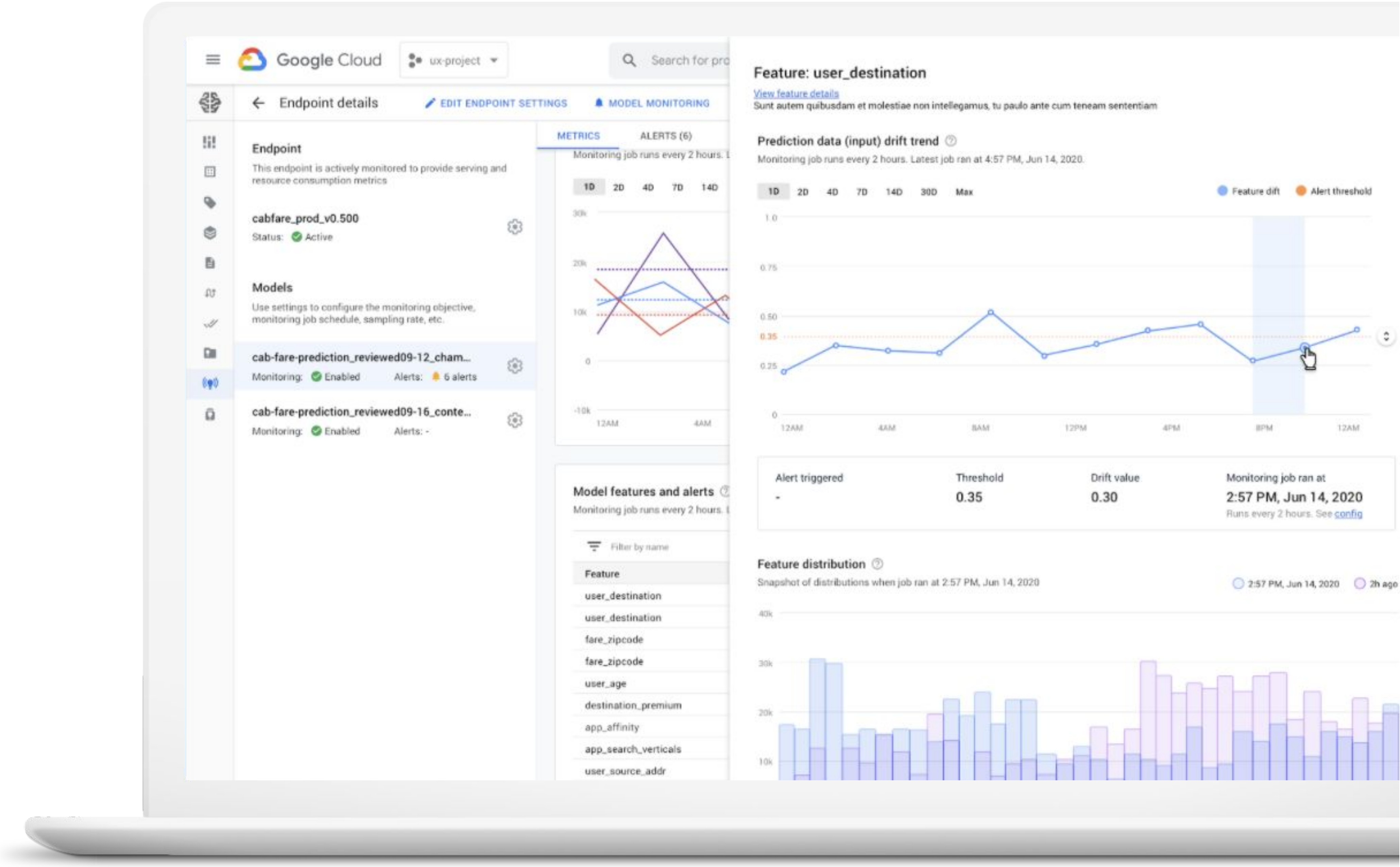
For model's predictive performance

Diagnose

Identify the cause for the deviation

Update Model

Trigger model re-training pipeline

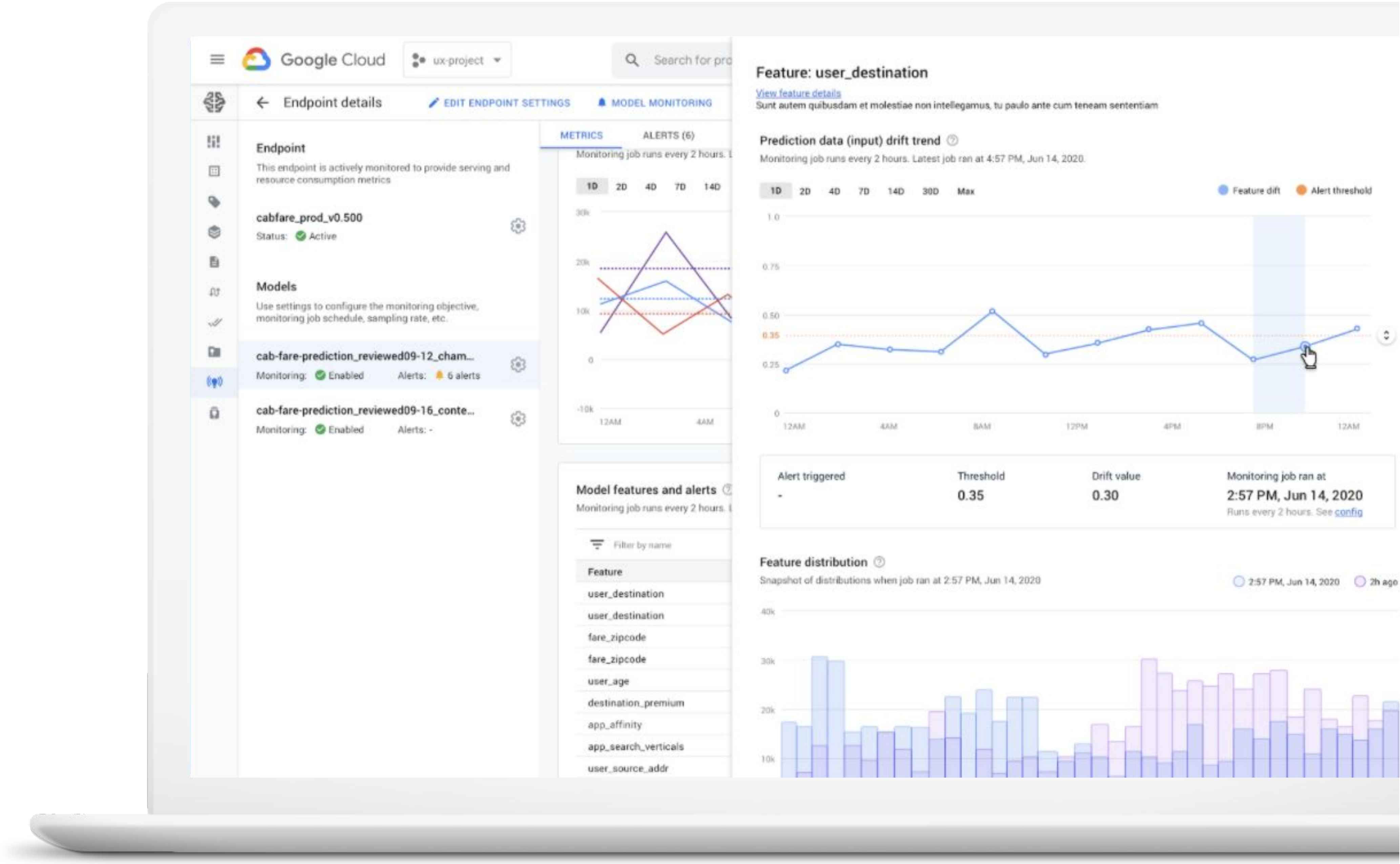


Drift and skew detection

Visualize drift over time

Easily **compare** latest feature distributions with baseline

Easily **set** or **update** alerting thresholds



How to detect Skew and Drift

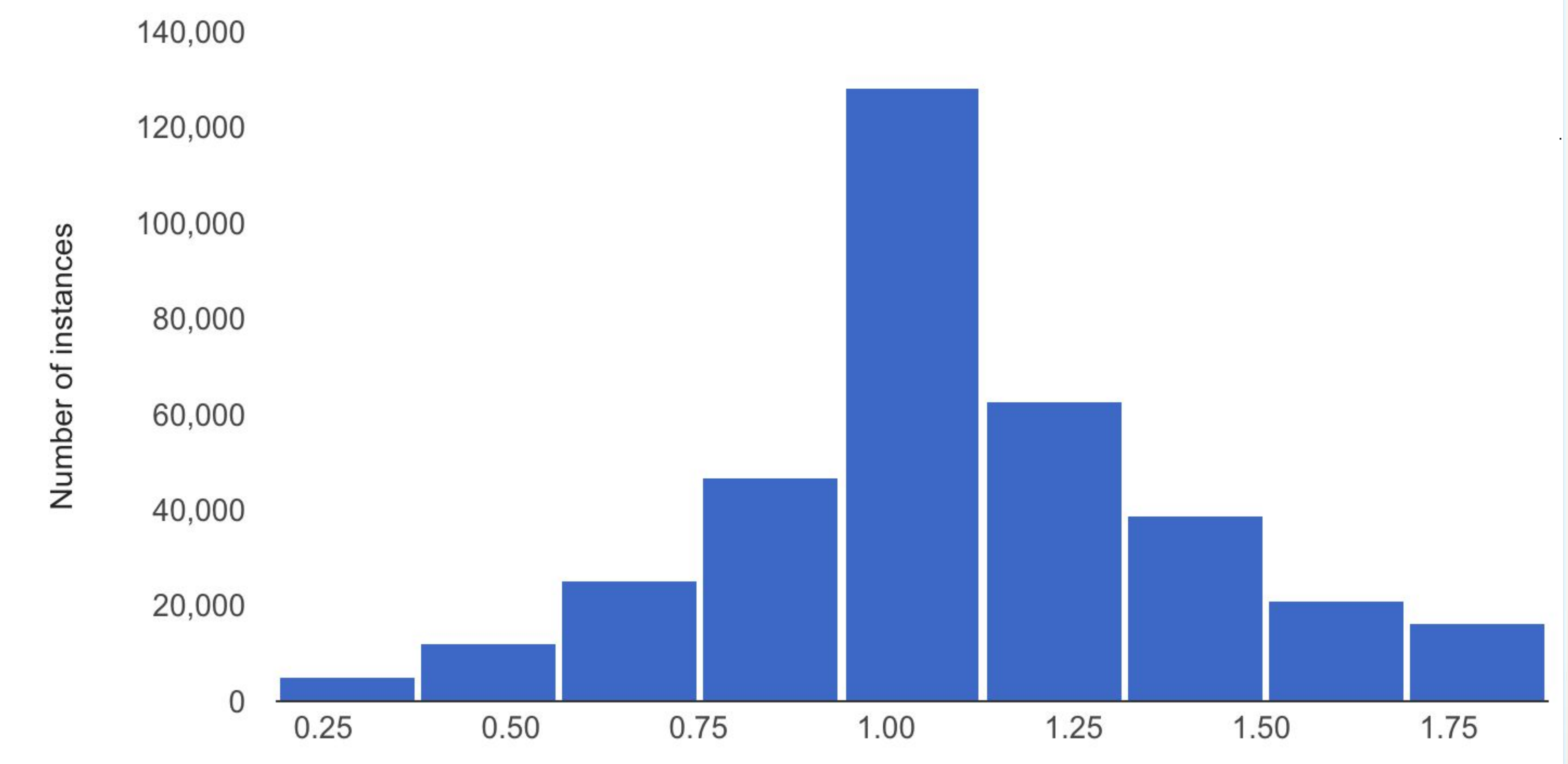
For Skew: set the baseline with stats of the training data

For Drift: set the baseline with the stats of the serving data in the past

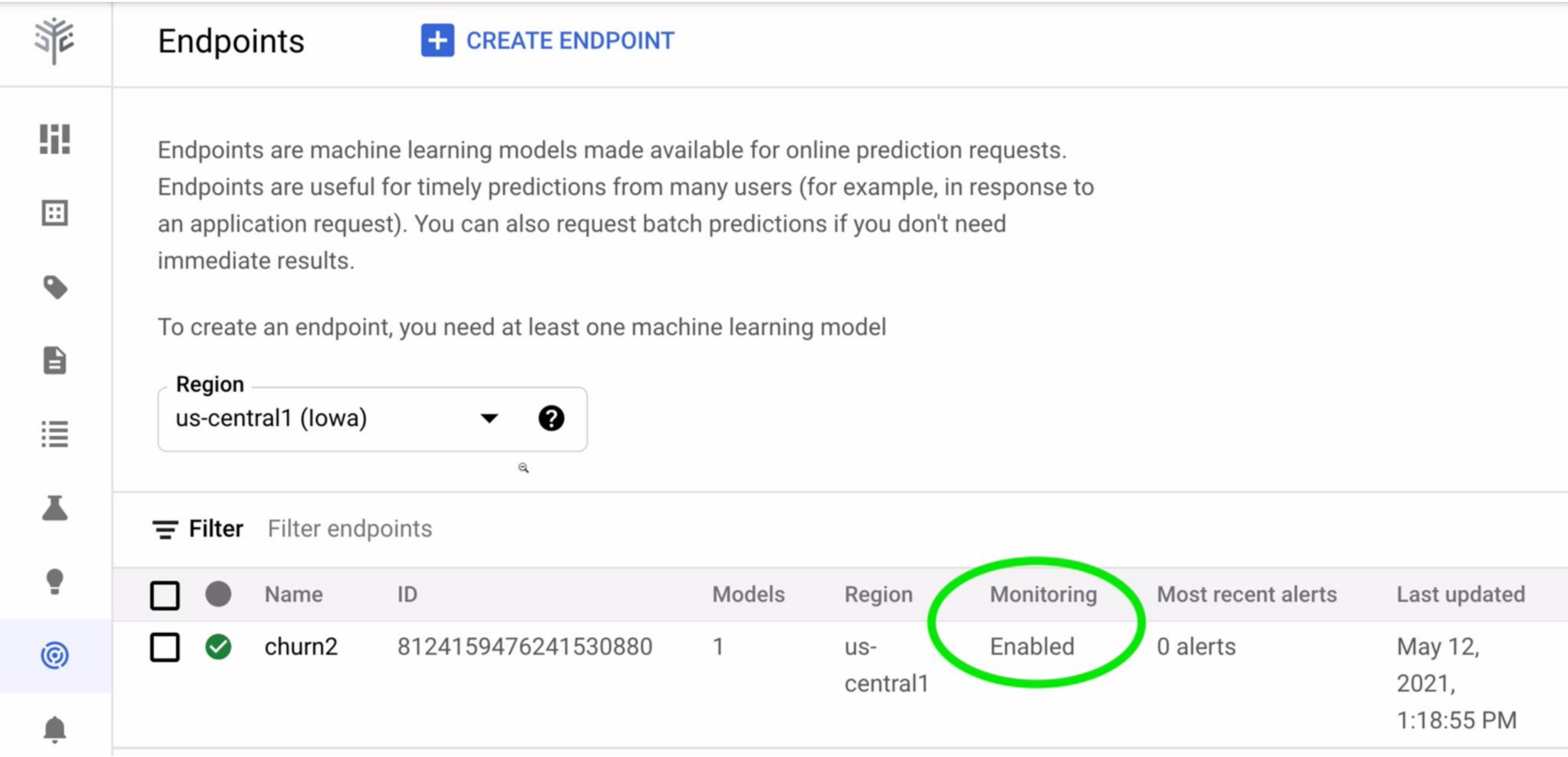
Measure:

JS divergence
(for numerical features)

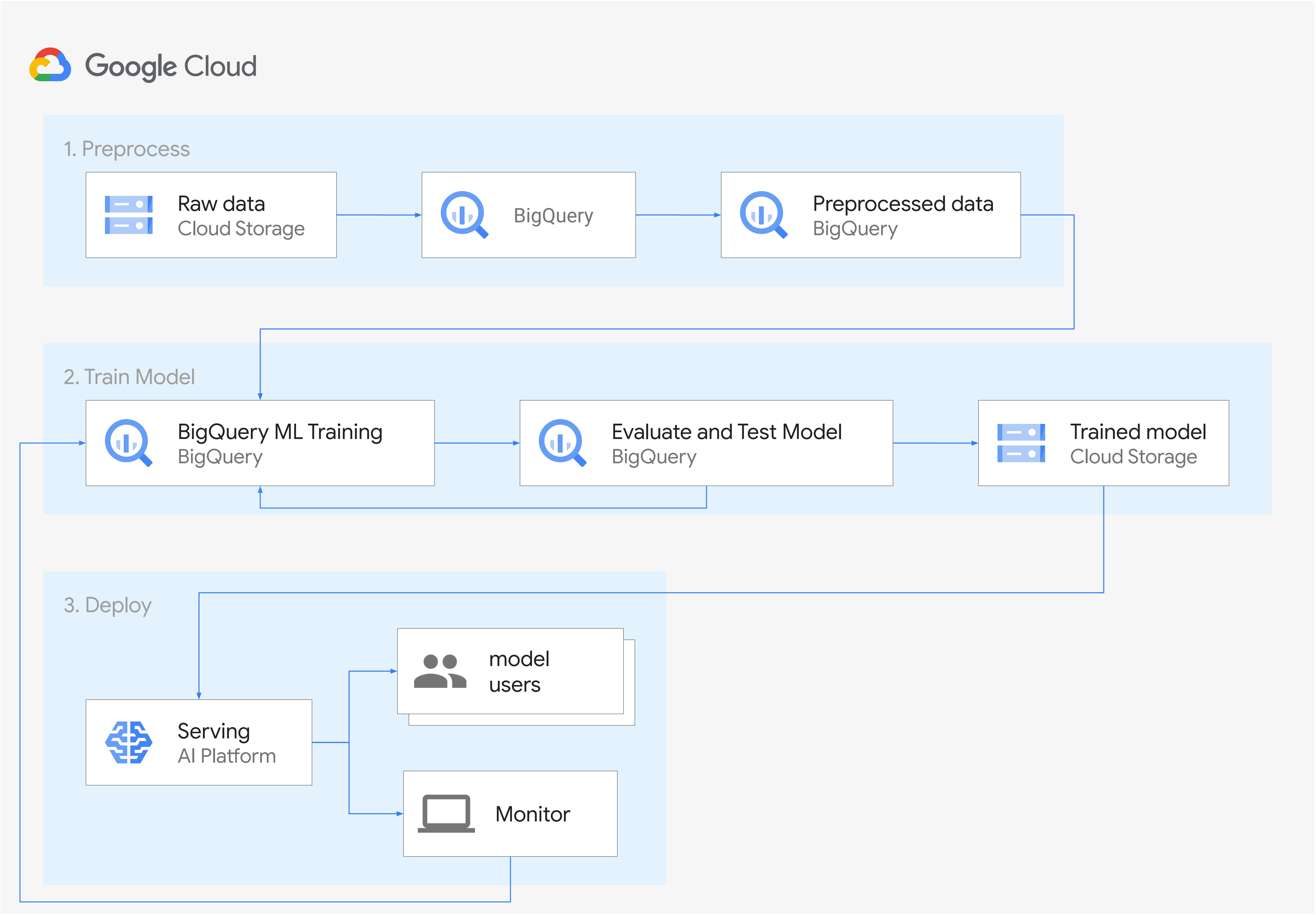
L-inf distance
(for categorical features)



04 Demo


Endpoints
CREATE ENDPOINT
Endpoints are machine learning models made available for online prediction requests. Endpoints are useful for timely predictions from many users (for example, in response to an application request). You can also request batch predictions if you don't need immediate results.
To create an endpoint, you need at least one machine learning model
Region
us-central1 (Iowa)
Filter
Filter endpoints
Name
ID
Models
Region
Monitoring
Most recent alerts
Last updated
churn2
8124159476241530880
1
us-central1
Enabled
0 alerts
May 12, 2021, 1:18:55 PM

Model Monitoring Demo



05

Next Steps

Thank you.