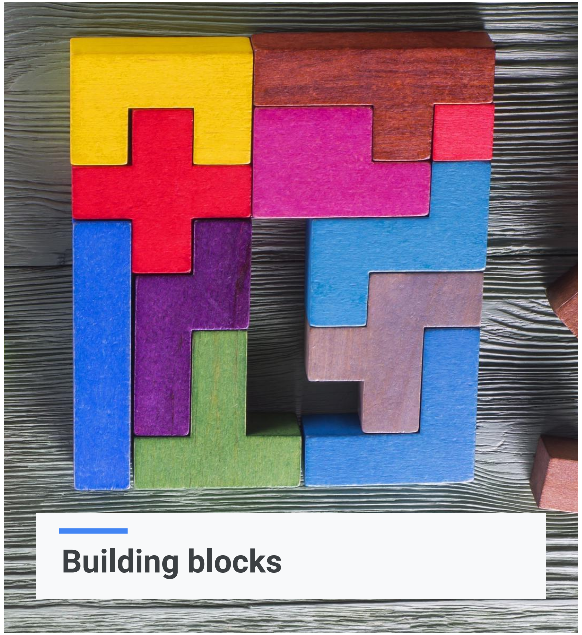
S91030 - Hybrid Machine Learning with the Kubeflow Pipelines and RAPIDS

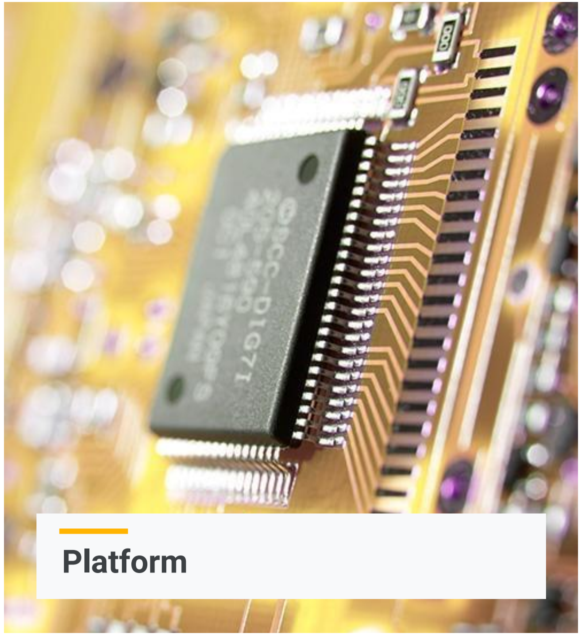
Sina Chavoshi
Technical Program Manager

Google Cloud

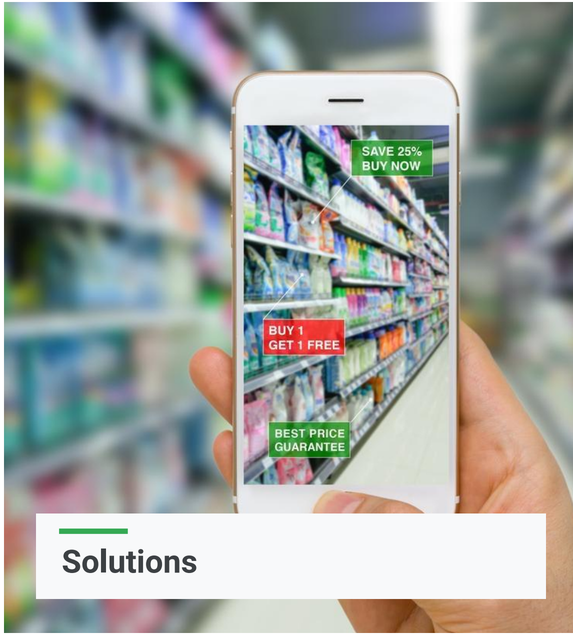
Cloud AI Strategy:
The right approach for the right problem



Building blocks



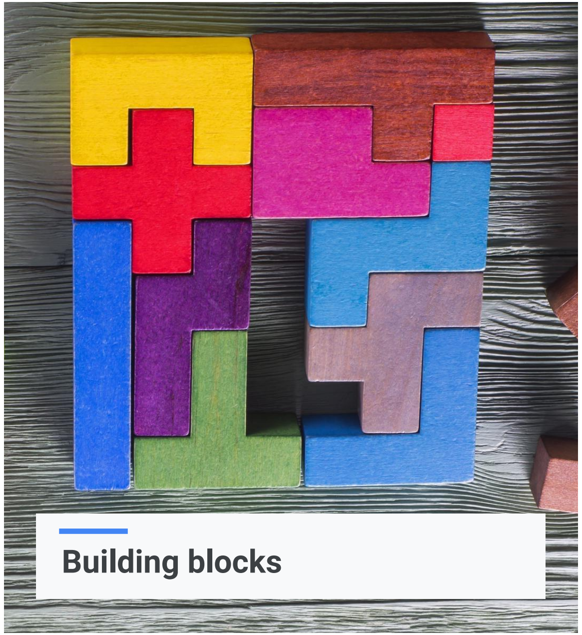
Platform



Solutions

Cloud AI Strategy:

The right approach for the right problem



Google Cloud

Building Blocks

Sight

Cloud Vision API
Image recognition and classification.

Cloud Video Intelligence API
Scene-level video annotation.

AutoML Vision BETA
Custom image classification models.

Language

Cloud Translation API
Language detection and translation.

Cloud Natural Language API
Text parsing and analysis.

AutoML Translation BETA
Custom domain-specific translation.

AutoML Natural Language BETA
Custom text classification models.

Conversation

Dialogflow Enterprise Edition
Build conversational interfaces.

Cloud Text-to-Speech API
Convert text to speech.

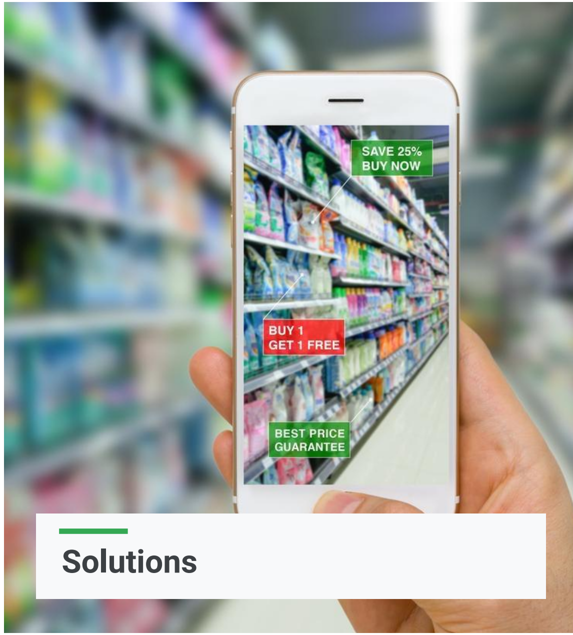
Cloud Speech-to-Text API
Convert speech to text.

Cloud AI Strategy:

The right approach for the right problem

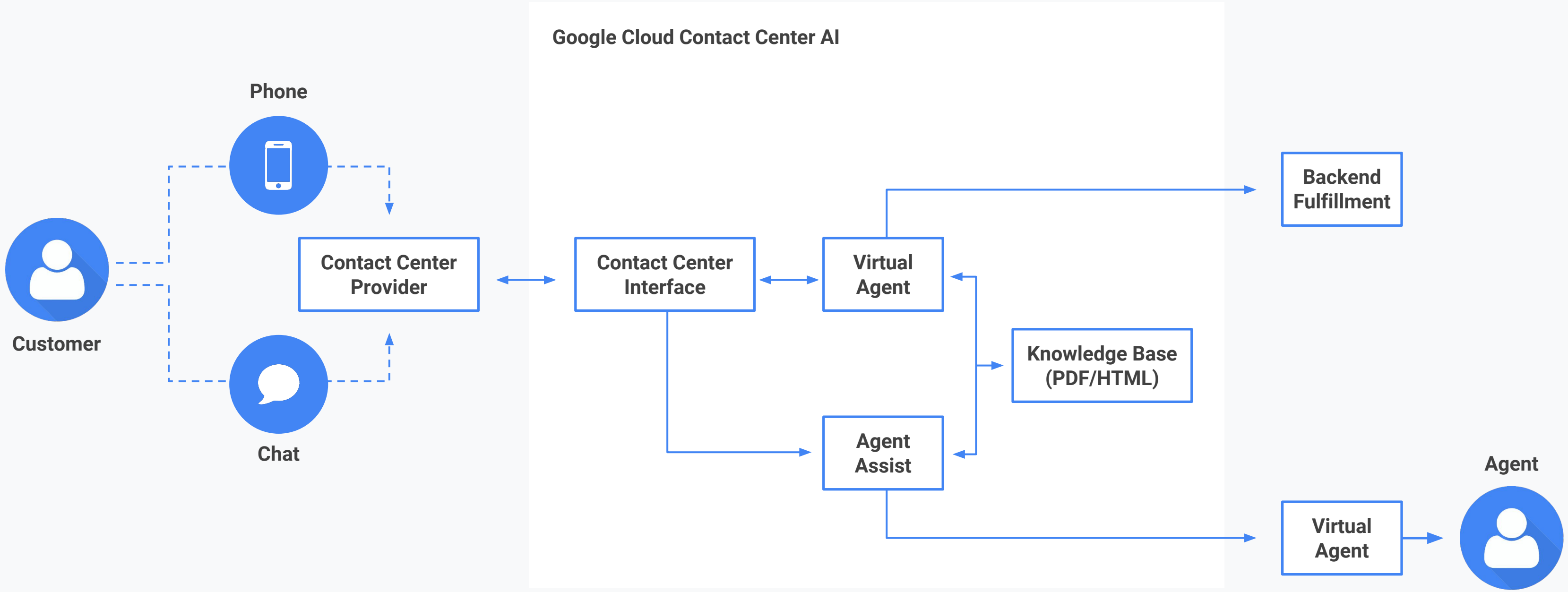
Building blocks

Platform

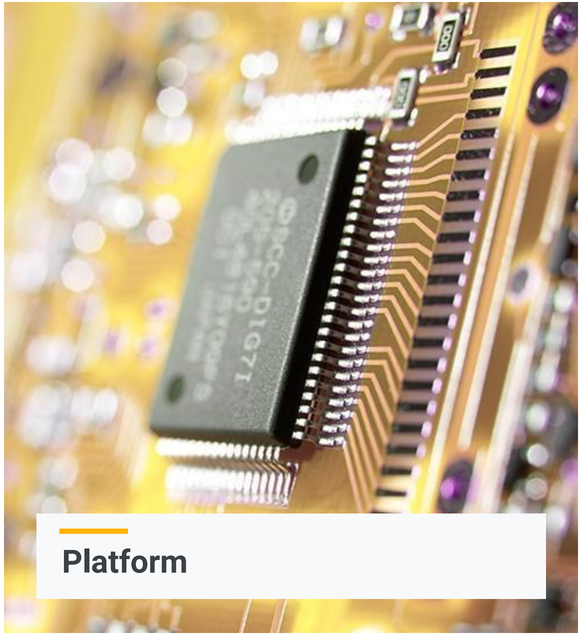


Solutions

Solutions / Contact Center



Cloud AI Strategy:
The right approach for the right problem



Google Cloud

Cloud AI Platform

Data pipeline

Cloud Dataprep

BigQuery

Cloud Dataflow

Cloud Dataproc

Model development

Cloud ML Engine

TensorFlow

Model deployment and management

Cloud ML Engine

Cloud Kubernetes Engine

TensorFlow

Kubeflow

Tools

Jupyter Notebooks

Services

ASL

Community

kaggle

Google Cloud

Building & deploying real-life ML applications is hard and costly because of lack of tooling that covers end-to-end ML development & deployment.

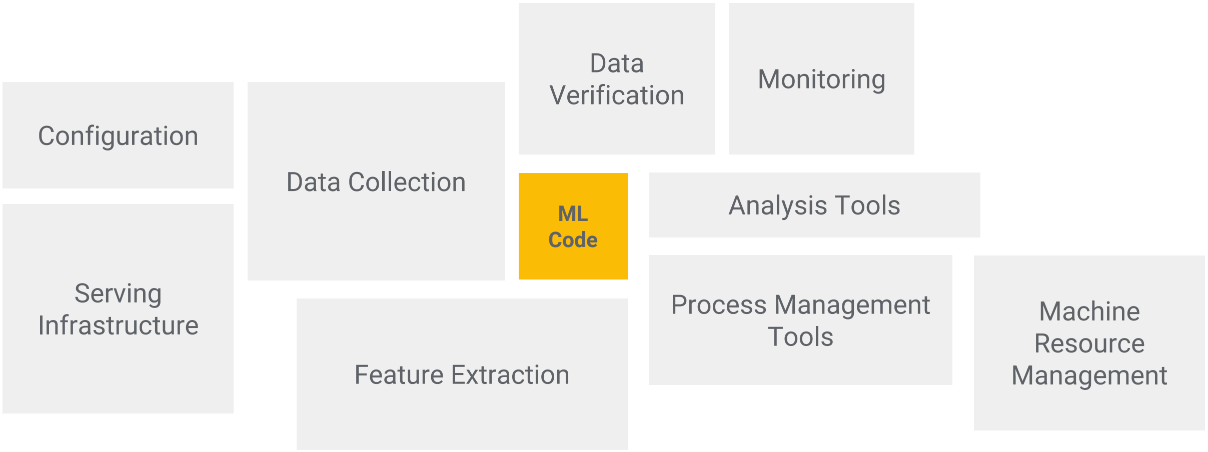
In addition to the actual ML...

TensorFlow

ML
Code

You have to worry about so much more.

TensorFlow



Source: Sculley et al.: Hidden Technical Debt in Machine Learning Systems

AI problems today

Problems

Deployment
Brittle, opinionated infrastructure that is hard to productionize and breaks between cloud and on-prem

Talent
Machine Learning expertise is scarce

Collaboration
Difficult to find, leverage existing solutions

Solutions

01 Kubeflow

02 Reusable pipelines

03 Google Cloud AI Hub

01: Kubeflow

Scalable ML services on Kubernetes

Easy to get started

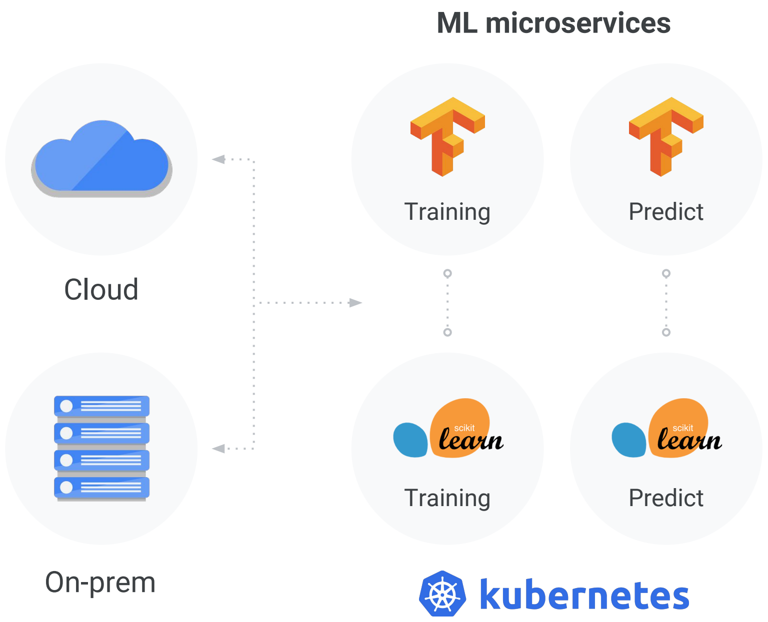
- Out-of-box support for top frameworks
  - – pytorch, caffe, tf and xgboost
- Kubernetes manages dependencies, resources

Swappable & scalable

- Library of ML services
- GPU support
- Massive scale

Meet customer where they are

- GCP
- On-prem with Cisco

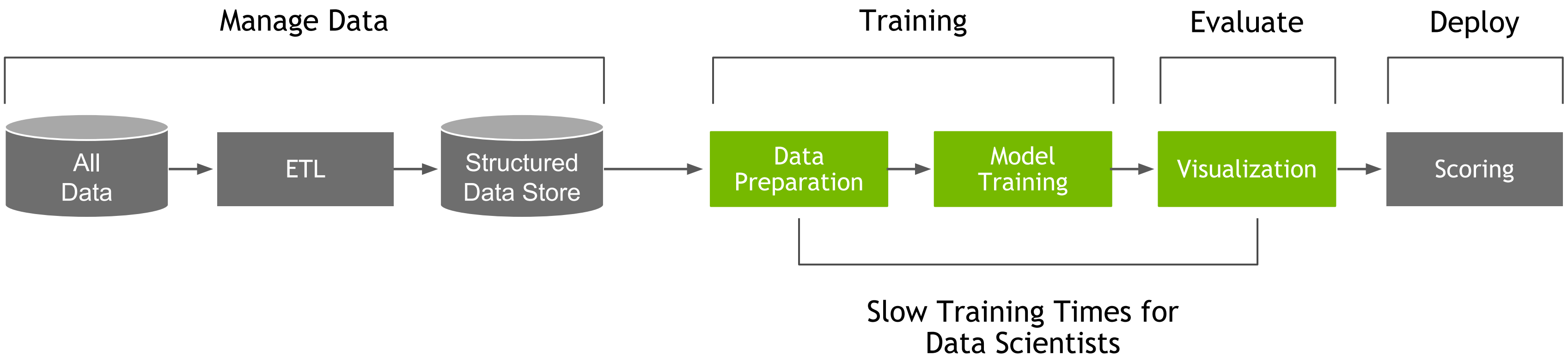


NVIDIA

RAPIDS

Product Overview

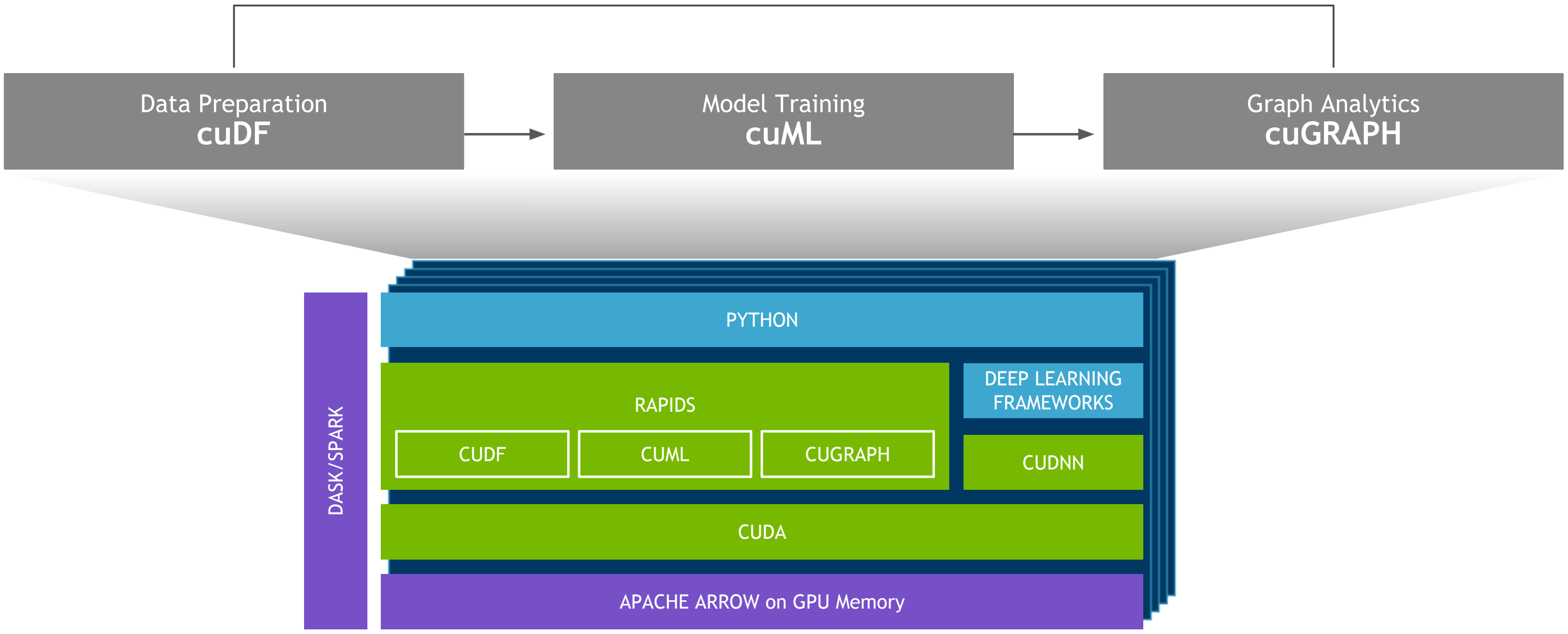
THE BIG PROBLEM IN DATA SCIENCE



Google Cloud

RAPIDS — OPEN GPU DATA SCIENCE

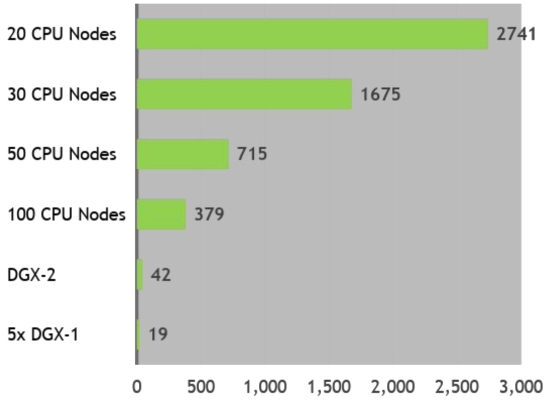
Software Stack Python



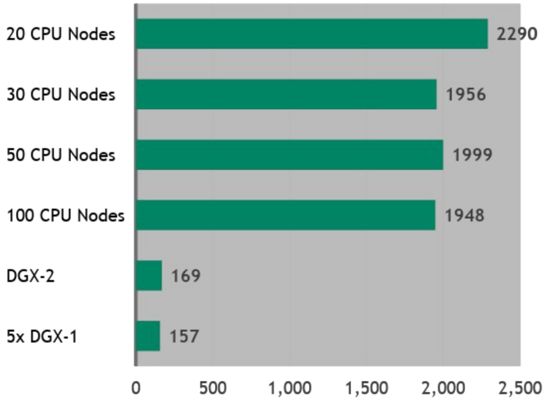
Google Cloud

BENCHMARKS

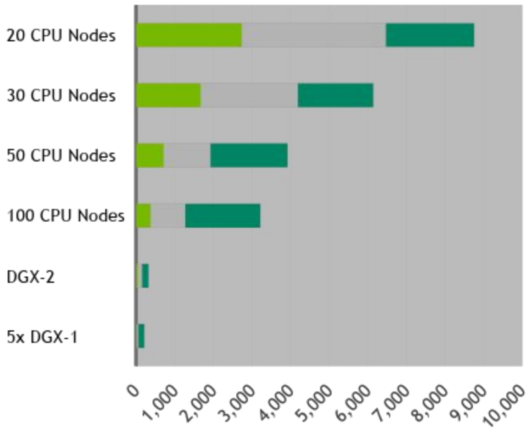
cuIO/cuDF — Load and Data Preparation



cuML — XGBoost



End-to-End



Time in seconds — Shorter is better

cuIO / cuDF (Load and Data Preparation) XGBoost Data Conversion

---

Benchmark

200GB CSV dataset; Data preparation includes joins, variable transformations.

CPU Cluster Configuration

CPU nodes (61 GiB of memory, 8 vCPUs, 64-bit platform), Apache Spark

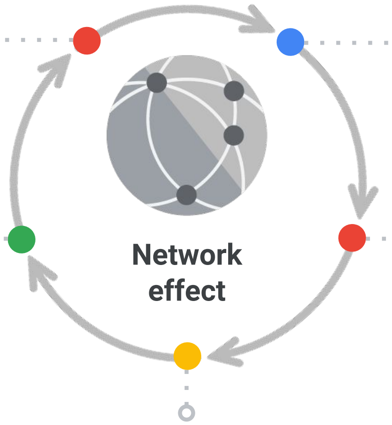
DGX Cluster Configuration

5x DGX-1 on InfiniBand network

Google Cloud

AI Hub & Pipelines: Fast & simple adoption of AI

The Flywheel of AI Adoption



1. Search & Discover
Find best-of-breed solutions on the AI Hub which leverage Cloud AI solutions

2. Deploy
Quick 1-click implementation of ML pipelines onto Google Cloud Platform .

3. Customize
Experiment and adjustment out-of-the-box pipelines to custom use cases.

4. Run in production
Deploy customized pipelines in production.

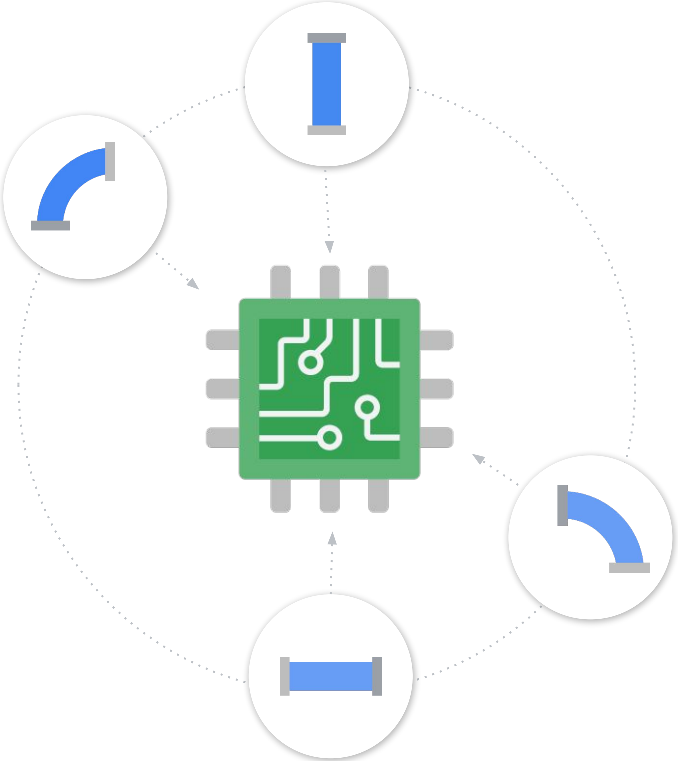
5. Publish
Upload & share pipelines running best within your org or publicly.

Google Cloud

02: Reusable Pipelines

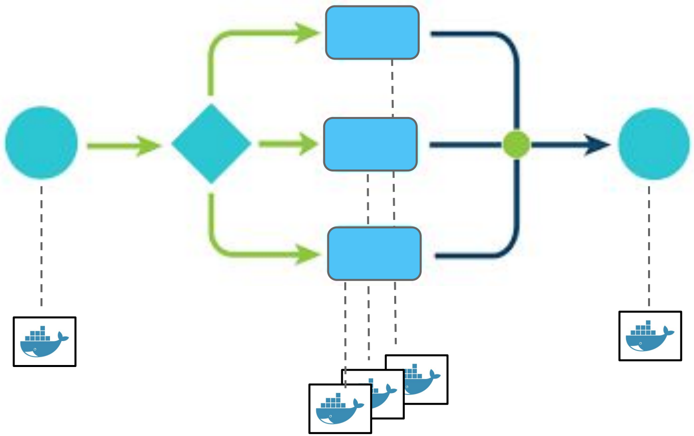
Enable developers to build custom ML applications by easily “stitching” and connecting various components.

- Reuse instead of reimplement or reinvent
- Discover, learn and replicate successful pipelines



What constitutes a Kubeflow Pipeline

- 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
- Specification of the sequence of steps
  - Specified via Python SDK
- Input Parameters
  - A “Job” = Pipeline invoked w/ specific parameters



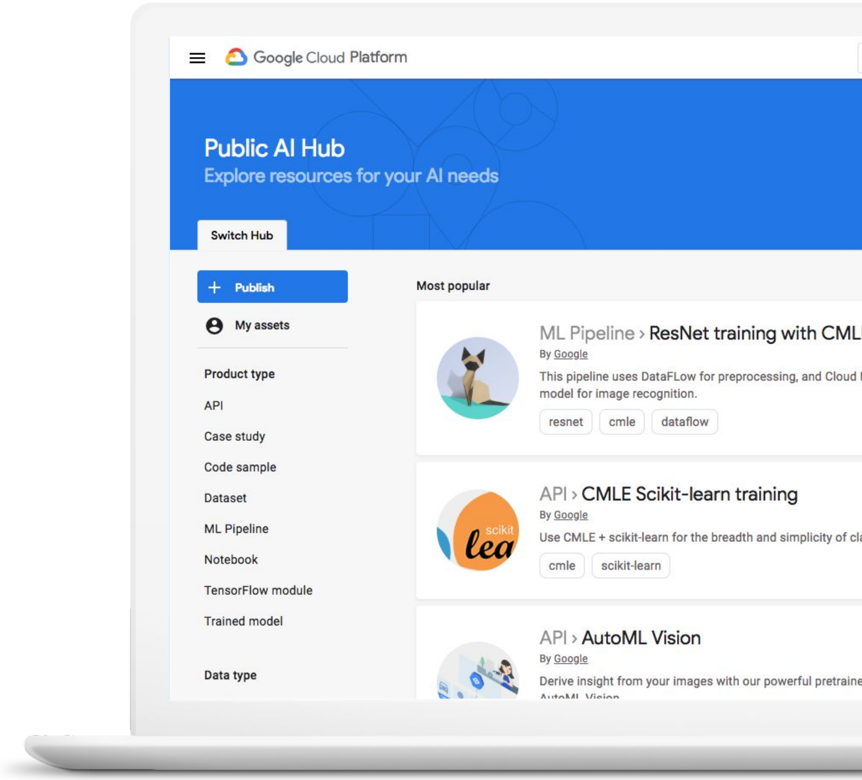
Google Cloud

03: AI Hub at a glance

1 **All AI content in one place**
Quick discovery of **plug & play** AI pipelines & other content built by teams across Google and by partners and customers.

2 **Fast & simple implementation of AI on GCP**
One-click deployment of AI pipelines via Kubeflow on **GCP as the go-to platform for AI** + hybrid & on premise.

3 **Enterprise-grade internal & external sharing**
Foster reuse by sharing deployable AI pipelines & other content privately within organizations & publicly.



Google Cloud

Mission

The one place for everything AI, from experimentation to production.

Google Cloud AI Hub

Public and private AI Hub

Public content

By Google

Unique AI assets by Google

AutoML, TPUs, kaggle
Cloud AI Platform, etc.

Research at Google

DeepMind

By partners

Created, shared & monetized by anyone

+ Private content

By customers

Content shared securely within and with other organizations

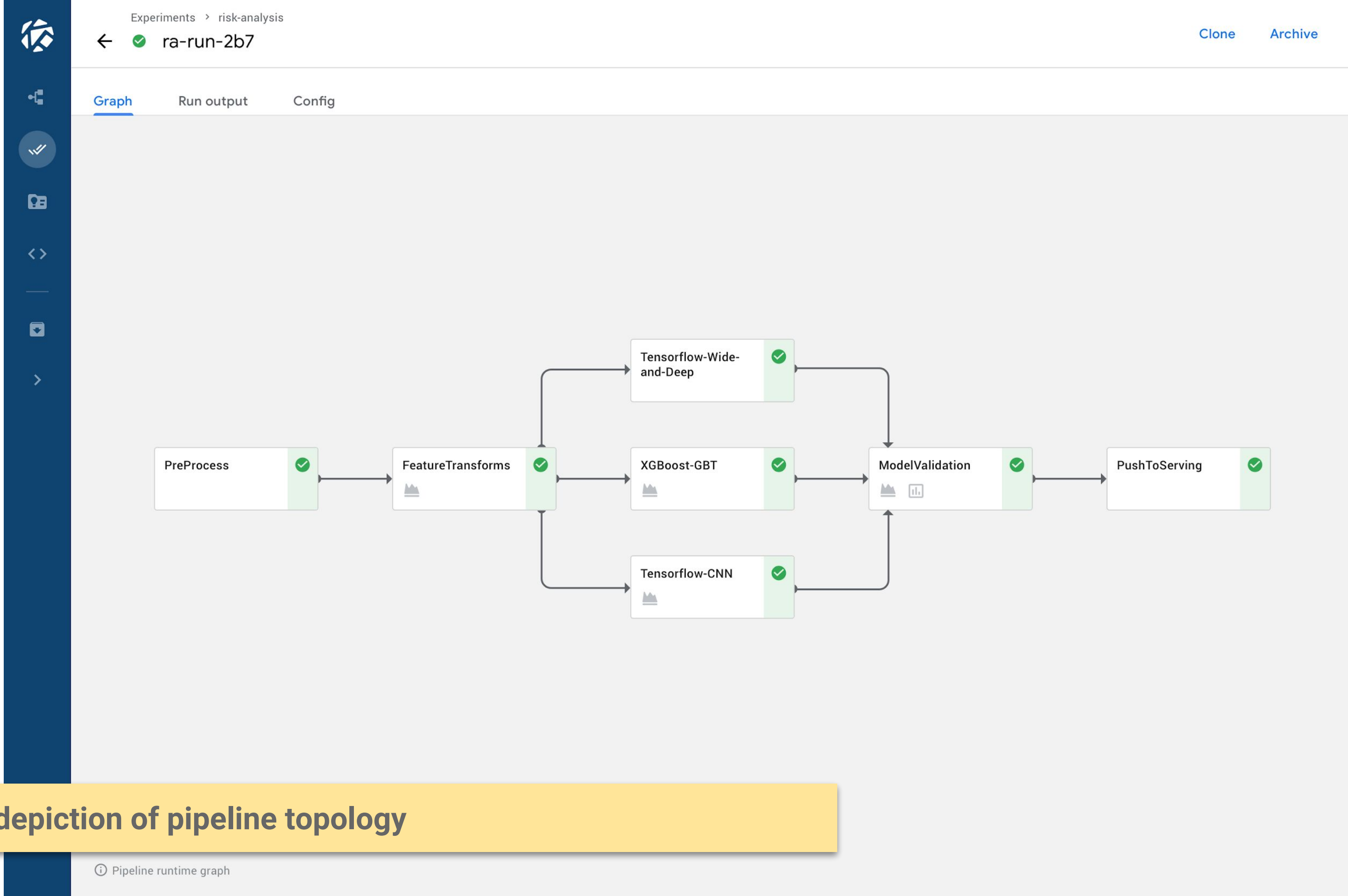
Kubeflow Pipelines enable

Workflow orchestration

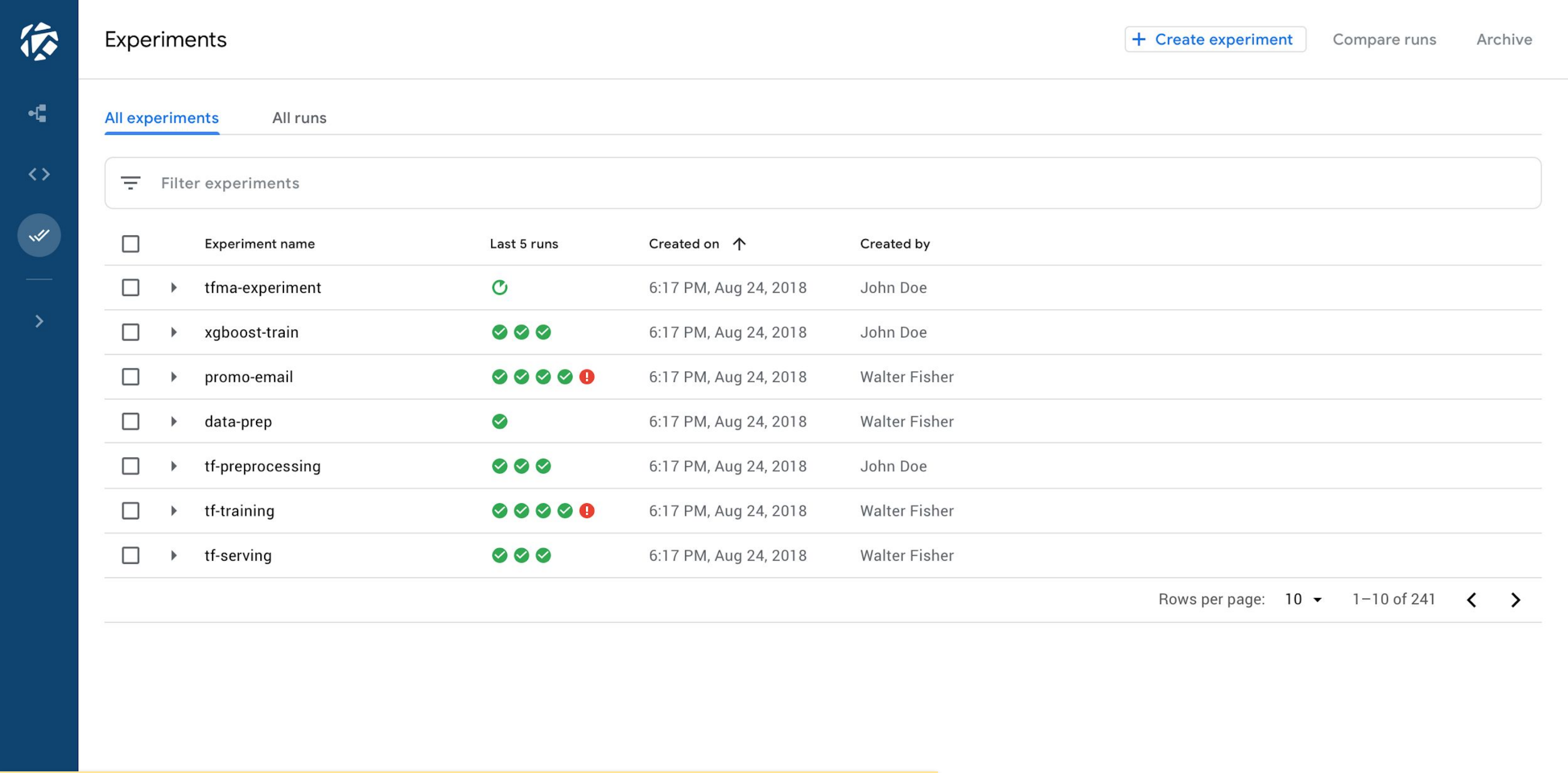
Rapid reliable experimentation

Share, re-use & compose

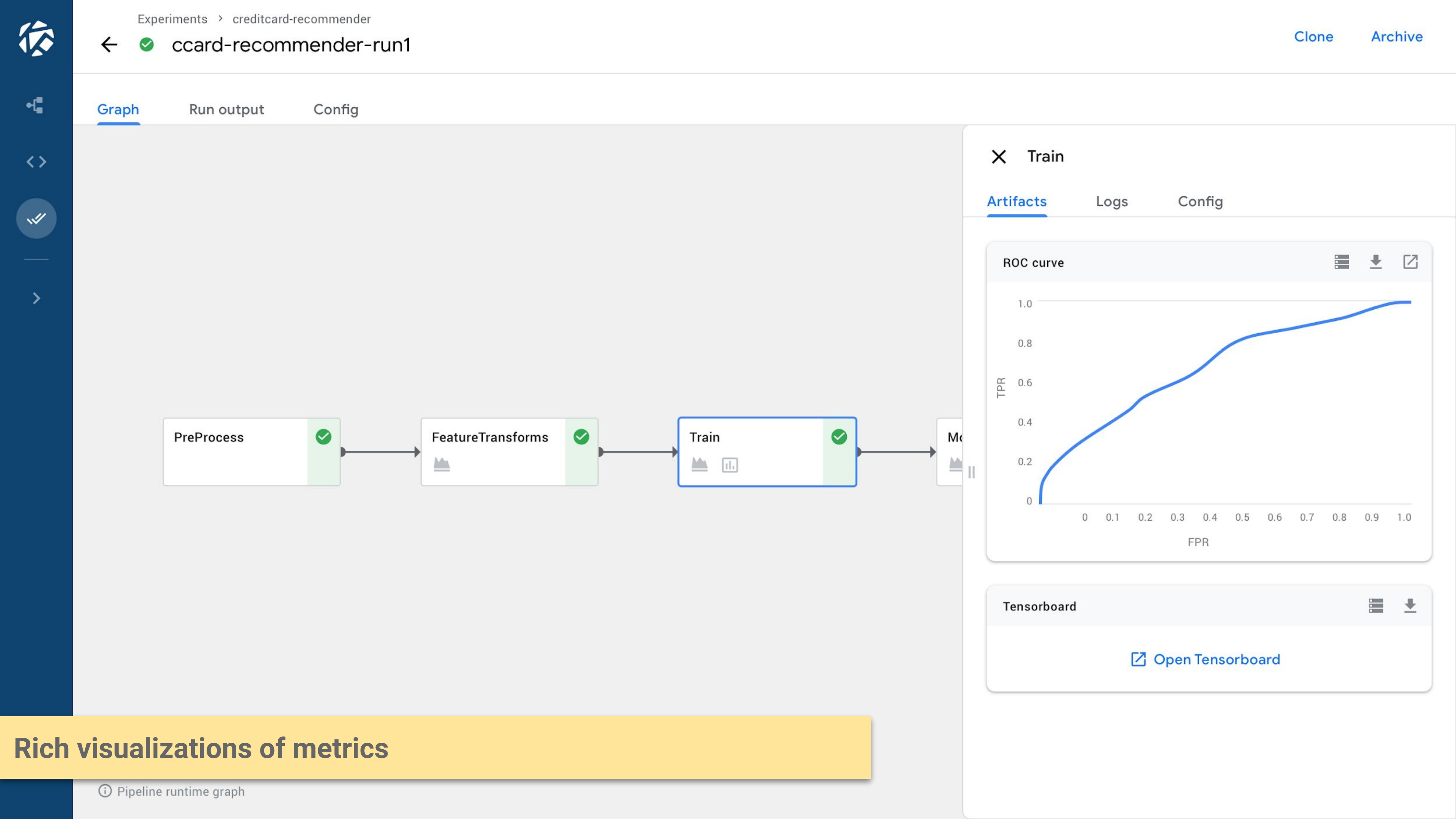
Demo

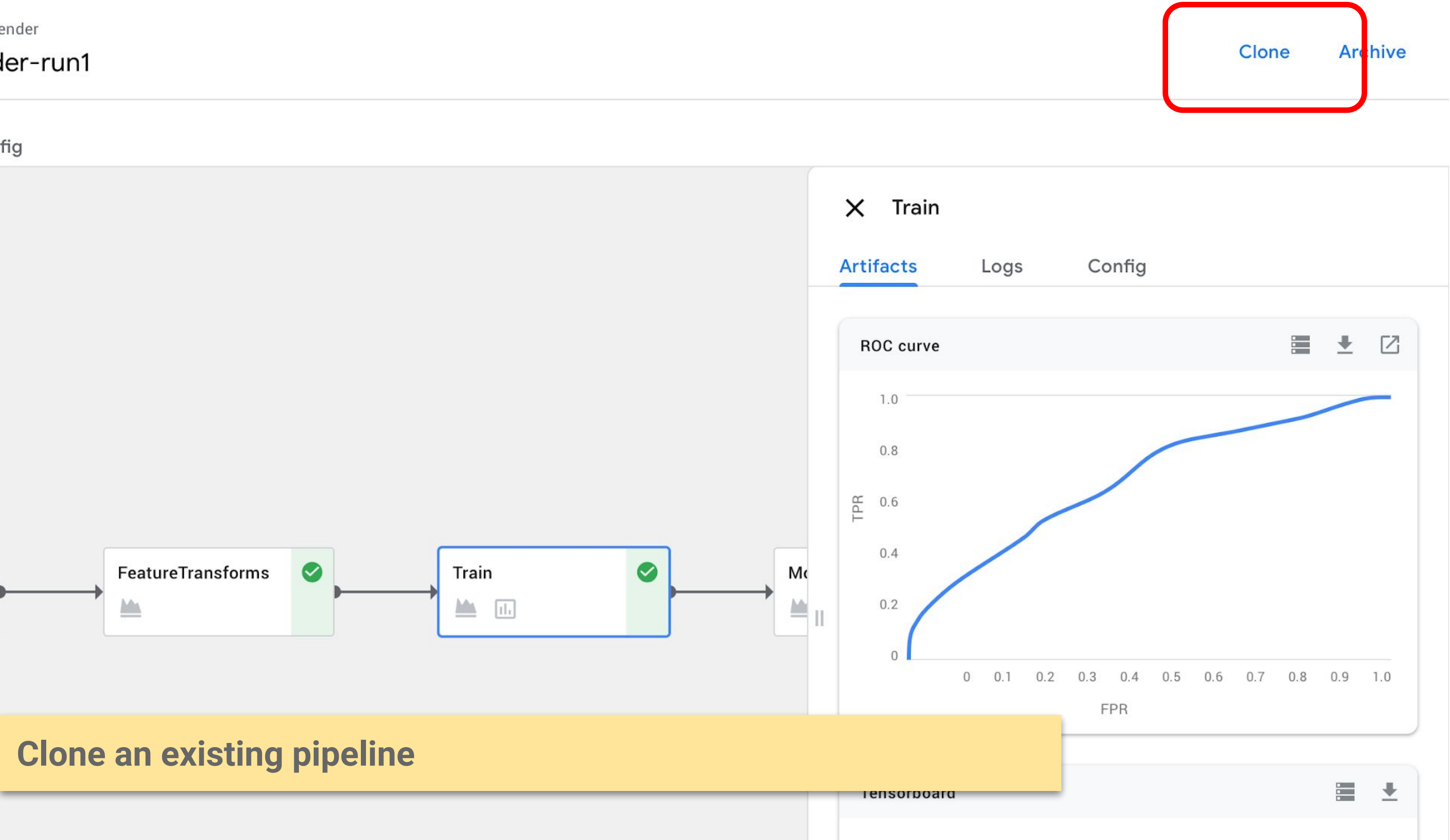


Visual depiction of pipeline topology

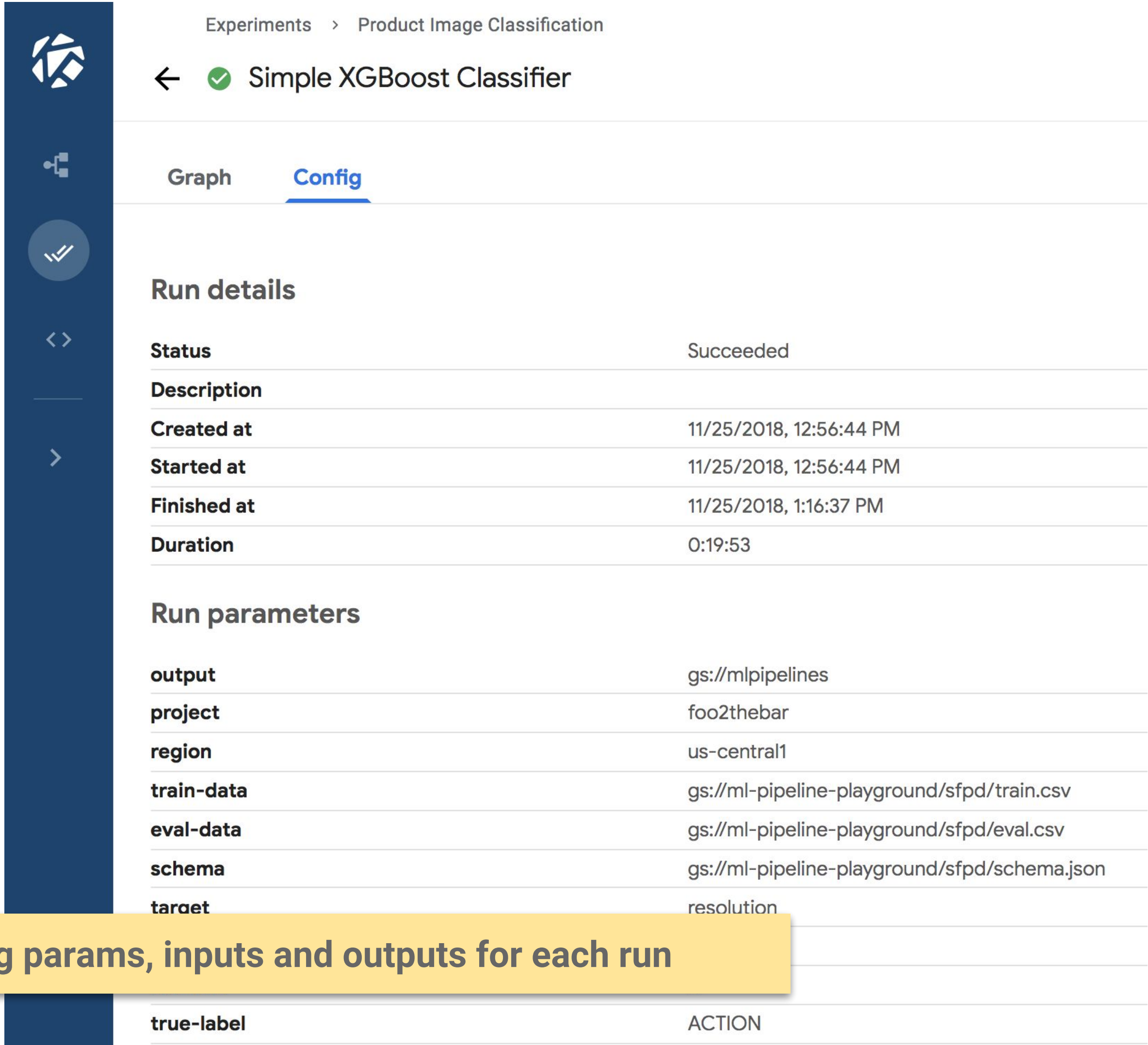


View all current and historical runs, grouped as “Experiments”


Experiments > creditcard-recommender
ccard-recommender-run1
Clone
Archive
Graph
Run output
Config
PreProcess
FeatureTransforms
Train
Mc
Train
Artifacts
Logs
Config
ROC curve
1.0
0.8
0.6
0.4
0.2
0
TPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
FPR
Tensorboard
Open Tensorboard
Rich visualizations of metrics
Pipeline runtime graph



Clone an existing pipeline



Access to all config params, inputs and outputs for each run

Run details

Pipeline*
xgboost training - confusion matrix
Choose

Run name*
product-recommender-model

Description (optional)
Train XBG model for product recommendation application.

Run parameters

Specify parameters required by the pipeline

output

project

region
us-central1

train-data
gs://ml-pipeline-playground/sfpd/train.csv

eval-data
gs://ml-pipeline-playground/sfpd/eval.csv

schema
gs://ml-pipeline-playground/sfpd/schema.json

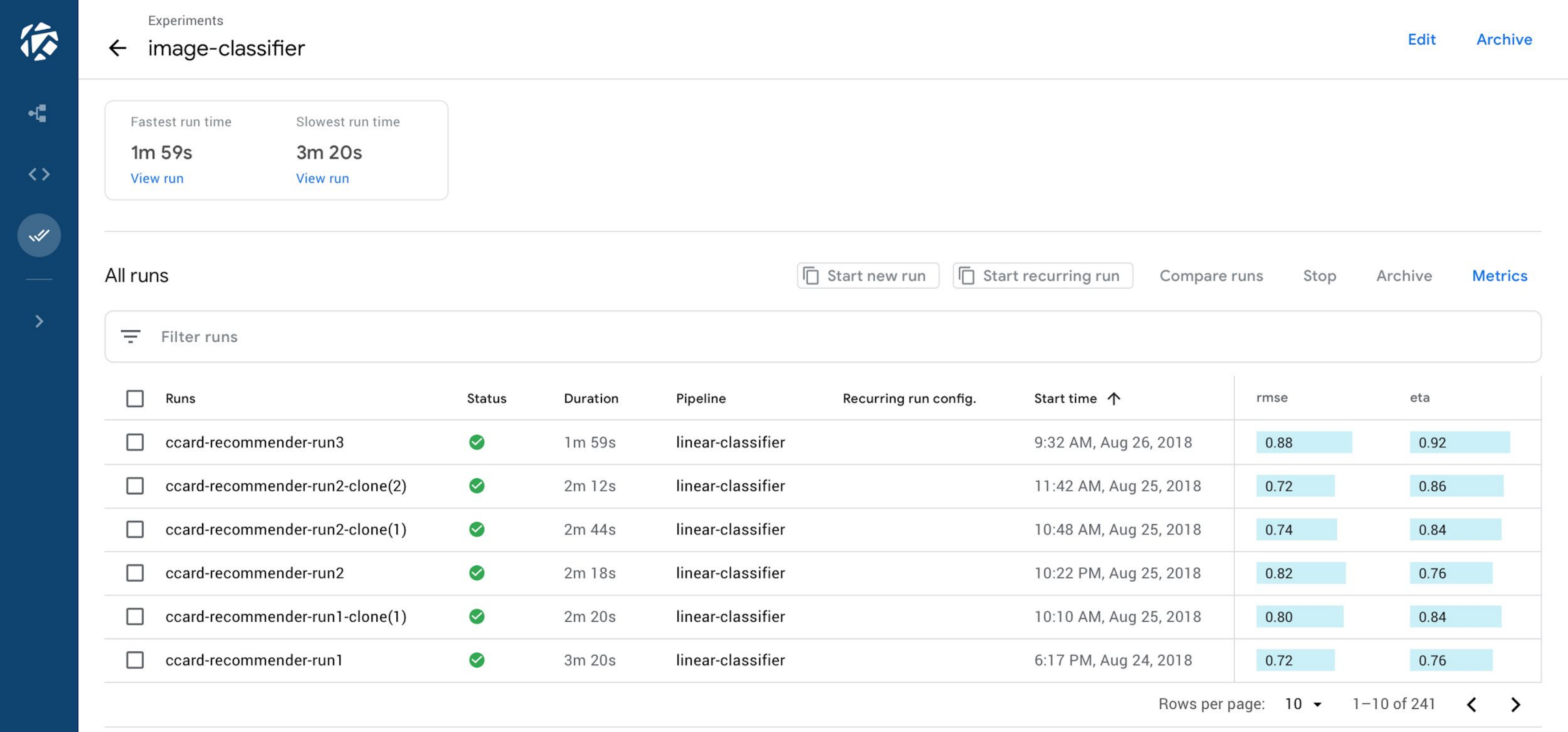
target
resolution

rounds
200

workers
2

true-label
ACTION

Update parameters and submit



Easy comparison of Runs

Experiments

image-classifier

Edit Archive

Fastest run time	Slowest run time
1m 59s	3m 20s
View run	View run

All runs

Start new run | Start recurring run | Compare runs | Stop | Archive | Metrics

Filter runs

Runs	Status	Duration	Pipeline	Recurring run config.	Start time ↑	rmse	eta
ccard-recommender-run3	✓	1m 59s	linear-classifier		9:32 AM, Aug 26, 2018	0.88	0.92
ccard-recommender-run2-clone(2)	✓	2m 12s	linear-classifier		11:42 AM, Aug 25, 2018	0.72	0.86
ccard-recommender-run2-clone(1)	✓	2m 44s	linear-classifier		10:48 AM, Aug 25, 2018	0.74	0.84
ccard-recommender-run2	✓	2m 18s	linear-classifier		10:22 PM, Aug 25, 2018	0.82	0.76
ccard-recommender-run1-clone(1)	✓	2m 20s	linear-classifier		10:10 AM, Aug 25, 2018	0.80	0.84
ccard-recommender-run1	✓	3m 20s	linear-classifier		6:17 PM, Aug 24, 2018	0.72	0.76

Rows per page: 10 1–10 of 241

Easy comparison of Runs

Experiments > creditcard-recommender

Compare runs

Run overview

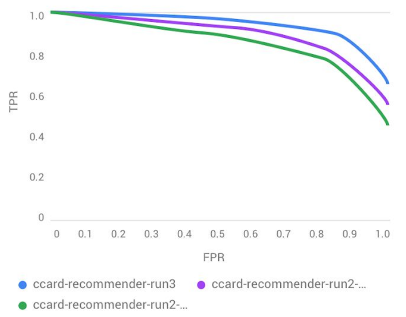
Show	Run name	Status	Pipeline	Duration
☑	ccard-recommender-run3	✔	linear-classifier	3m 20s
☑	ccard-recommender-run2-clone(2)	✔	linear-classifier	3m 20s
☑	ccard-recommender-run2-clone(1)	✔	linear-classifier	3m 20s

Parameters

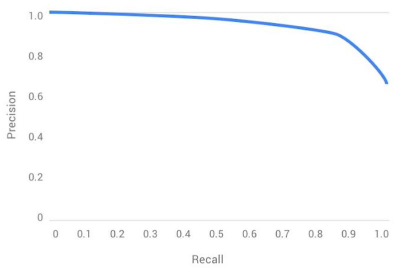
Metrics

Precision Recall

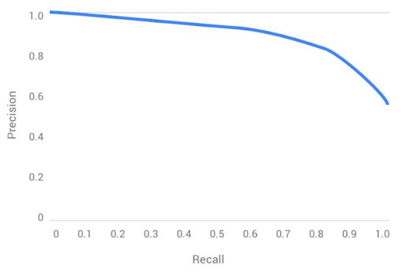
All selected runs



ccard-recommender-run3



ccard-recommender-run2-clone(2)

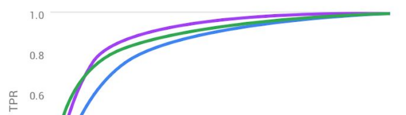


cc

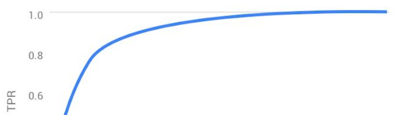
Precision

ROC curve

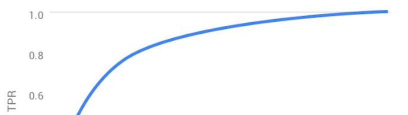
All selected runs



ccard-recommender-run3



ccard-recommender-run2-clone(2)



cc

TPR

That’s a wrap.

Google Cloud