TensorFlow

TensorFlow Extended (TFX)

(and a little bit of TensorFlow Lite)

@ BigDataX Singapore : 13-July-2019

Martin Andrews, ML GDE

TensorFlow

Martin Andrews

Google Developer Expert, Machine Learning

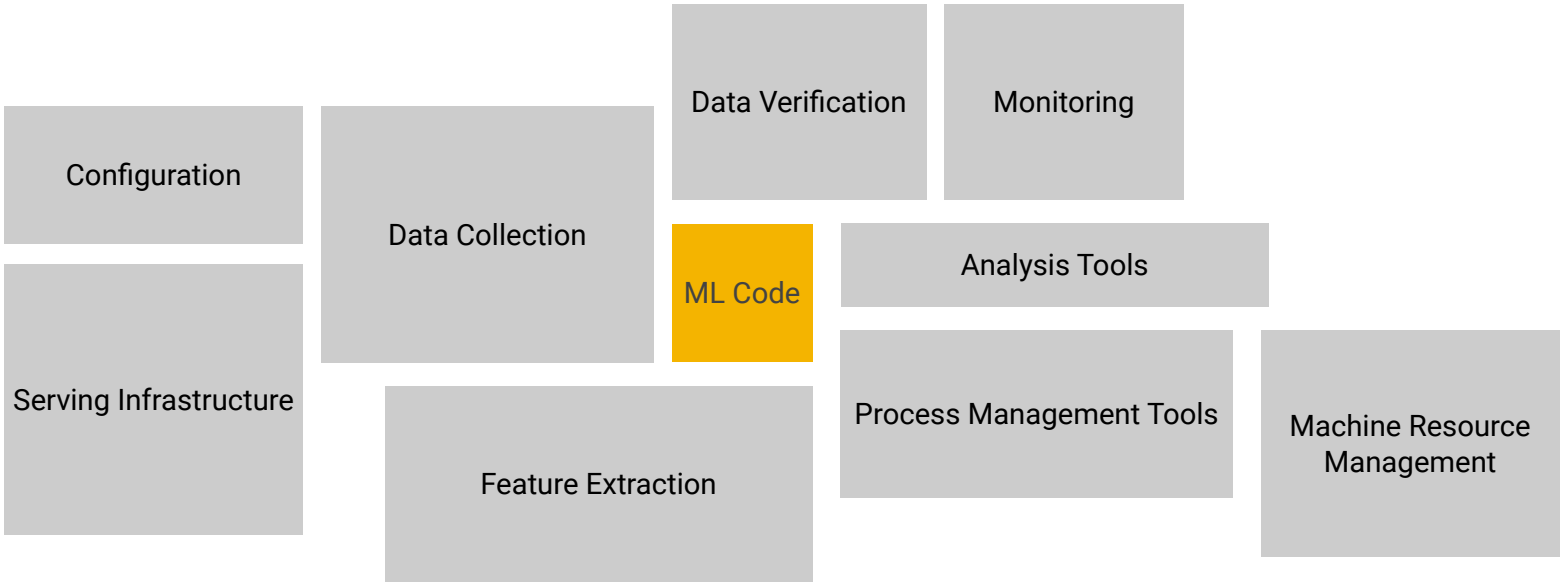
Red Dragon AI, Singapore

Outline

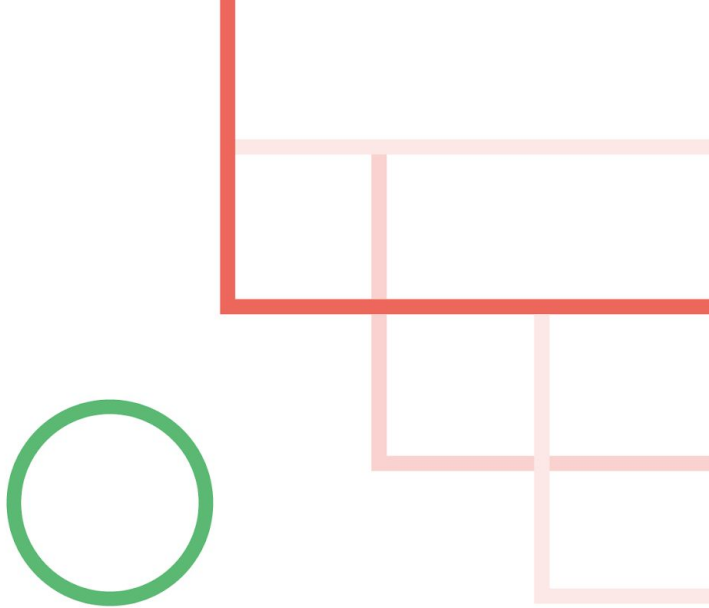
- Machine Learning for Production
  - Part of a Bigger Picture
- How the components are joined together
- What all the components do :
  - Data ingestion ...
  - ... ? ...
  - ... to serving (and TFlite)
- Wrap-up

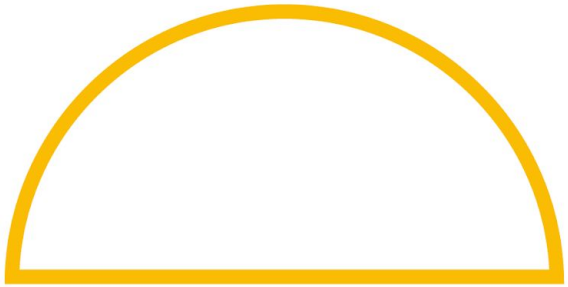
In addition to training a model ...

… a production solution requires so much more



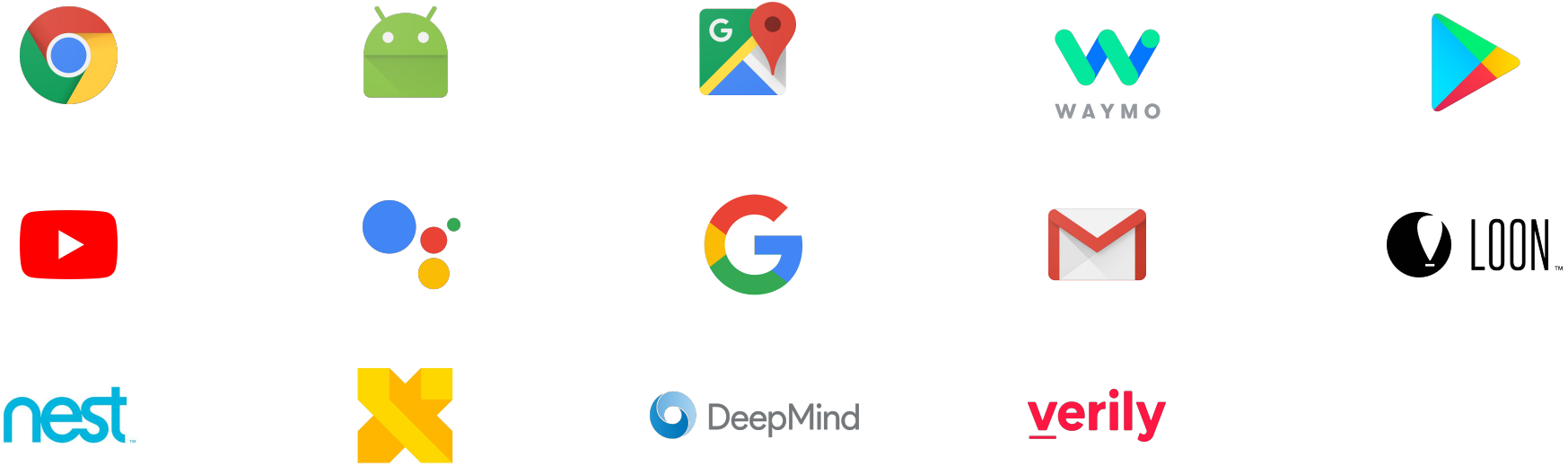
Tensorflow Extended (TFX)



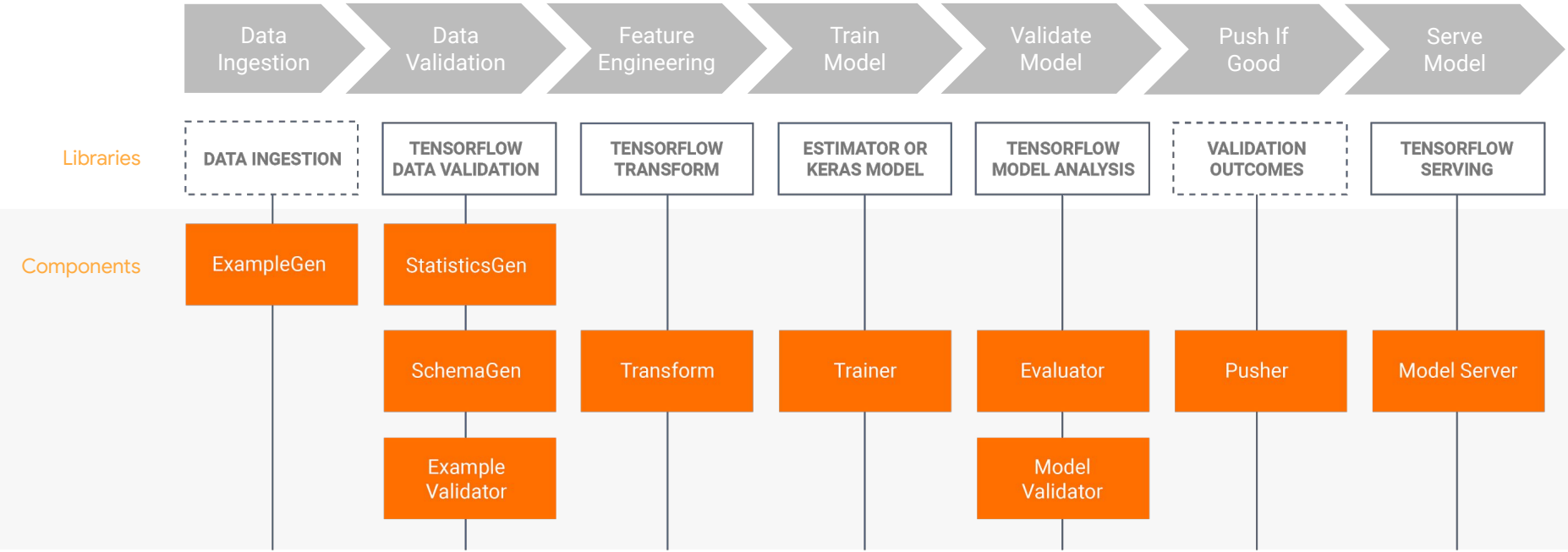


Tensorflow Extended (TFX)

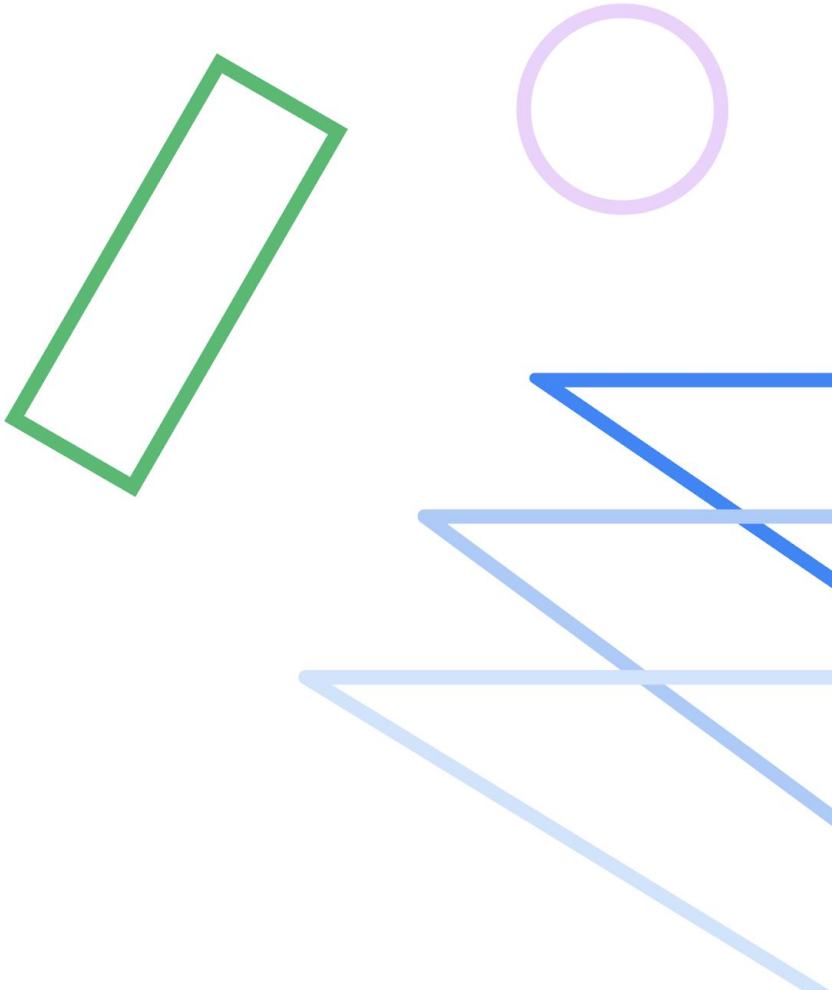
Powers Alphabet's most important bets and products

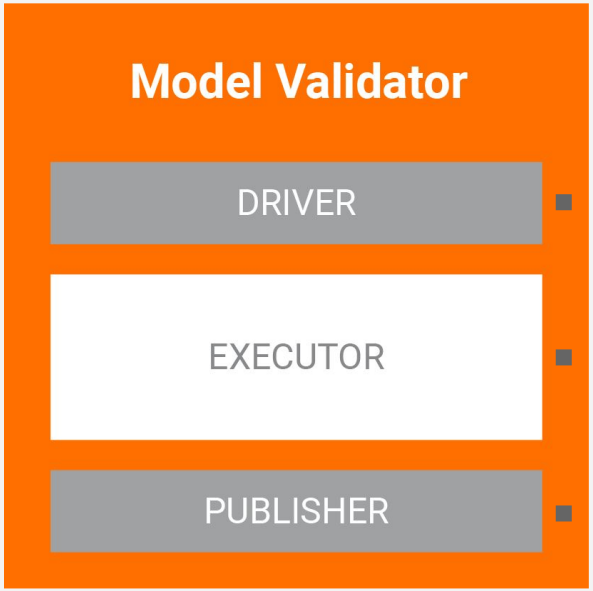


TFX Production Components



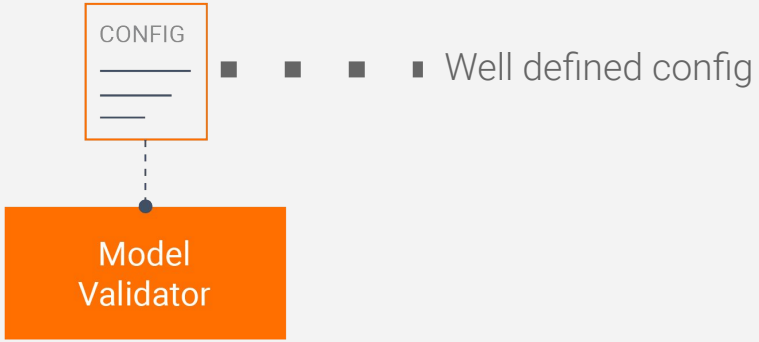
What is a Component?



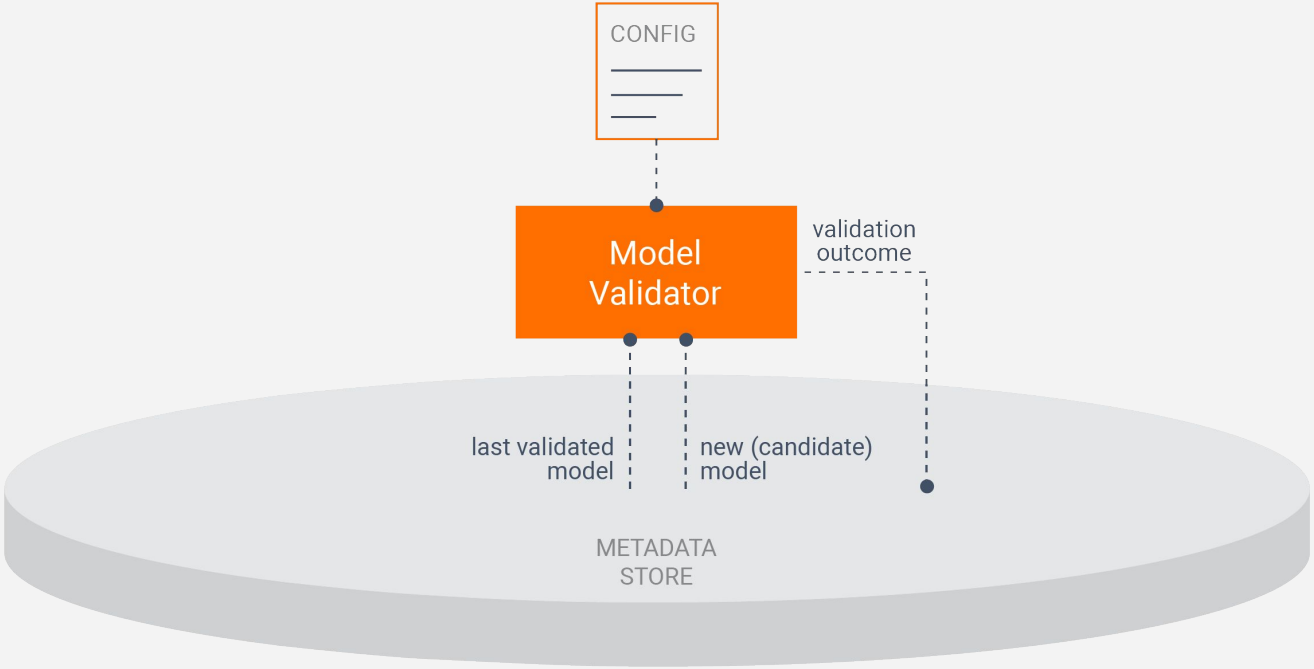


- Coordinates job execution
- Performs the work
- Updates ml.metadata

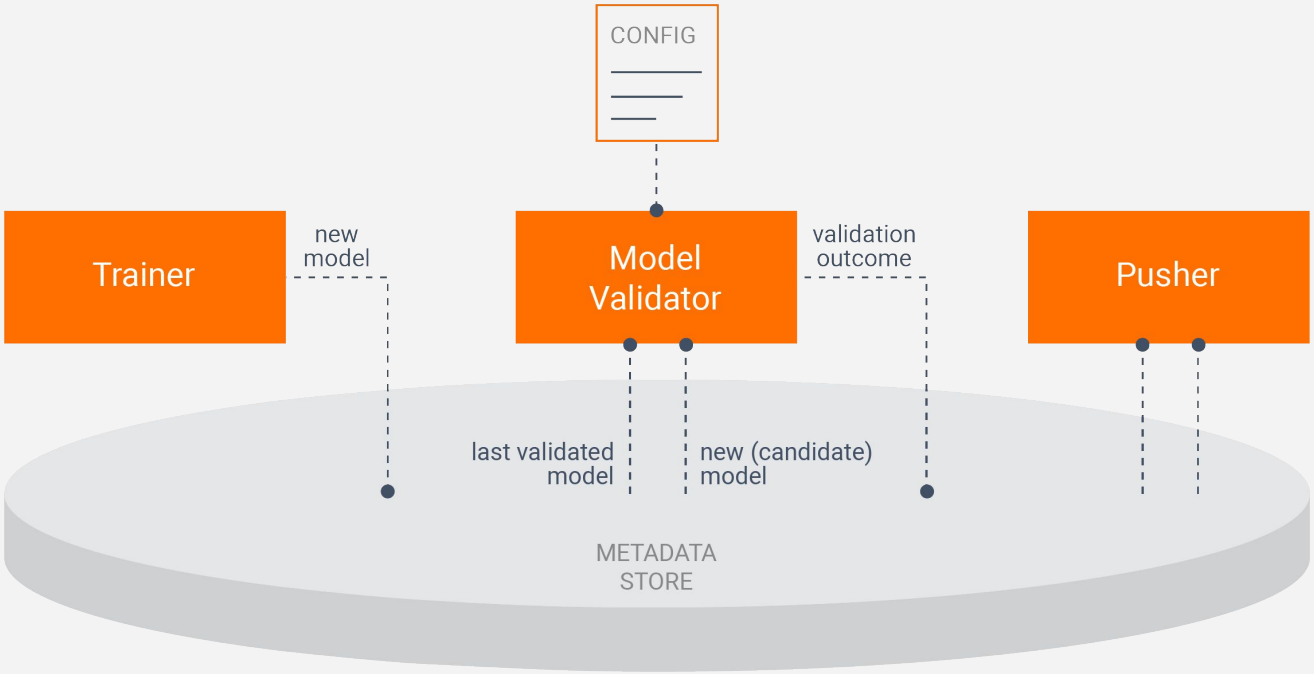
What makes a Component



What makes a Component?

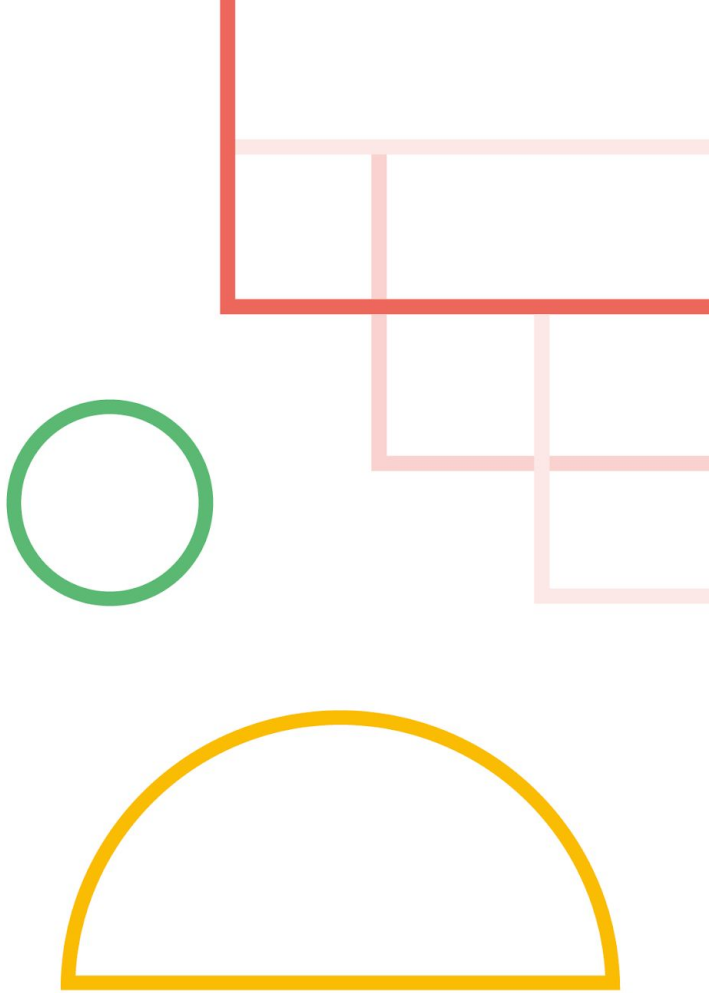


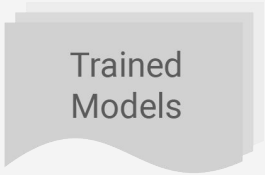
What makes a Component?



What makes a Component?

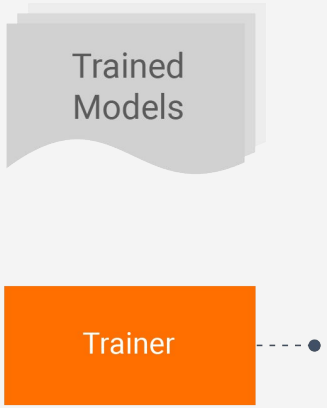
TFX: Metadata Store
What does it contain?





What is in Metadata Store?

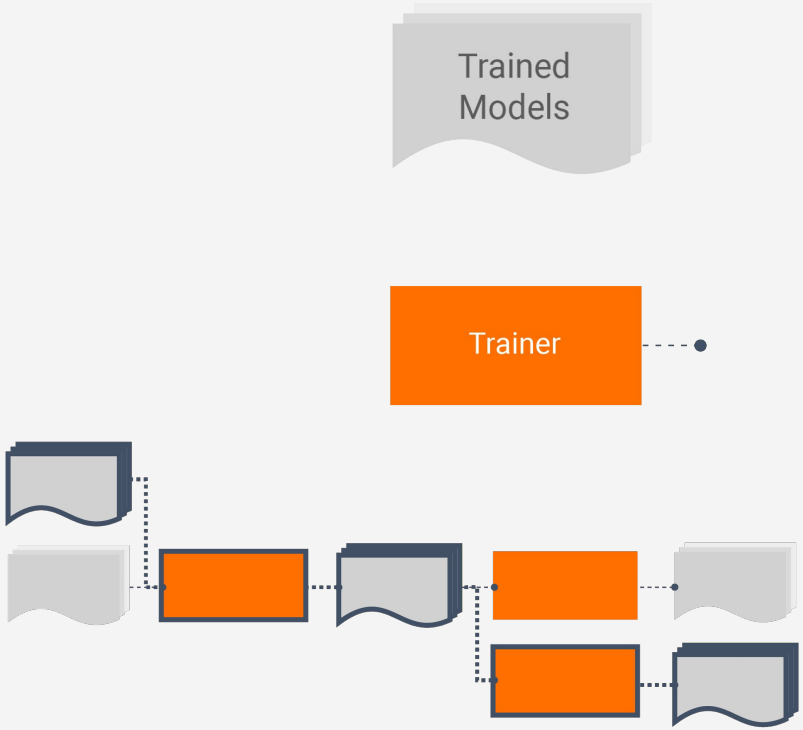
Type definitions of Artifacts and their Properties



What is in Metadata Store?

Type definitions of Artifacts and their Properties

Execution Records (Runs) of Components



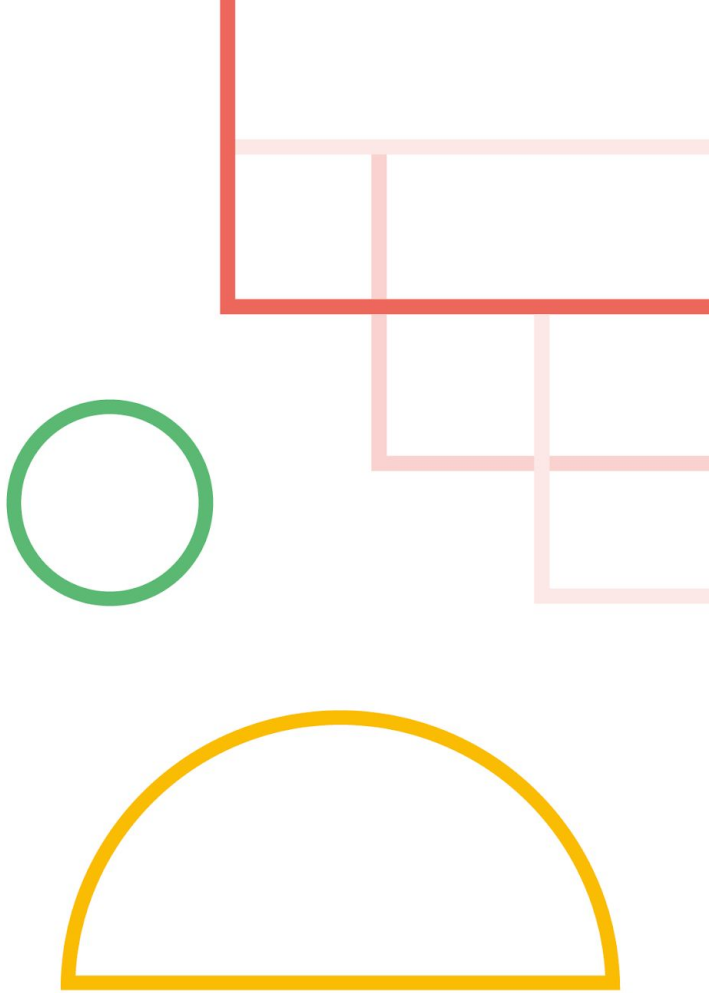
What is in Metadata Store?

Type definitions of Artifacts and their Properties

Execution Records (Runs) of Components

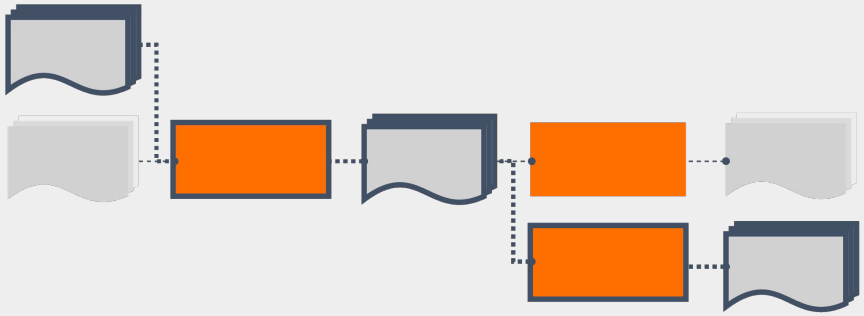
Data Provenance Across All Executions

Metadata-Powered Functionality



Metadata-Powered Functionality

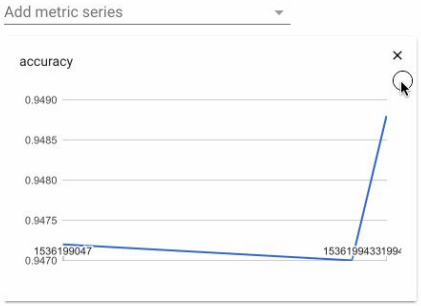
Find out which data a model was trained on



Metadata-Powered Functionality

Compare previous model runs

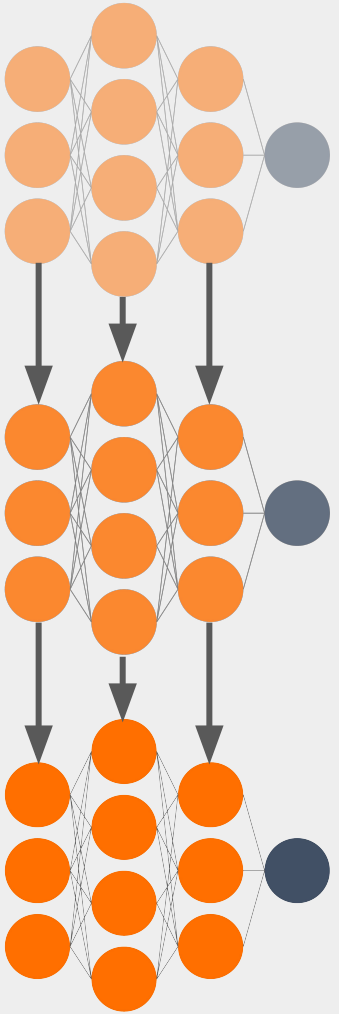
```
eval_results = tfma.make_eval_results([tfma_result_1, tfma_result_2, tfma_result_3],
                                      tfma.constants.MODEL_CENTRIC_MODE)
tfma.view.render_time_series(eval_results, OVERALL_SLICE_SPEC)
```



Model	Data	accuracy	accuracy_baseline	auc	auc_precision_recall	average_loss	label/mean	po
1536199479	data.csv	0.94880	0.94220	0.93168	0.98516	0.13980	0.94220	
1536199433	data.csv	0.94700	0.94220	0.93165	0.98170	0.13979	0.94220	
1536199047	data.csv	0.94720	0.94220	0.92914	0.99480	0.14103	0.94220	

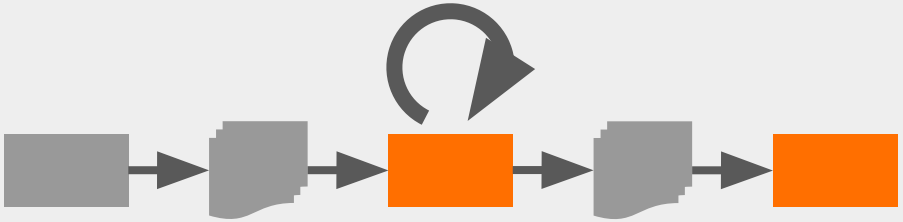
Metadata-Powered Functionality

Carry-over state from previous model runs

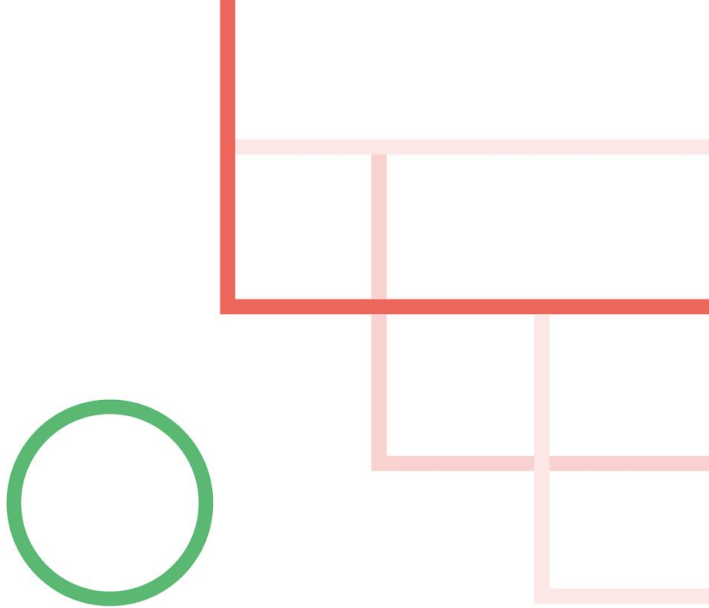


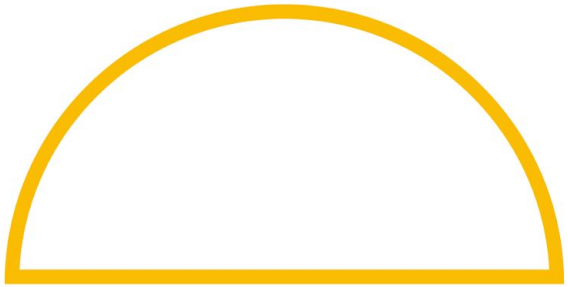
Metadata-Powered Functionality

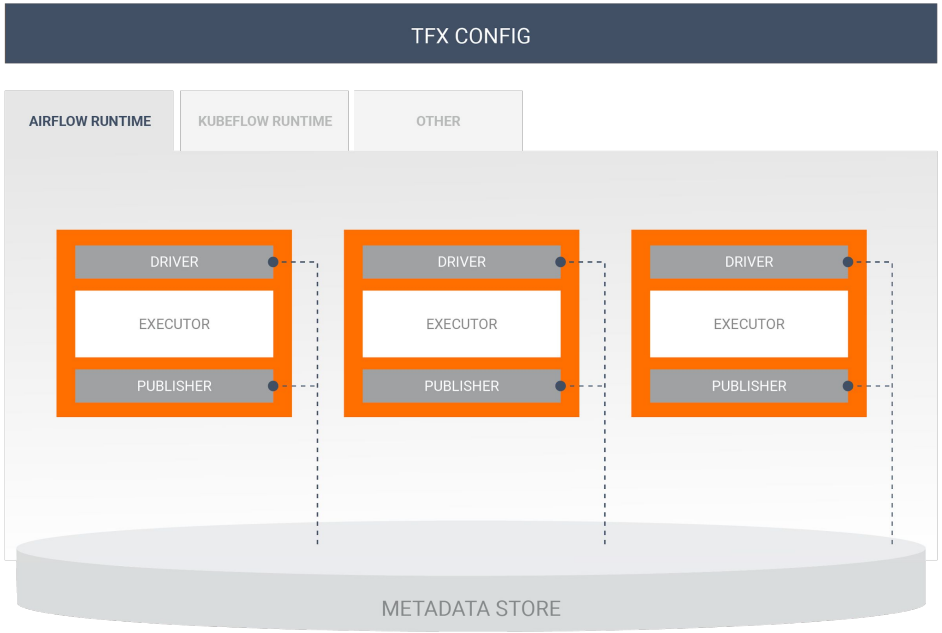
Re-use previously computed outputs



TFX Orchestration



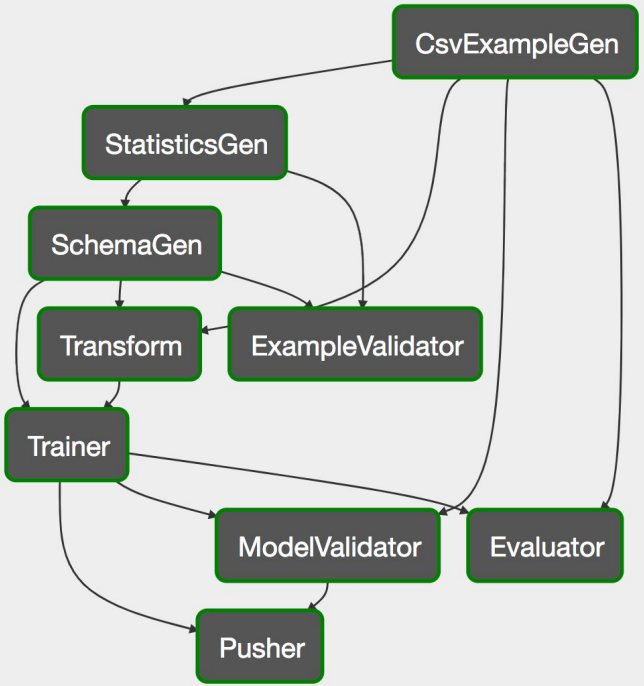
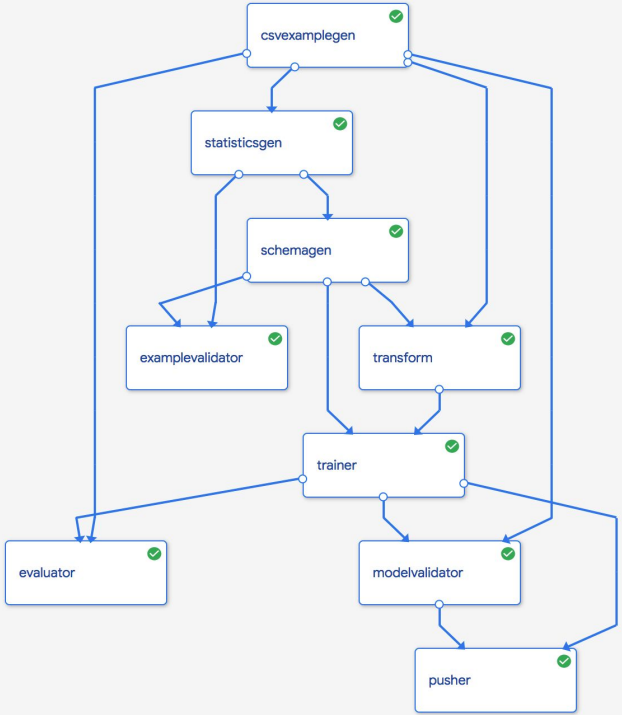




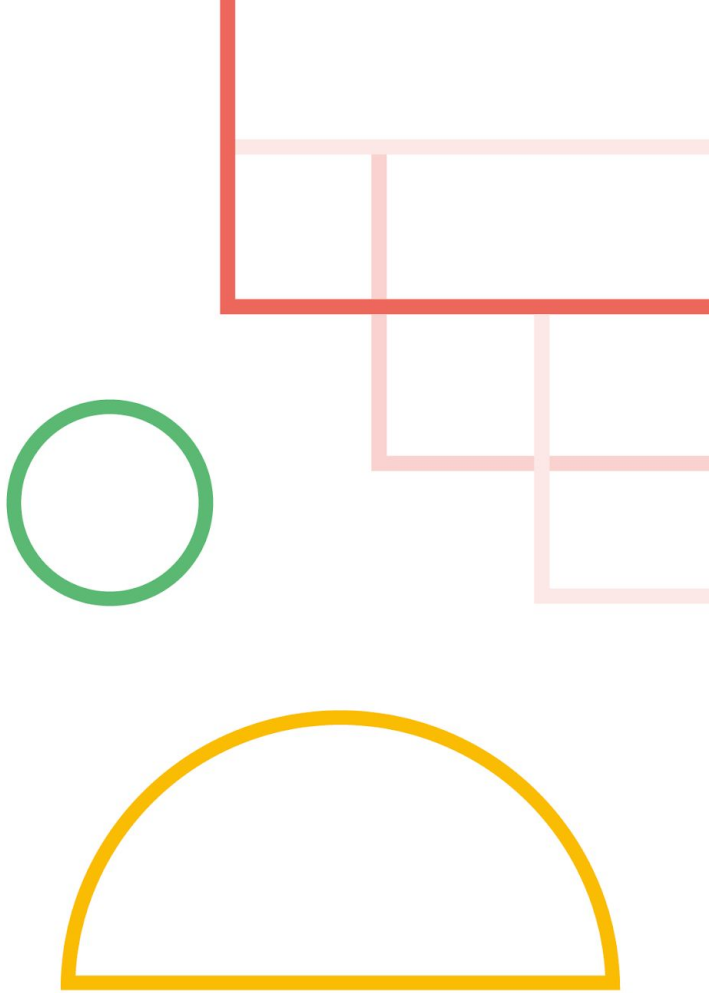
Bring your own Orchestrator

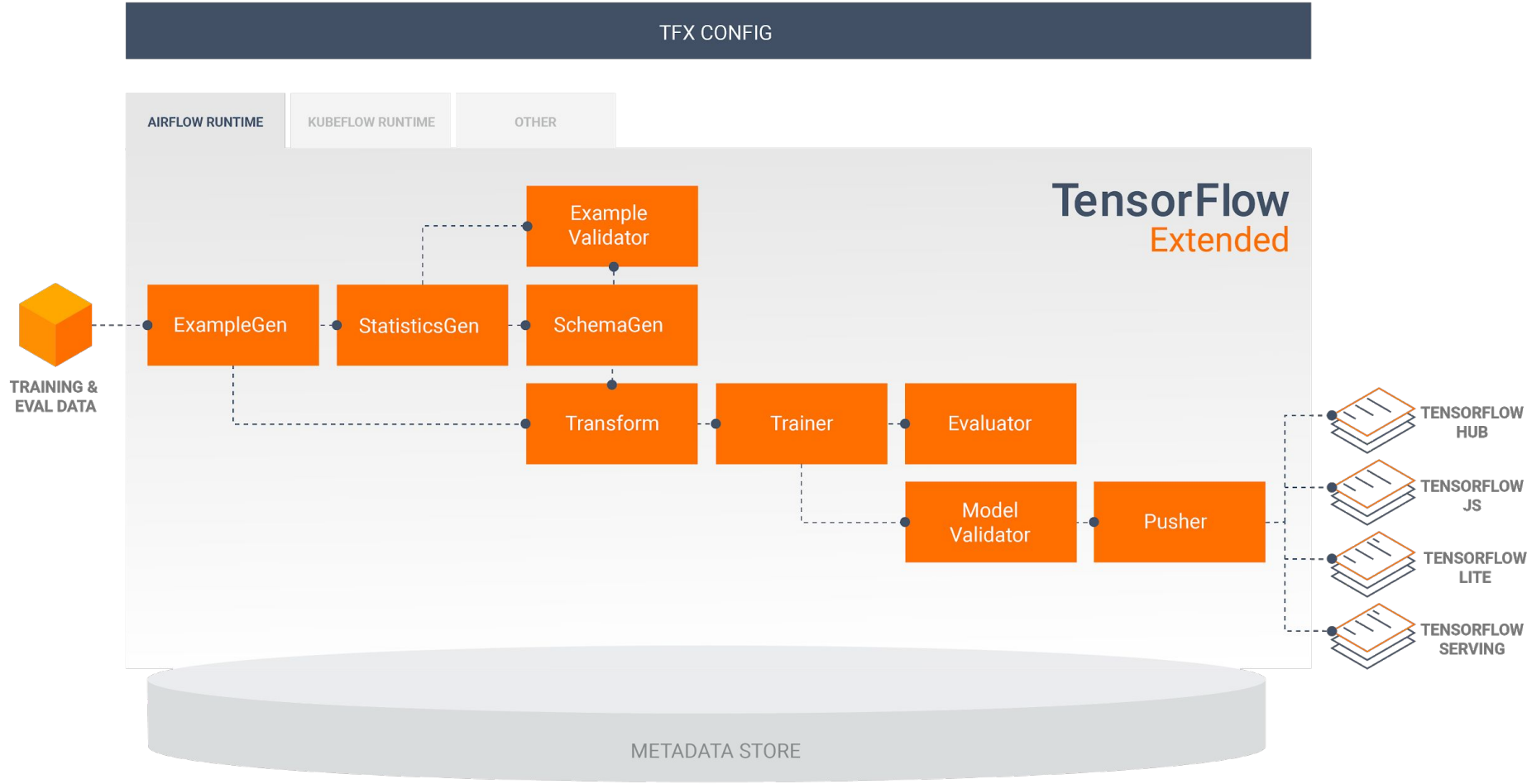
Flexible runtimes run components in the proper order using orchestration systems such as Airflow or Kubeflow

Orchestrators and DAGs

Airflow	Kubeflow Pipelines
	

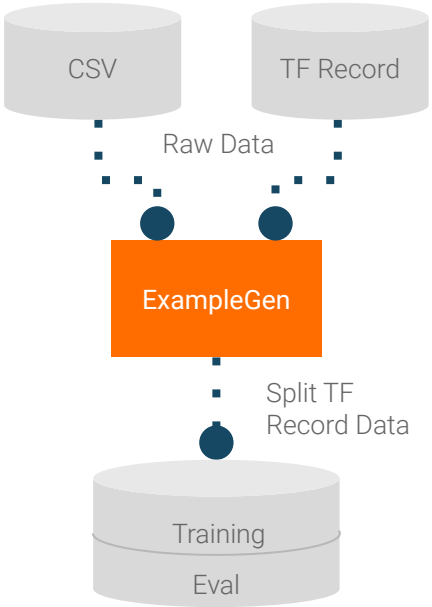
All the Components




TFX CONFIG
AIRFLOW RUNTIME
KUBEFLOW RUNTIME
OTHER
TensorFlow
Extended
Example Validator
ExampleGen
StatisticsGen
SchemaGen
TRAINING & EVAL DATA
Transform
Trainer
Evaluator
Model Validator
Pusher
TENSORFLOW HUB
TENSORFLOW JS
TENSORFLOW LITE
TENSORFLOW SERVING
METADATA STORE

Component: ExampleGen

Inputs and Outputs



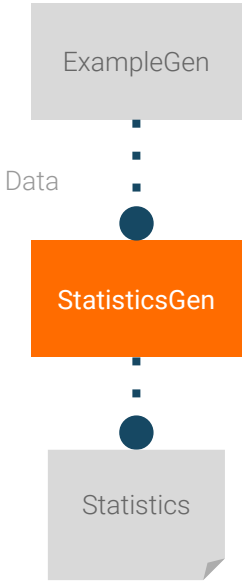
Configuration

```
examples = csv_input(os.path.join(data_root, 'simple'))

example_gen = CsvExampleGen(input_base=examples)
```

beam

Component: StatisticsGen

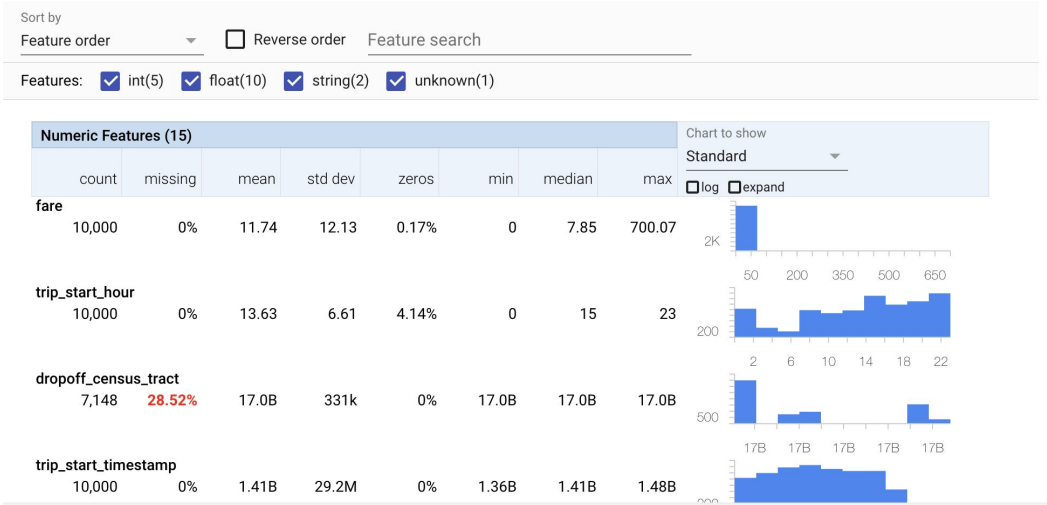
Inputs and Outputs



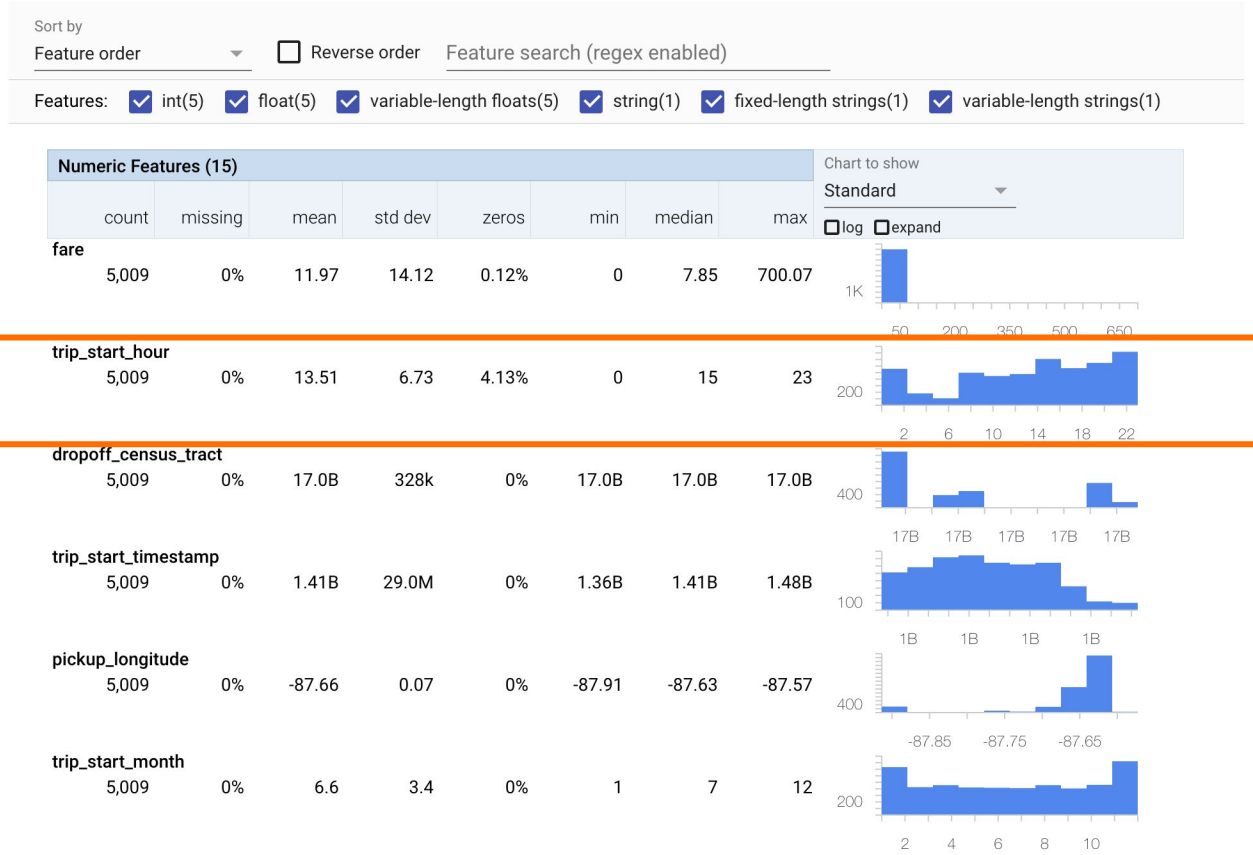
Configuration

```
statistics_gen =
  StatisticsGen(input_data=example_gen.outputs.examples)
```

Visualization

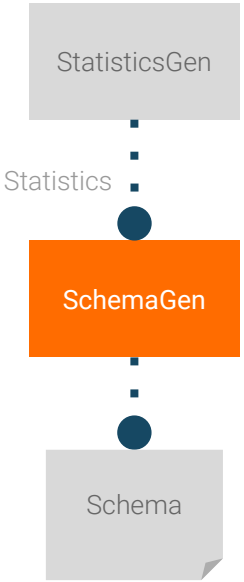


Analyzing Data with TensorFlow Data Validation



Component: SchemaGen

Inputs and Outputs



Configuration

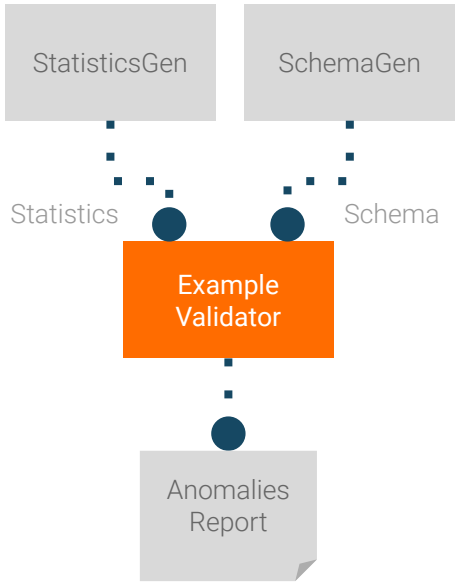
```
infer_schema = SchemaGen(stats=statistics_gen.outputs.output)
```

Visualization

Feature name	Type	Presence	Valency	Domain
'fare'	FLOAT	required	single	-
'trip_start_hour'	INT	required	single	-
'pickup_census_tract'	BYTES	optional		-
'dropoff_census_tract'	FLOAT	optional	single	-
'company'	STRING	optional	single	'company'

Component: ExampleValidator

Inputs and Outputs



Configuration

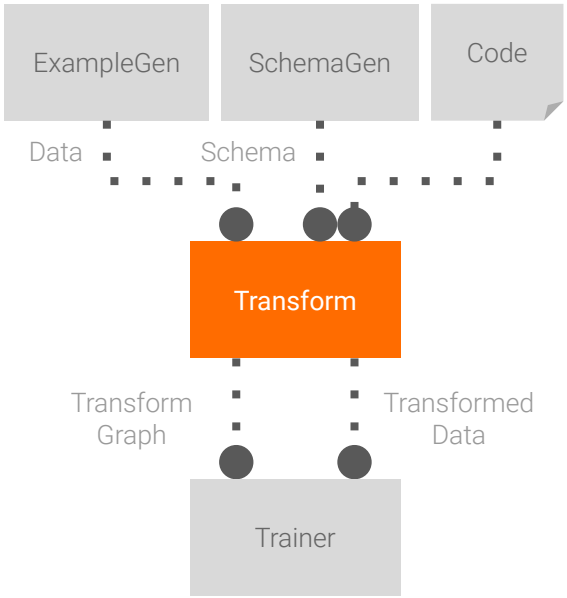
```
validate_stats = ExampleValidator(
    stats=statistics_gen.outputs.output,
    schema=infer_schema.outputs.output)
```

Visualization

Feature name	Anomaly short description	Anomaly long description
'payment_type'	Unexpected string values	Examples contain values missing from the schema: Prcard (<1%).
'company'	Unexpected string values	Examples contain values missing from the schema: 2092 - 61288 Sbeih company (<1%), 2192 - 73487 Zeymane Corp (<1%), 2192 - Zeymane Corp (<1%), 2823 - 73307 Seung Lee (<1%), 3094 - 24059 G.L.B. Cab Co (<1%), 3319 - CD Cab Co (<1%), 3385 - Eman Cab (<1%), 3897 - 57856 Ilie Malec (<1%), 4053 - 40193 Adwar H. Nikola (<1%), 4197 - Royal Star (<1%), 585 - 88805 Valley Cab Co (<1%), 5874 - Sergey Cab Corp. (<1%), 6057 - 24657 Richard Addo (<1%), 6574 - Babylon Express Inc. (<1%), 6742 - 83735 Tasha ride inc (<1%).

Component: Transform

Inputs and Outputs



beam

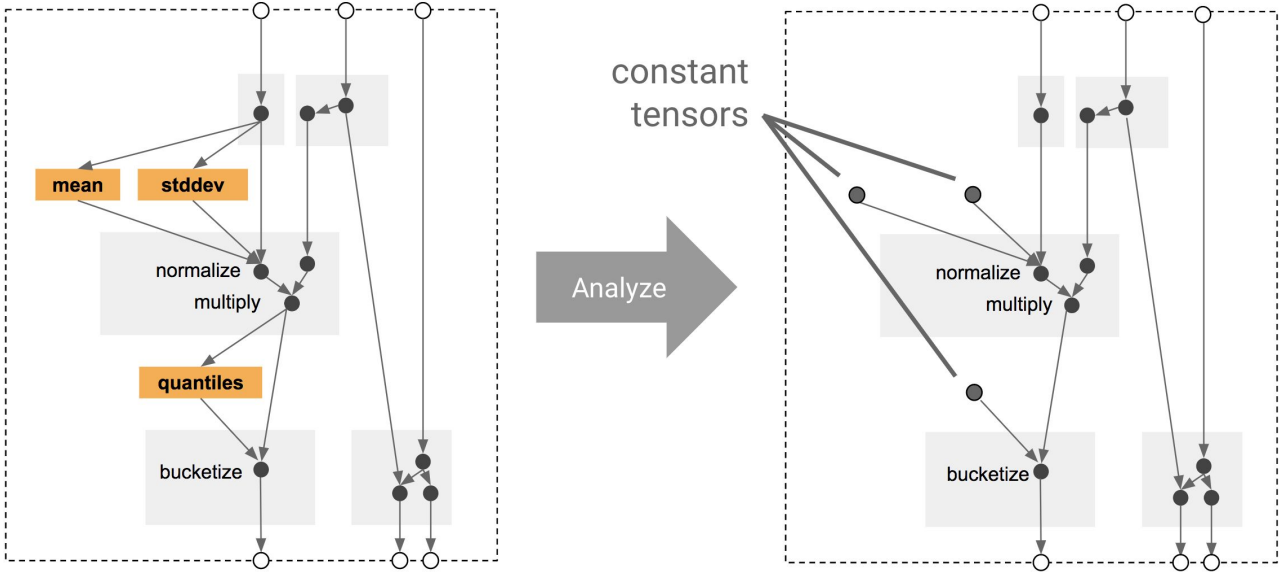
Configuration

```
transform = Transform(
    input_data=example_gen.outputs.examples,
    schema=infer_schema.outputs.output,
    module_file=taxi_module_file)
```

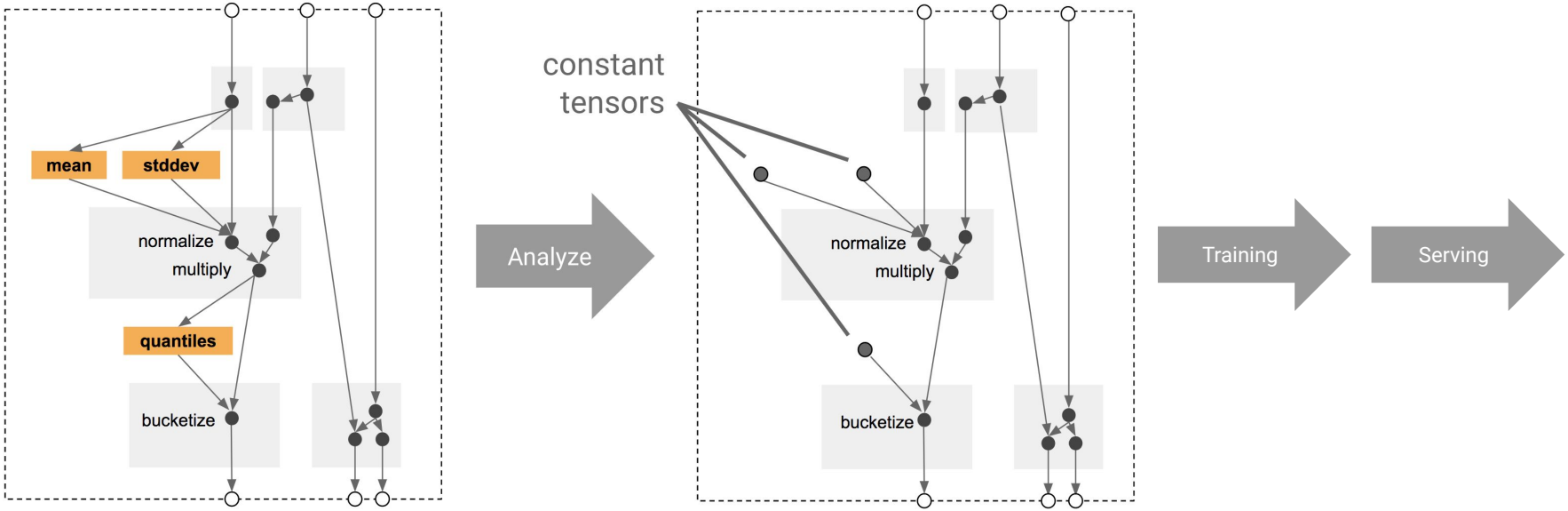
Code

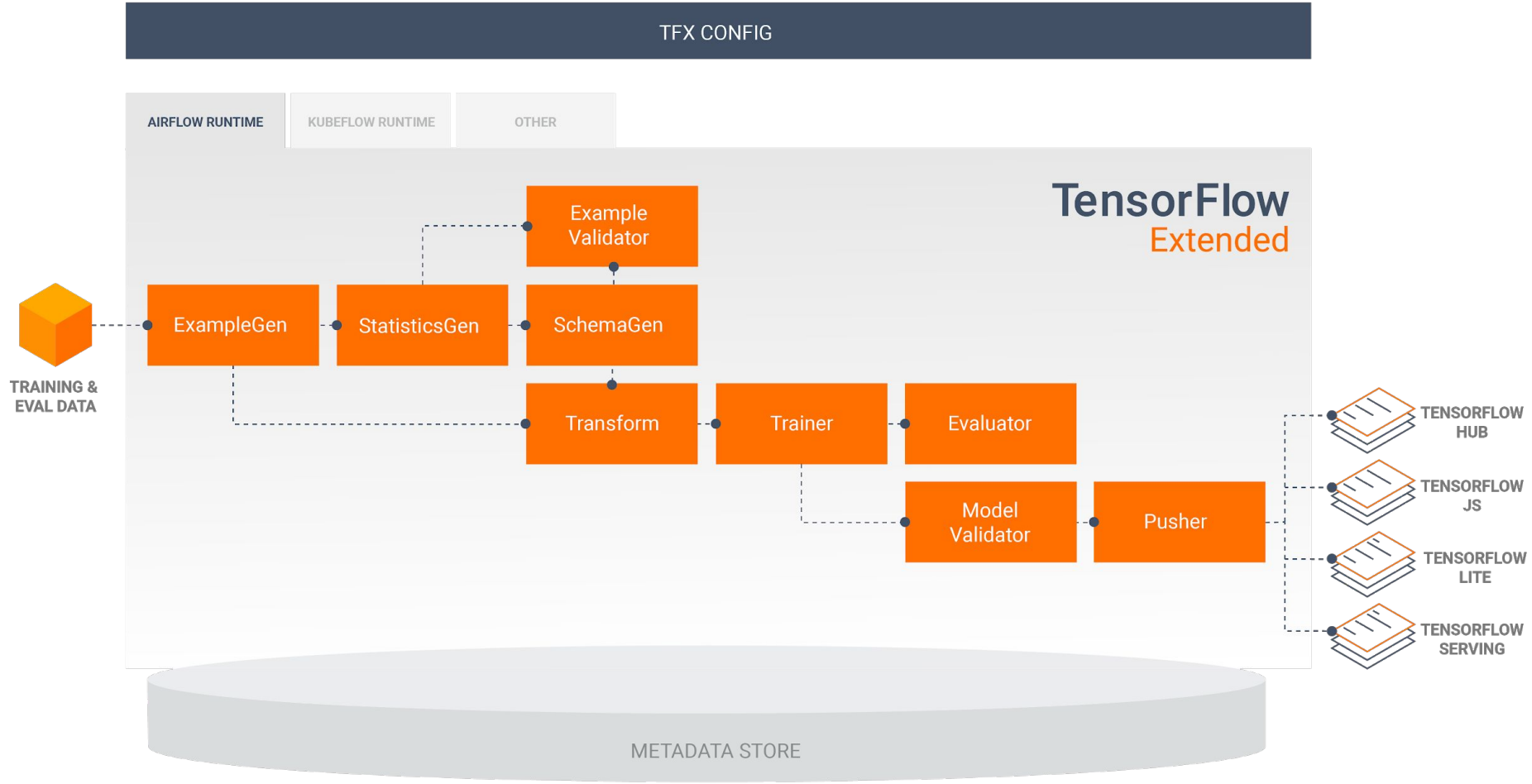
```
for key in _DENSE_FLOAT_FEATURE_KEYS:
    outputs[_transformed_name(key)] = transform.scale_to_z_score(
        _fill_in_missing(inputs[key]))
# ...

outputs[_transformed_name(_LABEL_KEY)] = tf.where(
      tf.is_nan(taxi_fare),
      tf.cast(tf.zeros_like(taxi_fare), tf.int64),
      # Test if the tip was > 20% of the fare.
      tf.cast(
          tf.greater(tips, tf.multiply(taxi_fare, tf.constant(0.2))), tf.int64))

# ...
```

Using TensorFlow Transform for Feature Engineering



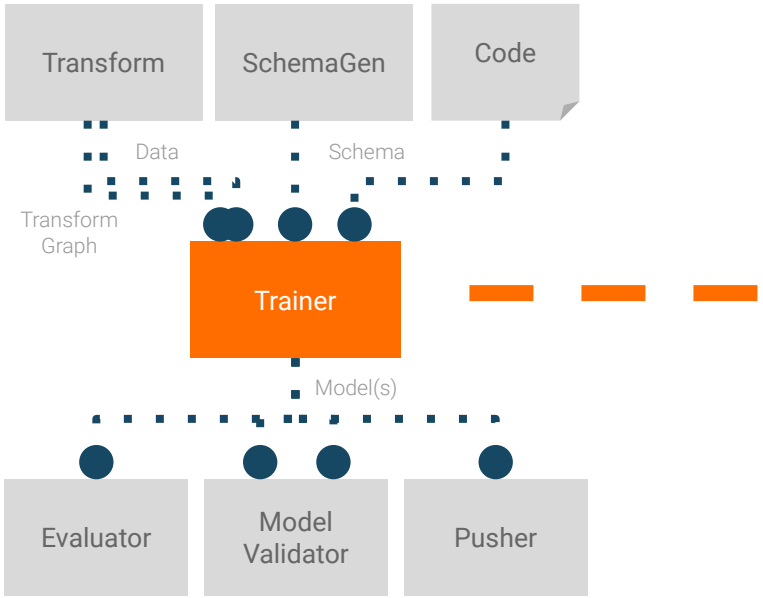
Using TensorFlow Transform for Feature Engineering




TFX CONFIG
AIRFLOW RUNTIME
KUBEFLOW RUNTIME
OTHER
TensorFlow
Extended
Example Validator
ExampleGen
StatisticsGen
SchemaGen
TRAINING & EVAL DATA
Transform
Trainer
Evaluator
Model Validator
Pusher
TENSORFLOW HUB
TENSORFLOW JS
TENSORFLOW LITE
TENSORFLOW SERVING
METADATA STORE

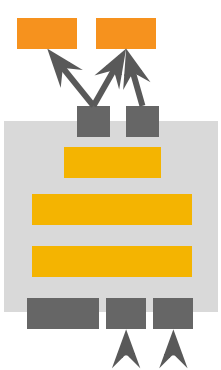
Component: Trainer

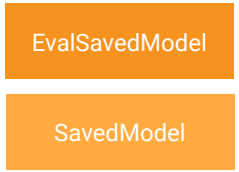
Inputs and Outputs



Highlight: SavedModel Format

Train, Eval, and Inference Graphs



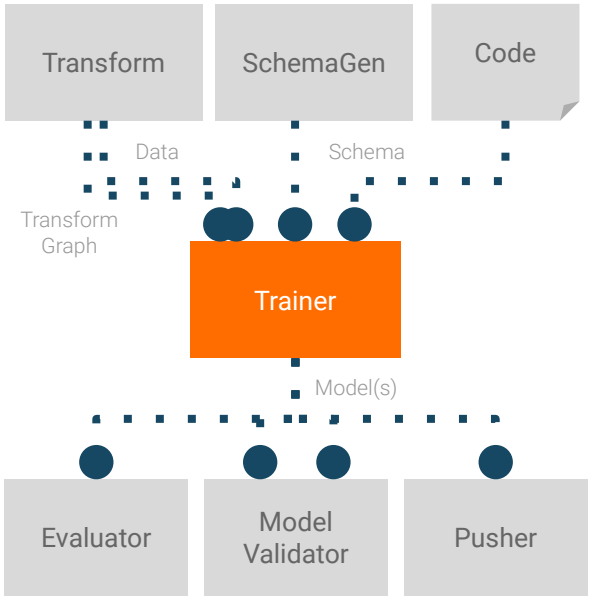


TensorFlow Model Analysis

TensorFlow Serving

Component: Trainer

Inputs and Outputs



Configuration

```
trainer = Trainer(
    module_file=taxi_module_file,
    transformed_examples=transform.outputs.transformed_examples,
    schema=infer_schema.outputs.output,
    transform_output=transform.outputs.transform_output,
    train_steps=10000,
    eval_steps=5000,
    warm_starting=True)
```

Code

Just TensorFlow :)

```
# Open up Tensorboard for model_id.
print(display_tensorboard(model_id))
```

```
http://your.host.name:53143
```

TensorBoard SCALARS GRAPHS DISTRIBUTIONS HISTOGRAMS PROJECTOR INACTIVE

Show data download links

Ignore outliers in chart scaling

Tooltip sorting method: default

Smoothing

0.6

Horizontal Axis

STEP RELATIVE WALL

Runs

Write a regex to filter runs

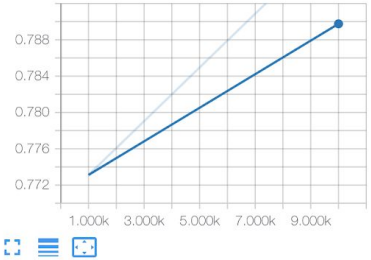
model_8/serving_model_dir

model_8/serving_model_dir/eval_chicago-taxi-eval

Filter tags (regular expressions supported)

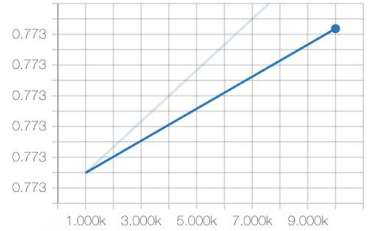
accuracy 1

accuracy



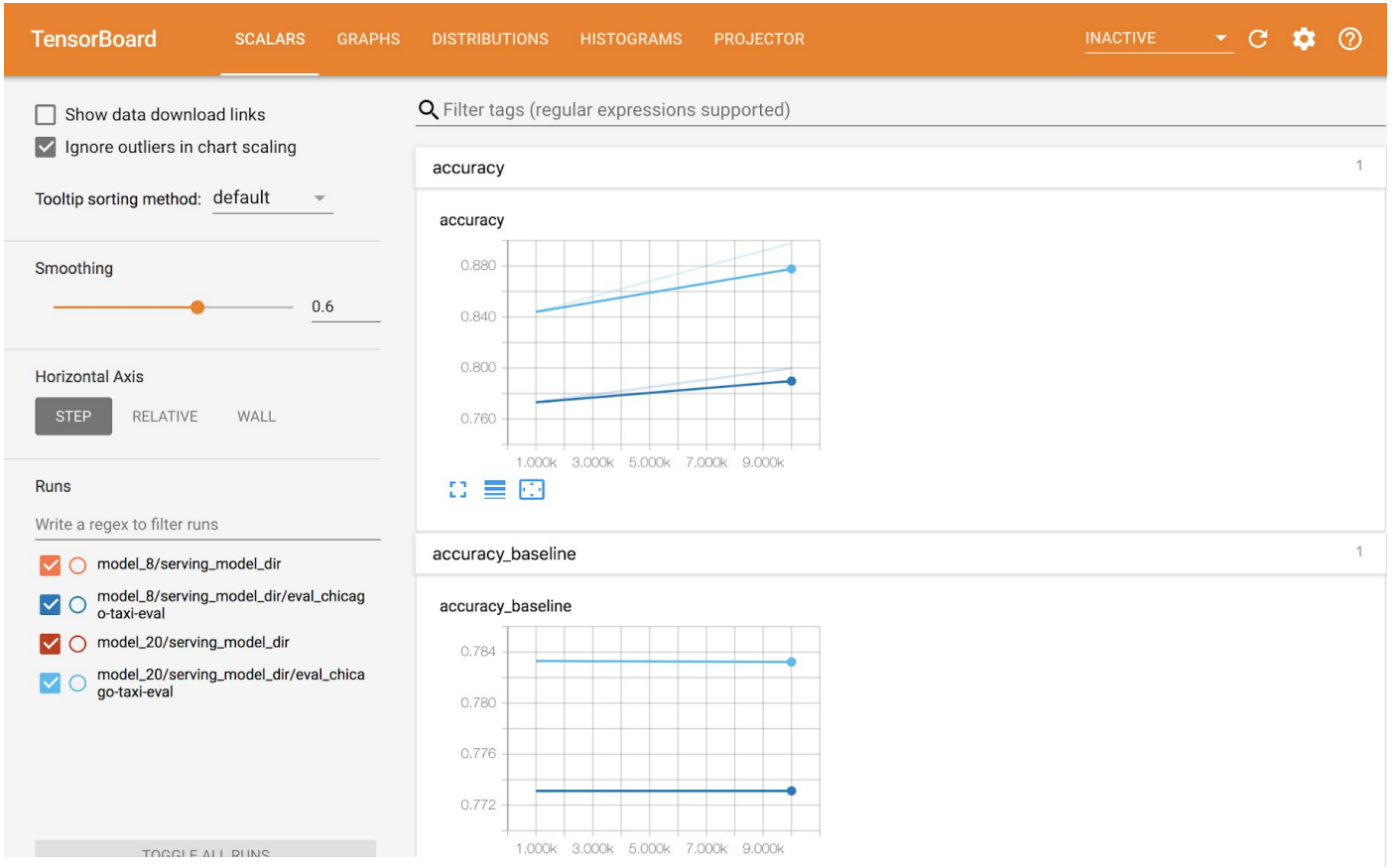
accuracy_baseline 1

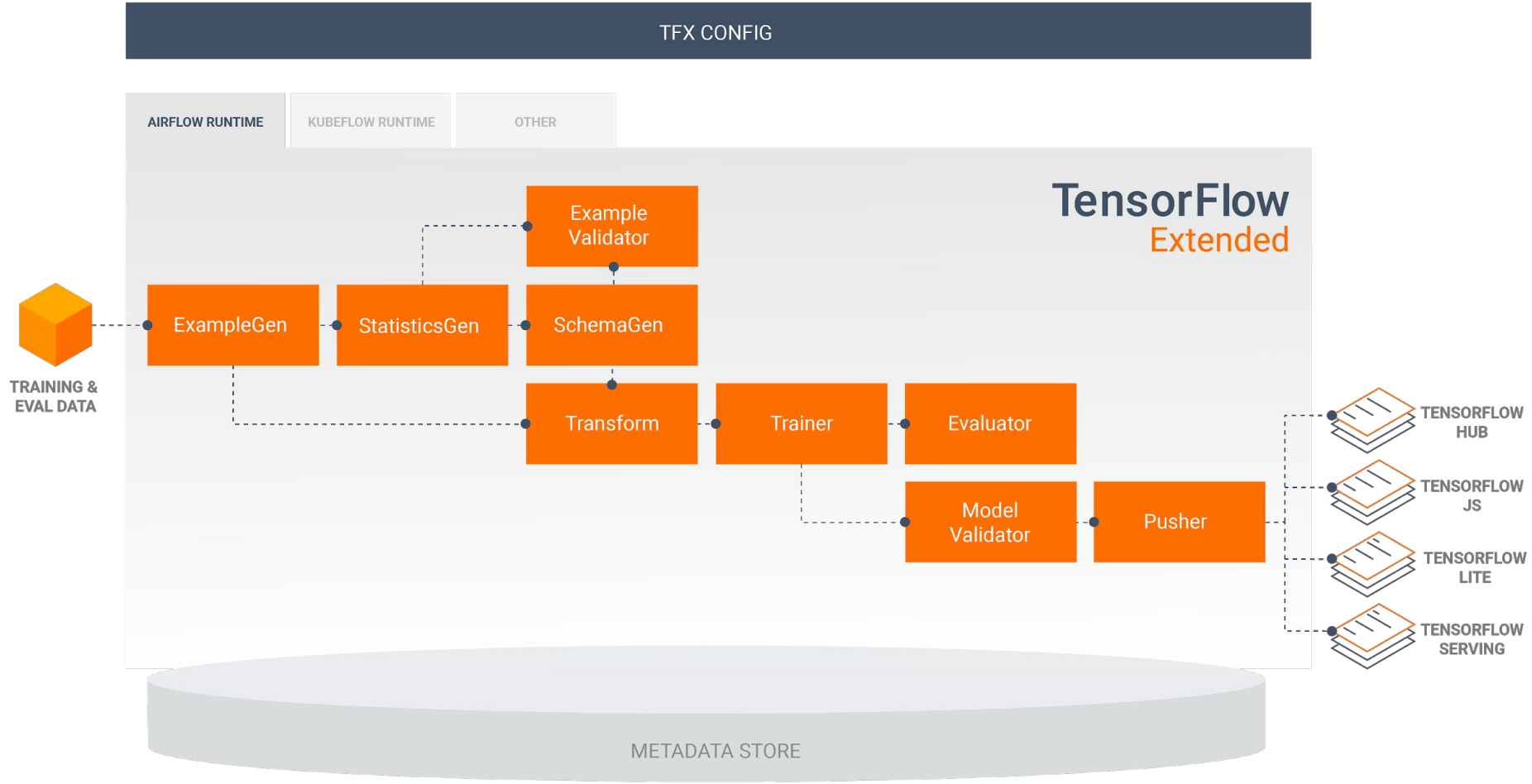
accuracy_baseline



```
# Compare Tensorboard metrics for different models.
if num_models > 1:
    print(display_tensorboard(model_id, other_model_id=other_model_id))
```

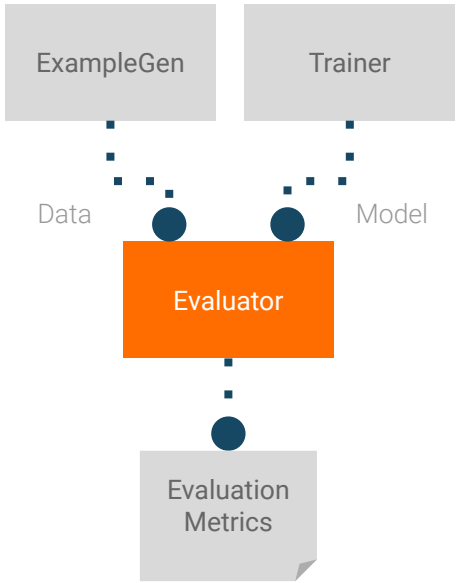
```
http://your.host.name:53230
```




TFX CONFIG
AIRFLOW RUNTIME
KUBEFLOW RUNTIME
OTHER
TensorFlow
Extended
Example Validator
ExampleGen
StatisticsGen
SchemaGen
TRAINING & EVAL DATA
Transform
Trainer
Evaluator
Model Validator
Pusher
TENSORFLOW HUB
TENSORFLOW JS
TENSORFLOW LITE
TENSORFLOW SERVING
METADATA STORE

Component: Evaluator

Inputs and Outputs

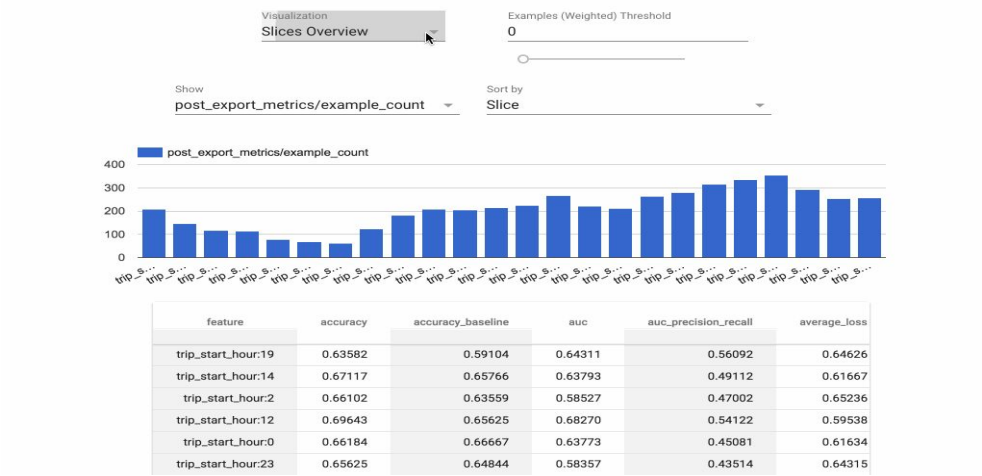


beam

Configuration

```
model_analyzer = Evaluator(
    examples=examples_gen.outputs.output,
    eval_spec=taxi_eval_spec,
    model_exports=trainer.outputs.output)
```

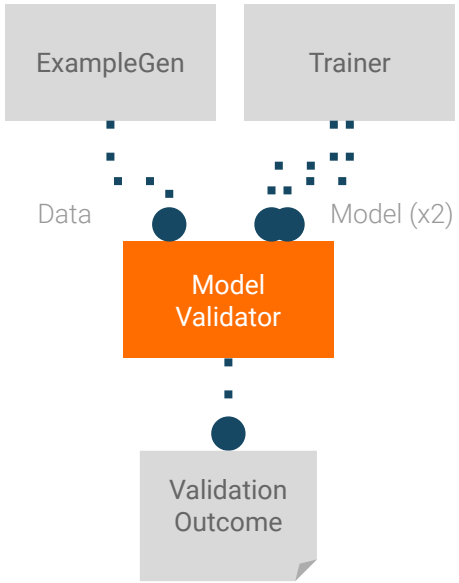
Visualization



feature	accuracy	accuracy_baseline	auc	auc_precision_recall	average_loss
trip_start_hour:19	0.63582	0.59104	0.64311	0.56092	0.64626
trip_start_hour:14	0.67117	0.65766	0.63793	0.49112	0.61667
trip_start_hour:2	0.66102	0.63559	0.58527	0.47002	0.65236
trip_start_hour:12	0.69643	0.65625	0.68270	0.54122	0.59538
trip_start_hour:0	0.66184	0.66667	0.63773	0.45081	0.61634
trip_start_hour:23	0.65625	0.64844	0.58357	0.43514	0.64315

Component: ModelValidator

Inputs and Outputs



Configuration

```
model_validator = ModelValidator(
    examples=examples_gen.outputs.output,
    model=trainer.outputs.output,
    eval_spec=taxi_mv_spec)
```

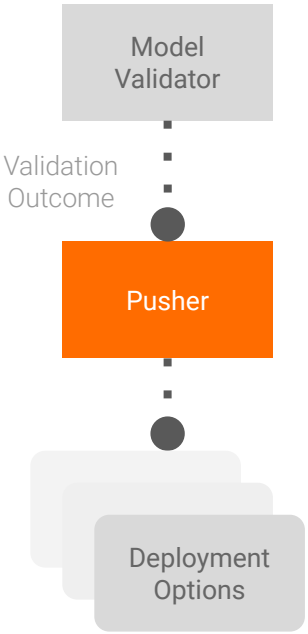
Configuration Options

- Validate using current eval data
- “Next-day eval”, validate using unseen data

beam

Component: Pusher

Inputs and Outputs



Configuration

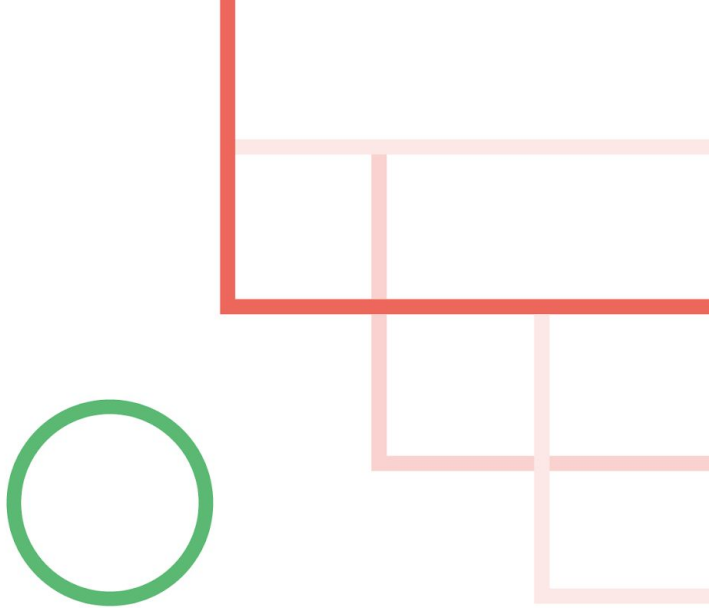
```
pusher = Pusher(
    model_export=trainer.outputs.output,
    model_blessing=model_validator.outputs.blessing,
    serving_model_dir=serving_model_dir)
```

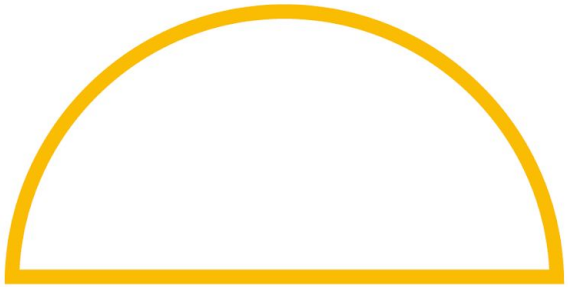
Block push on validation outcome

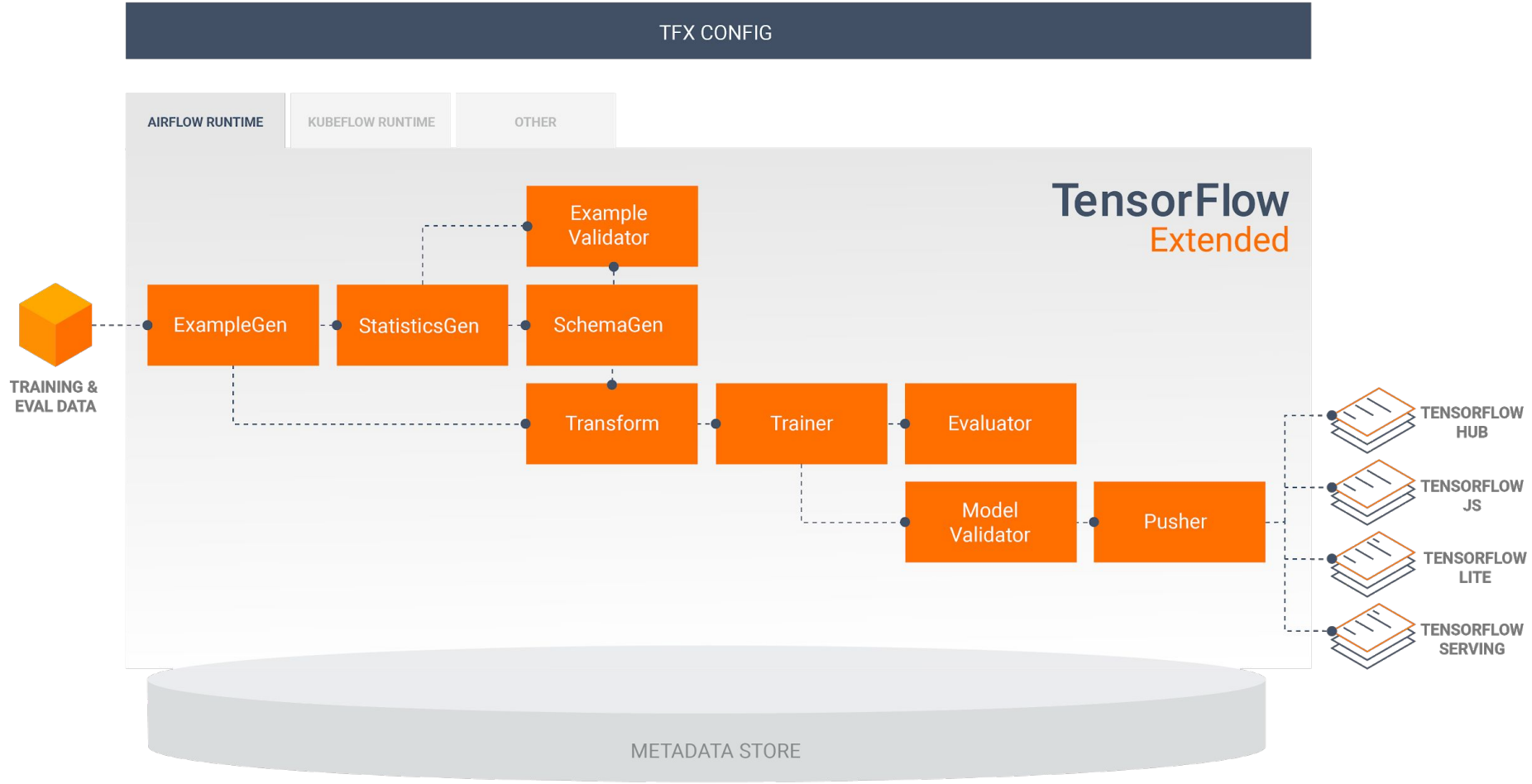
Push destinations supported today

- Filesystem (TensorFlow Lite, TensorFlow JS)
- TensorFlow Serving

Serve the Model !






TFX CONFIG
AIRFLOW RUNTIME
KUBEFLOW RUNTIME
OTHER
TensorFlow
Extended
Example Validator
ExampleGen
StatisticsGen
SchemaGen
TRAINING & EVAL DATA
Transform
Trainer
Evaluator
Model Validator
Pusher
TENSORFLOW HUB
TENSORFLOW JS
TENSORFLOW LITE
TENSORFLOW SERVING
METADATA STORE

TensorFlow Serving

Production-Ready

- Used for years at Google, millions of QPS
- Scale in minutes
- Dynamic version refresh

TensorFlow Serving

High-Performance

- Low-latency
- Request Batching
- Traffic Isolation

Deploy a REST API for your model in minutes ..

```
$ docker run -p 8501:8501 \
    -v '/path/to/savedmodel':/models/chicago_taxi
    -e MODEL_NAME=chicago_taxi -t tensorflow/serving
```

... using Docker ...

... or locally on your host ...

```
$ apt-get install tensorflow-model-server
$ tensorflow_model_server
    --port=8501
    --model_name=chicago_taxi
    --model_base_path='/path/to/savedmodel'
```

What is TensorFlow Lite?

TensorFlow Lite is a framework for deploying ML on mobile devices and embedded systems

More than 2B devices globally.

Have now deployed **TensorFlow Lite in production.**

Source: https://medium.com/tensorflow/recap-of-the-2019-tensorflow-dev-summit-1b5ede42da8d

Text

Classification

Prediction

Speech

Recognition

Text to Speech

Speech to Text

Image

Object detection

Object location

OCR

Gesture recognition

Facial modelling

Segmentation

Clustering

Compression

Super resolution

Audio

Translation

Voice synthesis

Content

Video generation

Text generation

Audio generation

Easy to get started

1 — 2 — 3 — 4

Jump start

Use our pretrained models or retrain

Custom model

Deploy your custom model

Performance

Benchmark, validate & accelerate your models.

Optimize

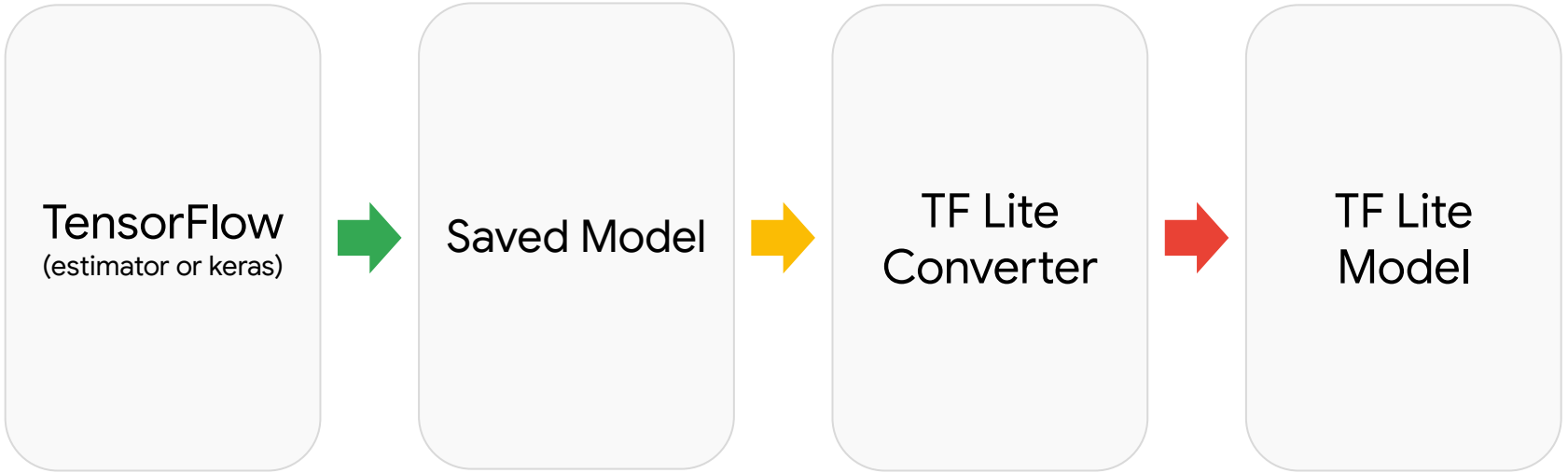
Try our Model Optimization Toolkit

TensorFlow Lite powers ML Kit

- ML Kit is out-of-the-box proprietary models that you can run on device

Converting your model

Custom Model



TensorFlow 1.X

```
import tensorflow as tf

converter =
tf.lite.TFLiteConverter.from_saved_model(saved_model_dir
)
tflite_model = converter.convert()
open("converted_model.tflite", "wb").write(tflite_model)
```

Custom Model

Conversion is sometimes hard

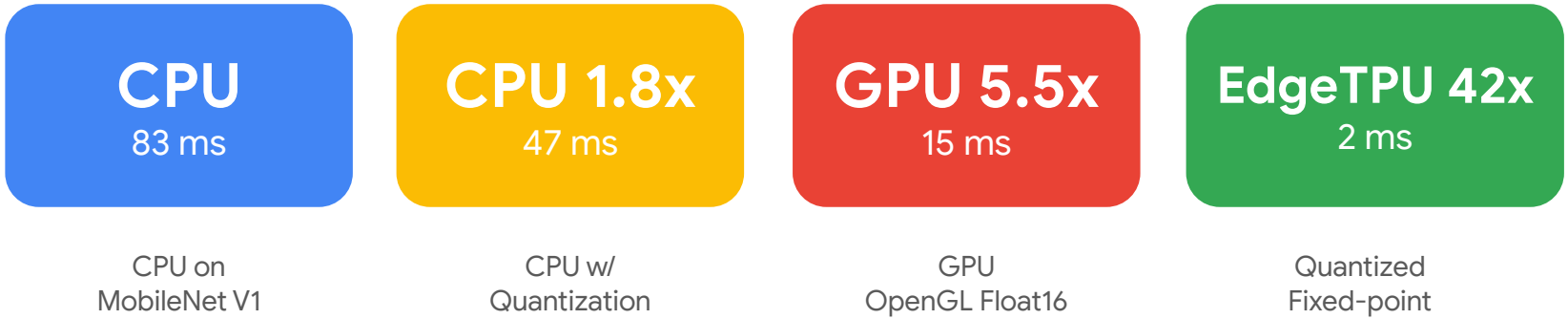
- Limited ops
- Unsupported semantics (i.e. control-flow in RNNs)

Performance

Get your models running as fast as possible

Goal: As fast as possible on all hardware (CPU, GPU, DSP, NPU)

Inference performance



MobileNet V1

Pixel 3 *- Single Threaded CPU*

Optimize

Quantization: Huge speedup and ~4x smaller size

Achieved by reducing the precision of weights and activations in your graph.

```
import tensorflow as tf

converter = tf.lite.TFLiteConverter.from_saved_model(saved_model_dir)
converter.optimizations = [tf.lite.Optimize.OPTIMIZE_FOR_SIZE]
tflite_quant_model = converter.convert()
```

Optimize

MCU

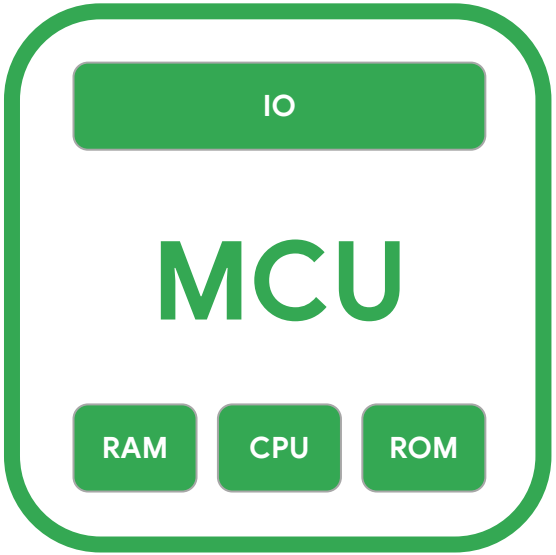
TensorFlow Lite for microcontrollers

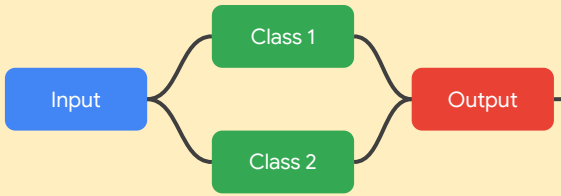
More than 150B microcontrollers exist globally today

What are they?

Small computer on a single circuit

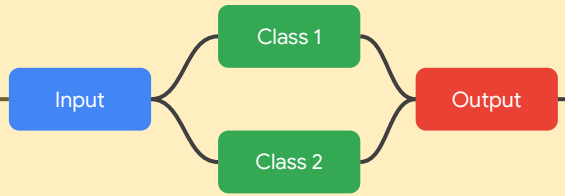
- No operating system
- Tens of KB of RAM & Flash
- Only CPU, memory & I/O peripherals





MCU

Is there any sound?



MCU

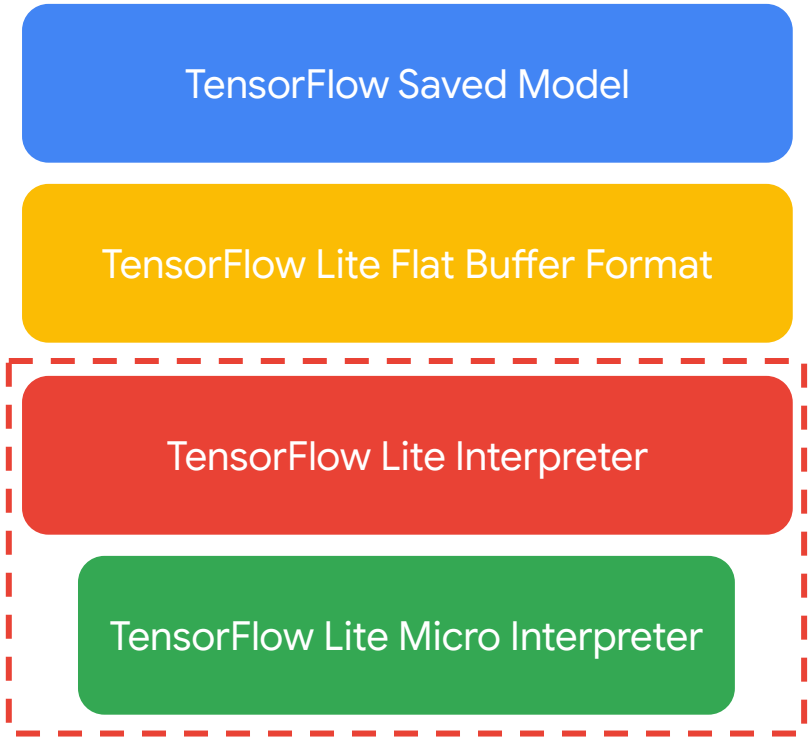
Is that human speech?

Deeper Network

Application Processor

TensorFlow Lite for microcontrollers

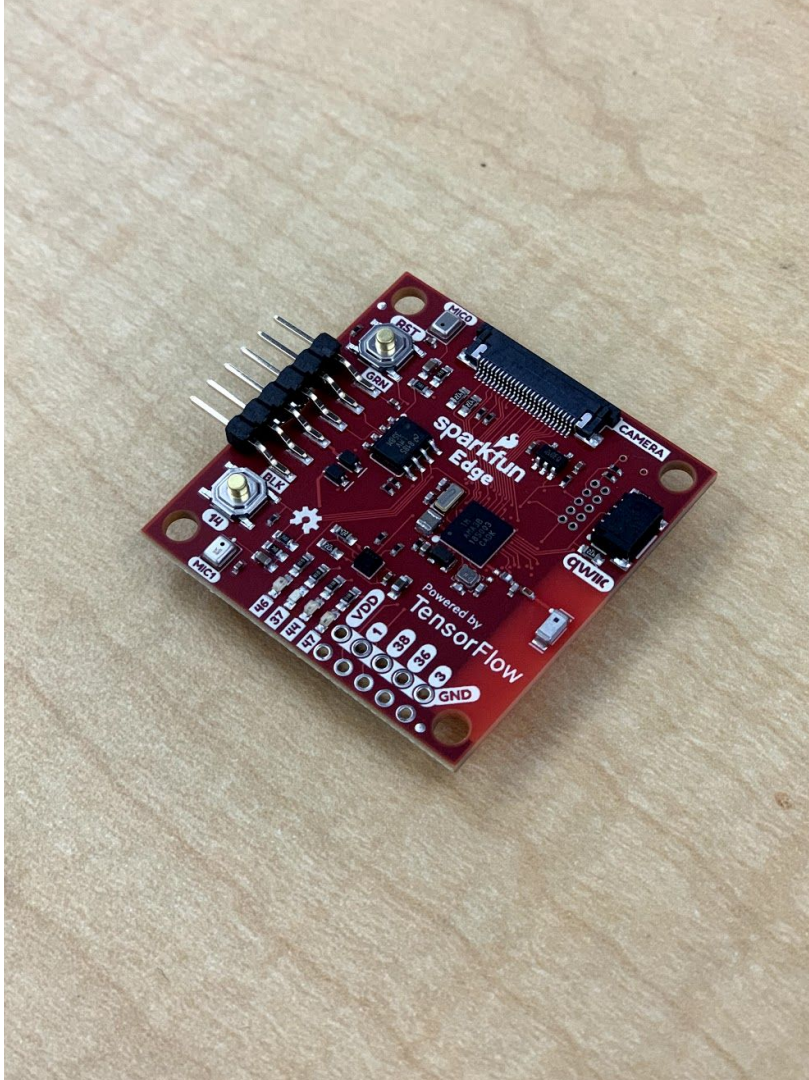
TensorFlow provides you **with a single framework** to deploy on Microcontrollers as well as phones



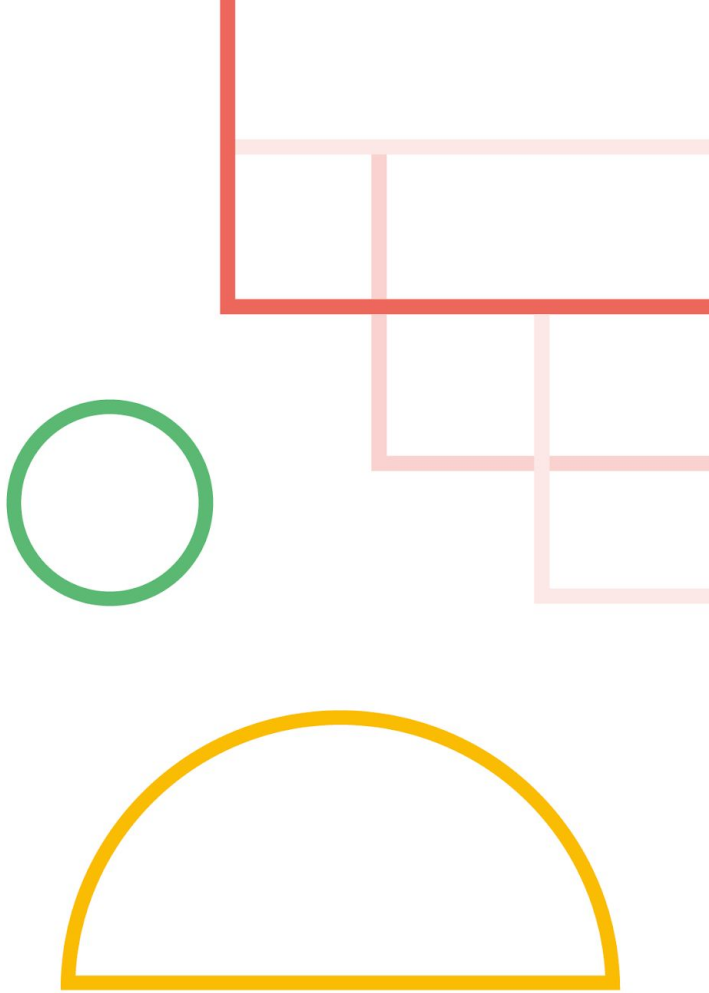
Example models

Available now on tensorflow.org

- Speech model (20KB)
- Image classifier (250KB) [Coming Soon]



Wrap-up



TensorFlow Extended (TFX)

Out-of-the-box components for your production model needs

Flexible orchestration and metadata

Extensible with custom components

Visit us at https://tensorflow.org/tfx and show us how you've used and extended TFX!

TFX | TensorFlow

https://www.tensorflow.org/tfx

TensorFlow Install Learn API Resources Community Why TensorFlow Search GitHub

For Production

Overview Tutorials Guide API

Apply to speak at TensorFlow World. Deadline April 23rd. Propose talk

TensorFlow Extended (TFX) is an end-to-end platform for deploying production ML pipelines

When you're ready to move your models from research to production, use TFX to create and manage a production

https://www.tensorflow.org/tfx

TensorFlow Lite

Serve models on mobile and embedded devices

Key features : Optimisation for speed and size

Makes TensorFlow ecosystem much more compelling

Visit us at https://tensorflow.org/tfx and show us how you've used and extended TFX!

Deploy machine learning models on mobile and IoT devices

TensorFlow Lite is an open source deep learning framework for on-device inference.

See the guide

Guides explain the concepts and components of TensorFlow Lite.

See examples

Explore TensorFlow Lite Android and iOS apps.

See models

Easily deploy pre-trained models.

How it works

Pick a model

Pick a new model or retrain an existing one.

Convert

Convert a TensorFlow model into a compressed flat buffer with the TensorFlow Lite Converter.

Deploy

Take the compressed .tflite file and load it into a mobile or embedded device.

Optimize

Quantize by converting 32-bit floats to more efficient 8-bit integers or run on GPU.

Deep Learning MeetUp Group

The Group :

- MeetUp.com / TensorFlow-and-Deep-Learning-Singapore
- > 3,900 members

The Meetings :

- Next : Date TBA, hosted at Google
  - Something for Beginners
  - Something from the Bleeding Edge
  - Lightning Talks

RED DRAGON AI

Deep Learning Courses in Singapore

Jumpstart Course : Two days in-person + One day online

- Hands-on with real model code
- Build your own Project

Other Modules:

- Advanced Computer Vision; Advanced NLP; Self-supervised ...

Each 'module' includes :

- In-depth instruction, by practitioners
- 70%-100% funding via IMDA for SG/PR

Red Dragon AI : Intern Hunt

Opportunity to do Deep Learning “all day”

Key Features :

- Work on something cutting-edge (+ publish!)
- Location : Singapore (SG/PR FTW) and/or Remote

Action points :

- Need to coordinate timing...
- Contact Martin or Sam via LinkedIn

Questions?