Vinicius Caridá

MLOps - Continuous Delivery And Automation Pipelines In Machine Learning

Vinicius Caridá, Ph.D.

Advanced Analytics Manager, Itaú Unibanco

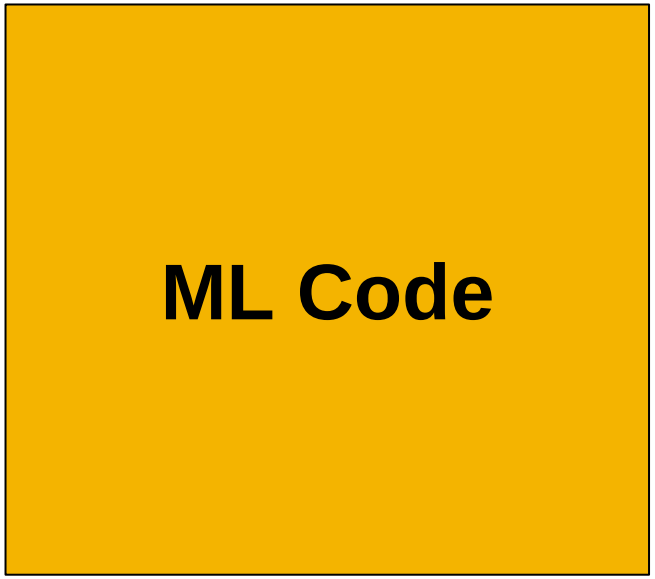
MBA Professor, FIAP

Google Developer Expert – Machine Learning

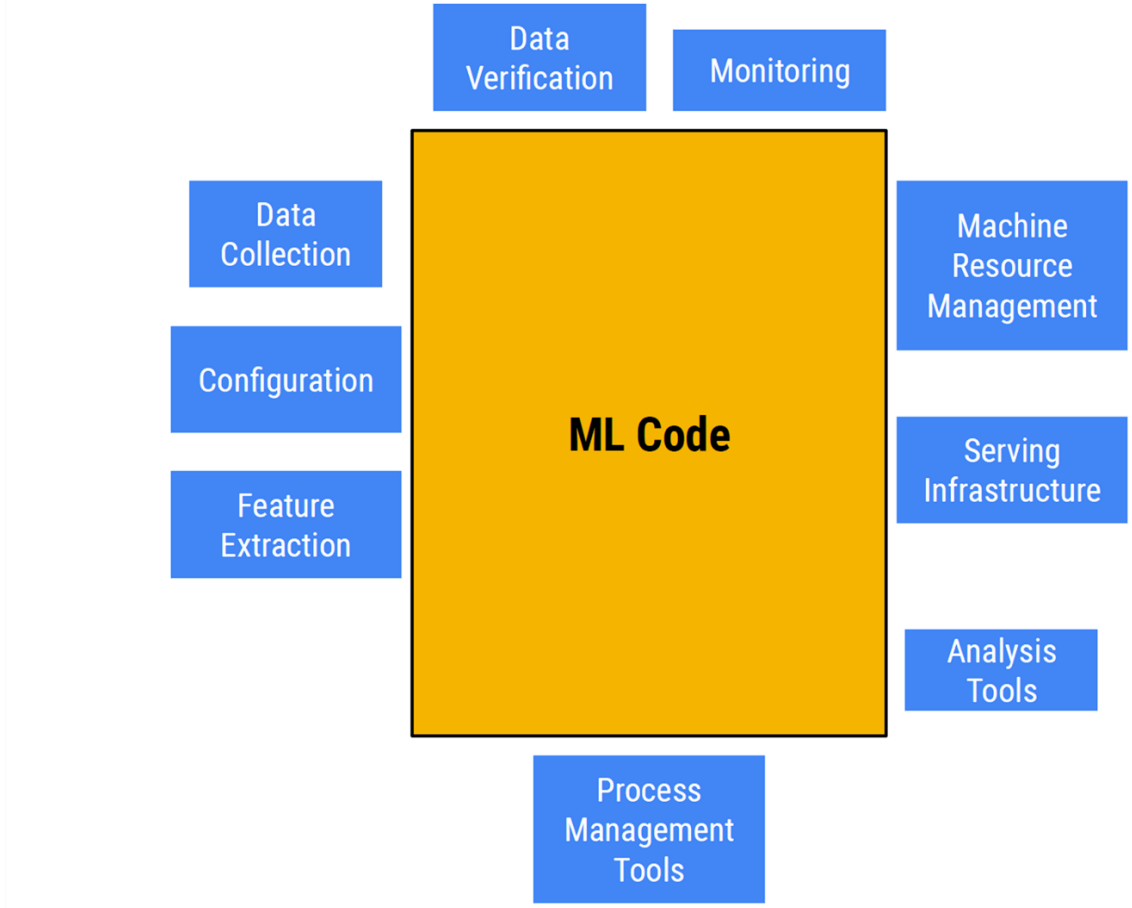
@vfcarida

Production Machine Learning

In addition to training an amazing model ...

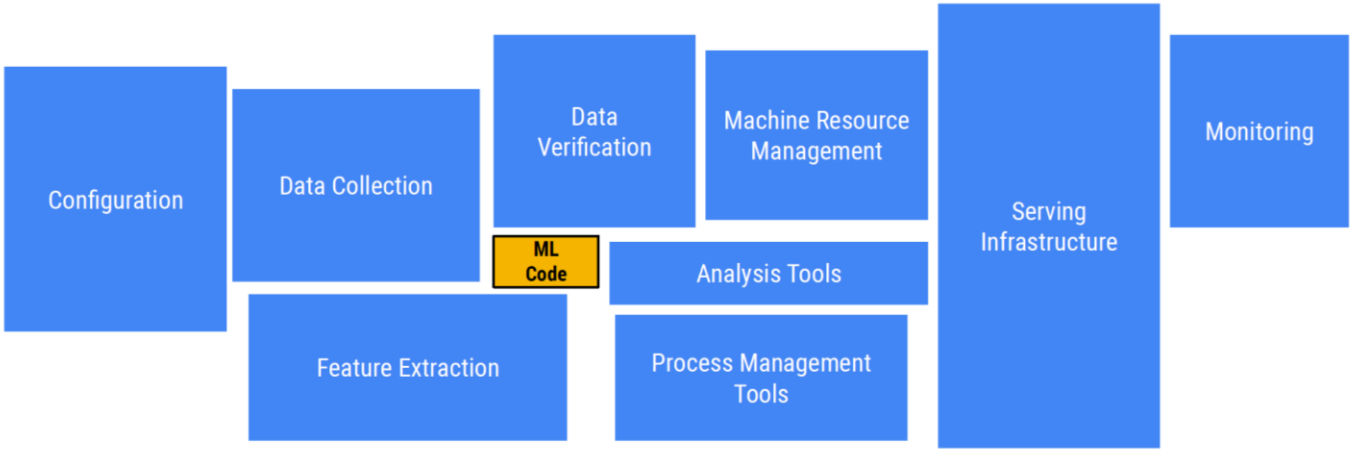


Production Machine Learning



Production Machine Learning

Reality: ML requires DevOps; lots of it



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

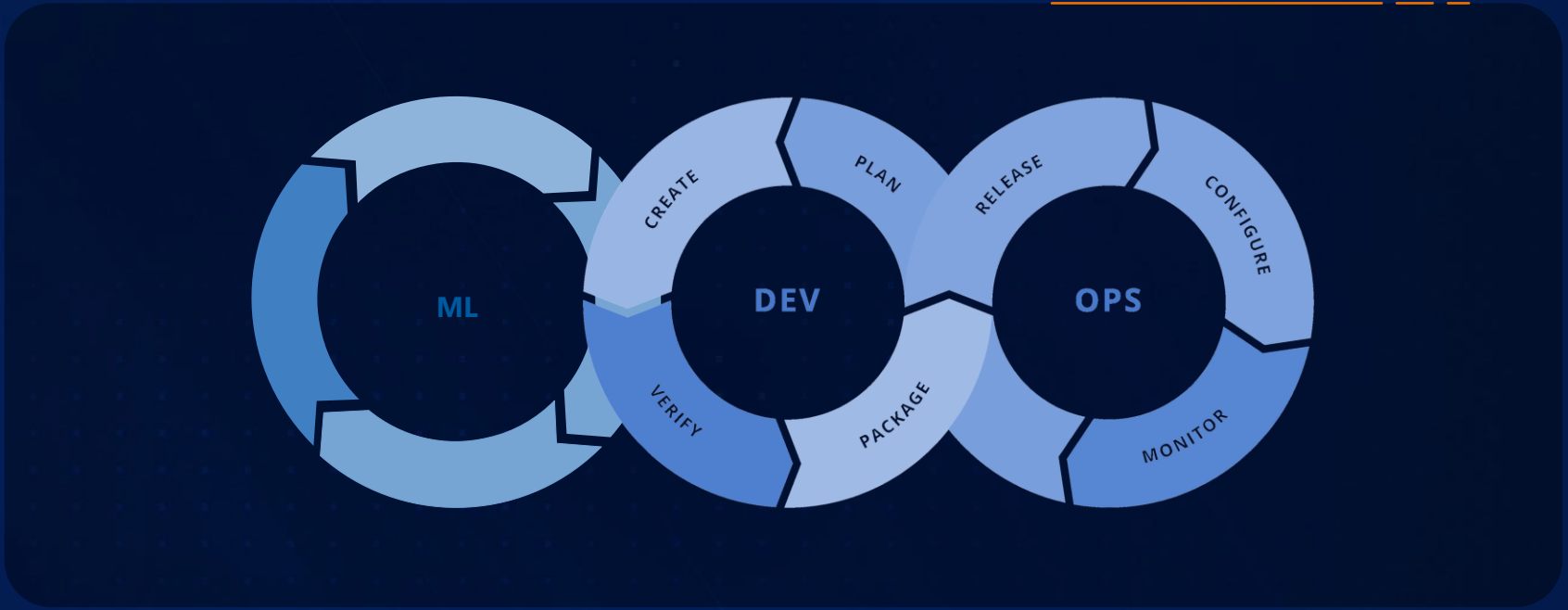
Como fazer ~~pão~~ um modelo?

What is **ML Ops**

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

ML Ops is to strongly advocate **automation and monitoring** at all steps of ML system construction, from integration, testing, releasing to deployment and infrastructure management.

MLOps = ML + DEV + OPS



Experiment
Data Acquisition
Business Understanding
Initial Modeling

Develop
Modeling + Testing
Continuous Integration
Continuous Deployment

Operate
Continuous Delivery
Data Feedback Loop
System + Model Monitoring

Production Machine Learning

Modern Software Development

- Scalability
- Extensibility
- Configuration
- Consistency & Reproducibility
- Modularity
- Best Practices
- Testability
- Monitoring
- Safety & Security

Production Machine Learning

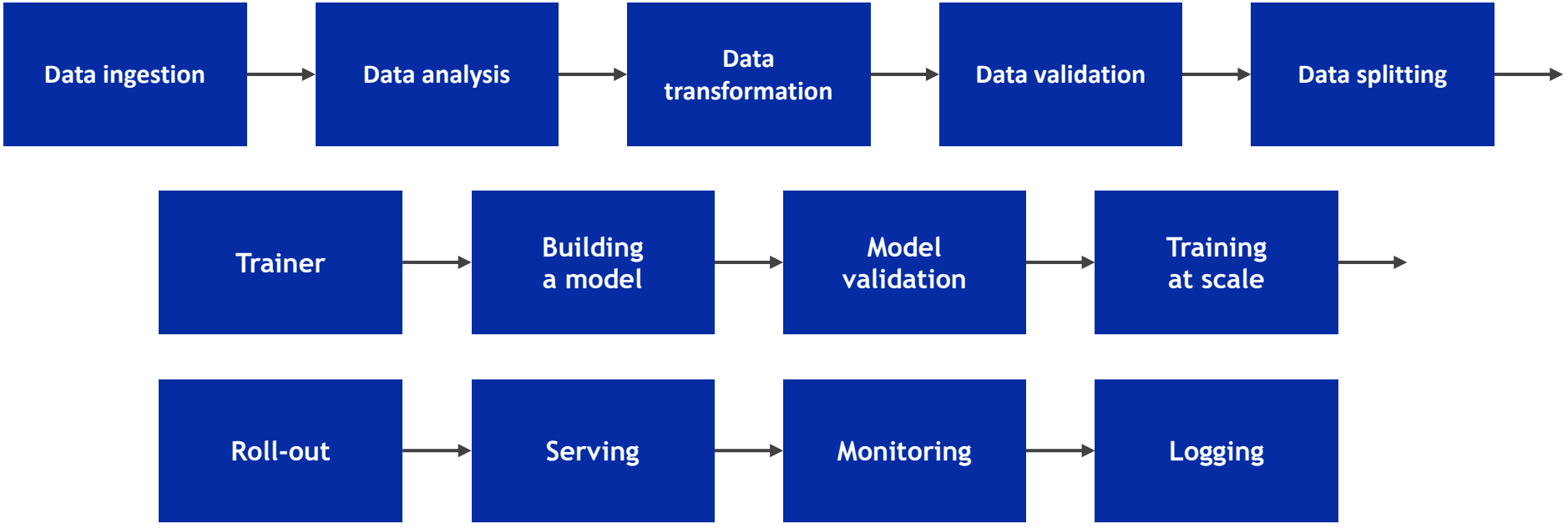
Modern Software Development

- Scalability
- Extensibility
- Configuration
- Consistency & Reproducibility
- Modularity
- Best Practices
- Testability
- Monitoring
- Safety & Security

Machine Learning Development

- Labeled data
- Feature space coverage
- Minimal dimensionality
- Maximum predictive data
- Fairness
- Rare conditions
- Data lifecycle management

Production Machine Learning



If ML is a rocket engine,
data is the fuel

Launching is easy, operating is hard

ML Deploy Platforms

- Uber - Michelangelo
- AirBnB - Bighead
- Facebook - FB Learner
- Lyft - Lyft Learn
- Data Robot - ParallelM

Production Machine Learning

“Hidden Technical Debt in Machine Learning Systems”

NIPS 2015

http://bit.ly/ml-techdebt

"Depending on a ML superhero"

A ML superhero is:

ML Researcher

Data engineer

Infra and Ops engineer

Product Manager

A partner to execs

From PoC to production

Google – TFX Production Components

TFX ML system overview

The following diagram shows how the various TFX libraries are integrated to compose an ML system.

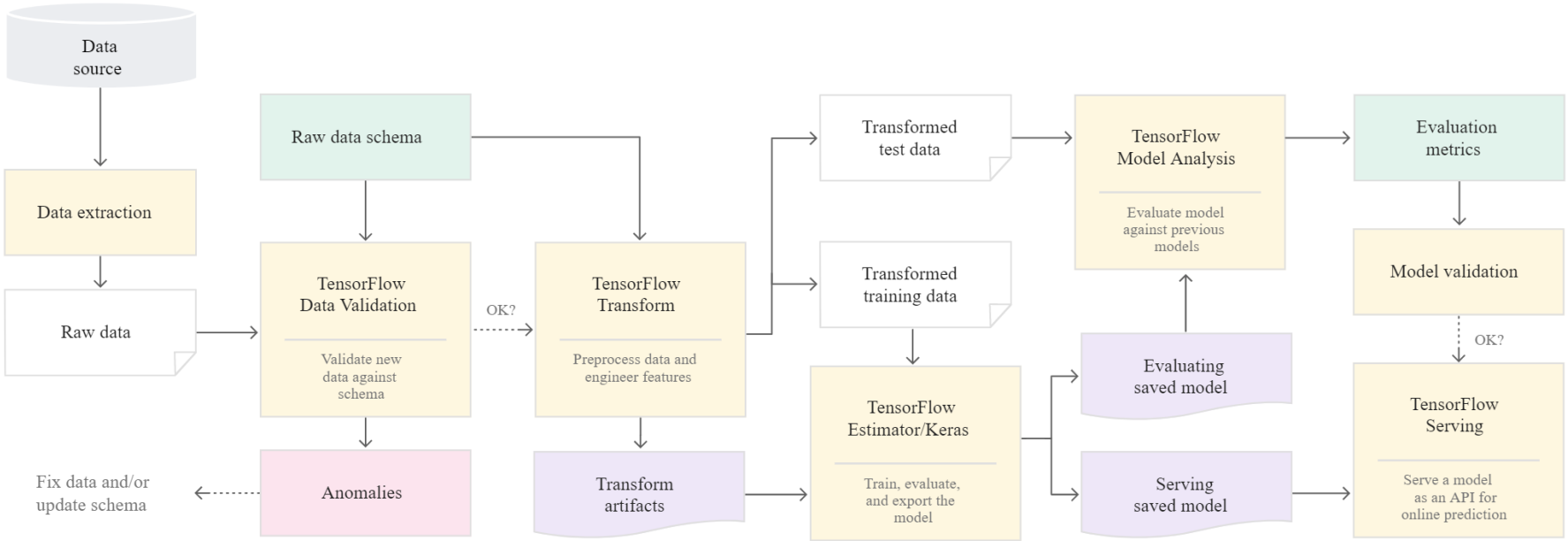
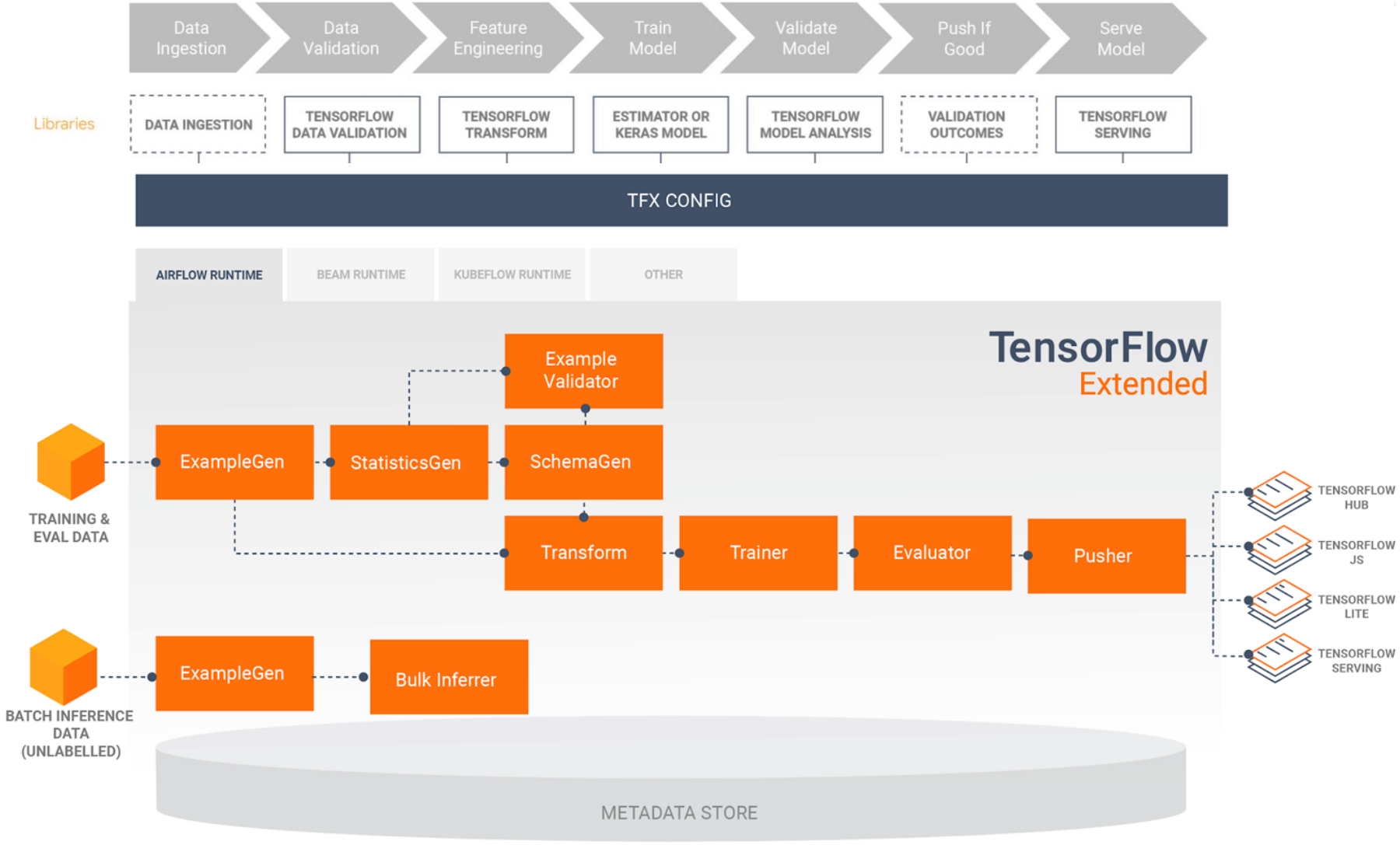


Figure 2. A typical TFX-based ML system.

https://cloud.google.com/solutions/machine-learning/architecture-for-mlops-using-tfx-kubeflow-pipelines-and-cloud-build

Google – TFX Pipeline Example



Solution

TFX is an end-to-end machine learning platform for TensorFlow.

TensorFlow Extended (TFX) is a TensorFlow-based general-purpose machine learning platform implemented at Google. We've already open sourced some TFX libraries with the rest of the system to come. For an overview of TFX, read our KDD'2017 paper.

TensorFlow Transform

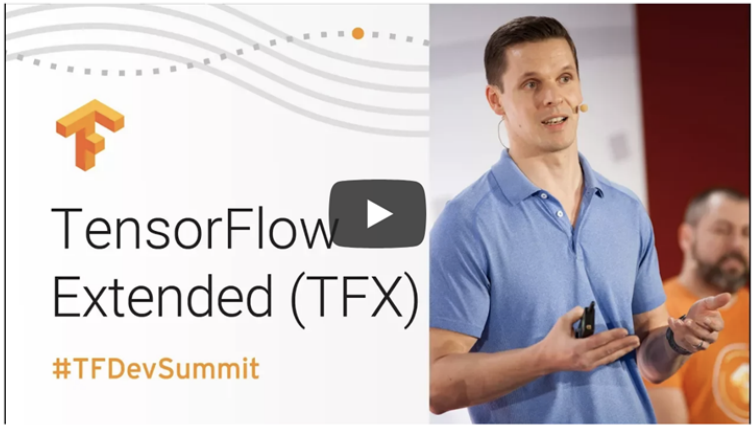
Perform full-pass analyze phases over data to create transformation graphs that are consistently applied during training and serving.

TensorFlow Model Analysis

A collection of libraries and visualization components to compute full-pass and sliced model metrics over large datasets, and analyze them in a notebook.

TensorFlow Serving

A flexible, high-performance serving system for machine learning models, designed for production environments.



See an end-to-end demo of how the TFX tools fit together.

Example: ML pipeline

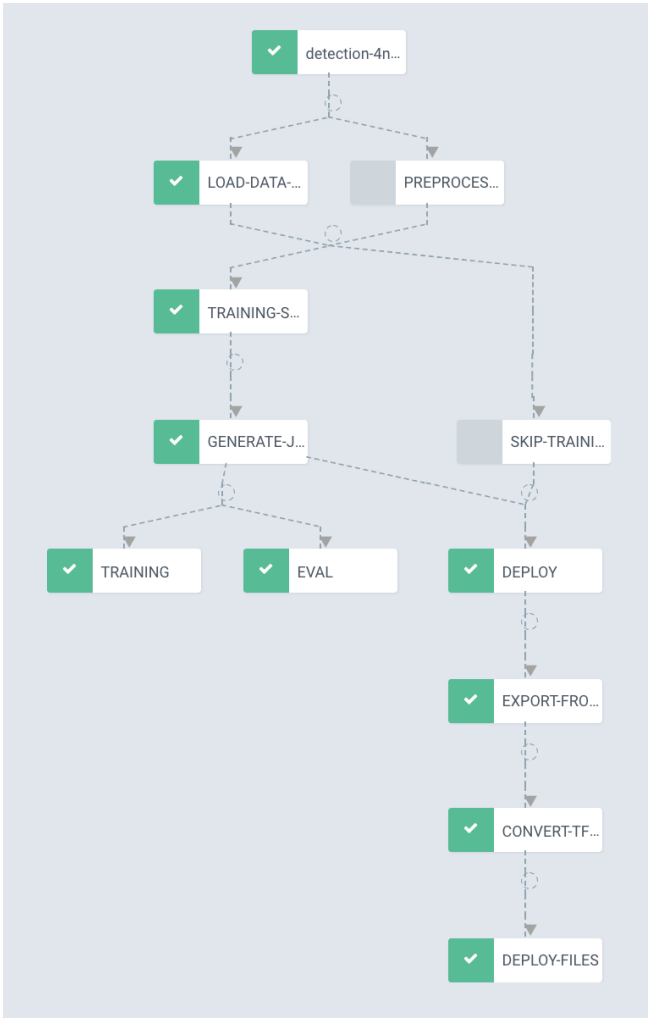
ML pipeline workflow

Loading training data

Training/Eval

Convert to TFLite

Deploy TFLite files to mobile devices



Introducing Argo—A Container-Native Workflow Engine for Kubernetes

Microsoft - Azure MLOps

Asset management & orchestration services to help manage the ML lifecycle.

ML Experimentation

Training Services

- Script Execution
- AutoML
- HyperDrive

Run History

Compute

- ML Compute
- ML VMs
- Databricks

Storage

- Blob Store
- Files
- Data Lake

ML Ops

Asset Management

- Environments
- Code
- Datasets
- Models

Orchestration Services

- ML Pipelines
- Model Packaging
- Model Validation **
- Model Deployment **
- ML Audit Trail **
- Azure DevOps ML Extension **

ML Inference

Real-Time
Azure Kubernetes Service

Batch
Azure ML Compute

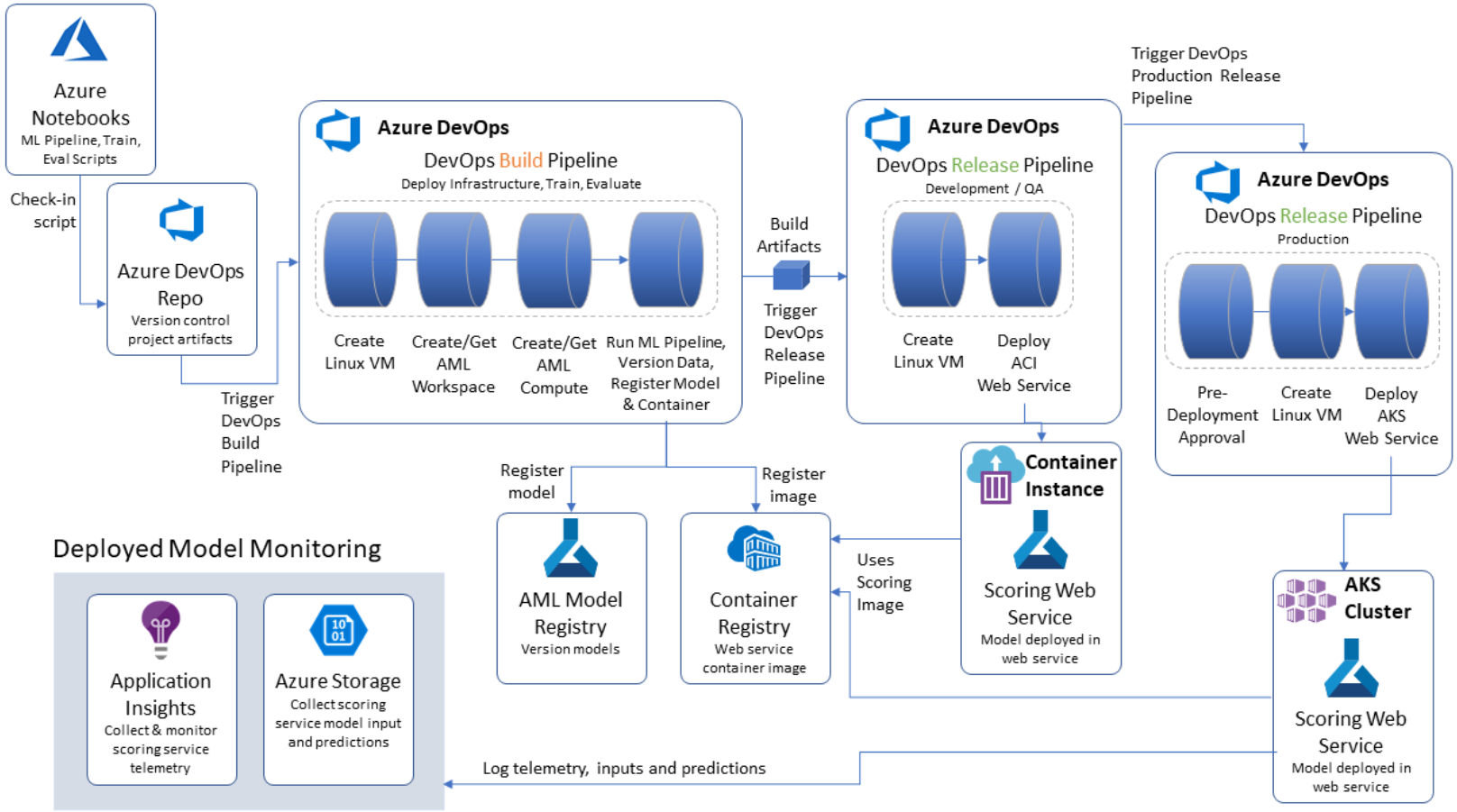
Edge
Azure IoT Hub

Monitoring

- Experimentation Monitoring
- Inference Monitoring
- ML Data Drift **

Microsoft - Azure MLOps

Asset management & orchestration services to help manage the ML lifecycle.



References:

[1] Machine Learning: the high interest credit card of Technical Debt, D. Sculley et al.

[2] Rules of Machine Learning, Martin Zinkevich

[3] TFX: A TensorFlow based production-scale machine learning platform, Denis Bayor et al.

[4] Introducing TensorFlow Model Analysis, Clemens Mewald

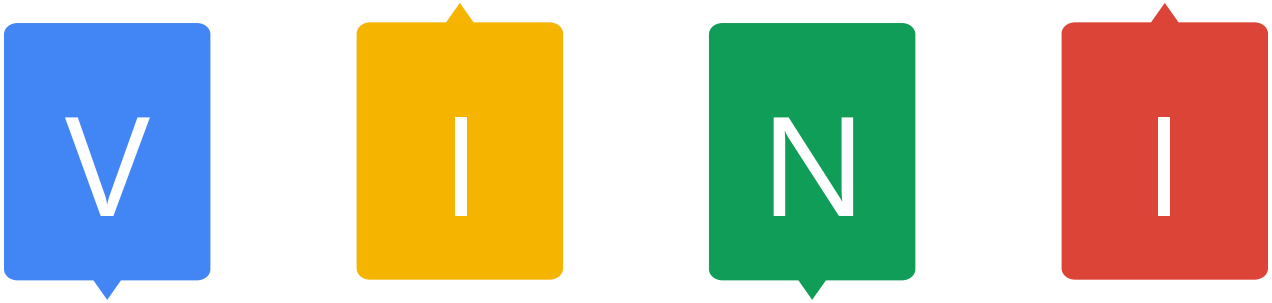
[5] https://cloud.google.com/blog/products/ai-machine-learning/itau-unibanco-how-we-built-a-cicd-pipeline-for-machine-learning-with-online-training-in-kubeflow, Vinicius Caridá

[6] https://towardsdatascience.com/ml-ops-machine-learning-as-an-engineering-discipline-b86ca4874a3f, Cristiano Breuel

[7] MLOps no Azure:

- Microsoft MLOps
- Microsoft Cloud Workshop – MLOps Hands-On lab
- MLOps with Python and Azure Machine Learning Services
- MLOps with Python and Azure Machine Learning Services and Databricks for model training

Thanks !



Vinicius Fernandes Caridá
vfcarida@gmail.com

@Vinicius Caridá

@Vinicius Caridá

@vfcarida

@vfcarida