Introduction to AI Platform Pipelines
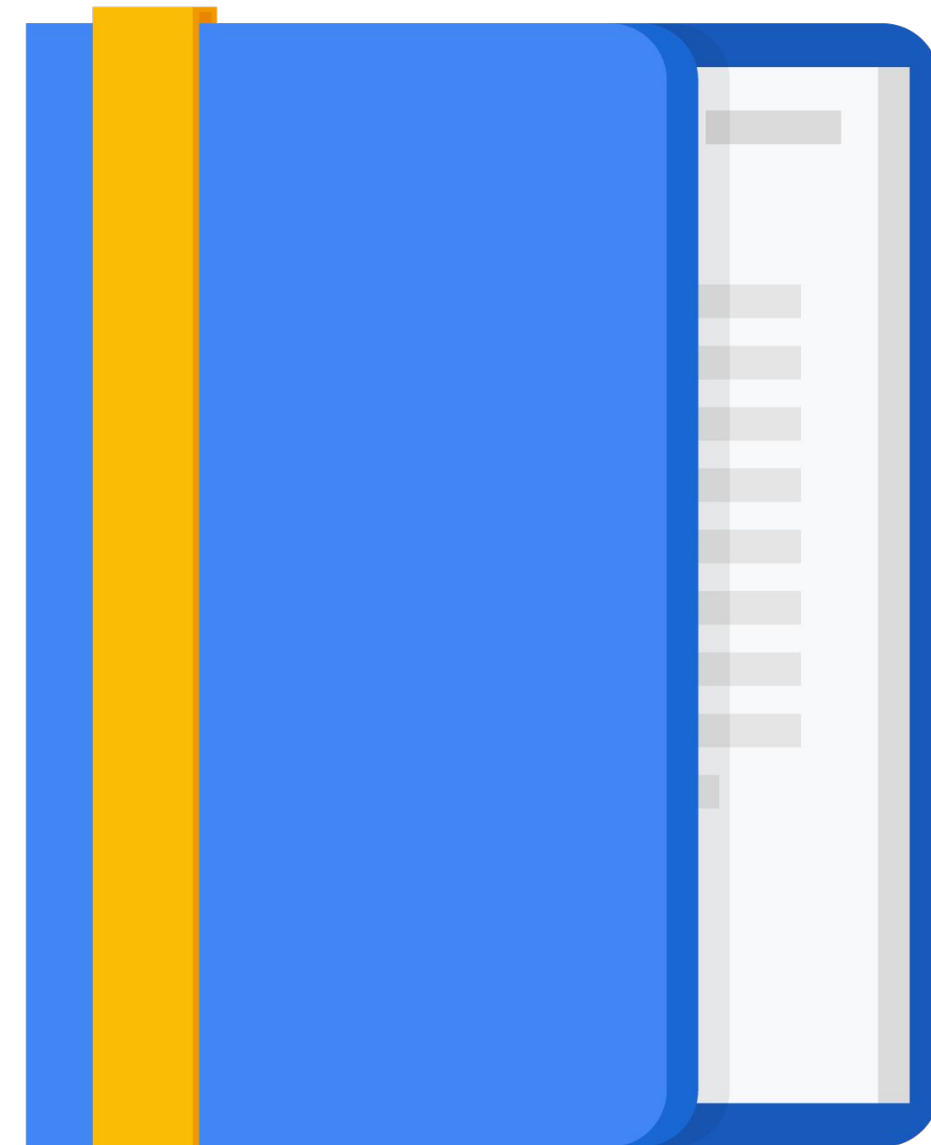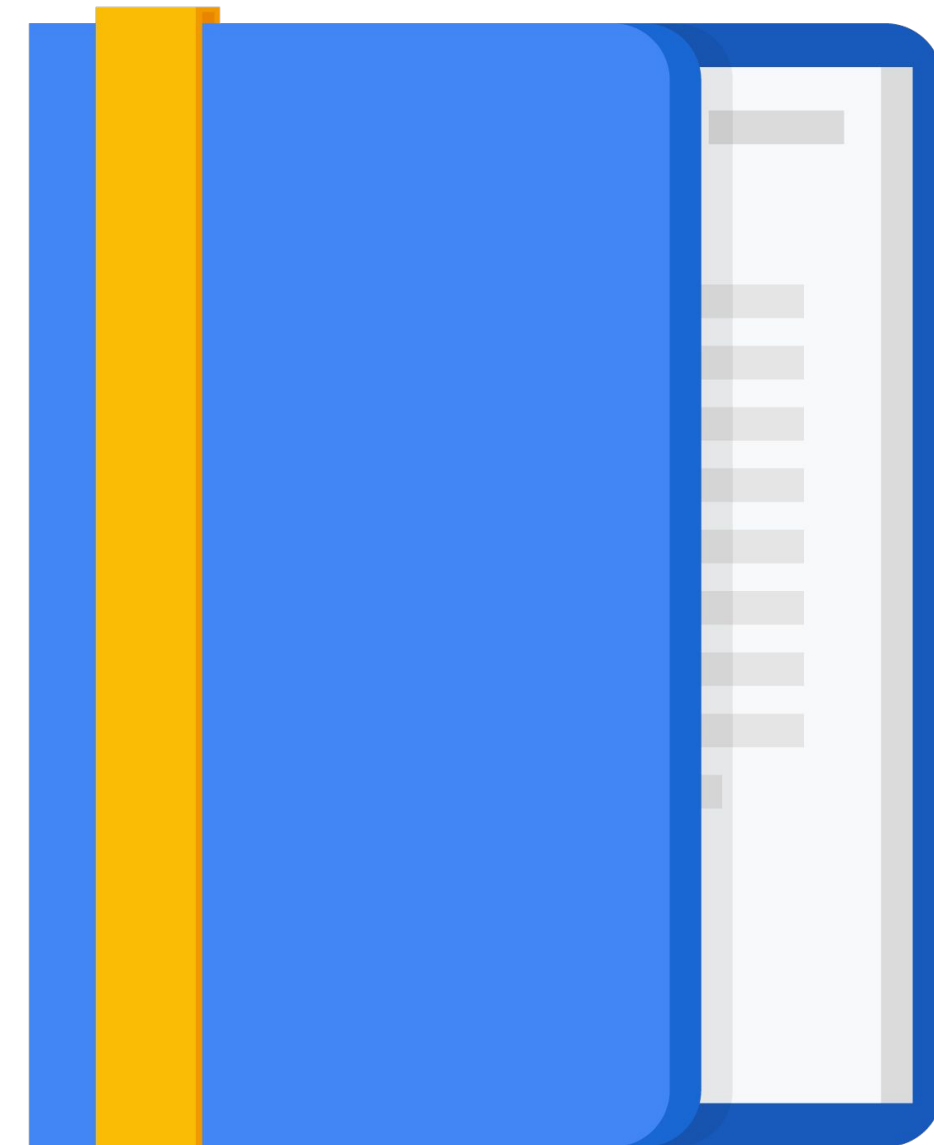
Nitin Aggarwal
Technical PM, Cloud AI

Agenda

- Overview

- Introduction to AI Platform Pipelines

- AI Platform Pipelines: Concepts

- AI Platform Pipelines: When to Use?

- AI Platform Pipelines: Ecosystem
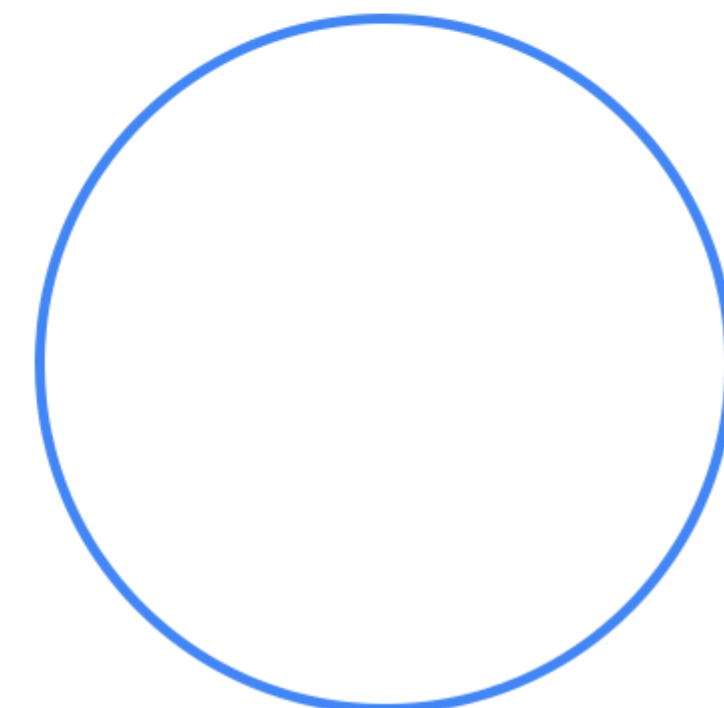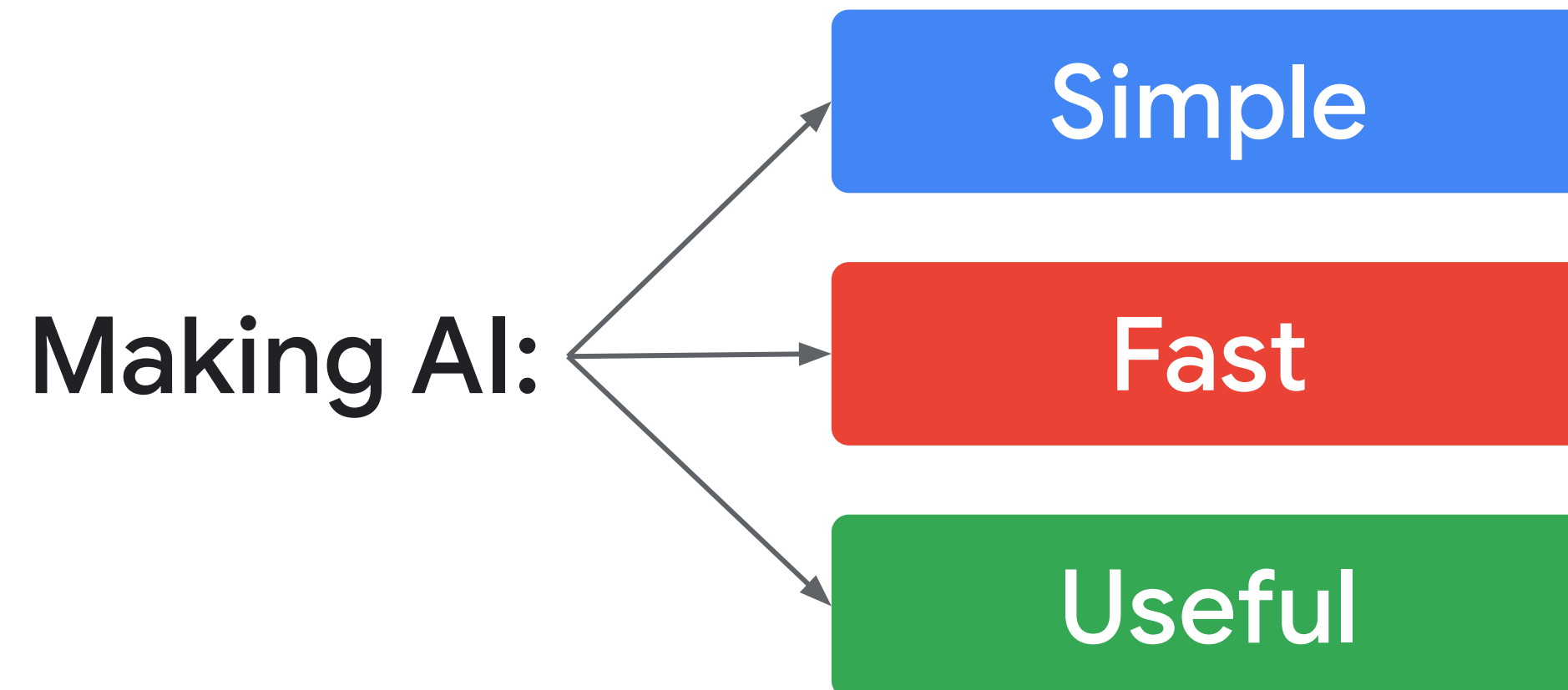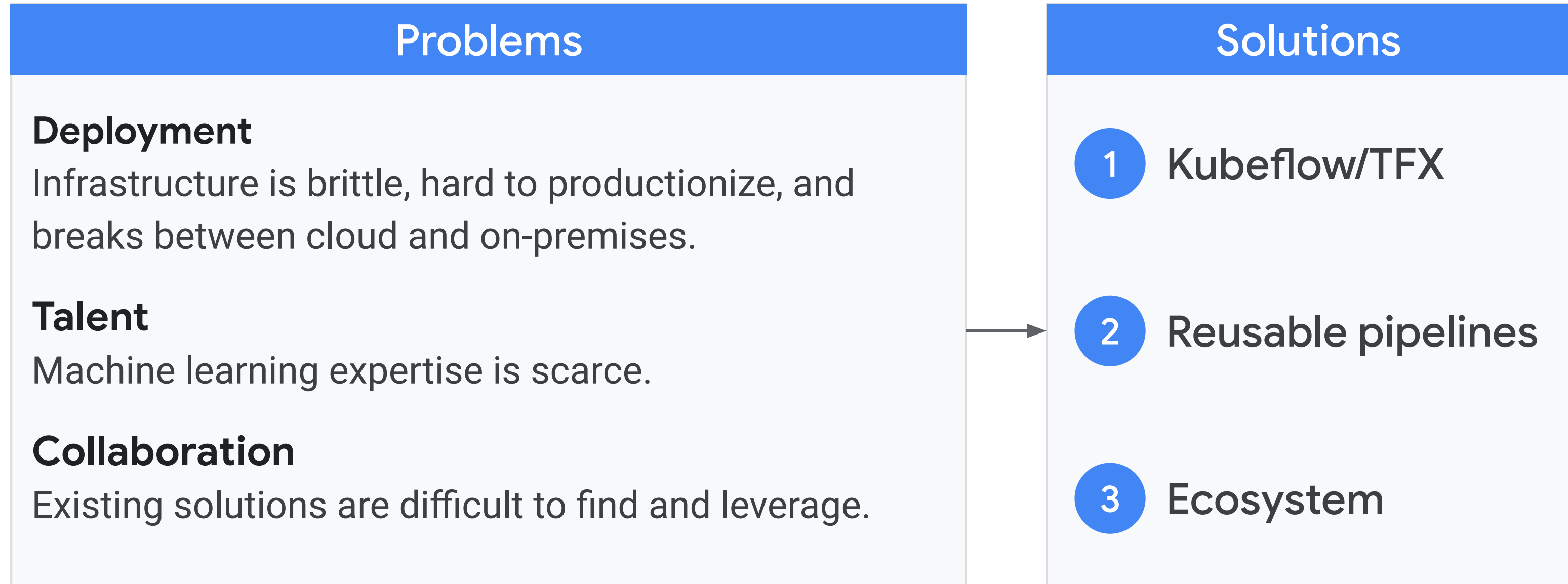
Agenda

- Overview

- Introduction to AI Platform Pipelines

- AI Platform Pipelines: Concepts

- AI Platform Pipelines: When to Use?

- AI Platform Pipelines: Ecosystem

Making AI: → Simple

→ Fast

→ Useful

Few common AI problems

Problems	Solutions
Deployment Infrastructure is brittle, hard to productionize, and breaks between cloud and on-premises. **Talent** Machine learning expertise is scarce. **Collaboration** Existing solutions are difficult to find and leverage.	1 Kubeflow/TFX 2 Reusable pipelines 3 Ecosystem

1 Kubeflow/TFX scalable ML services on Kubernetes

Easy to get started
- Kubeflow: Out-of-box support for top frameworks
  - pytorch, caffe, tf, and xgboost
- TFX: Google best practices on TF
- Kubernetes manages dependencies and resources

Swappable and scalable
- Library of ML services
- GPU support
- Massive scale

Meet customers where they are
- Google Cloud
- On-premises

ML microservices

Cloud

On-premises

Training   Predict

Training   Predict

Kubernetes
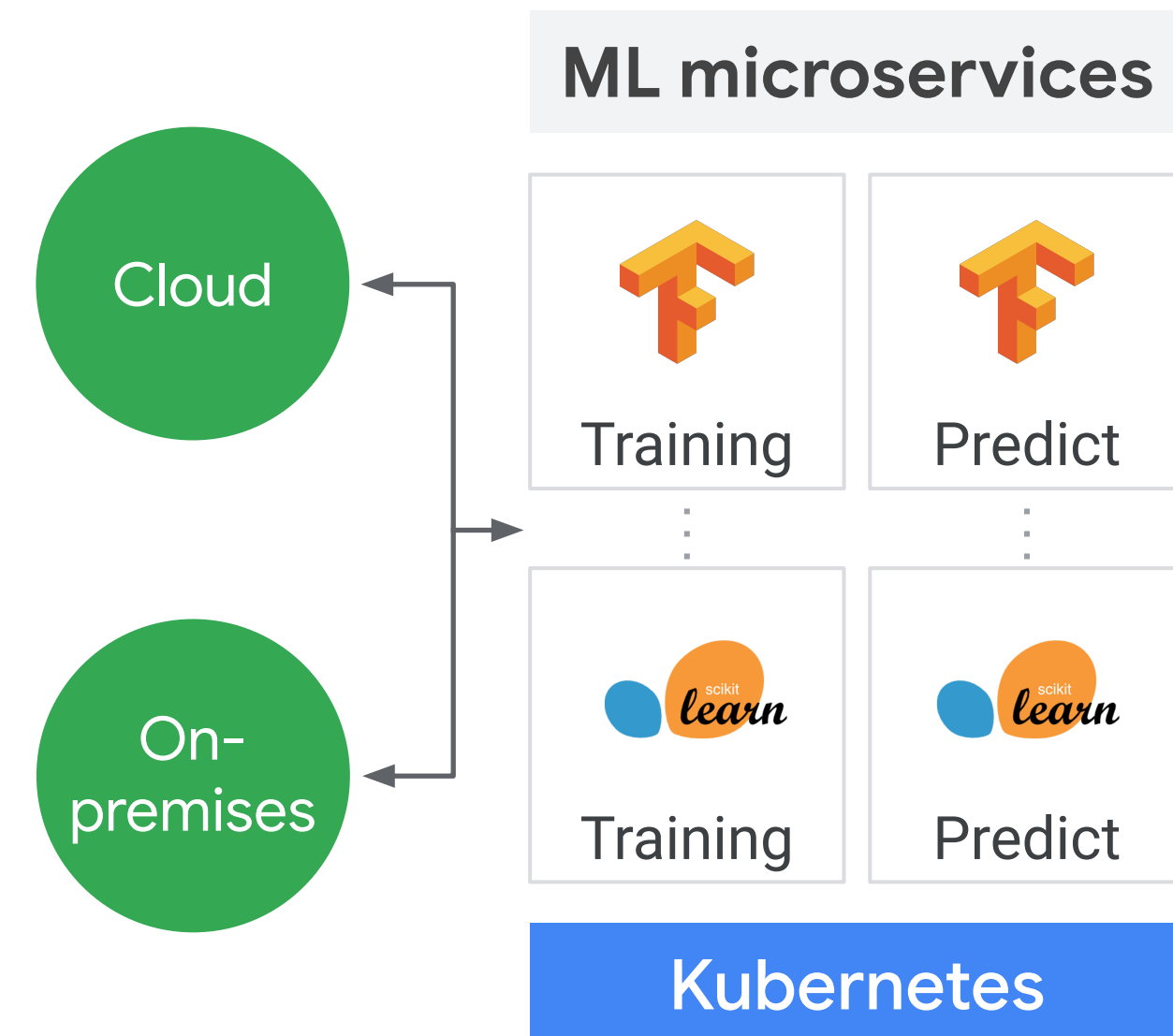
2 Reusable pipelines

Enable developers to build custom ML applications by easily "stitching" and connecting various components, like LEGO.

- Reuse instead of reimplement or reinvent.

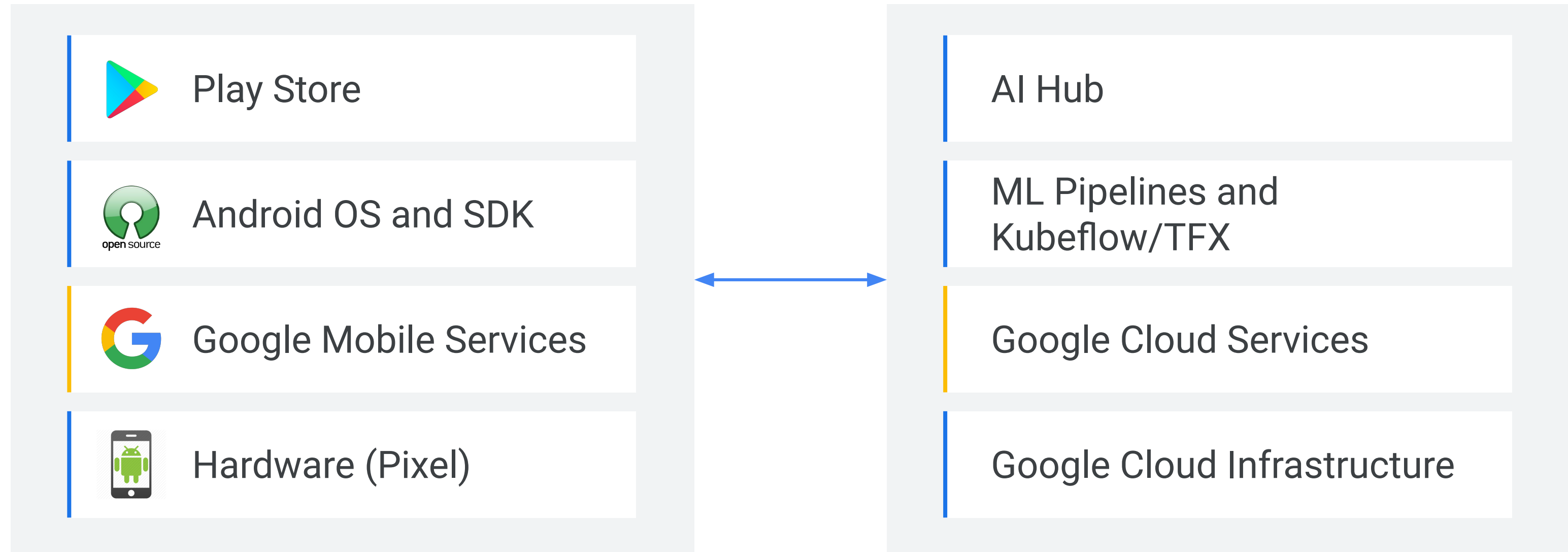- Discover, learn, and replicate successful pipelines.
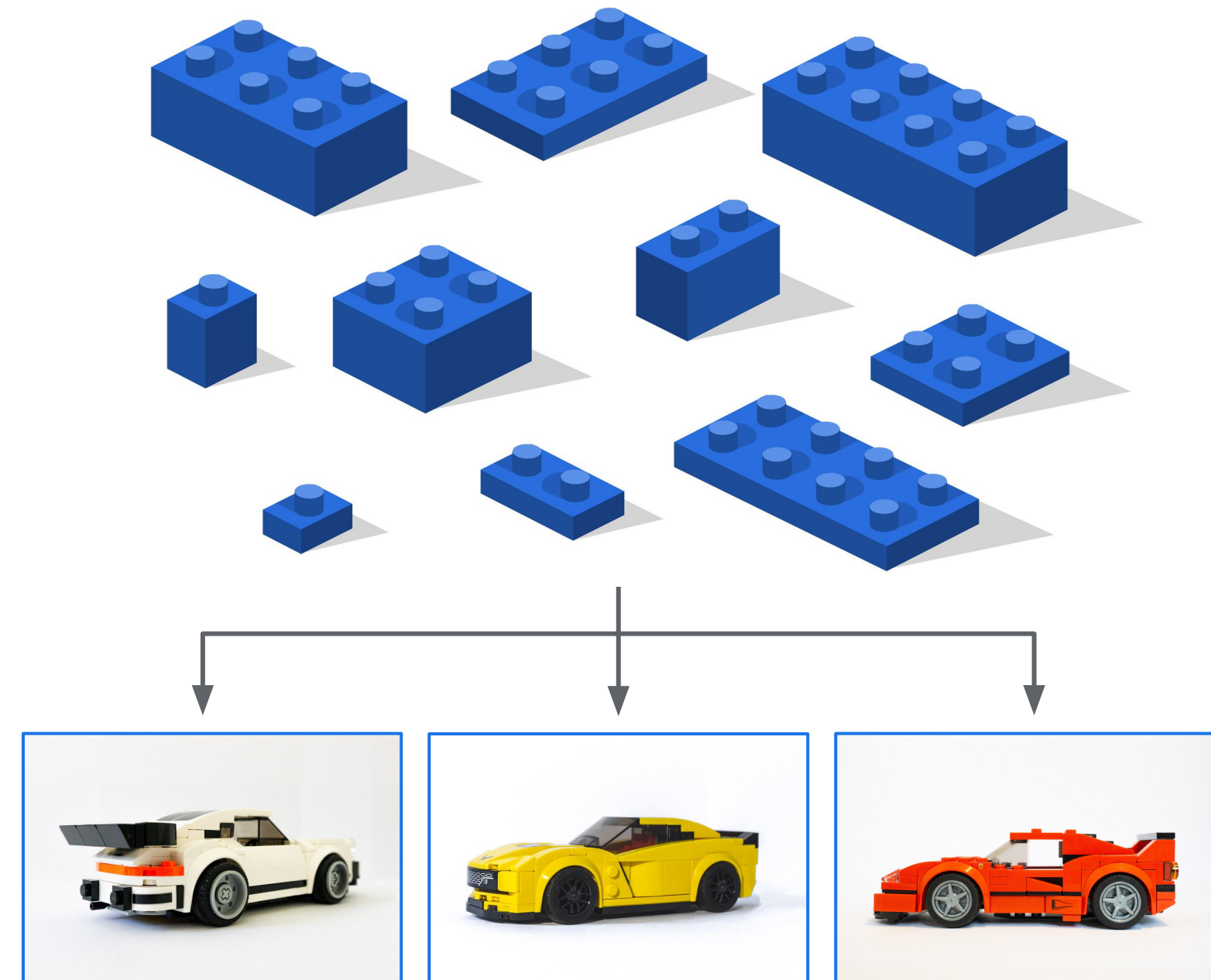
3 Ecosystem

AI Hub at a glance

1 All AI content in one place
- Quick discovery of plug and play AI pipelines and other content build by teams across Google and by partners and customers

2 Fast and simple implementation of AI on Google Cloud
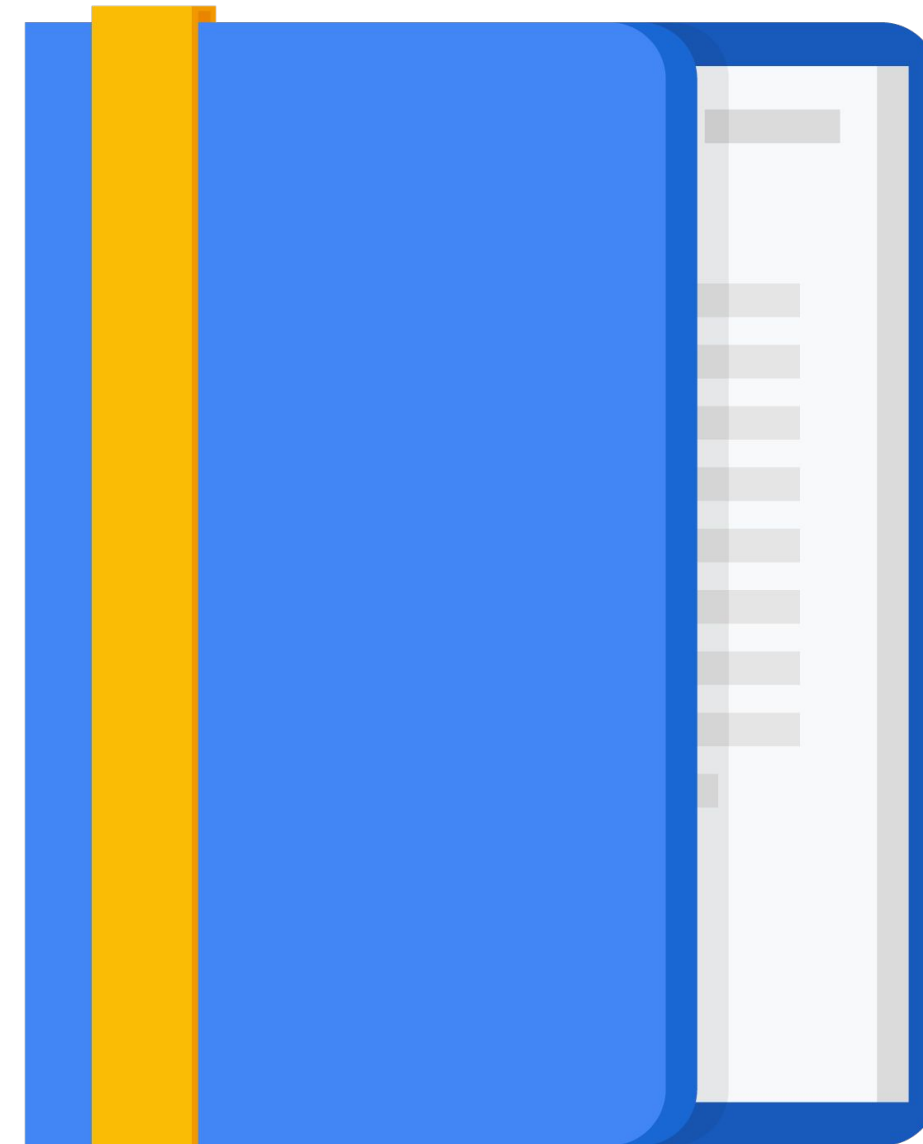- One-click deployment of AI pipelines via Kubeflow on Google Cloud
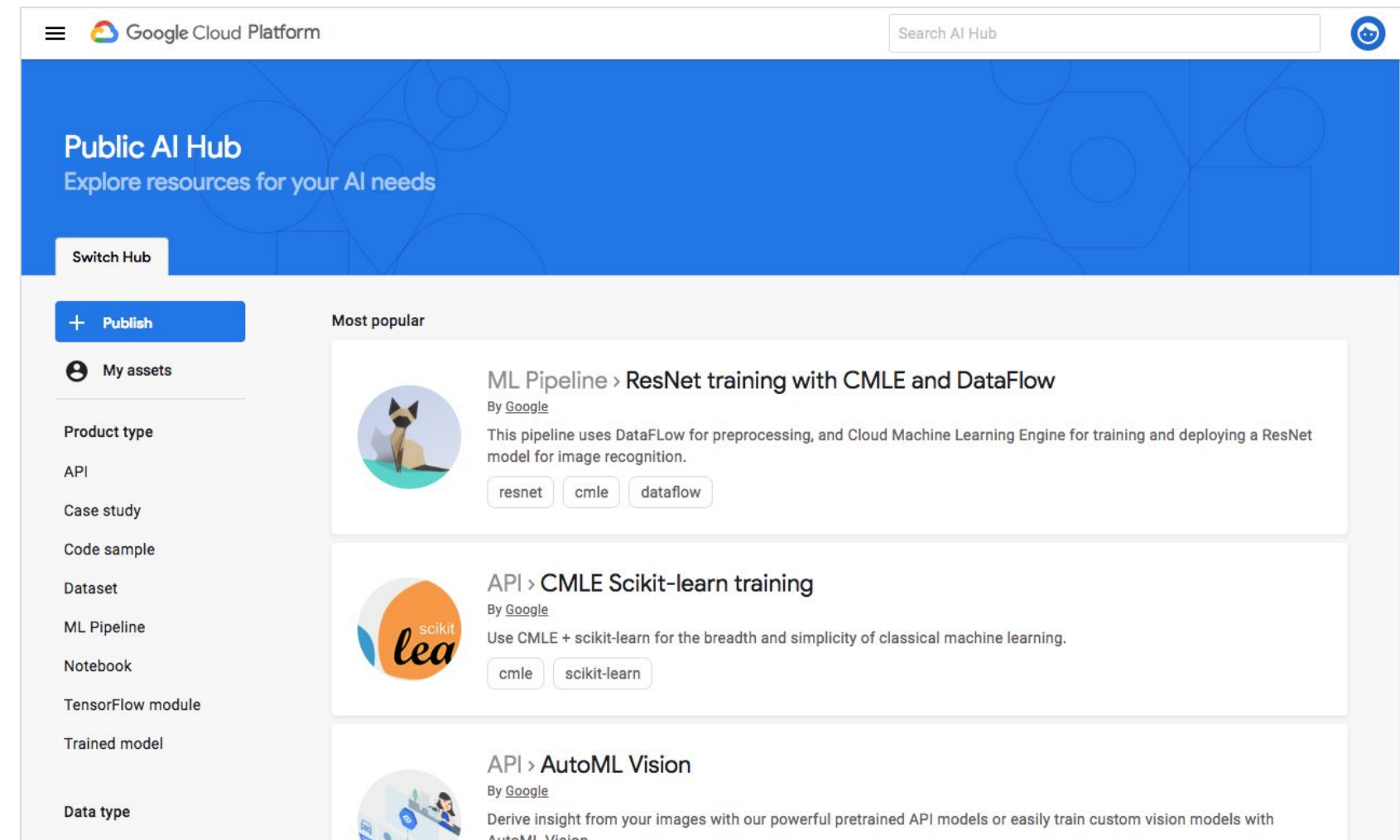
3 Enterprise-grade internal & external sharing
- Foster reuse by sharing deployable AI pipelines and other content privately within organizations and publicly

Android ecosystem analogy

Play Store	AI Hub
Android OS and SDK	ML Pipelines and Kubeflow/TFX
Google Mobile Services	Google Cloud Services
Hardware (Pixel)	Google Cloud Infrastructure

Agenda

- Overview

- Introduction to AI Platform Pipelines

- AI Platform Pipelines: Concepts

- AI Platform Pipelines: When to Use?

- AI Platform Pipelines: Ecosystem

What is an ML pipeline?

ML pipelines are portable, scalable ML workflows based on containers.

You can use ML pipelines to:

- Apply MLOps strategies to automate repeatable processes.

- Experiment by running an ML workflow with different sets of hyperparameters, number of training steps or iterations, etc.

- Reuse a pipeline's workflow to train a new model.

Understanding pipeline components

- Self-contained sets of code to perform various operations

- Composed of set of input/output parameters and location of container image

- Container image: Package that includes the component's executable code and a definition of the environment that the code runs in

Example pipeline component

Container image location

gcr.io/example-components/...

Inputs

train_data=gs:examples/train.csv,...
eval_data=gs:examples/train.csv,...

Outputs

output=output.txt
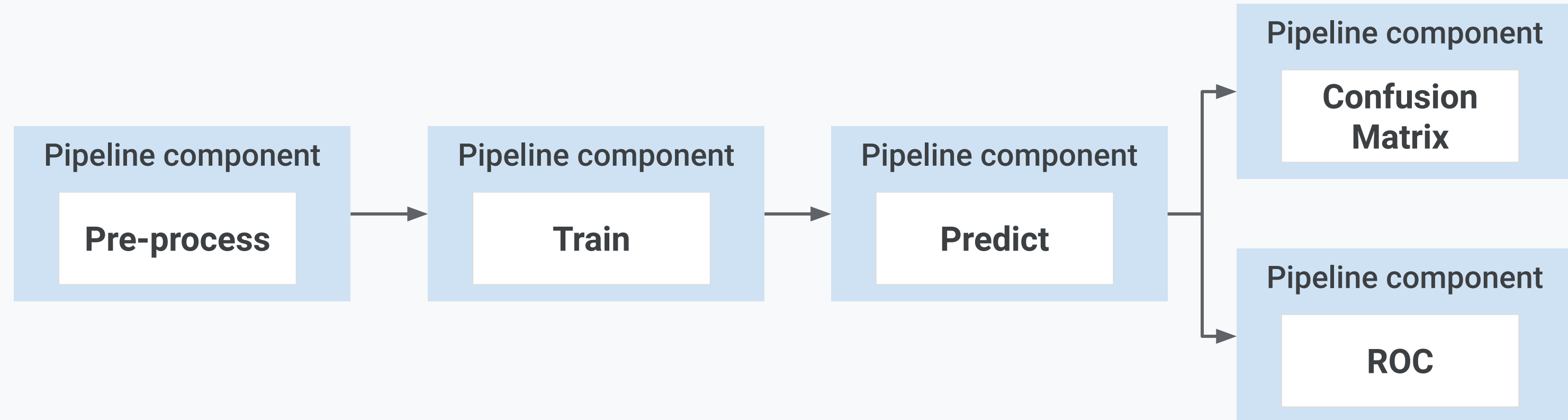
Container image registry

Component container image

Understanding a pipeline workflow

Example pipeline

Understanding a pipeline workflow

Example pipeline

Pipeline component

Pre-process

Pipeline component

Train

Pipeline component

Predict

Pipeline component

Confusion Matrix

Pipeline component

ROC

Understanding a pipeline workflow

Understanding a pipeline workflow

Example pipeline

Understanding a pipeline workflow

Example pipeline

Pipeline component
Pre-process

Pipeline component
Train

Pipeline component
Predict

Pipeline component
Confusion Matrix

Pipeline component
ROC

Understanding a pipeline workflow

Understanding a pipeline workflow

Example pipeline

Pipeline component

Pre-process

→

Pipeline component

Train

→

Pipeline component

Predict

Pipeline component

Confusion Matrix

Pipeline component

ROC
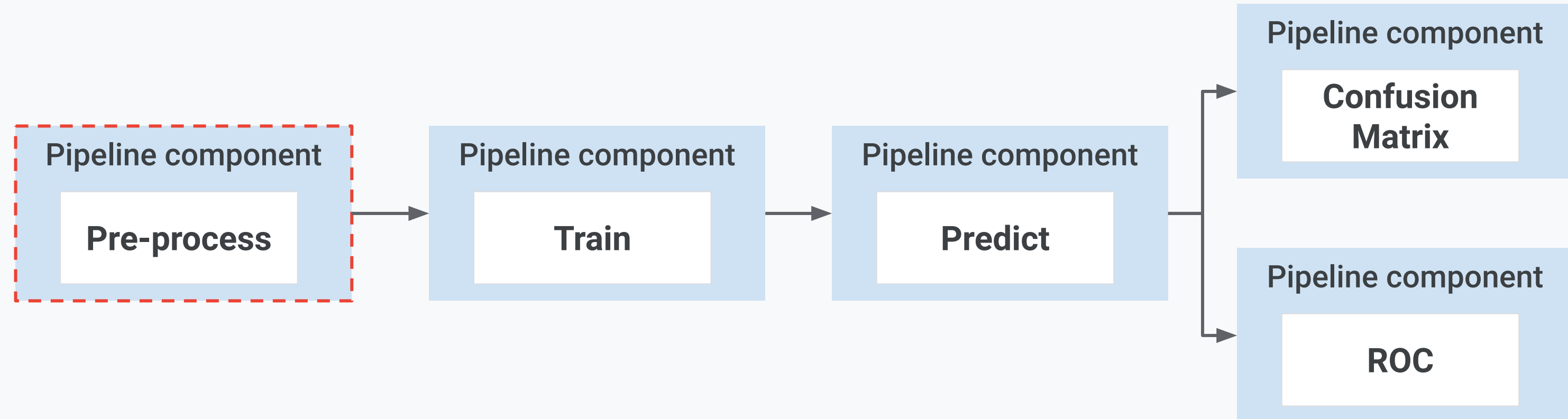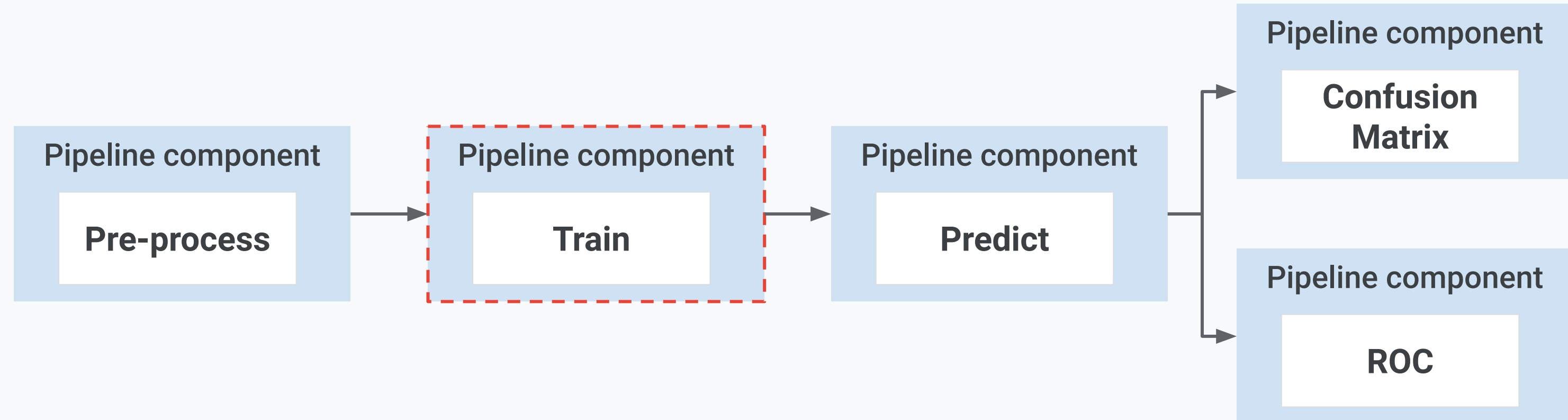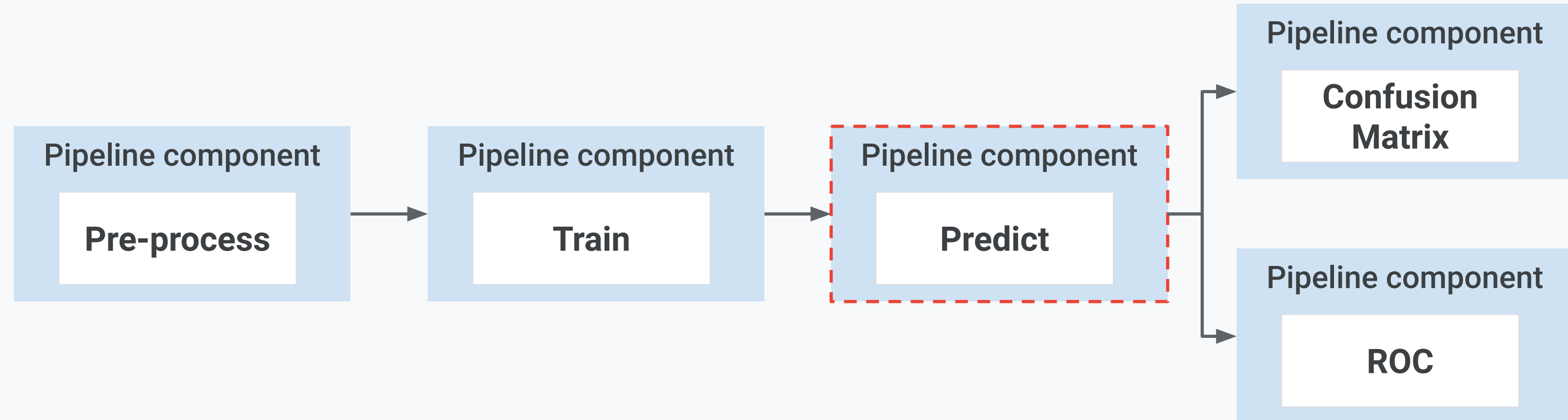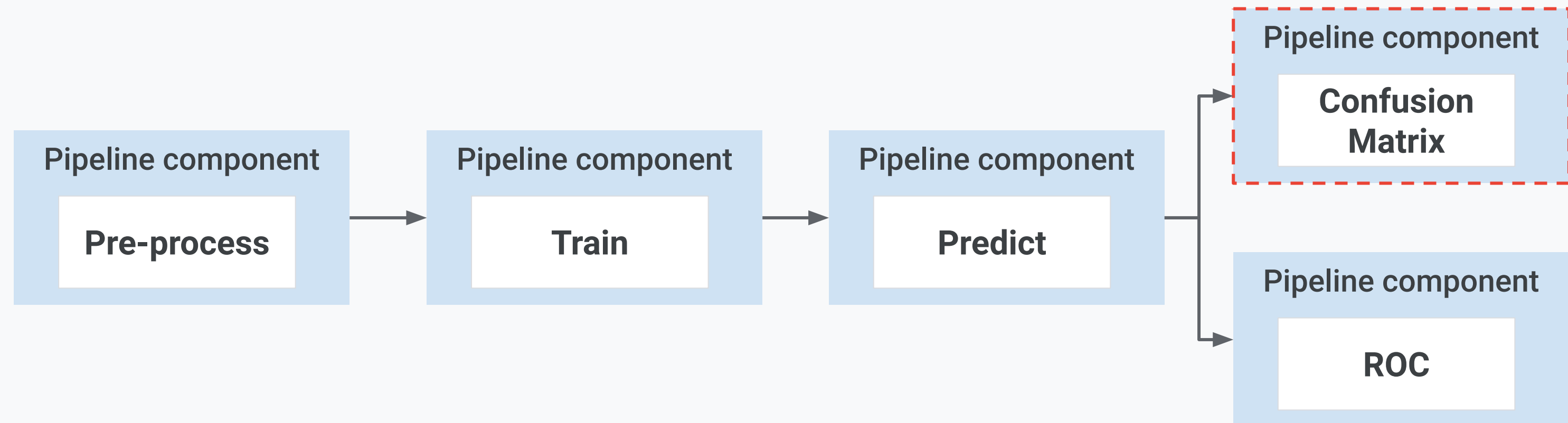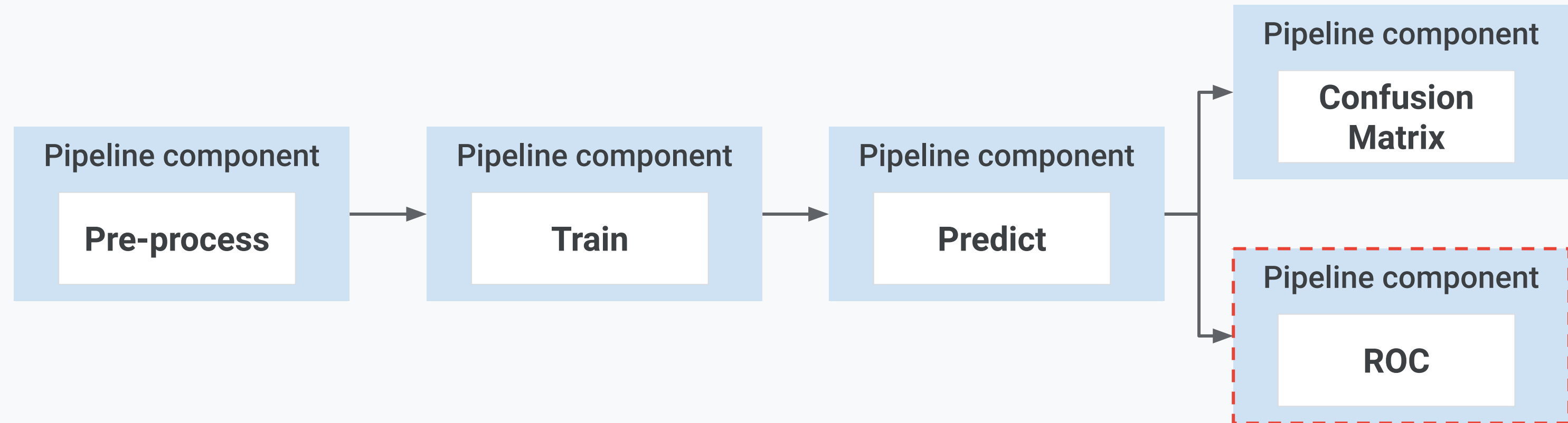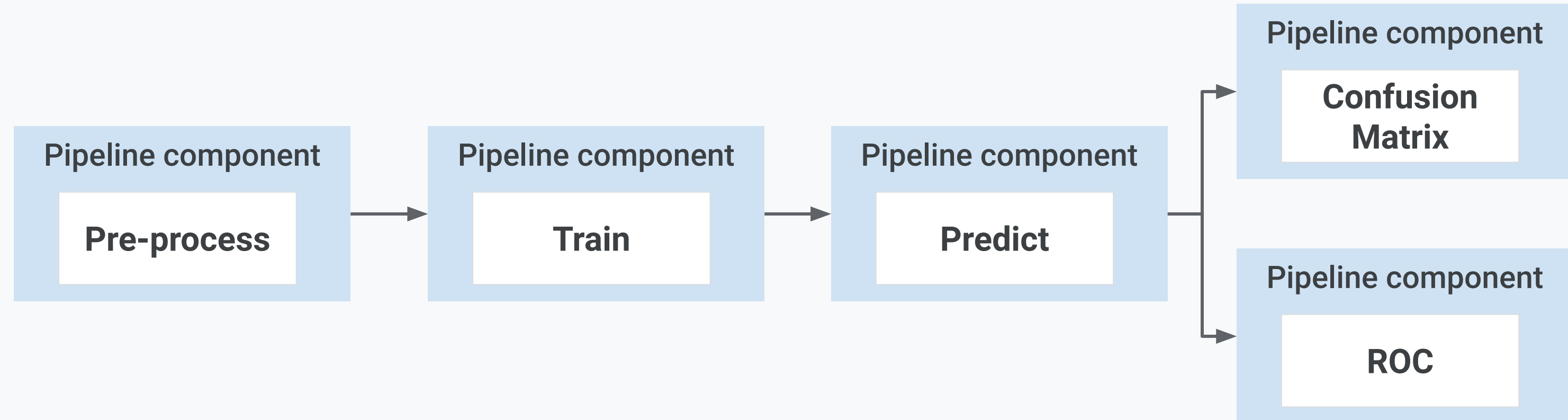
Current process for building an MLOps pipeline

- Set up a Google Kubernetes Engine (GKE) cluster.

- Create a Cloud Storage bucket for storing data.

- Install Kubeflow pipelines.

- Set up port forwarding.

- Create a process to share the pipeline with your team.

Can we automate these processes to make MLOps a seamless and easy experience?
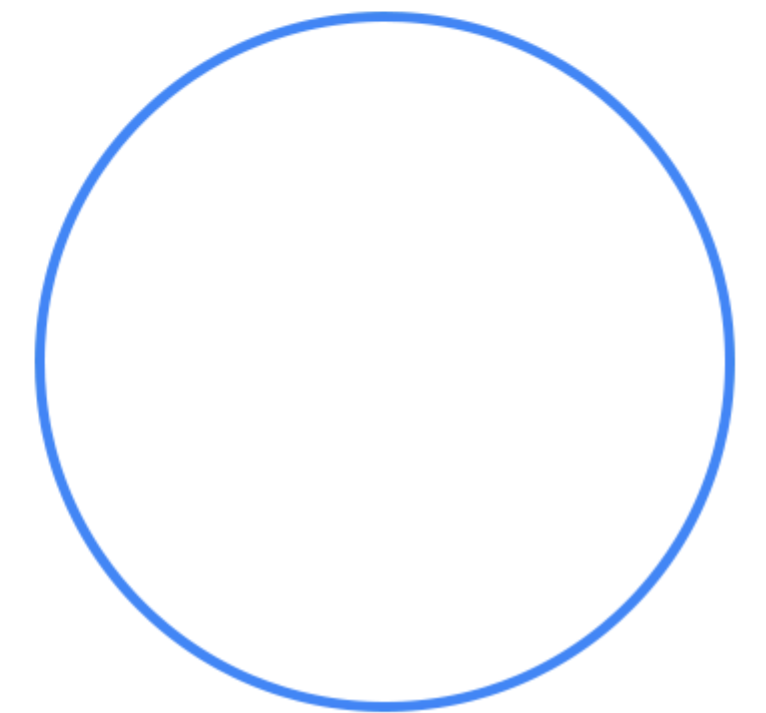
A new product was needed to deploy robust, repeatable machine learning pipelines along with monitoring, auditing, version tracking, and reproducibility and deliver an enterprise-ready, easy-to-install, secure execution environment for your ML workflows.
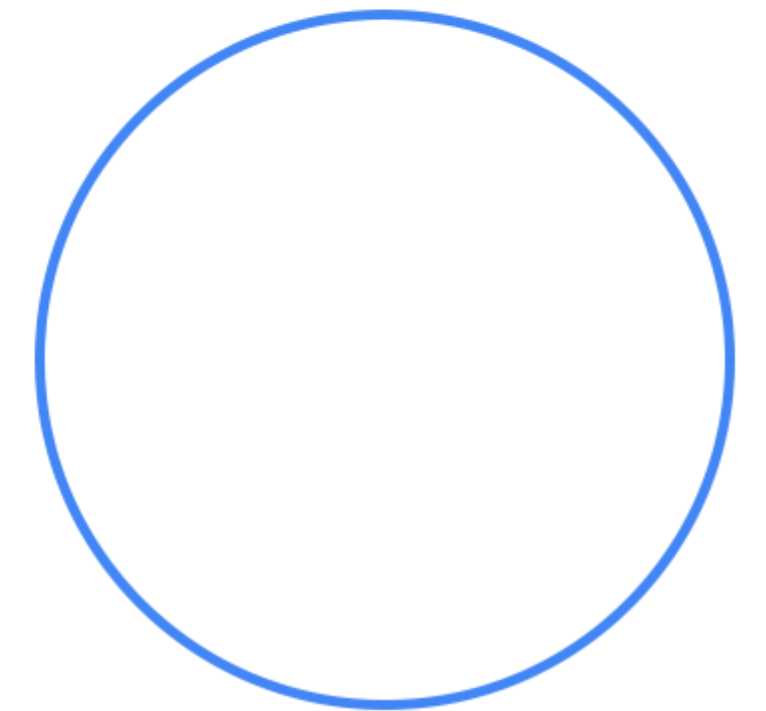
What is AI Platform Pipelines?

- One-click installation via the Google Cloud Console

- Enterprise features for running MLOps workloads

- Seamless integration with Google Cloud managed services

- Prebuilt pipeline *components* (pipeline steps) for ML workflows

- Easy customization for new components

Two major parts of an AI Platform Pipelines instance

Enterprise-ready infrastructure

For deploying and running ML workflows with tight integration with Google Cloud services

Pipeline ecosystem

For building, debugging, and sharing the pipeline and components
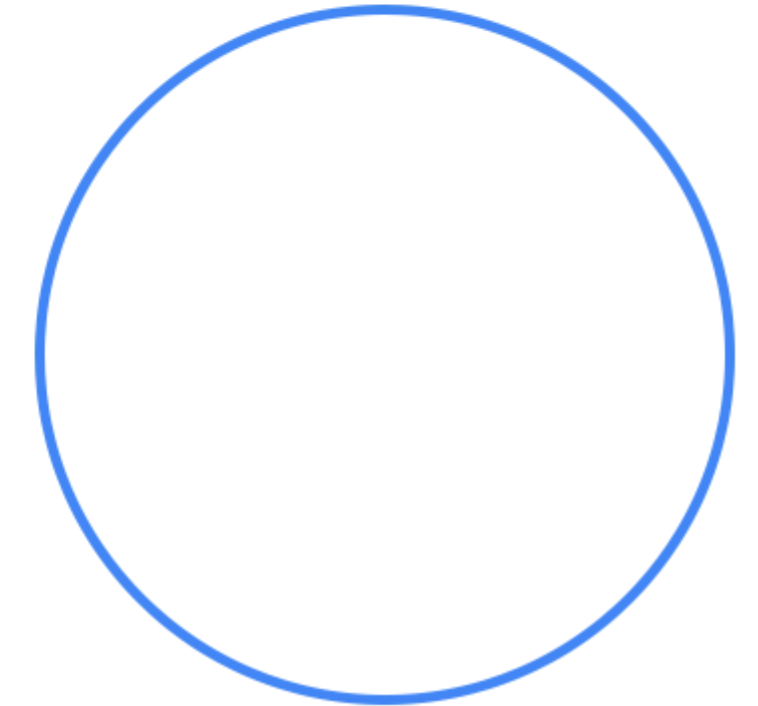
Key benefits of using AI Platform Pipelines

1. Easy installation

   - Click one button on the Google Cloud Console

   - Automated creation of Google Kubernetes Engine (GKE) cluster

   - Customization of existing GKE cluster for deployment

   - Easy to delete and reinstall to retain persisted previous state while updating the pipeline version

| | AI Platform Pipelines BETA | | | | | + New Instance | Refresh | Delete |

⇝ Filter						
Status	**Name** ↑	**Zone**	**Version**	**Cluster**	**Namespace**	
☐ ✅	Kuberflow-pipeline-1　OPEN PIPELINES DASHBOARD	us-central1-a	0.2.4	cluster-5	default	⚙ SETTINGS
☐ ✅	Kuberflow-pipeline-2　OPEN PIPELINES DASHBOARD	us-west1-a	0.2.5	mpdev-verify	kfp	⚙ SETTINGS

Key benefits of using AI Platform Pipelines

2. Easy authentication process
   - Fully secure and provides authenticated access to Pipelines UI
   - No need to set up port forwarding
   - Easy to share with team members
   - Easy to access through REST API service
   - Seamless performance of using Pipelines SDK from Notebooks
     - Define pipeline
   - Schedule run job
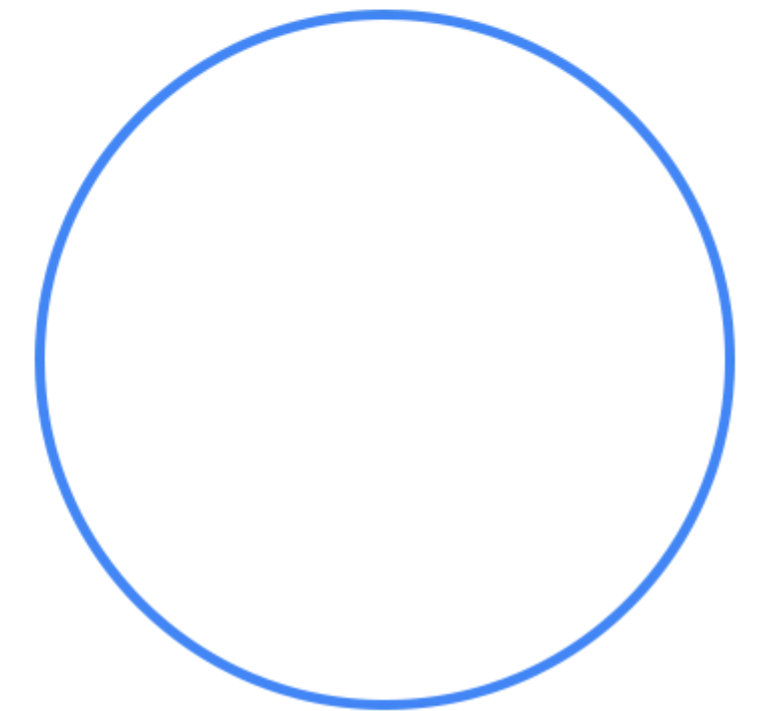
Agenda

- Overview

- Introduction to AI Platform Pipelines

- AI Platform Pipelines: Concepts

- AI Platform Pipelines: When to Use?
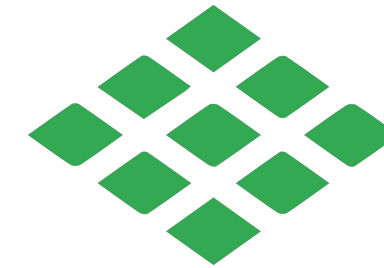
- AI Platform Pipelines: Ecosystem

AI Platform Pipelines tech stack

User model	Custom components	Prebuilt components

Notebooks

TFX Pipeline template	Custom Pipeline

TFX SDK	

Pipelines and Lineage UI

KFP SDK

GKE

Kubeflow Pipelines

Argo Workflow Engine

Pipeline steps

Metadata Store MLMD

Authorized access to Google Cloud Services

Dataflow	AI Platform Training	BigQuery
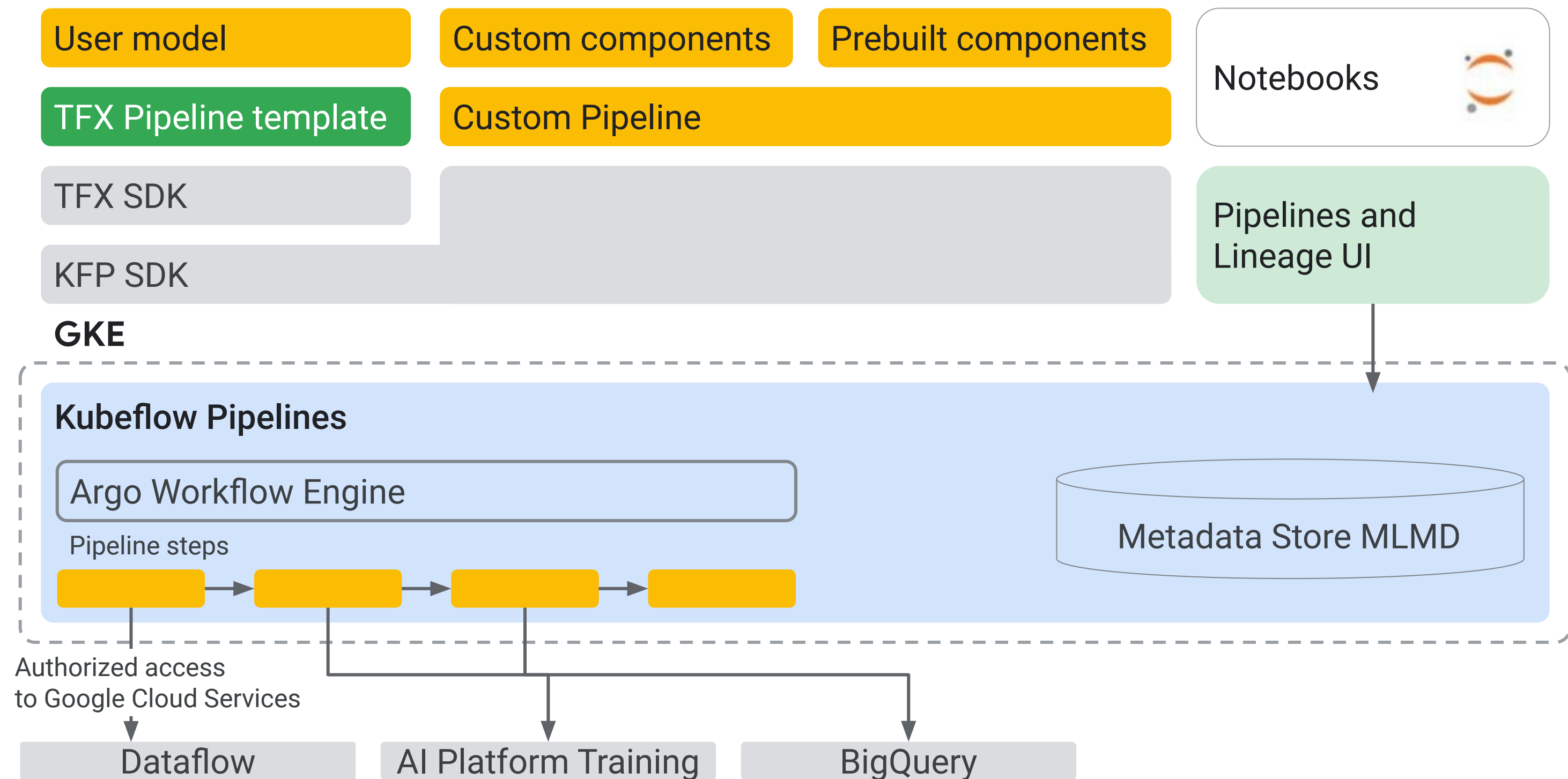
AI Platform Pipelines implementation strategy

Kubeflow Pipelines	TensorFlow Extended (TFX)
Through Kubeflow Pipelines SDK	**Through TFX SDK**

Through Kubeflow Pipelines SDK

- Lower-level ML framework−agnostic implementation
- Enables direct control of Kubernetes resource control
- Simple sharing of containerized components
- Use it for fully custom pipelines

Through TFX SDK

- Higher-level abstraction
- Prescriptive but customizable components with pre-defined ML types
- Brings Google best practices for robust/scalable ML workloads
- Use it for E2E TF-based pipeline with customizable data pre-processing and training code

Some features of AI Platform Pipelines

1. Build your own ML pipeline with TFX examples

Some features of AI Platform Pipelines

2. Pipeline versioning

Pipelines			+ Upload pipeline	Refresh	Delete

	Pipeline name	Description	Uploaded on ↓
▸	[Tutorial] DSL - Control struct...	**Source code** Shows how to use conditional execution and exit...	2/20/2020, 3:28:12 PM
▸	[Tutorial] Data passing in pyt...	**Source code** Shows how to pass data between python comp...	2/20/2020, 3:28:11 PM
▾	[Demo] TFX - Taxi Tip Predict...	**Source code GCP Permission Requirements**. Example pipeline that does classification...	2/20/2020, 3:28:10 PM

Version name	Uploaded on ↓
TFX - Taxi Tip Prediction Model Trainer_version_at_2020-03-03T15:44:30.197Z	3/3/2020, 7:55:03 AM
[Demo] TFX - Taxi Tip Prediction Model Trainer	2/20/2020, 3:28:10 PM

1 of 200

	Pipeline name	Description	Uploaded on
▸	[Demo] XGBoost - Training...	**Source code GCP Prmission requirements**. A trainer that does end-to-end distributed tra...	2/20/2020, 3:28:09 PM

Some features of AI Platform Pipelines

3. Artifact tracking

Artifacts					
Pipeline / Workspace	Name	ID	Type	URL	Created at
		1	ExternalArtifact	gs://ml-pipeline-play...	2/20/2020, 5:1...
▼ taxi_pipeline_with_parameters	examples	2	Examples	gs://ml-pipeline-play...	2/20/2020, 5:1...
	statistics	3	ExampleStatistics	gs://ml-pipeline-play...	2/20/2020, 5:1...
	schema	4	Schema	gs://ml-pipeline-play...	2/20/2020, 5:1...
	anomalies	5	ExampleAnomalies	gs://ml-pipeline-play...	2/20/2020, 5:1...
	transform_gra...	6	TransformGraph	gs://ml-pipeline-play...	2/20/2020, 5:1...
	transformed_e...	7	Examples	gs://ml-pipeline-play...	2/20/2020, 5:1...

Some features of AI Platform Pipelines

3. Artifact tracking

Artifacts

← model

Overview   Lineage Explorer

Type: Model

URI

gs://aju-pipelines/tfx_taxi_simple/85265540-6a06-4969-a49e-1f65741878be/Trainer/model/7

Properties

Custom Properties

name	pipeline_name	producer_component	state
model	taxi_pipeline_with_parameters	Trainer	published

Some features of AI Platform Pipelines

4. Lineage tracking
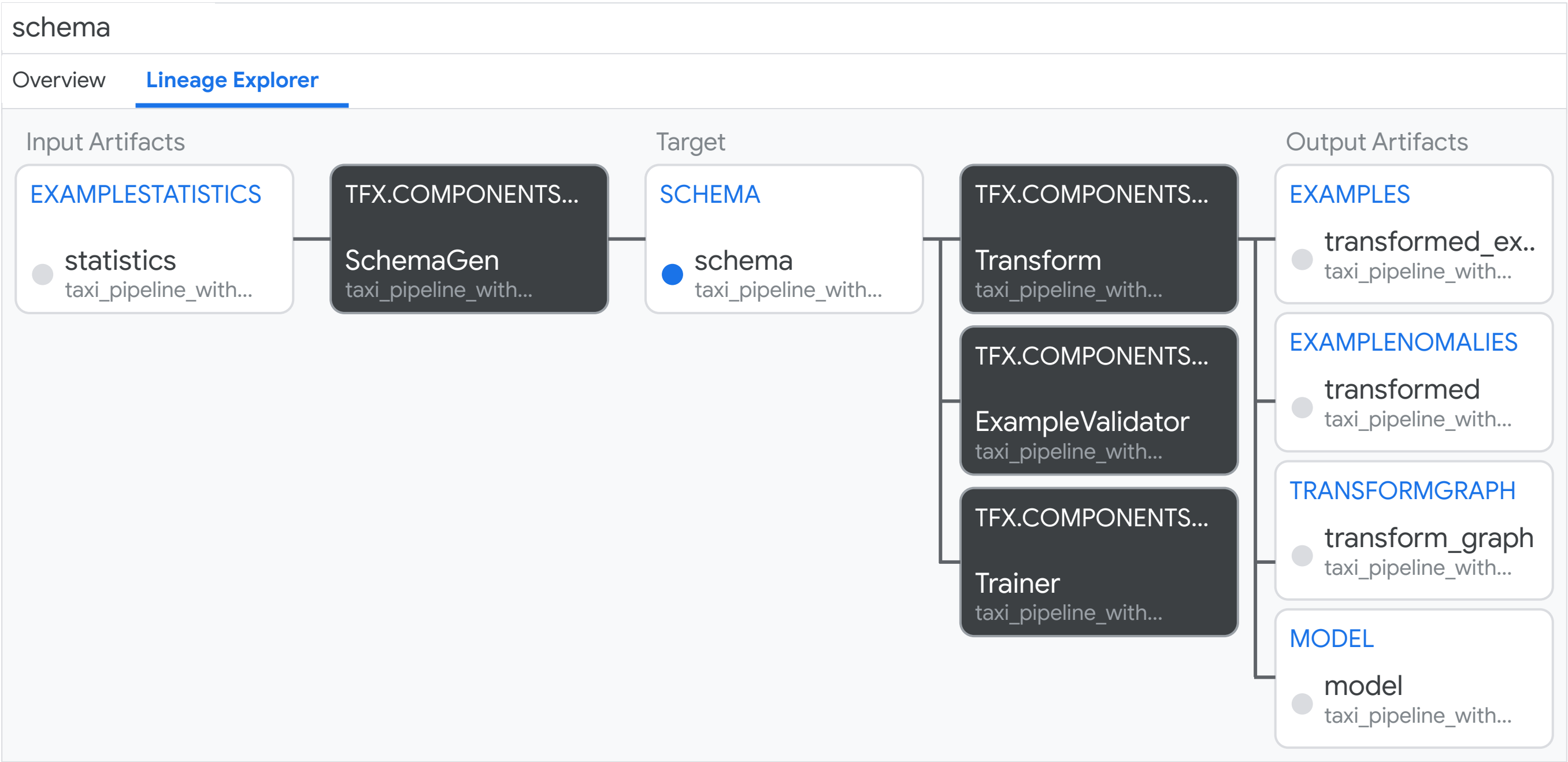
Agenda

- Overview

- Introduction to AI Platform Pipelines

- AI Platform Pipelines: Concepts

- AI Platform Pipelines: When to Use?

- AI Platform Pipelines: Ecosystem

What does AI Platform Pipelines enable?
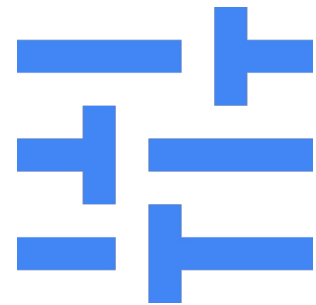
Workflow orchestration

Rapid, reliable, repeatable experimentation

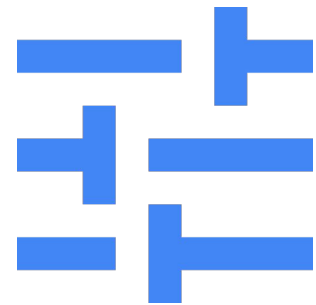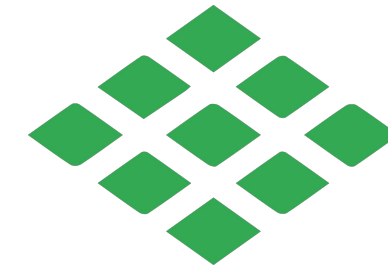Share, re-use, and compose

What does AI Platform Pipelines enable?
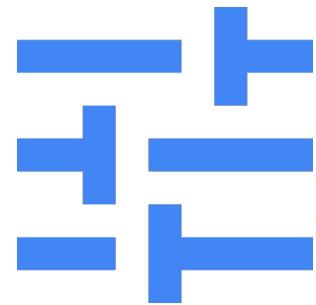
Workflow orchestration

Rapid, reliable, repeatable experimentation
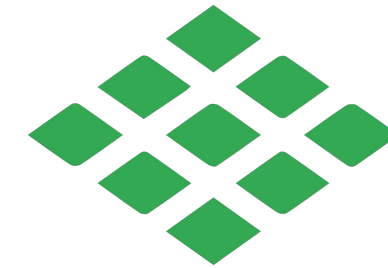
Share, re-use, and compose

# Visual depiction of pipeline topology
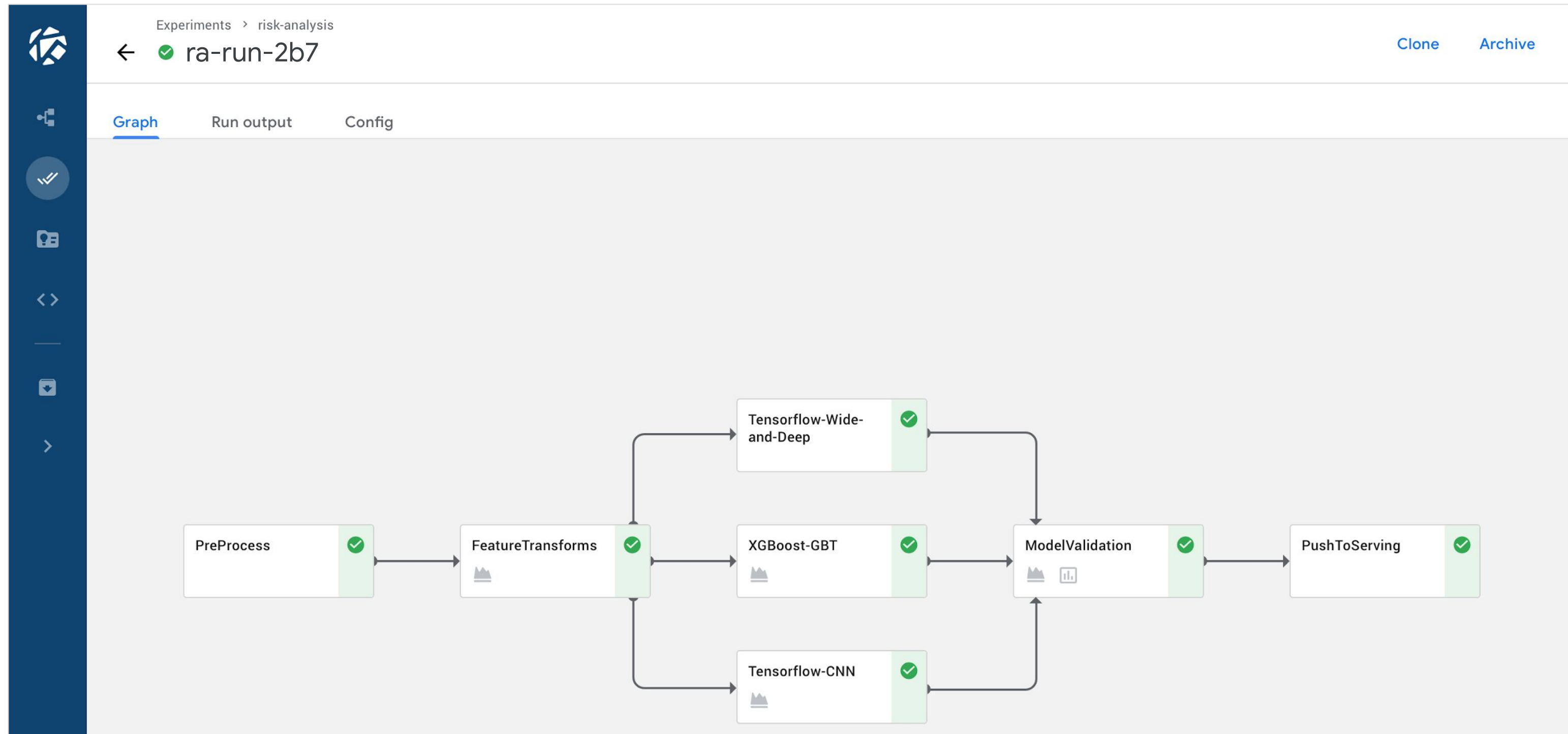
Rich visualizations of metrics
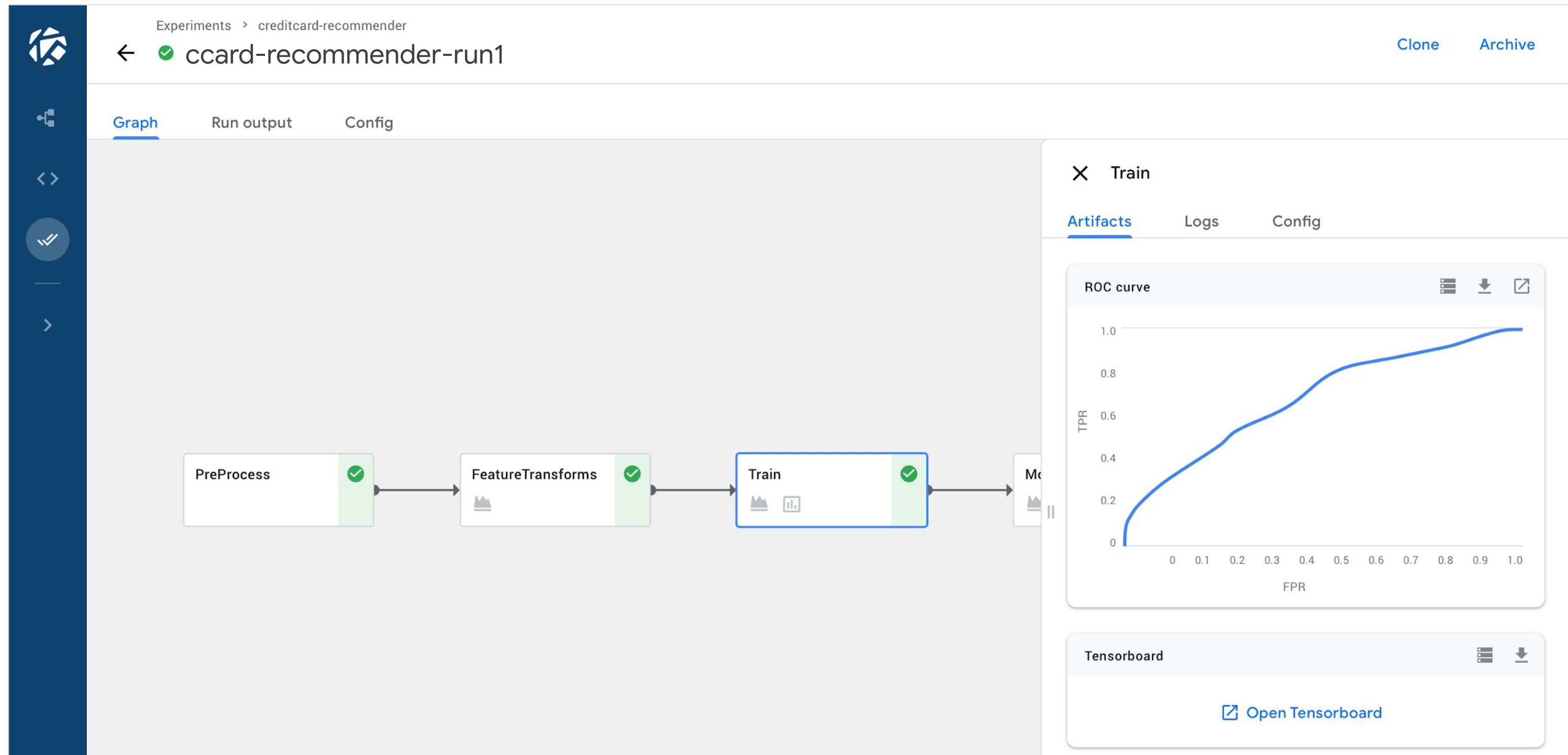
Access to all config params, inputs, and outputs for each run

What constitutes an AI Platform Pipelines instance?

- Containerized implementations of ML tasks
  - Containers provide portability, repeatability, and encapsulation.
  - A task can be single node or *distributed*.
  - A containerized task can invoke other services, like CAIP, Dataflow, or Dataproc.
- Specification of the sequence of steps
  - Specified via Python SDK
- Input parameters
  - A "job" = A pipeline invoked with specific parameters
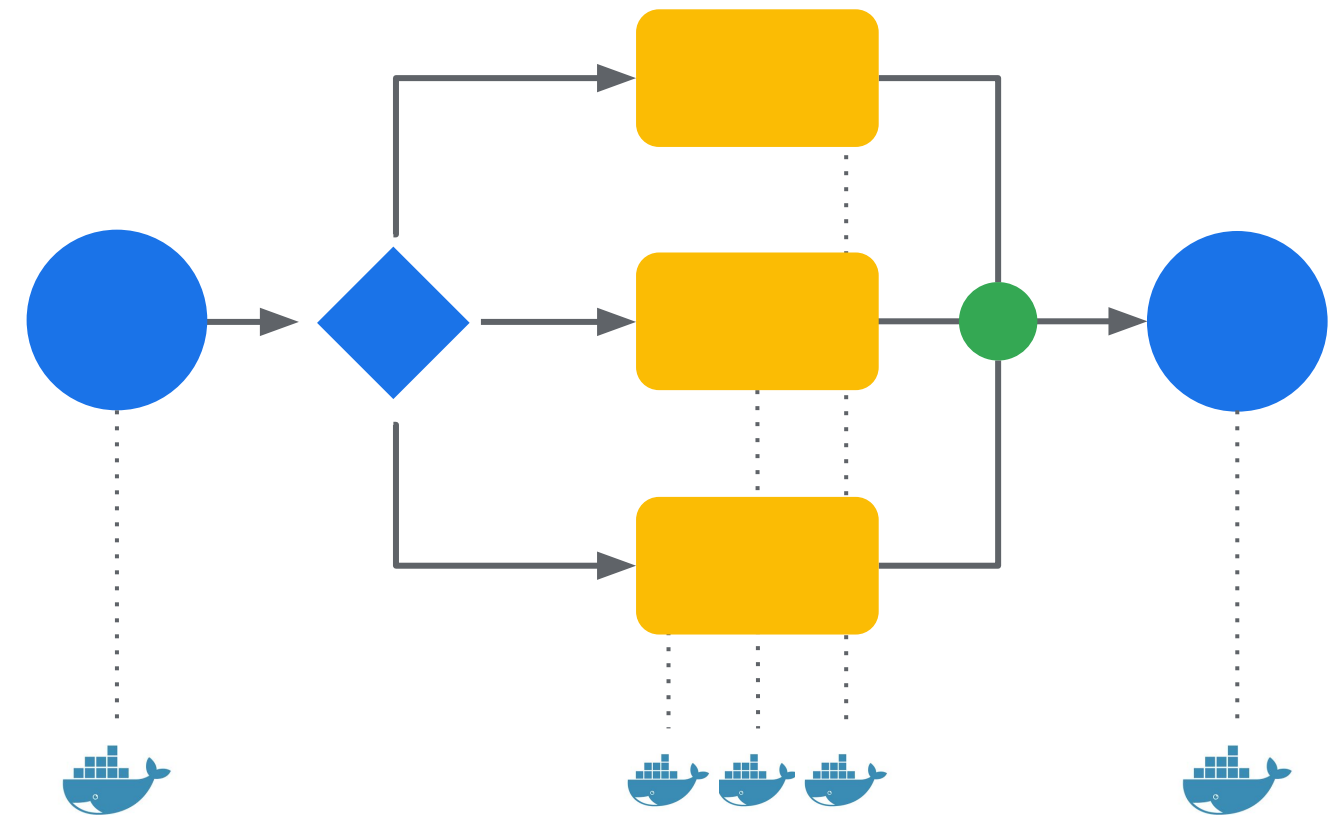
TFX + Kubeflow Pipelines

Robust extensible platform and
best-in-class components for
typical stages of good ML workflow
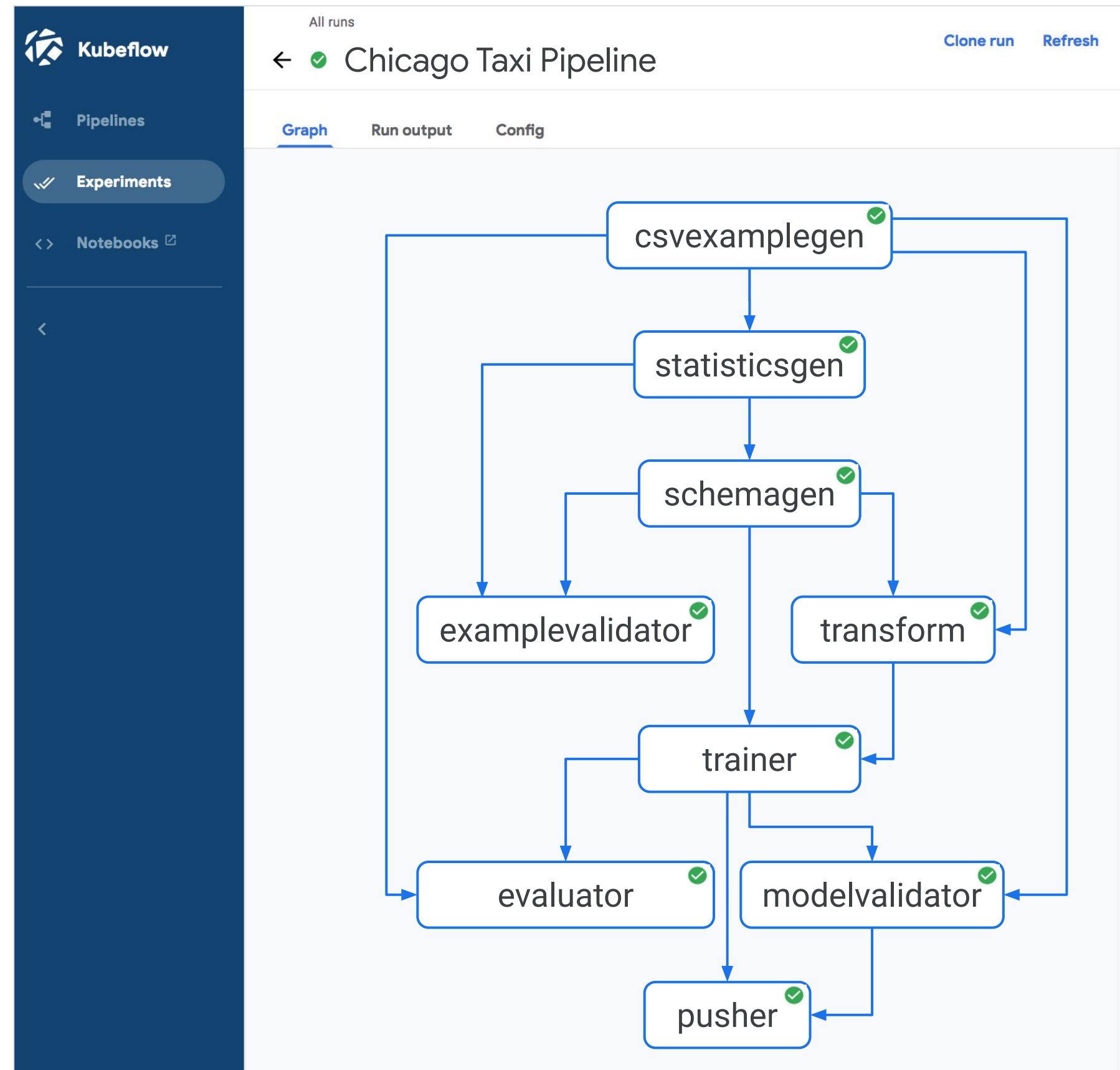
What does AI Platform Pipelines enable?
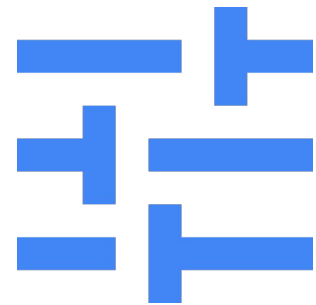
Workflow orchestration

Rapid, reliable, repeatable experimentation
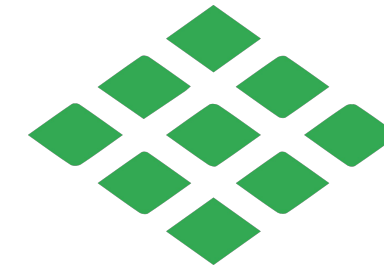
Share, re-use, and compose

Rapid, reliable experimentation

Clone

Edit

Rapidly iterate
on ideas

Submit
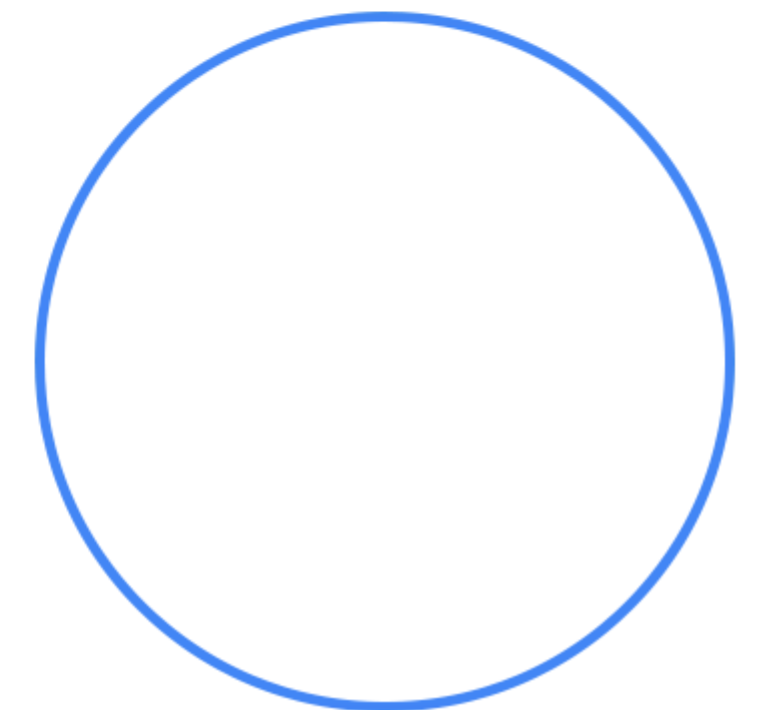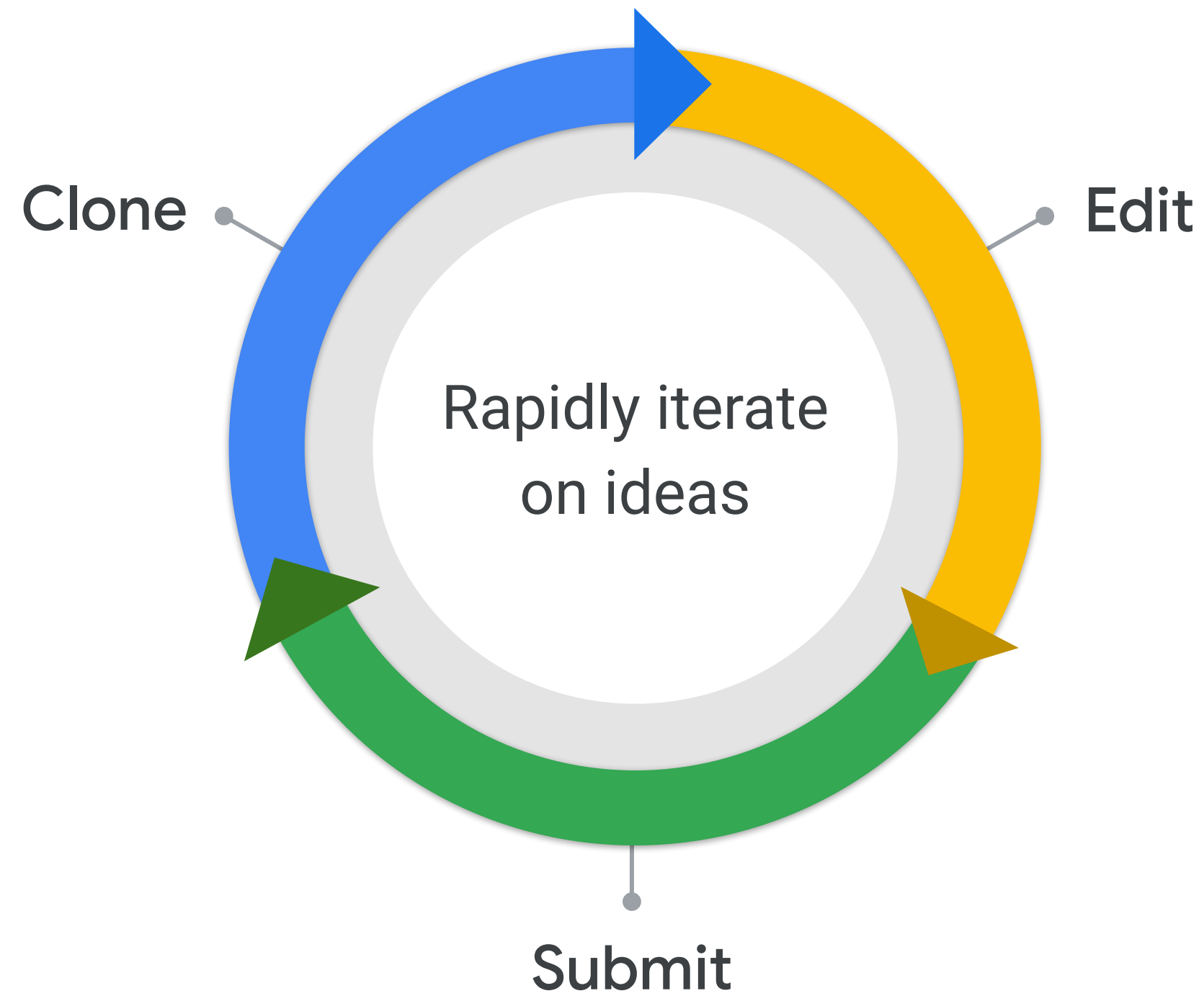
View all current and historical runs grouped as "Experiments"

Select any Run to see all params and metrics
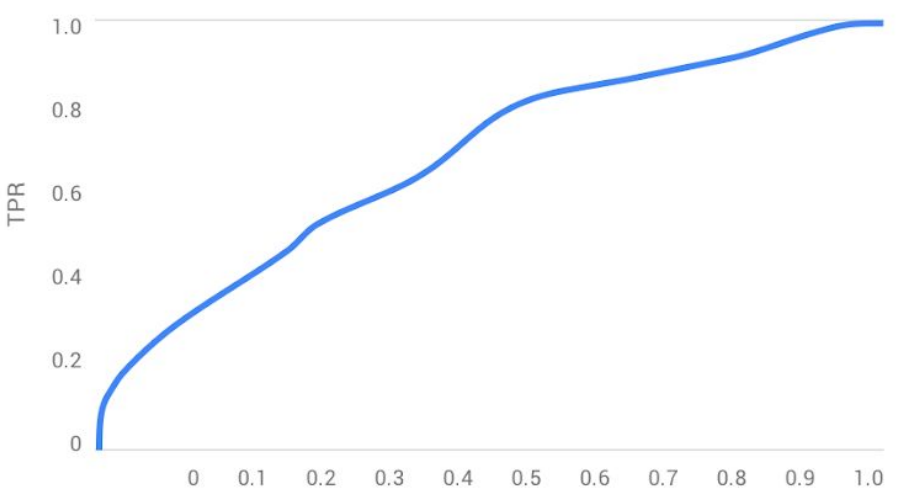
Clone an existing pipeline
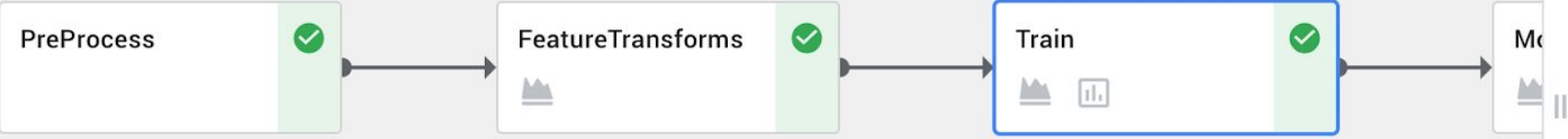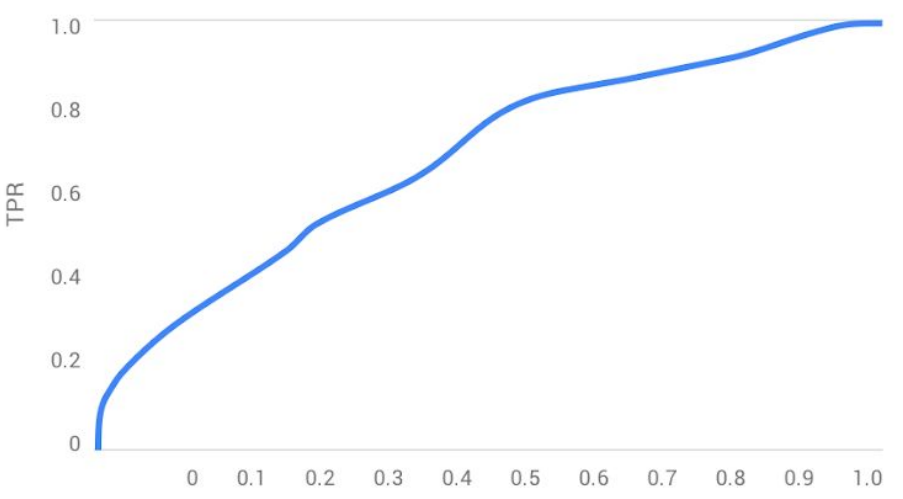
What does AI Platform Pipelines enable?

Workflow orchestration

Rapid, reliable, repeatable experimentation

Share, re-use, and compose

Share, re-use, and compose

- Re-use instead of re-implement
- Easy to compose or swap
  - Standard artifact types
  - Standard interfaces
  - Conventions and best practices

A lot of reusable components **+** Composable pipelines
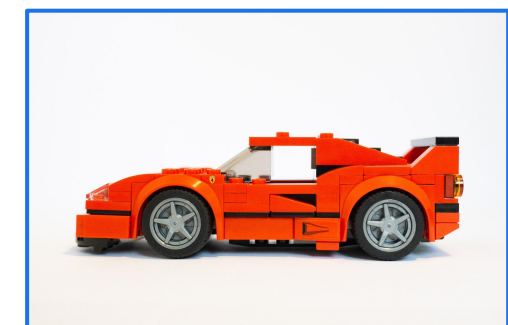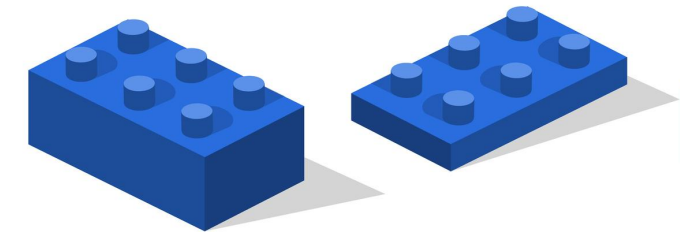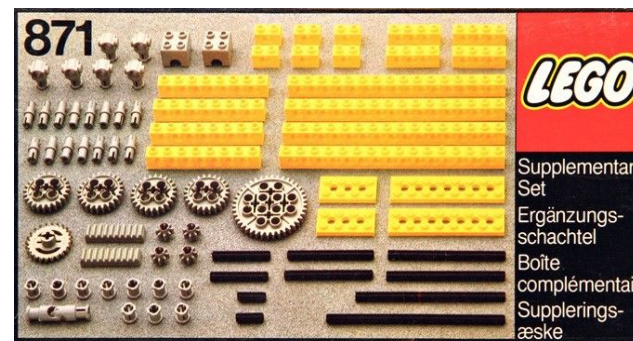
Agenda

- Overview

- Introduction to AI Platform Pipelines

- AI Platform Pipelines: Concepts

- AI Platform Pipelines: When to Use?

- AI Platform Pipelines: Ecosystem
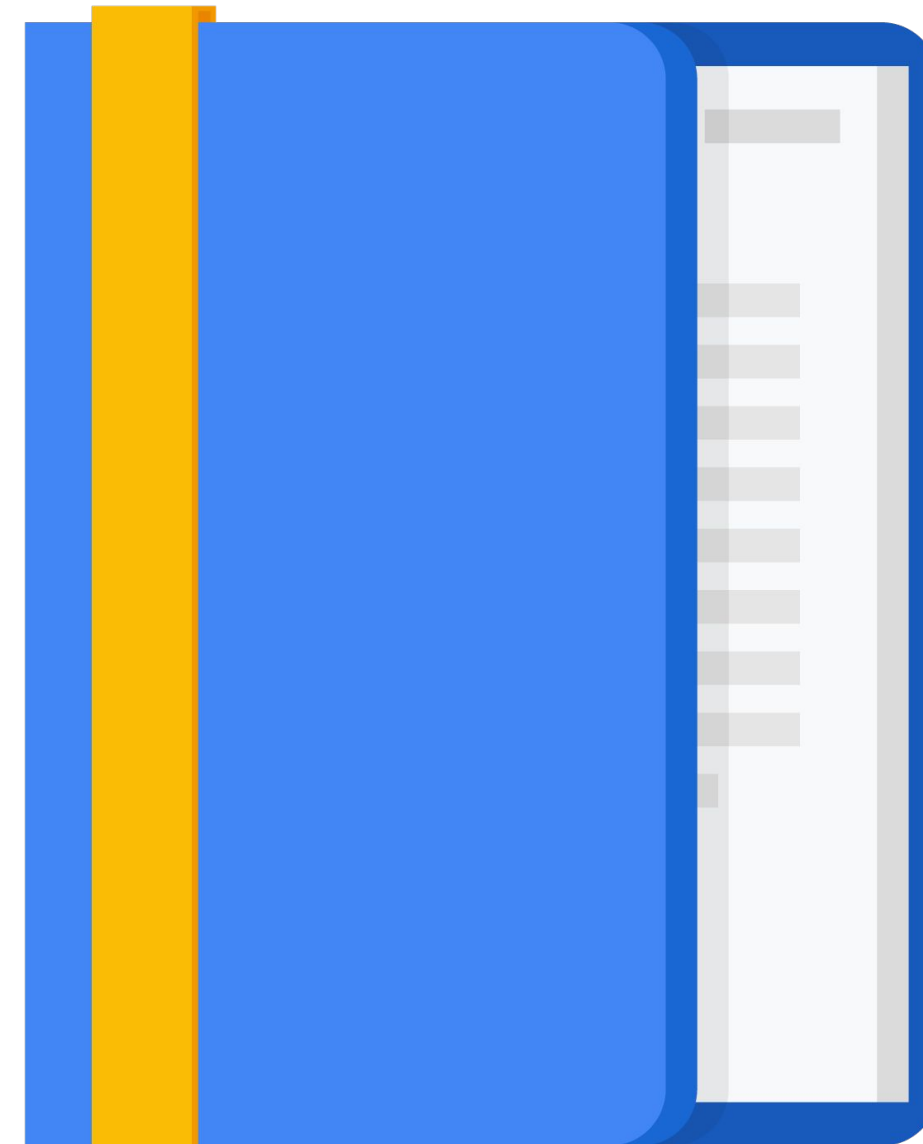
AI problems today

Talent	Ecosystem	Flexibility
Lack of expertise in ML	Difficult to find and leverage existing solutions	Brittle, opinionated infrastructure that breaks between cloud and on-premises
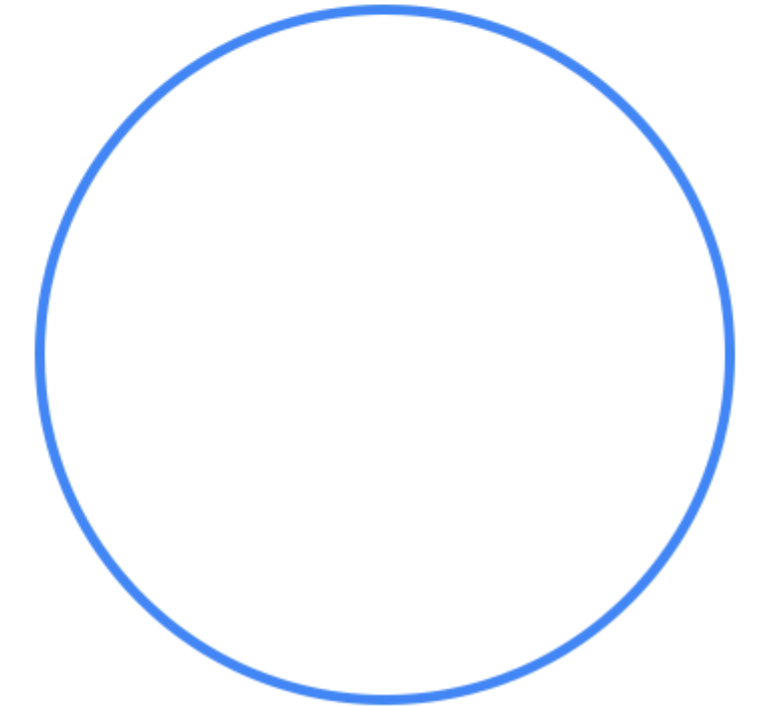
AI problems today

Talent	Ecosystem	Flexibility
Lack of expertise in ML	Difficult to find and leverage existing solutions	Brittle, opinionated infrastructure that breaks between cloud and on-premises

One of the most important factors to address
for better collaboration and business impact
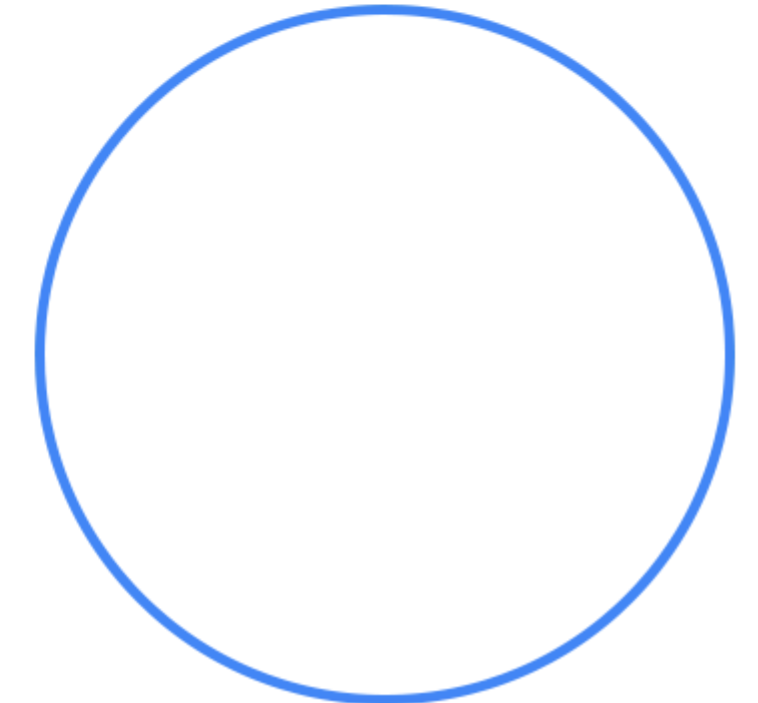
Reusable pipelines: Force multiplier for data scientists

TODAY

SWE
- Connect to database ○
- Create features and labels ○

SWE + Data Scientist
- Create train/test datasets ○
- Train model ○
- HP tuning ○
- Publish model ○

SWE
- Business logic ○
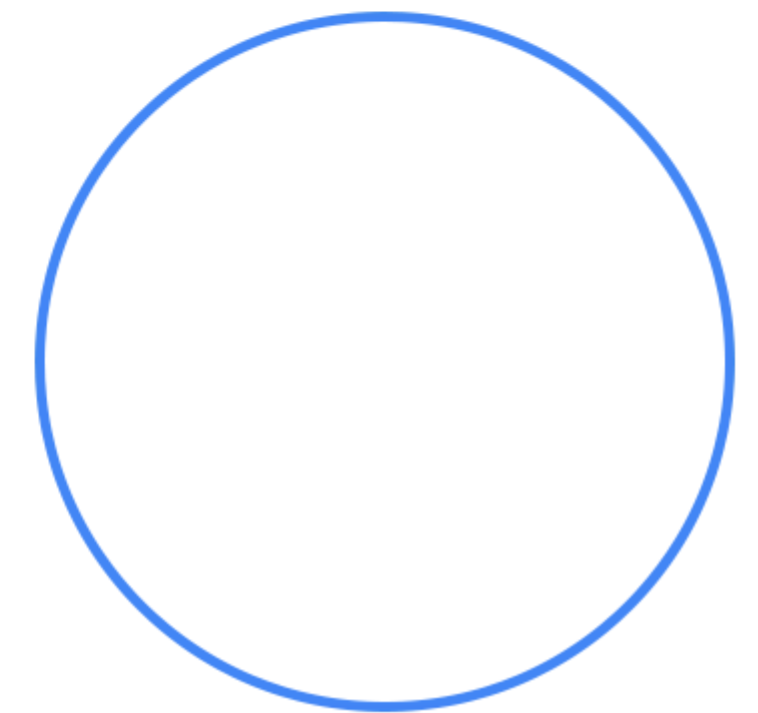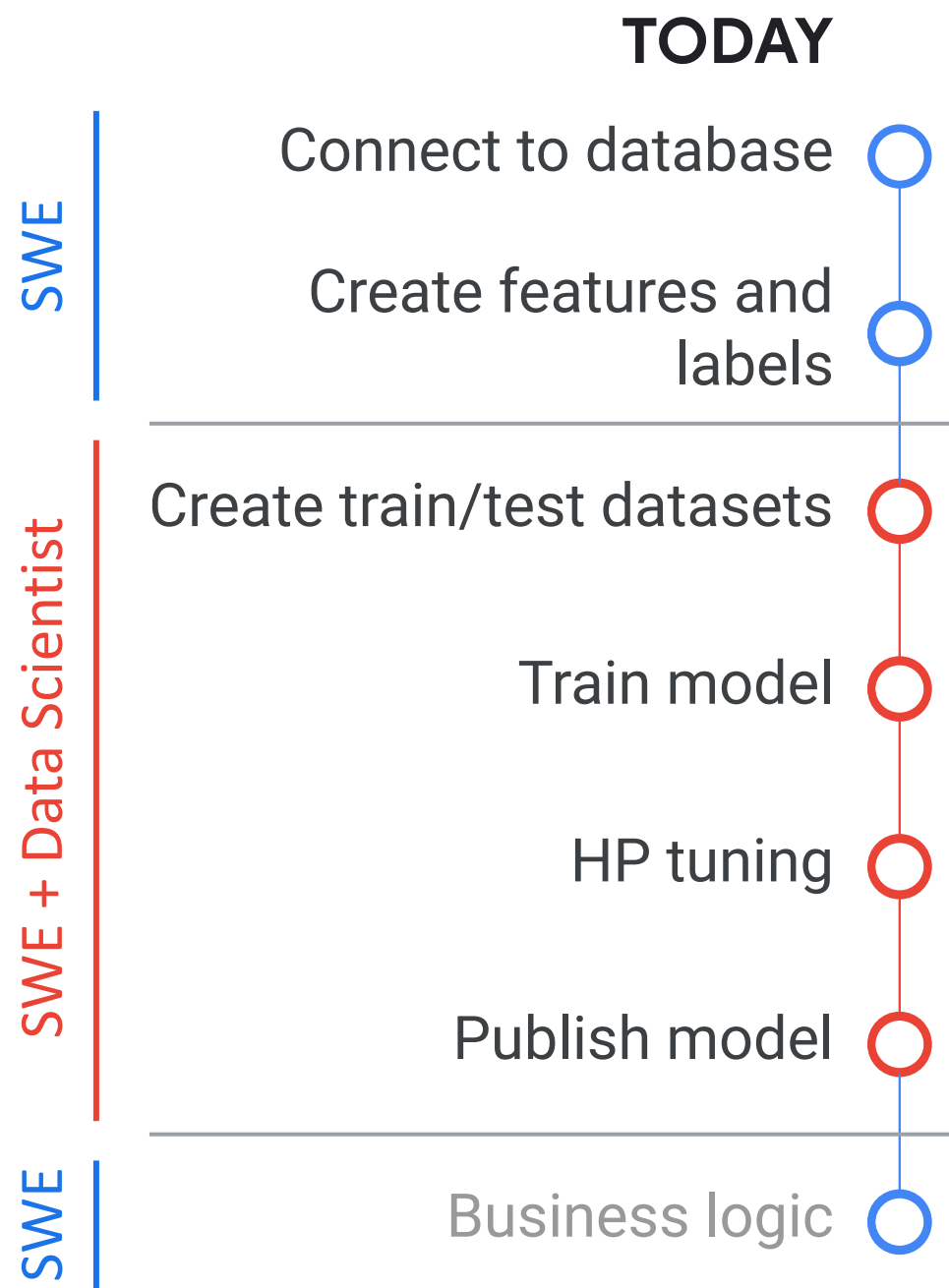
Reusable pipelines: Force multiplier for data scientists
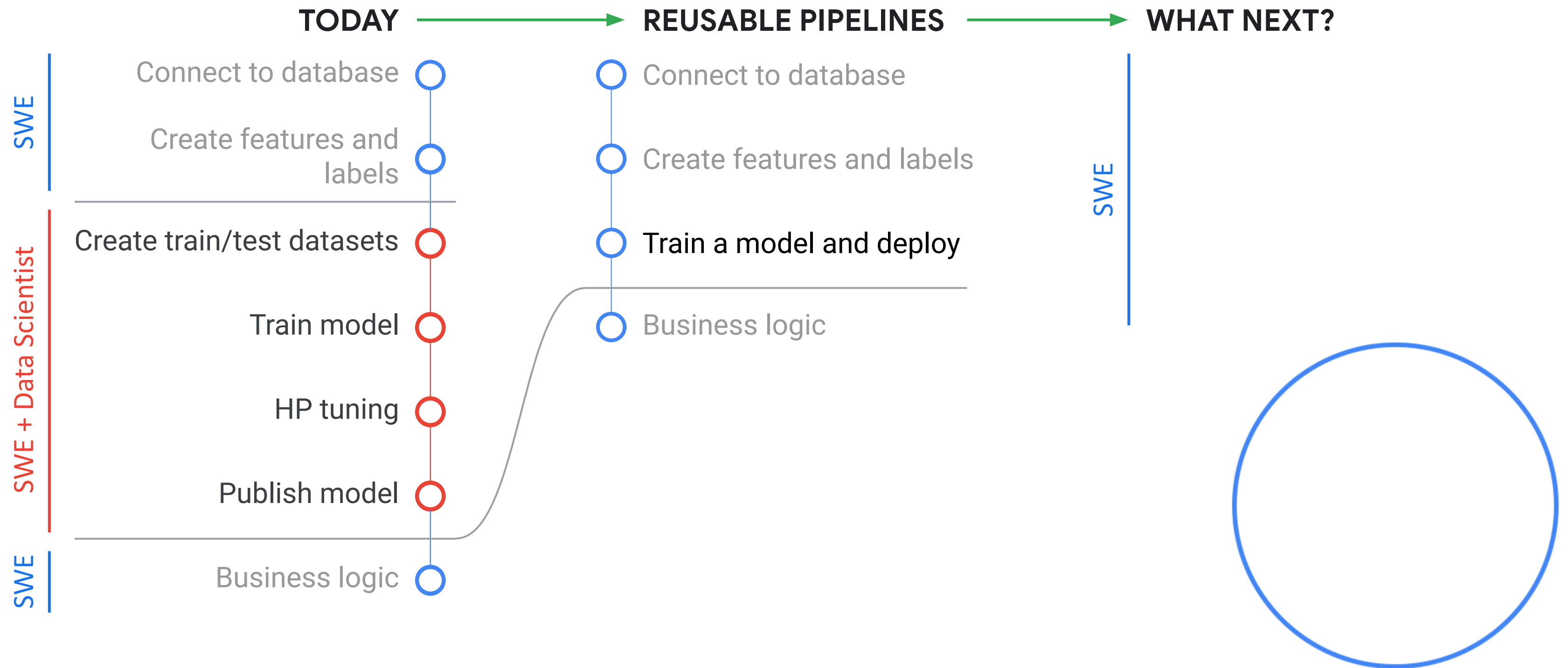
Mission

The one place for everything AI,
from experimentation to production

AI Hub

Public Content

+ Private Content

By Google

Unique AI assets by Google

By Partners

Created, shared, and monetized by anyone

By Customers

Content shared securely within and with other organizations

AutoML, TPUs, **kaggle** Cloud AI Platform, etc.

Research at **Google**

DeepMind
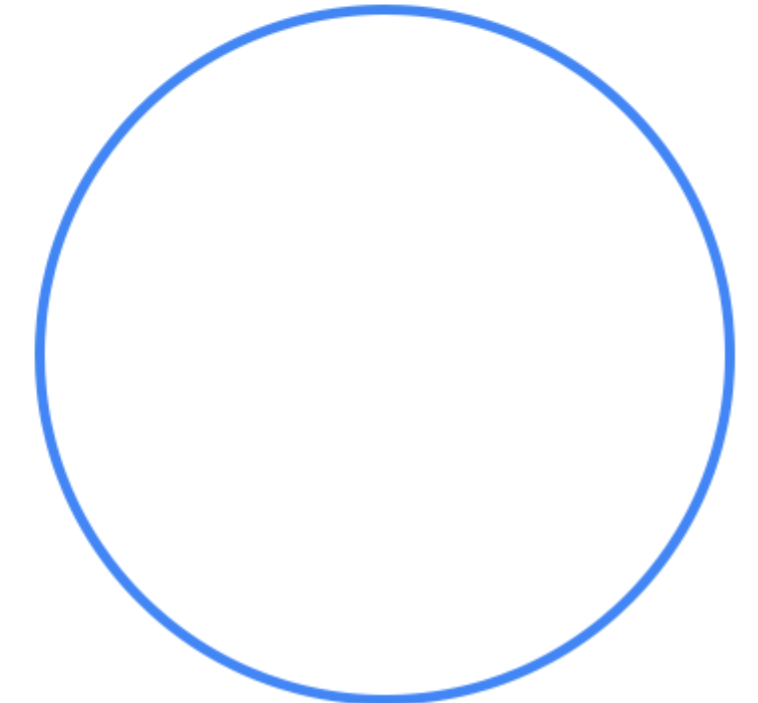
AI Hub and Pipelines: Fast and simple adoption of AI

The flywheel of AI adoption

1. SEARCH and DISCOVER: Find best-of-breed pipelines on the hub that leverage Cloud AI solutions (AutoML, GPU, TPU, CMLE, etc.).

5. PUBLISH: Upload and share pipelines that run best on Google Cloud with your org or publicly.

Network Effect
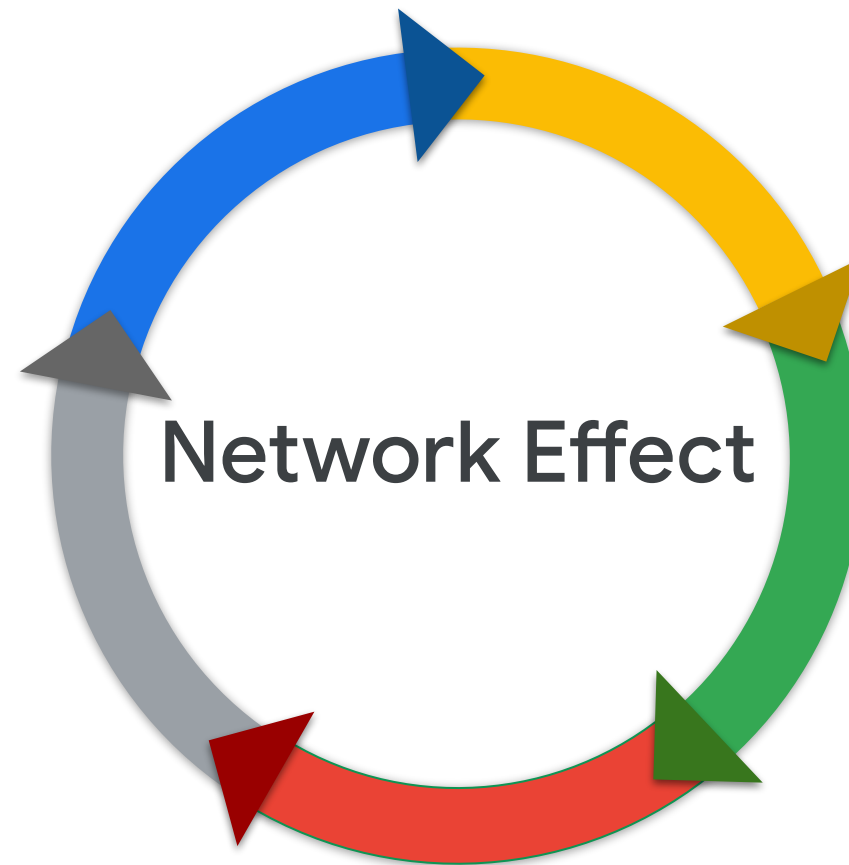
2. DEPLOY: Quick 1-click implementation of ML pipelines onto Google Cloud/GKE.

3. CUSTOMIZE: Experiment and adjust out-of-the-box pipelines to custom use cases via pipelines UI.

4. RUN IN PRODUCTION: Deploy customized pipelines in production on Google Cloud.

Let's start building an AI Platform Pipeline

Agenda

- Start building an AI Platform Pipelines Instance

- Run an example pipeline
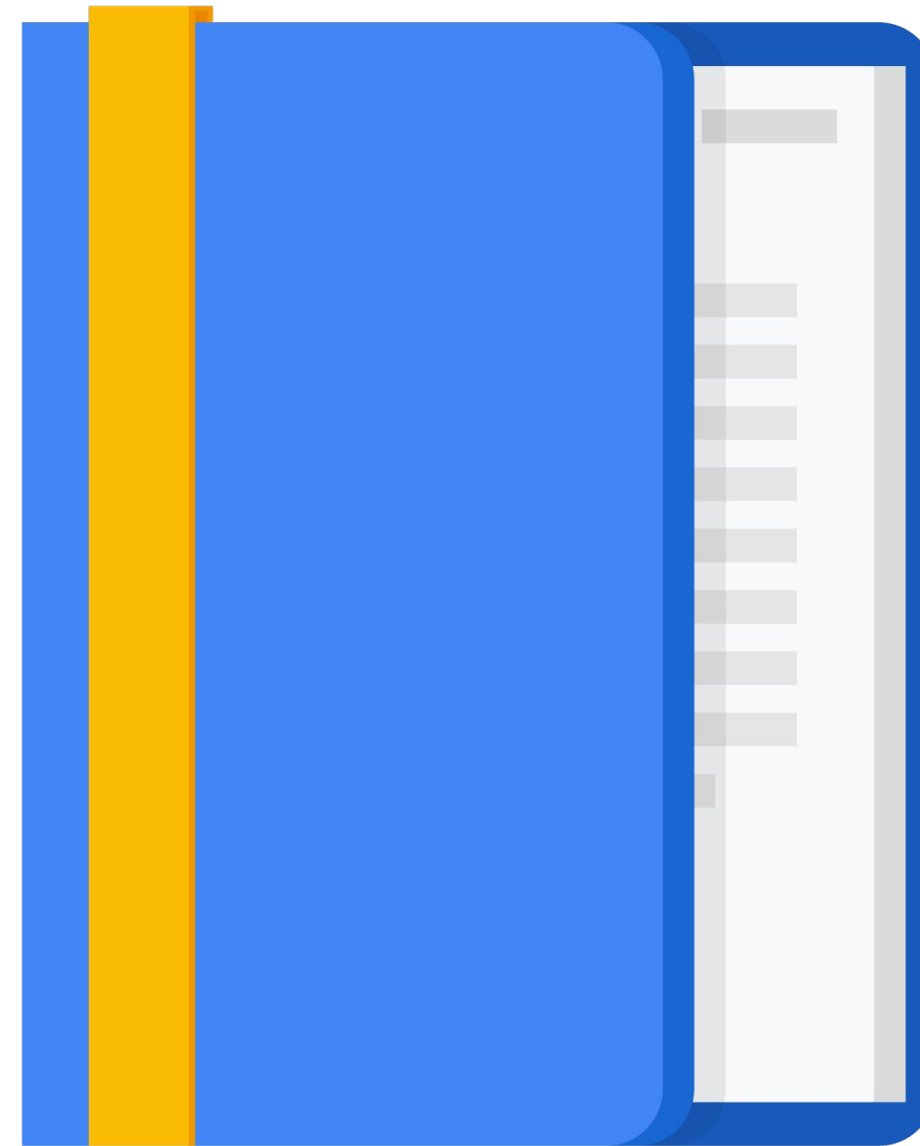
Start with your GCP project

1. Open Cloud shell

2. Edit IAM policy

gcloud projects get-iam-policy <you-project-name> --flatten="bindings[].members"
--format="table(bindings.role, bindings.members)" --filter="bindings.role:roles/container.admin OR
bindings.role:roles/viewer"
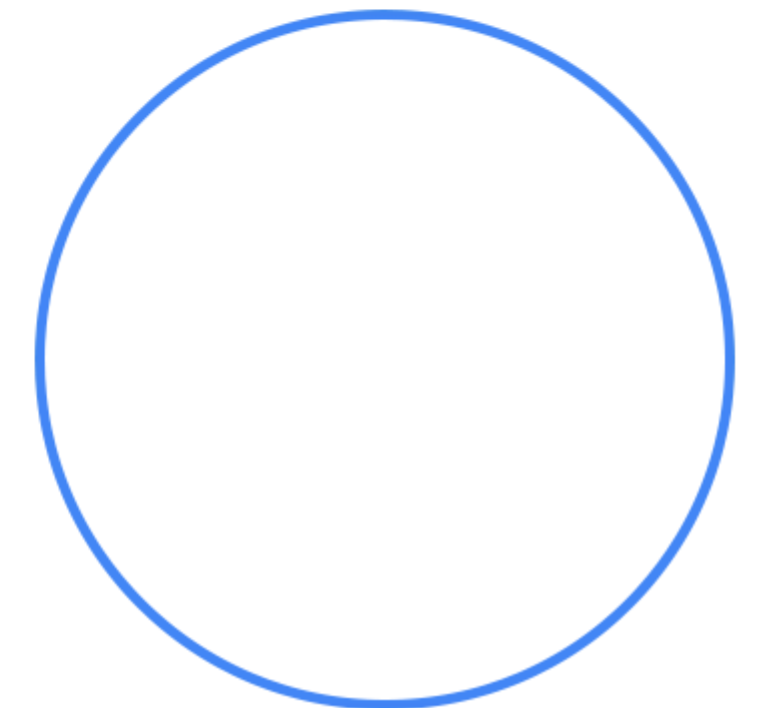
Start with your GCP project

3. Open AI platform Pipelines
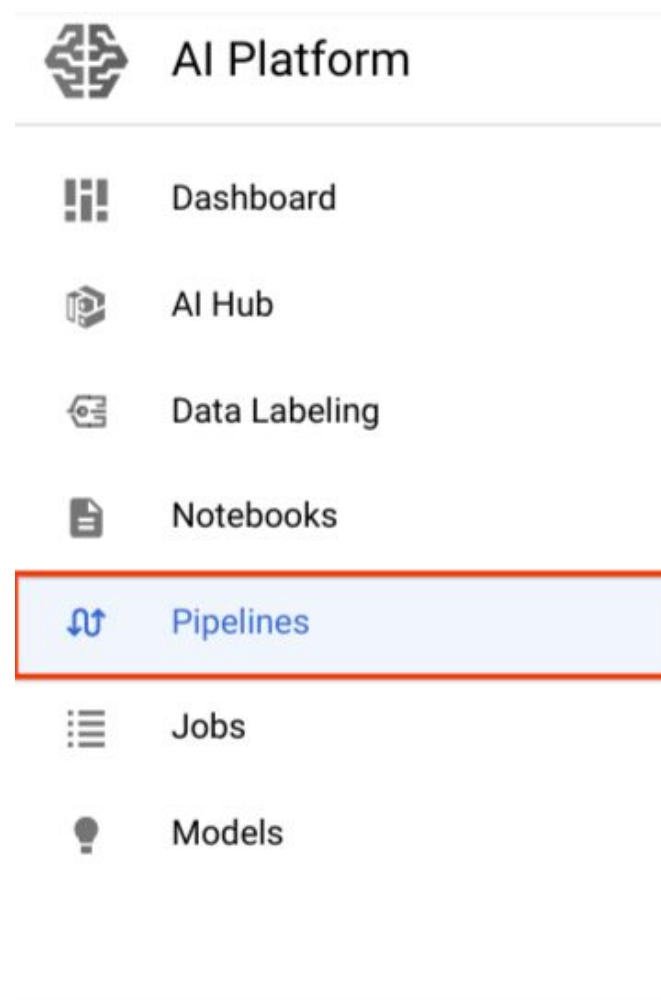
4. Create a new instance
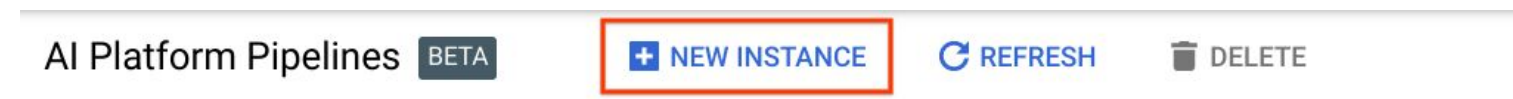
5. This will open Kubeflow pipeline page. Click "*Configure*".
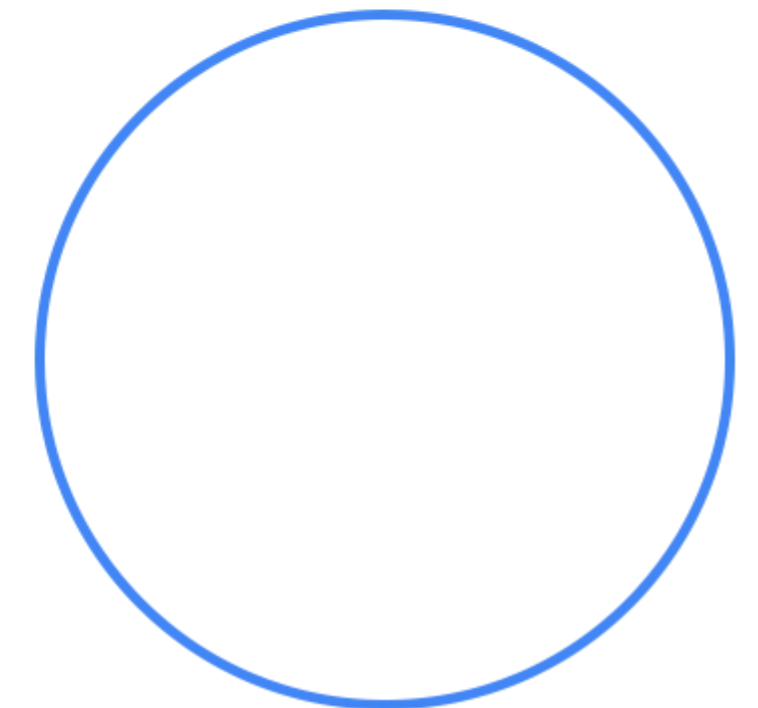
Start with your GCP project

6. Select any region for the cluster and check "access to Cloud APIs" for default access. Click "*Create cluster*". This will automatically create cluster for you in the selected region. This may take a moment to create an instance.

Your app will use compute instances managed in a logical grouping called a "cluster", which will be configured in a way that's great for getting started with Kubernetes. For more options, visit the Kubernetes engine cluster creation page.

Cluster zone ⓘ

us-central1-a ▾

☑ Allow access to the following Cloud APIs ⓘ
https://www.googleapis.com/auth/cloud-platform

ⓘ All applications in cluster will have the access without additional settings. Learn more ↗

Create cluster

or Select an existing cluster
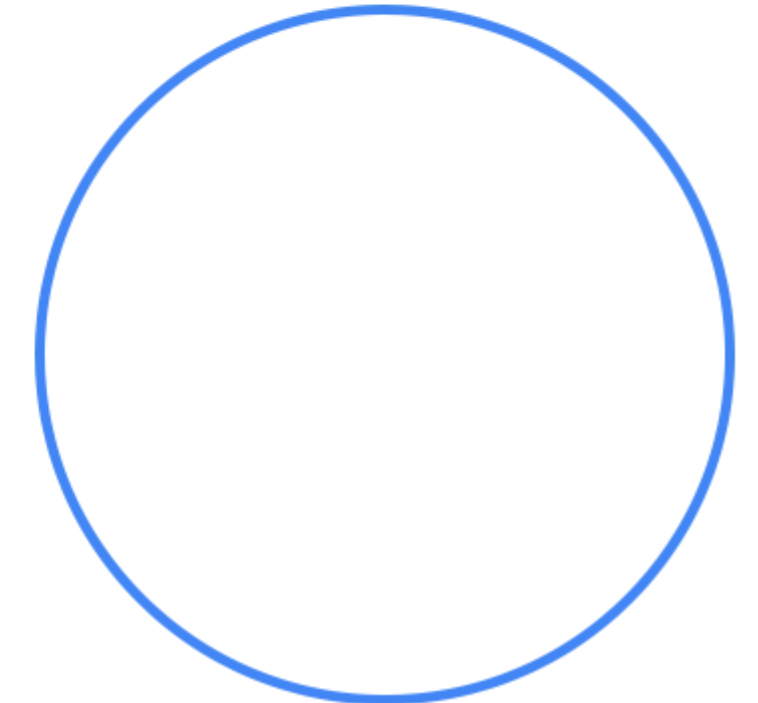
Start with your GCP project

7. After the cluster is created, Select "default" namespace and your App instance name my-first-pipeline. This step will take some time.
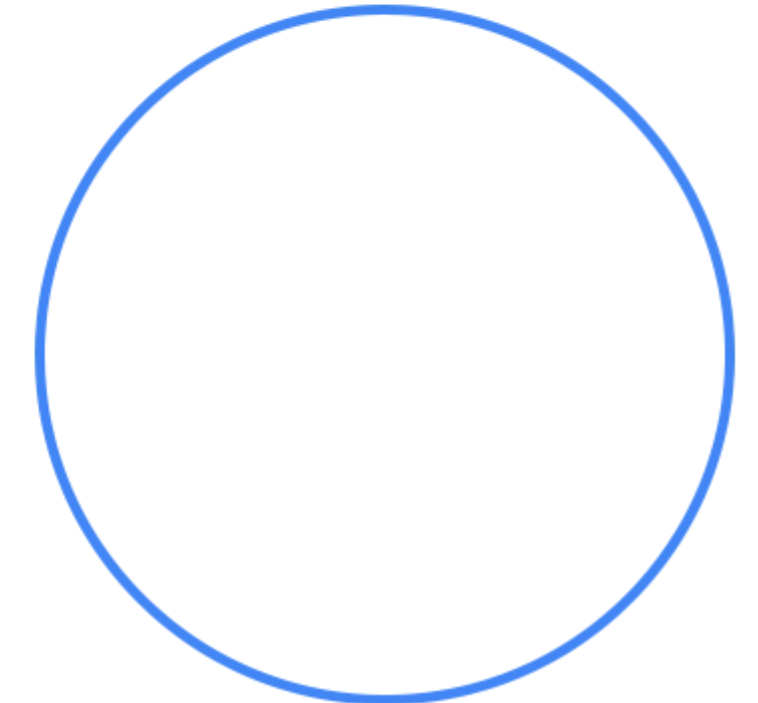
Start with your GCP project

8. This will create App instance as well as AI Platform pipeline instance. Click on "*Open Pipelines Dashboard*"
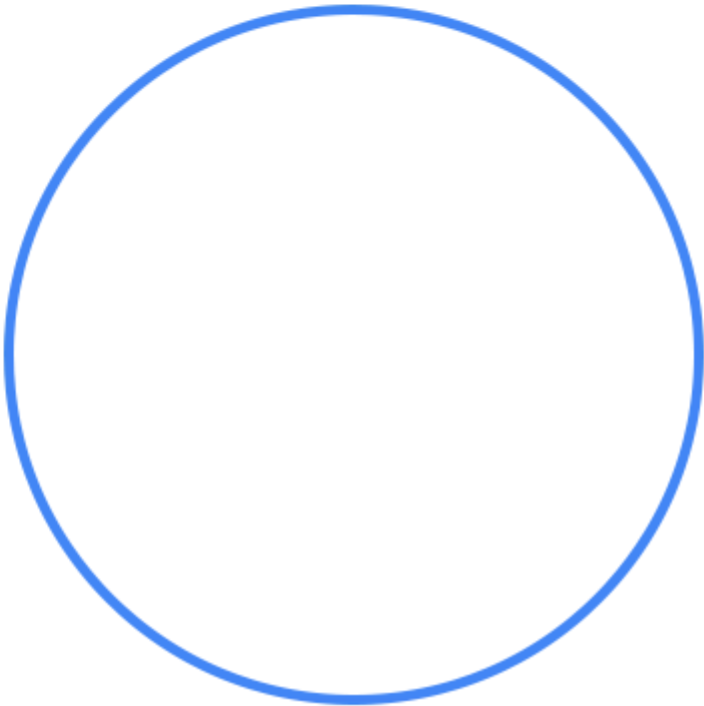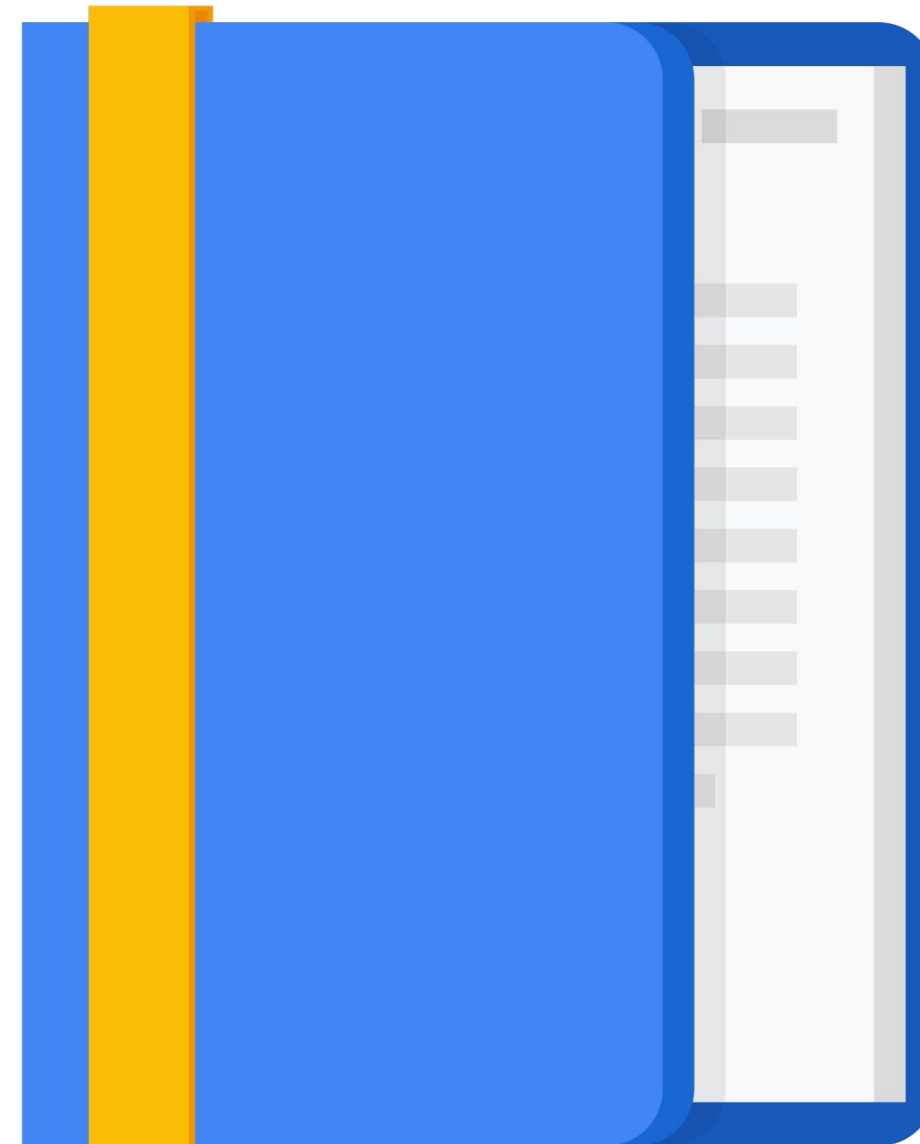
Start with your GCP project

9. This will open AI Platform Pipelines Dashboard. Click "*Pipelines*"

Agenda

- Start building an AI Platform Pipelines Instance

- Run an example pipeline

Start with an example pipeline

1. Let's start with an example pipeline "*[Demo] TFX - Taxi Tip Prediction Model Trainer*"

2. To run or schedule the pipeline, click Create run. A form where you can enter the run details opens.

3. You will need to add Run details that includes:
   a. Pipeline
   b. Pipeline version
   c. Run name
   d. Description
   e. Experiment

Start with an example pipeline

Start with an example pipeline

Final output from the experiment pipeline