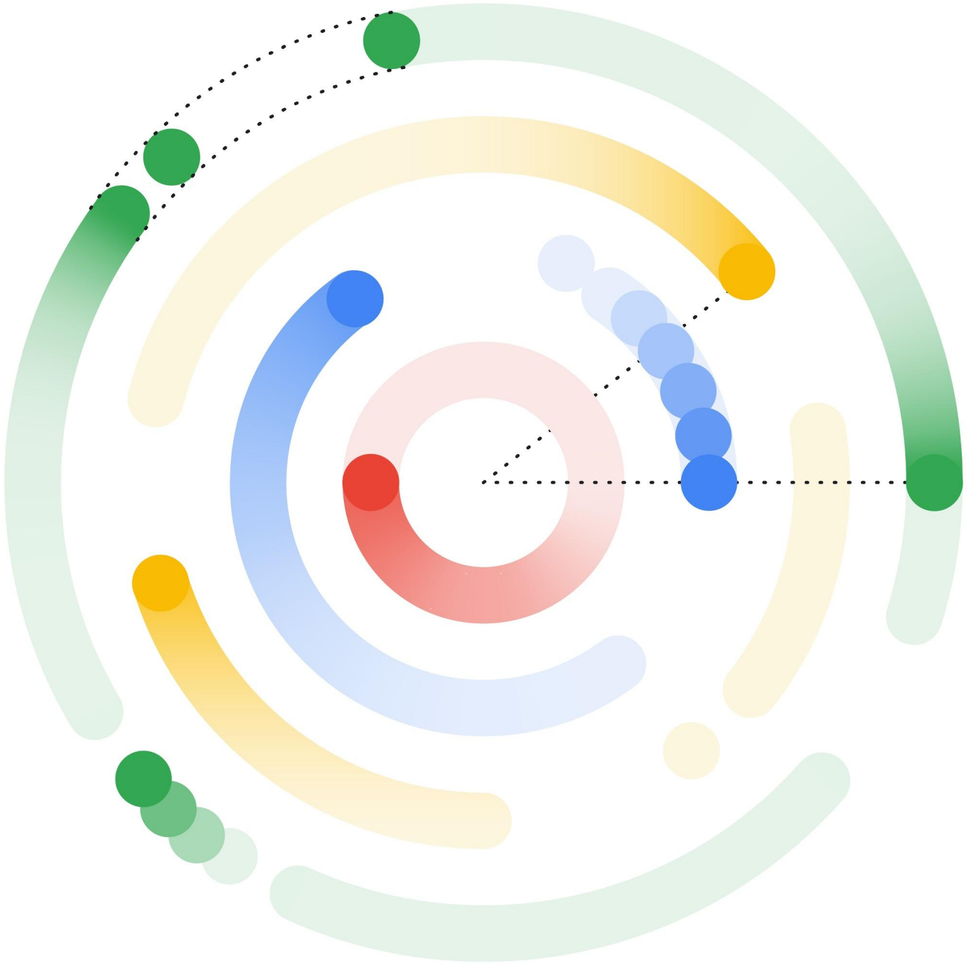
Google Cloud

Building an End-to-End Pipeline

Google Cloud Applied ML Summit

Solving for the future.



06/10/21

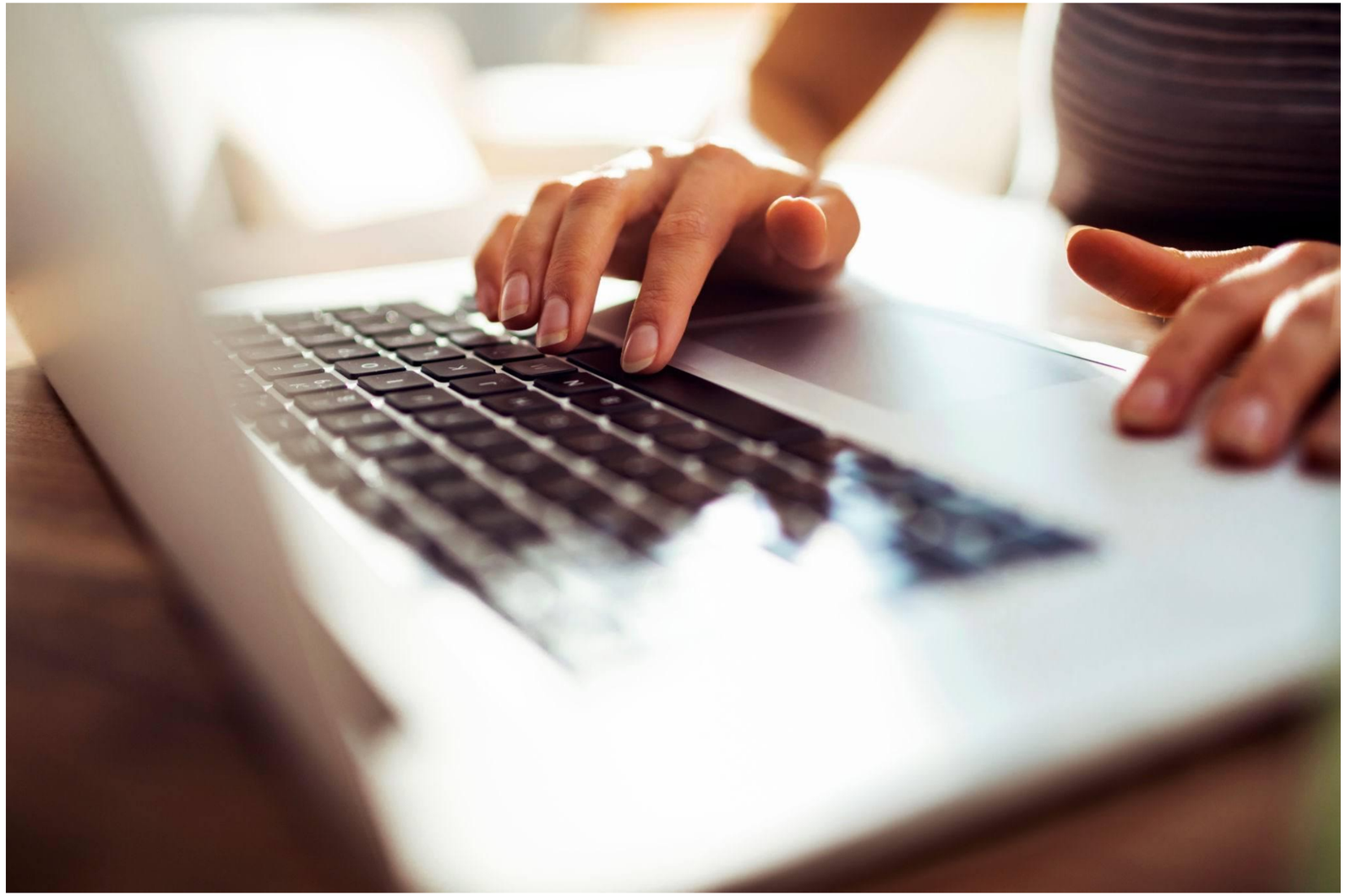
Amy Unruh
Staff Developer Advocate
Google Cloud

Agenda

01 MLOps on Vertex AI

02 Introducing Vertex Pipelines

03 Vertex Pipelines demos



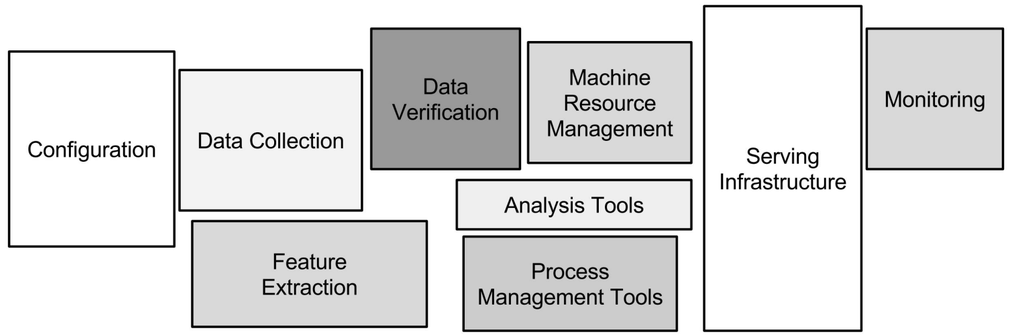
Building the first proof-of-concept version of a machine learning system can be pretty easy...

But when you try to productionize...

- What was built as the prototype is only a small piece of what you need to pay attention to
- Problems show up when you try to scale out, and keep a system in long-term continuous operation

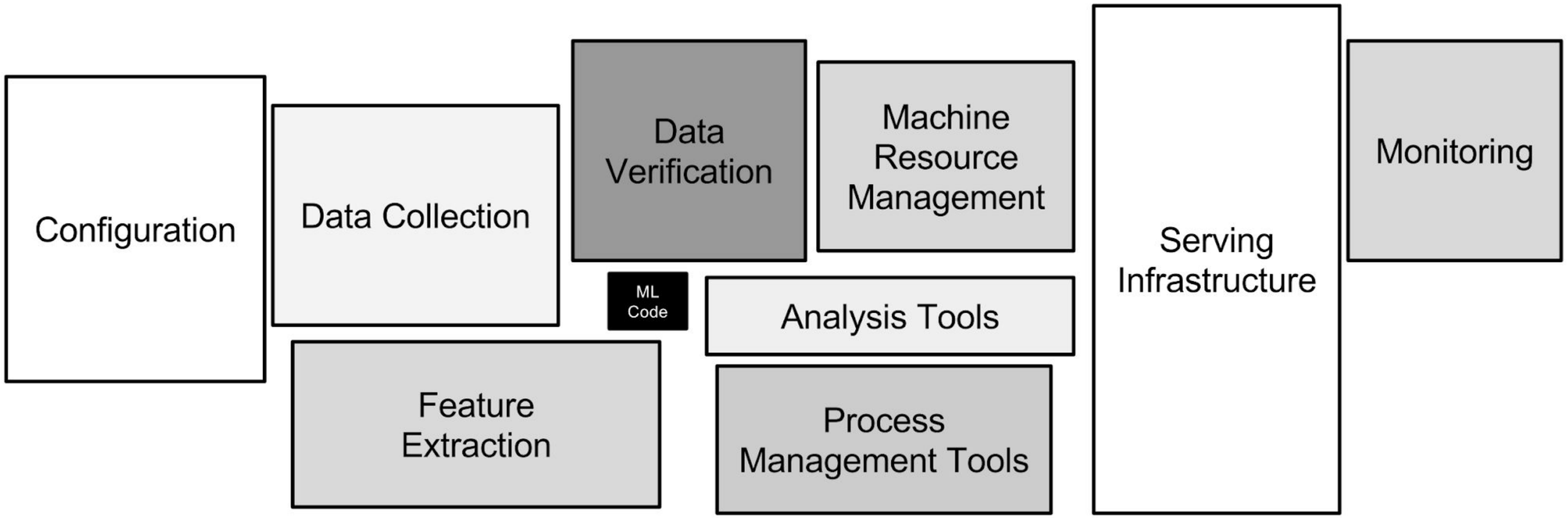
Perception

ML Code



Credit: Hidden Technical Debt of Machine Learning Systems, D. Sculley, et al.

Reality



Credit: Hidden Technical Debt of Machine Learning Systems, D. Sculley, et al.

Why do things become harder? (an incomplete list)

- **data cleaning** and **processing** is hard at scale
- **scaling** out training and serving; infrastructure issues
- **tracking, monitoring,** and **reproducibility** requirements
  - model or data **drift**
  - training/serving **skew**
- **access control** issues, **security** requirements
- (and lots more)

MLOps

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

A canonical ML workflow

Training

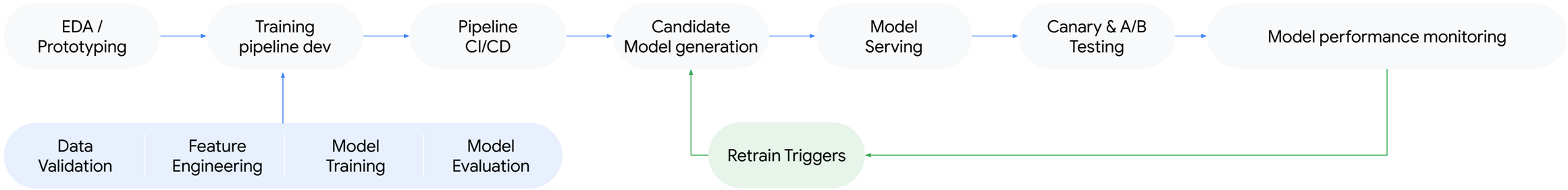
1 Experimentation

2 (Re) Training

Serving

3 Model Deployment

4 Continuous Monitoring

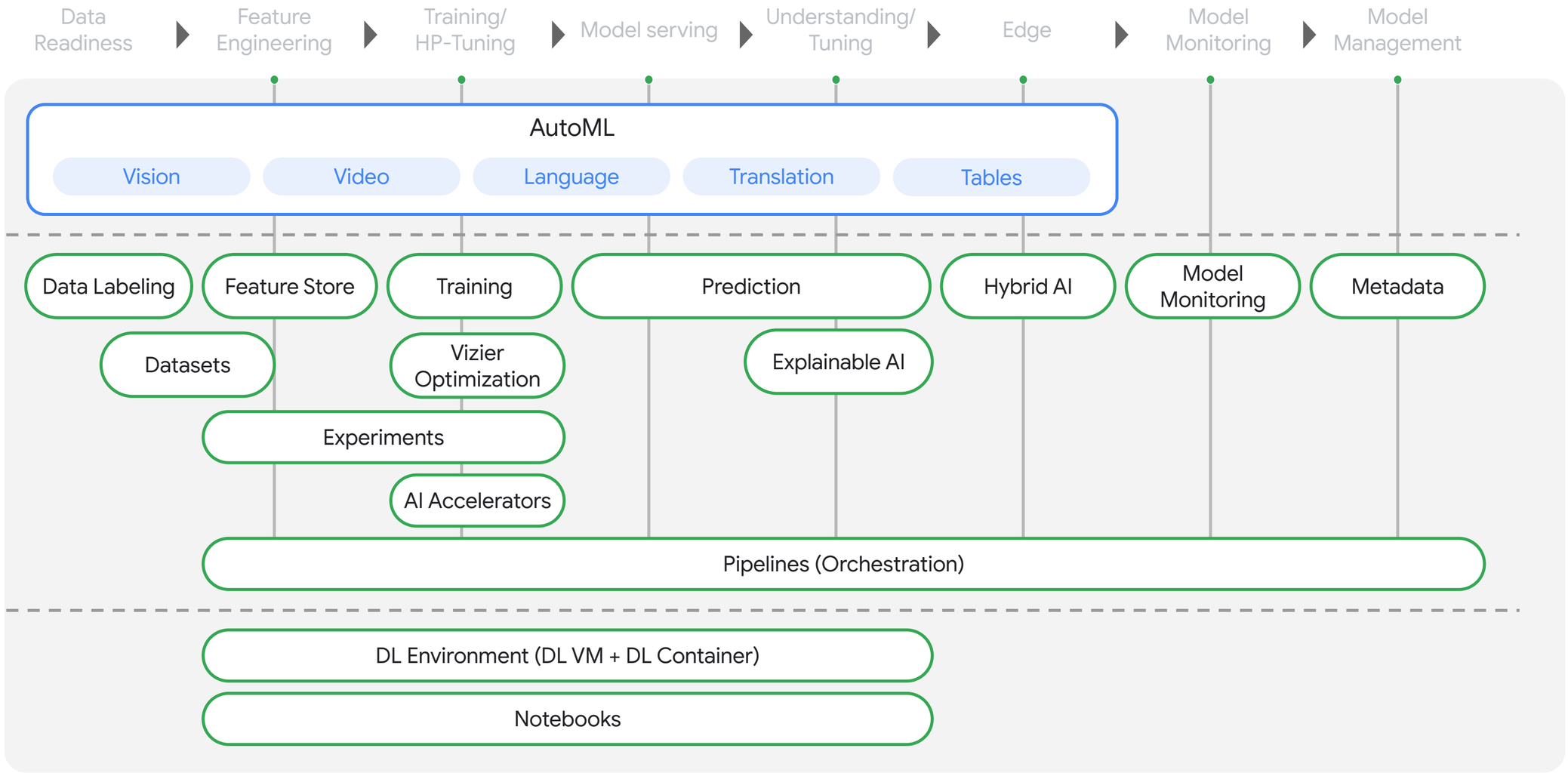


Model Management & Governance

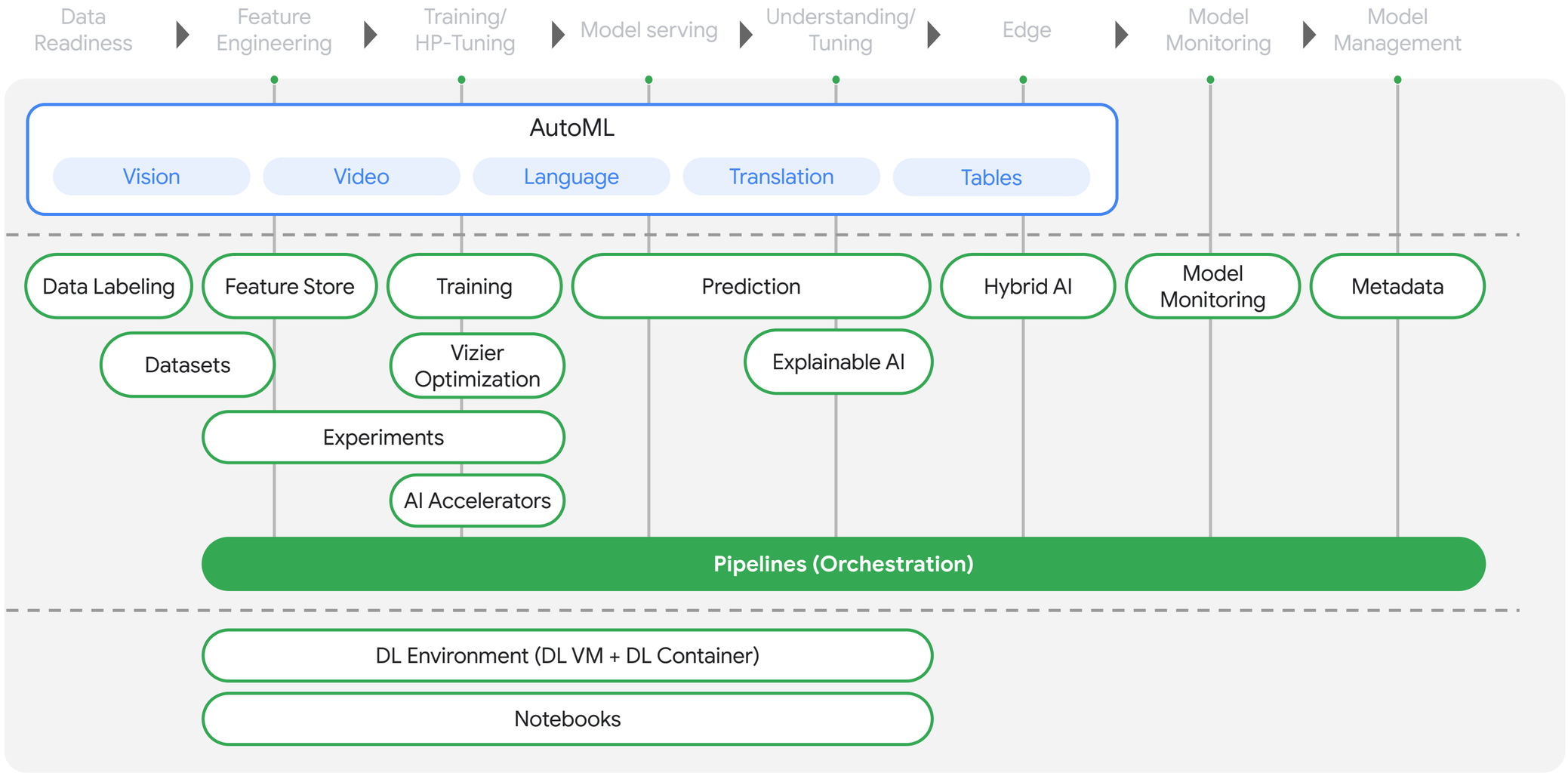
Model Registry | Model Cards & Reporting | Model Provenance | Compliance

Vertex AI is a
managed ML platform
for every practitioner
to speed the rate of
experimentation and
accelerate deployment
of AI models.

What's included in Vertex AI?



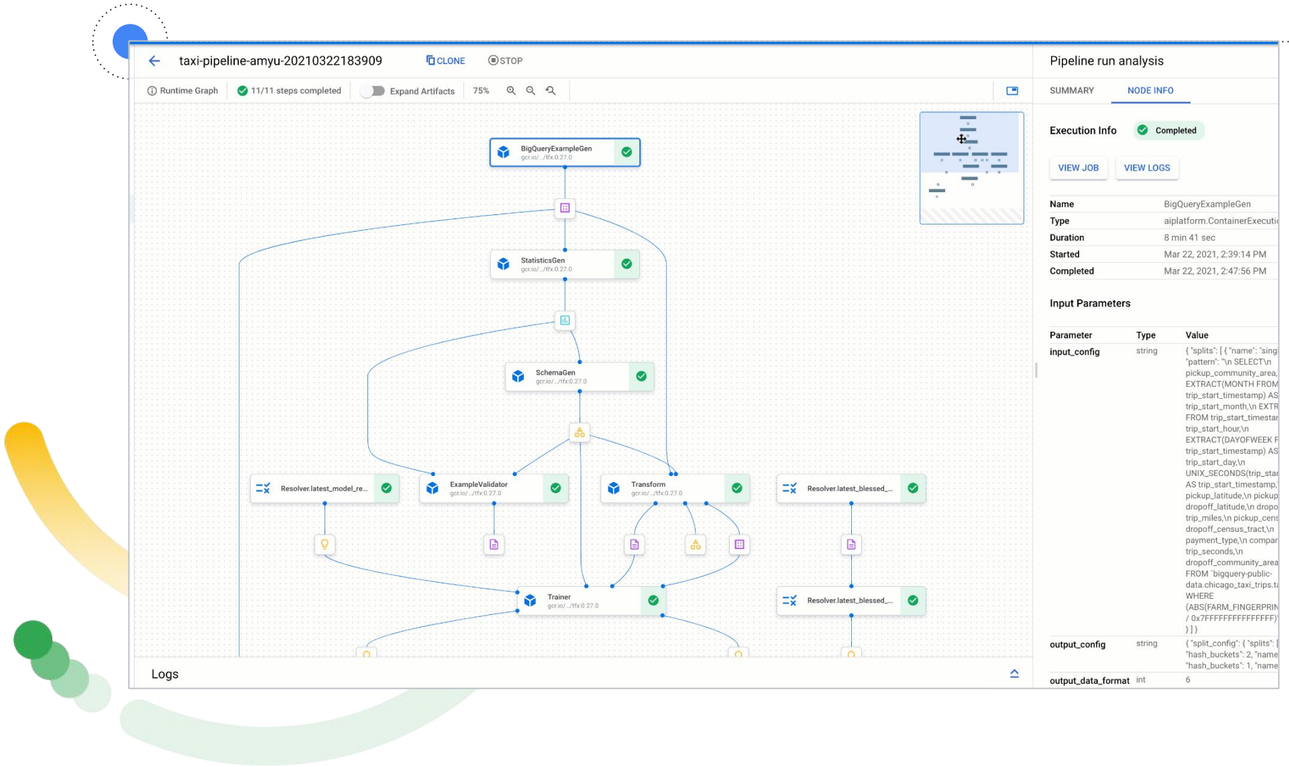
Pipelines are the backbone of production ML systems



Vertex Pipelines

Automated, **scalable**, **serverless**, **cost-effective**: pay only for what you use

Build pipelines with familiar **open-source Python SDKs** like TFX and Kubeflow Pipelines



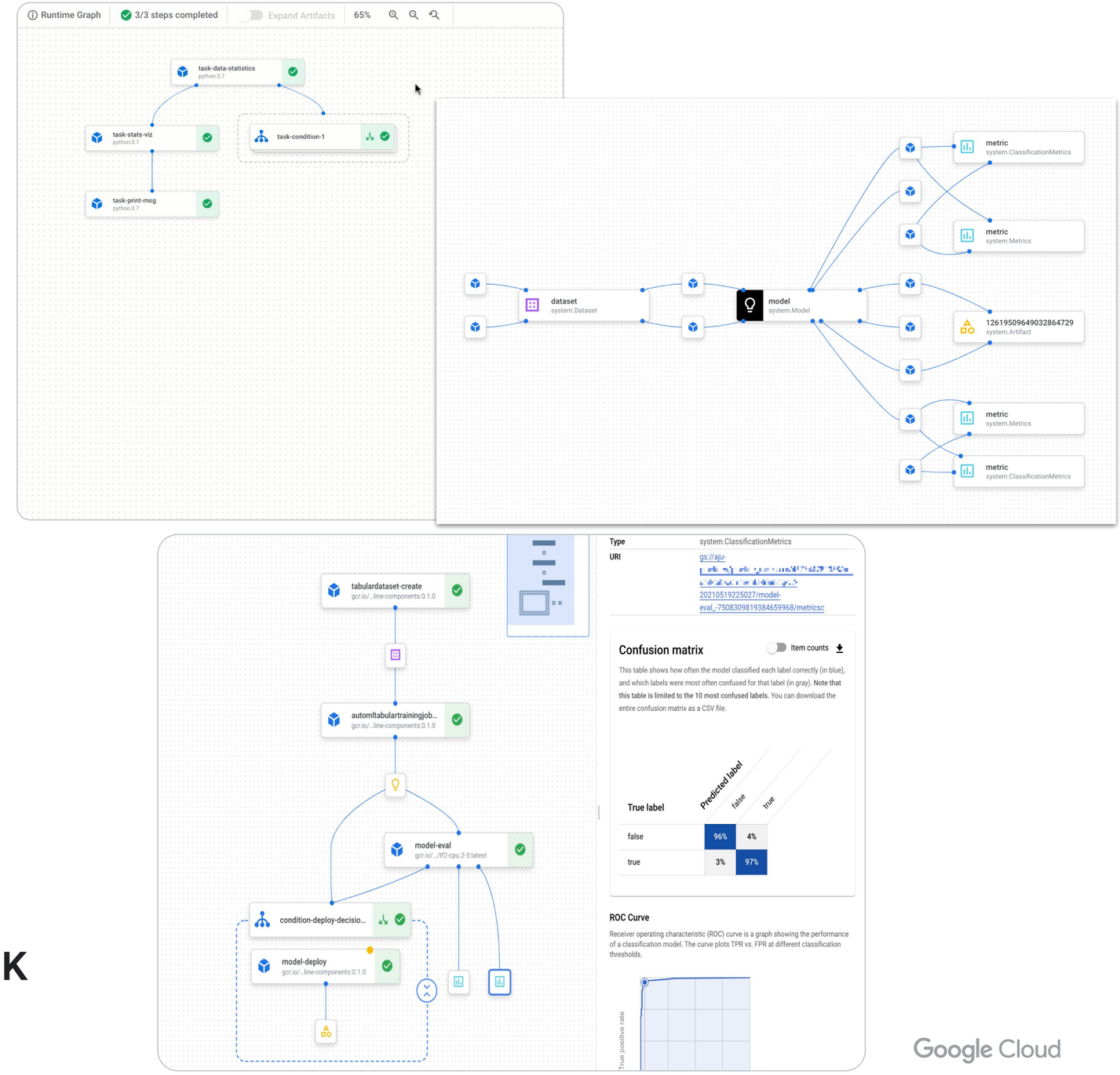
Vertex Pipelines

Add conditional logic and branches to your pipeline

Automatically log **metadata** for every artifact produced by the pipeline

Track **artifacts**, **lineage**, **metrics**, and **execution** across your ML workflow

Support for **Cloud IAM, VPC-SC,** and **CMEK**



Google Cloud

Kubeflow Pipelines (KFP) and TensorFlow Extended (TFX) SDKs

KFP: https://www.kubeflow.org/docs/pipelines/

TFX: https://www.tensorflow.org/tfx

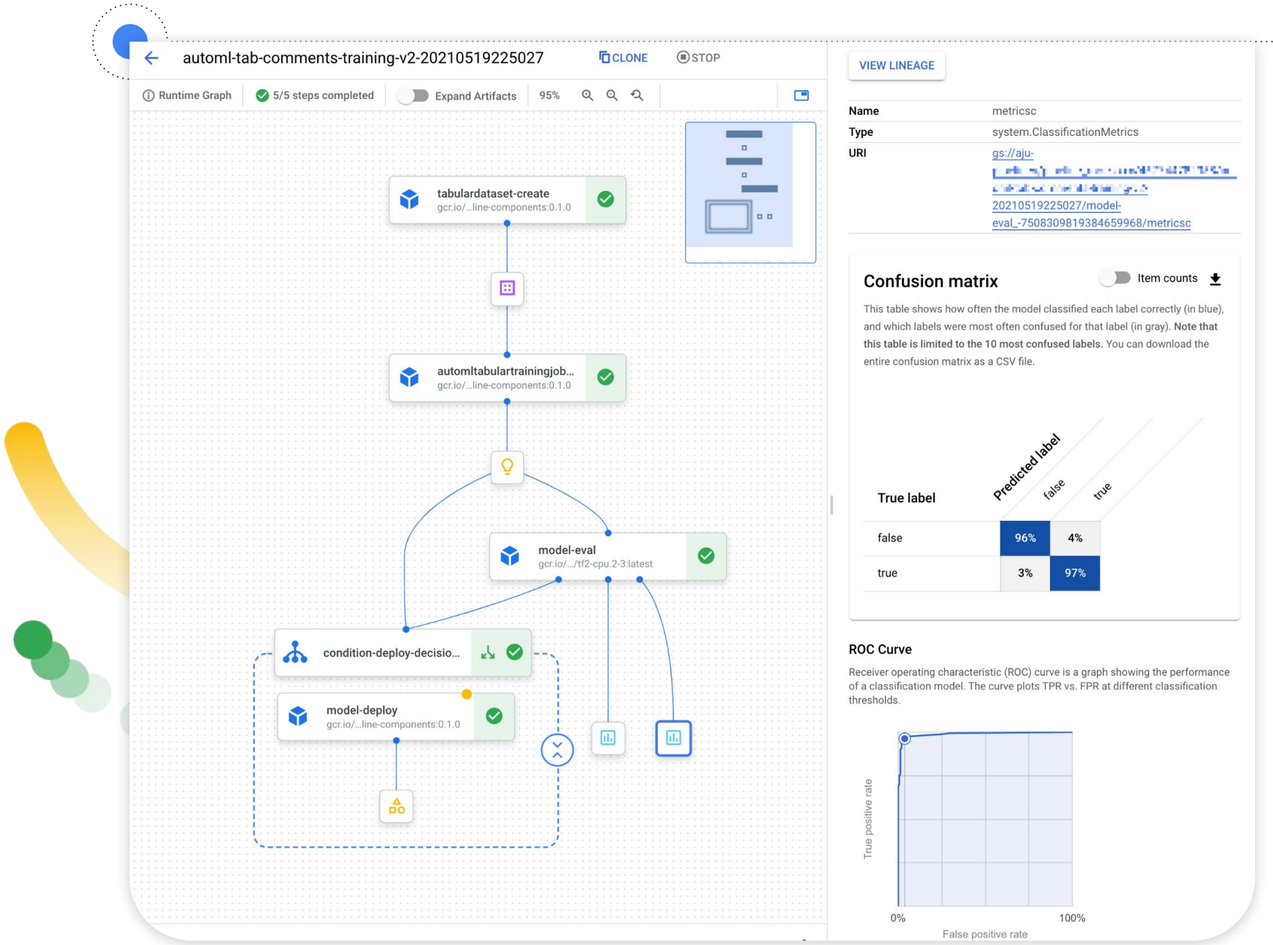
Both **open-source** SDKs are supported by both **KFP OSS**/Hosted KFP and **Vertex Pipelines**.

Both SDKs support use of both prebuilt and custom *components* (pipeline step definitions).

Demo and code!

Some Demo Highlights

- GCP prebuilt pipeline components
- Pipeline control flow
- Python-function-based components
- Metrics and visualization support
- Step-level caching



Vertex Pipelines: Key capabilities

Python SDKs

Data Scientist friendly Python SDKs

Serverless and Scalable

Run as many pipelines on as much data as you want.

Metadata and lineage

Store metadata for every artifact produced by the pipeline.

Monitoring UIs and APIs

Track and debug pipelines executions

Security

Supports Cloud IAM, VPC-SC, and CMEK.

Cost-effective

Only pay for the pipelines you run and the resources they use

Vertex Pipelines Resources

- Documentation: cloud.google.com/vertex-ai/docs/pipelines
- Examples and tutorials: github.com/GoogleCloudPlatform/ai-platform-samples/tree/master/ai-platform-unified/notebooks/unofficial/pipelines
- Google Cloud Pipeline Components: github.com/kubeflow/pipelines/tree/master/components/google-cloud
- KFP v2: https://www.kubeflow.org/docs/components/pipelines/sdk/v2/

Google Cloud

Thank you!