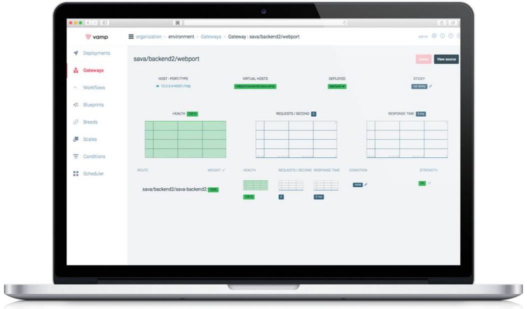
@vlesierse @joerg_schad

A Tale of Two Worlds
Canary-Testing for Both ML Models and Microservices

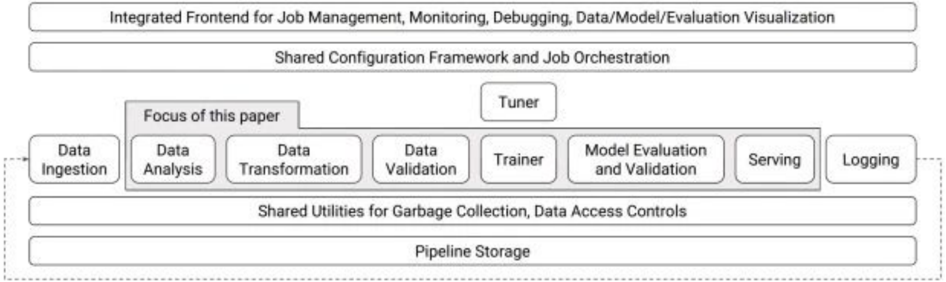


sava/backend2/webport

HOST - PORT/TYPE: 10.0.0.4:40001/http
VIRTUAL HOSTS: webport.backend2.sava.vamp
DEPLOYED: deployed ✓
STICKY: not sticky

HEALTH 100 % | REQUESTS / SECOND 0 | RESPONSE TIME 0 ms

ROUTE	WEIGHT	HEALTH	REQUESTS / SECOND	RESPONSE TIME	CONDITION	STRENGTH
sava/backend2/sava-backend2:	100%	100 %	0	0 ms	none	0%



Integrated Frontend for Job Management, Monitoring, Debugging, Data/Model/Evaluation Visualization

Shared Configuration Framework and Job Orchestration

Tuner

Focus of this paper

Data Ingestion

Data Analysis

Data Transformation

Data Validation

Trainer

Model Evaluation and Validation

Serving

Logging

Shared Utilities for Garbage Collection, Data Access Controls

Pipeline Storage

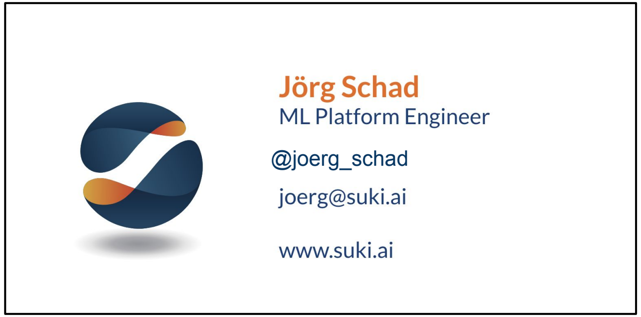
vamp

ArangoDB

Vincent Lesierse

Technical Product Manager

- **Vamp.io**
- **@vlesierse**



Jörg Schad
ML Platform Engineer

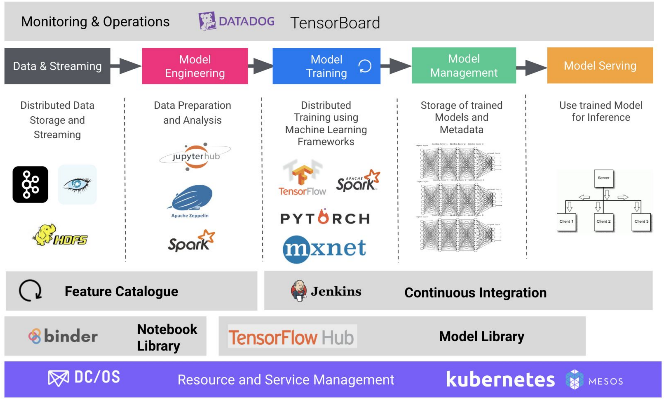
@joerg_schad

joerg@suki.ai

www.suki.ai

- Suki.ai
- Mesosphere
- PhD Distributed DB Systems

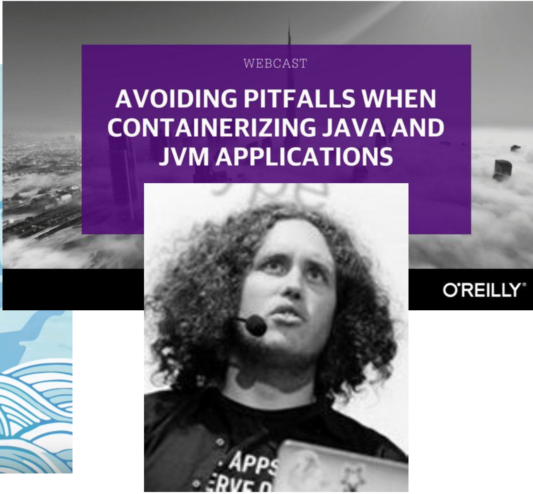
- @joerg_schad



KubeCon
CloudNativeCon
China 2018

Operating Deep Learning Pipelines Anywhere Using Kubeflow

Jörg Schad & Gilbert Song, Mesosphere



WEBCAST

AVOIDING PITFALLS WHEN CONTAINERIZING JAVA AND JVM APPLICATIONS

A tale of two worlds

It was the best of times -where Microservices allowed broke our code in manageable pieces,

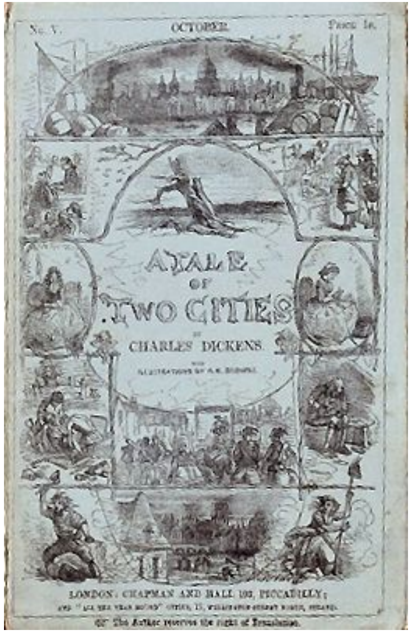
it was the worst of times -where we actually had to deploy all those pieces,

it was the age of wisdom - where Machine Learning allowed for cool Models,

it was the age of foolishness - where we suffered to productionize these models,

it was the epoch of belief, it was the epoch of incredulity,

it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.[2]



A tale of two worlds

Deployment of Microservices

TensorFlow Extended

Deployment of Machine Learning Models

Microservice Deployments

Challenges

- Blue/green….
- Canary Releases
- A/B Testing

Solutions

- Service Meshes (Istio)
- Vamp

Vamp’s history

Our mission: Provide an easy-to-use and powerful solution to transform software from ideas into value as efficiently as possible, using cloud-native technologies.

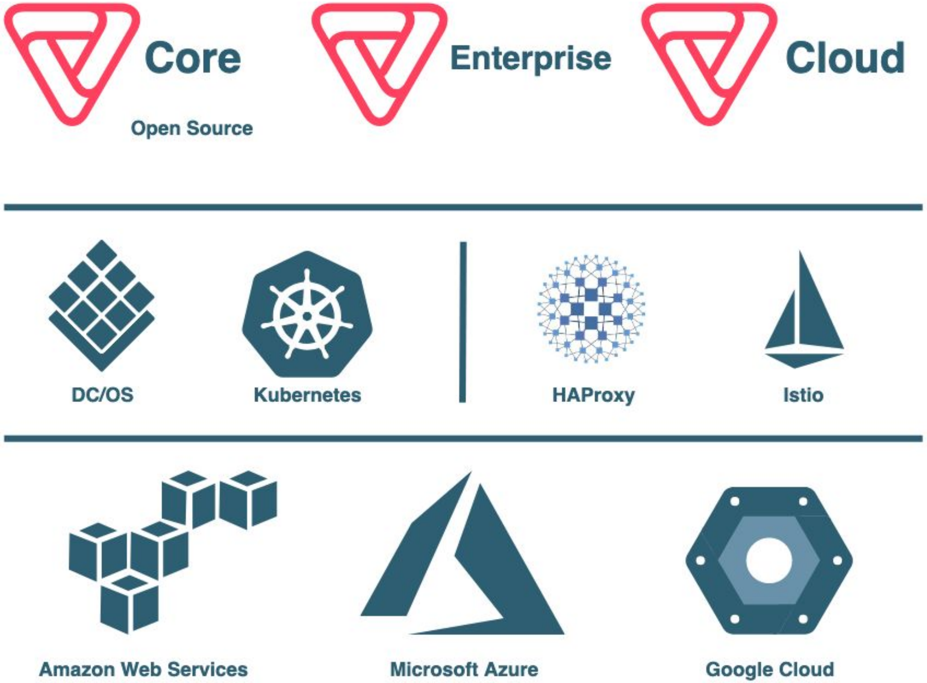
Important: Releasing is (much) more than technical deployment. Continuous validation requires observability and “golden metrics”. Processes are crucial. Tools and technologies are means to an end.

2014: canary testing & releasing and SLA-based auto-scaling on top of Mesos+Marathon, using HAProxy for networking (service-discovery/load-balancing/mesh/ingress)

2016: added Kubernetes support

2019: added Istio support

Vamp



Core

Open Source

Enterprise

Cloud

DC/OS

Kubernetes

HAProxy

Istio

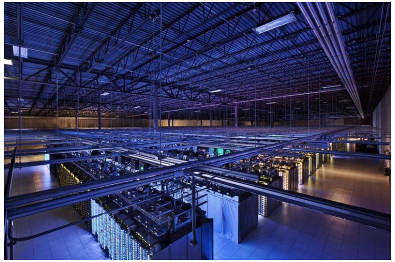
Amazon Web Services

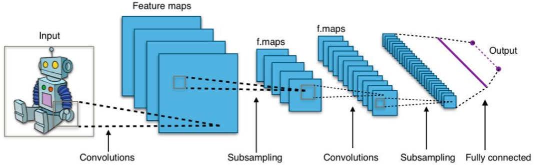
Microsoft Azure

Google Cloud

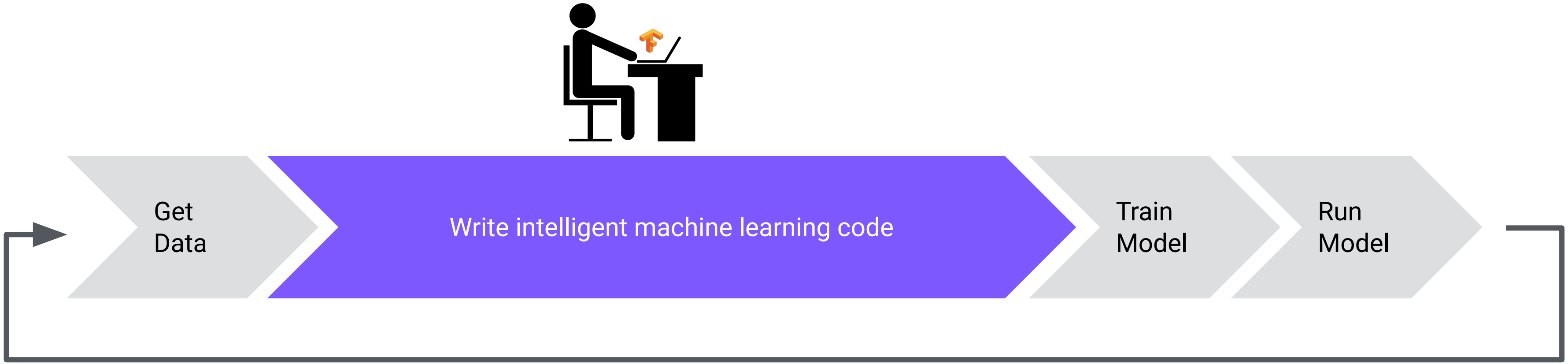
Machine Learning….

Dataset
Free
Data
Websites
Here
Datascience
Use
Big
Top
Projects
Science
Sas
Python
Gather
Best
Excel
Part
Un
Lists
World
Curated
Bank
Kaggle




Input
Feature maps
f.maps
f.maps
Output
Convolutions
Subsampling
Convolutions
Subsampling
Fully connected

What you want to be doing



Get Data

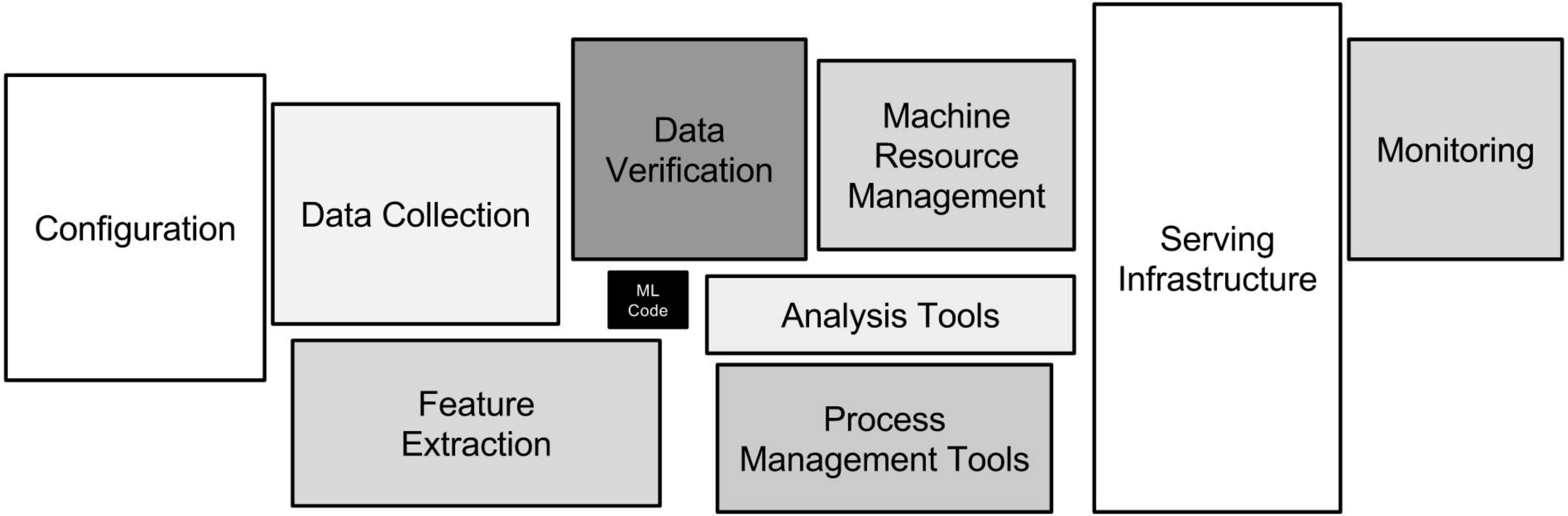
Write intelligent machine learning code

Train Model

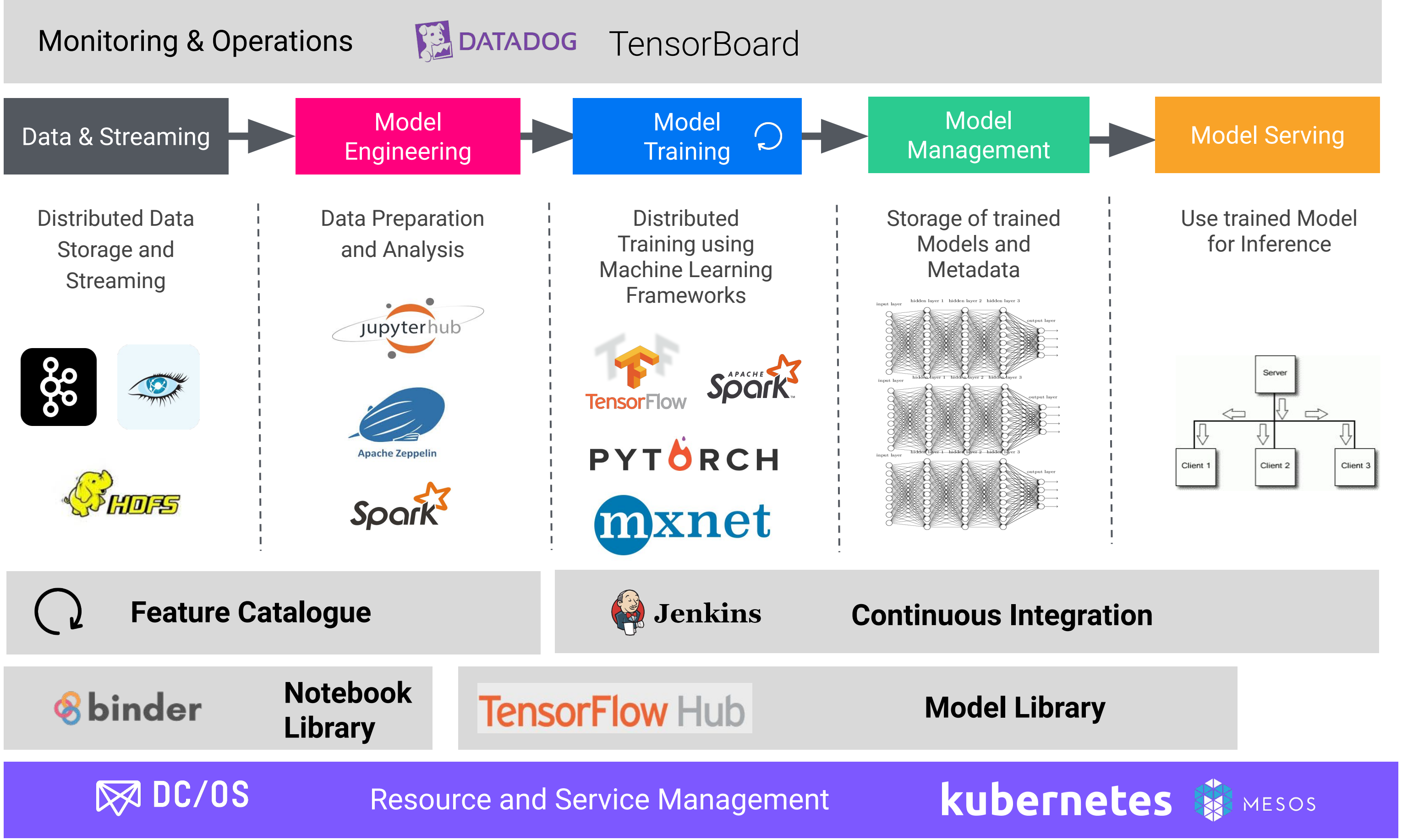
Run Model

Repeat

What you’re actually doing



Sculley, D., Holt, G., Golovin, D. et al. Hidden Technical Debt in Machine Learning Systems



Monitoring & Operations DATADOG TensorBoard

Data & Streaming → Model Engineering → Model Training → Model Management → Model Serving

Data & Streaming: Distributed Data Storage and Streaming

HDFS

Model Engineering: Data Preparation and Analysis

jupyterhub

Apache Zeppelin

Spark

Model Training: Distributed Training using Machine Learning Frameworks

TensorFlow

APACHE Spark

PYTORCH

mxnet

Model Management: Storage of trained Models and Metadata

Model Serving: Use trained Model for Inference

Feature Catalogue

Jenkins **Continuous Integration**

binder **Notebook Library**

TensorFlow Hub **Model Library**

DC/OS Resource and Service Management kubernetes MESOS

Challenge: Persona(s)

The Rise of the *DataOps Engineer*

Combines two key skills:

- Data science
- Distributed systems engineering

The equivalent of *DevOps* for *Data Science*

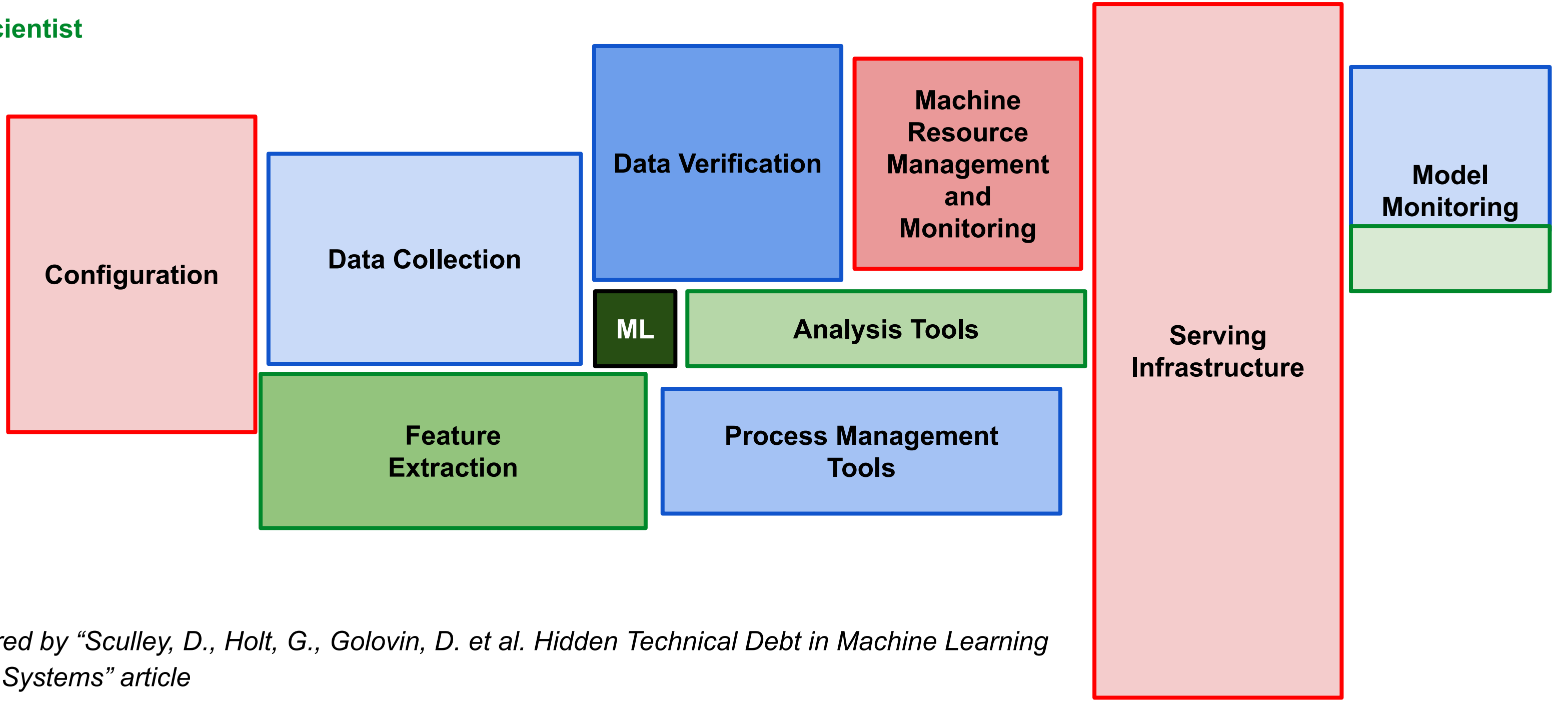
- **Build** automation software to run machine learning systems
- **Operate** systems so they're available, scalable, and performant
- **Evangelize** tools and best practices among data scientists

Division of Labor

System Admin/ DevOps

Data Engineer/DataOps

Data Scientist



Configuration

Data Collection

Data Verification

Machine Resource Management and Monitoring

ML

Analysis Tools

Feature Extraction

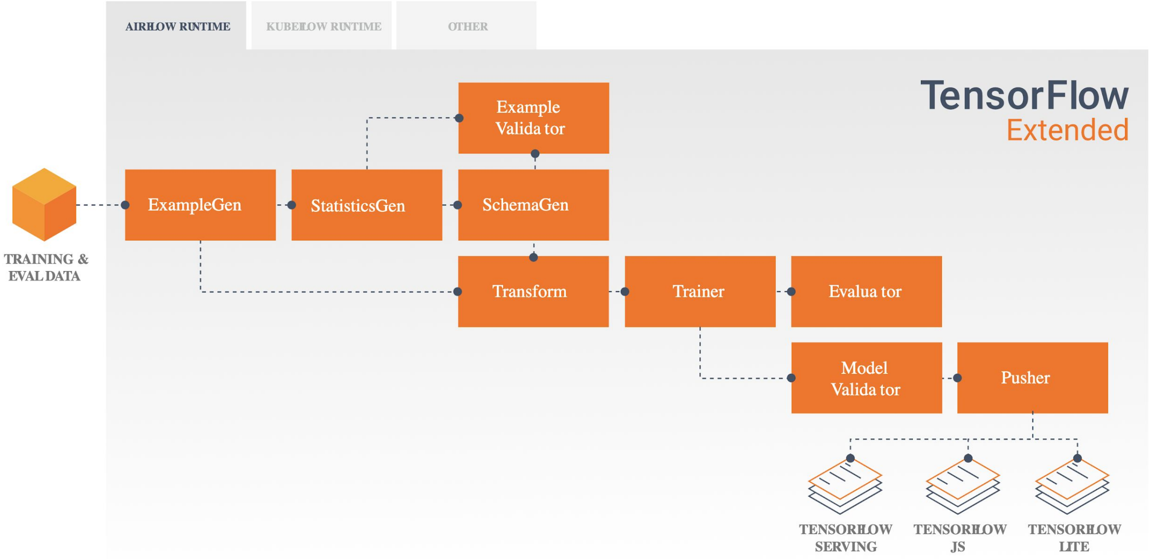
Process Management Tools

Serving Infrastructure

Model Monitoring

Inspired by “Sculley, D., Holt, G., Golovin, D. et al. Hidden Technical Debt in Machine Learning Systems” article

TensorFlow Extended

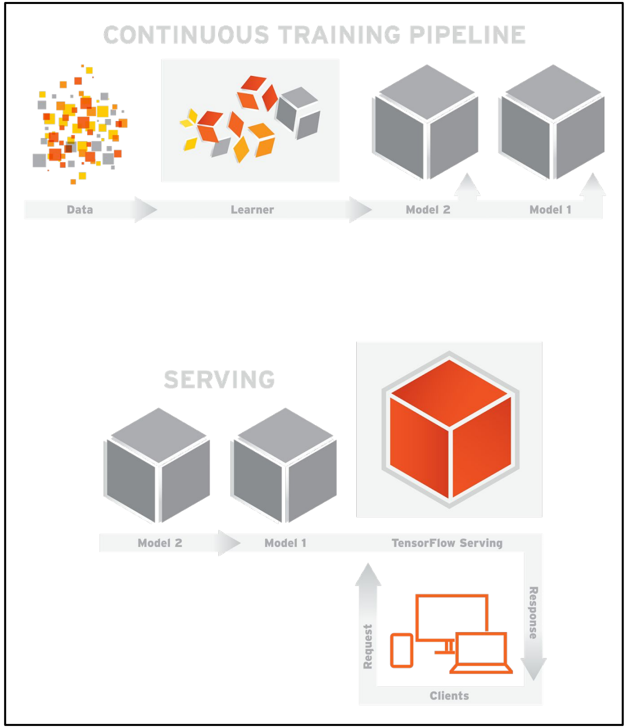


http://www.kdd.org/kdd2017/papers/view/tfx-a-tensorflow-based-production-scale-machine-learning-platform

https://www.youtube.com/watch?v=fPTwLVCq00U

Challenge: Serving Environment

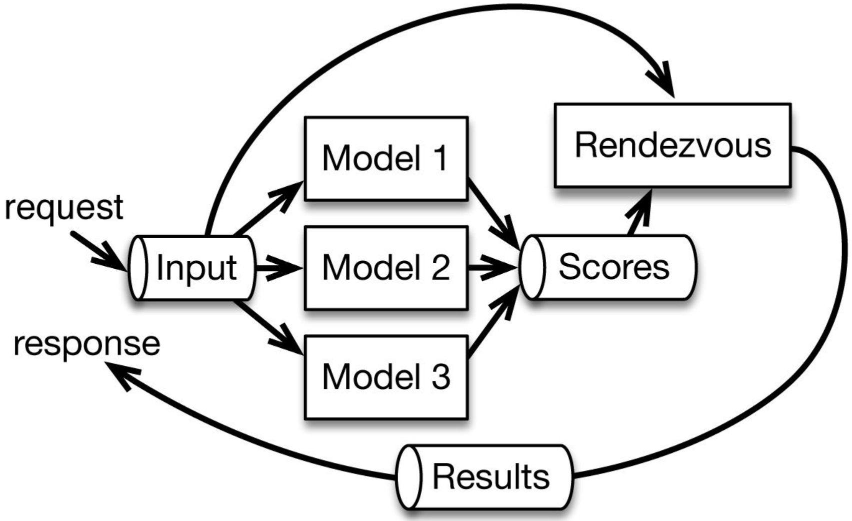
- How to Deploy Models?
  - Zero Downtime
  - Canary
- Multiple Models?
  - Testing
  - Different Scenarios
  - Updating models
  - Shadow models
  - A/B Canary testing
- Common Metadata



https://ai.googleblog.com/2016/02/running-your-models-in-production-with.html

Challenge: Serving Environment

- How to Deploy Models?
  - Zero Downtime
  - Canary
- Multiple Models?
  - Testing



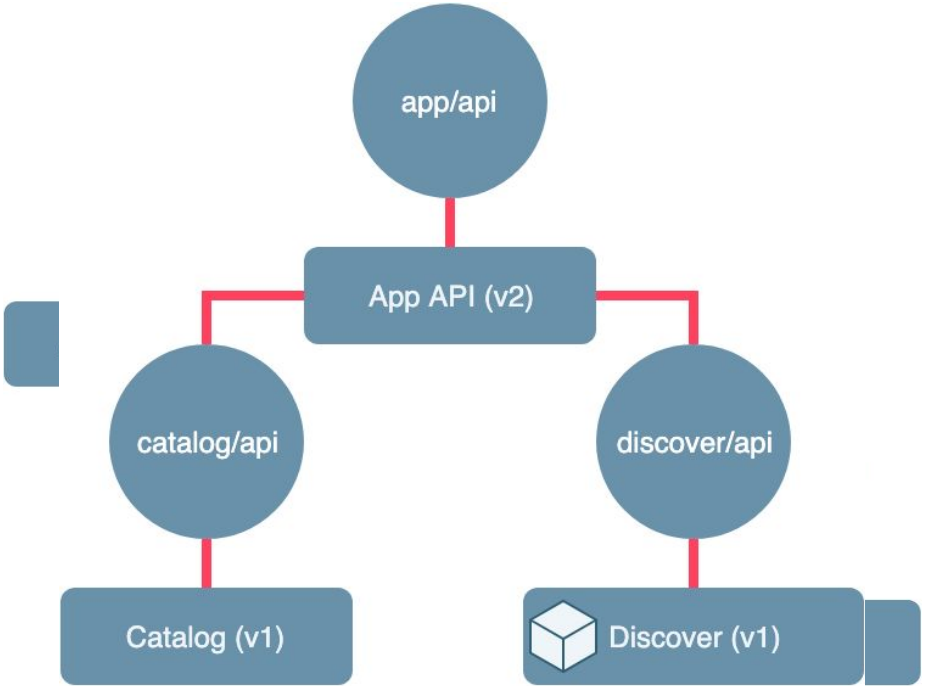
https://mapr.com/ebooks/machine-learning-logistics/ch03.html

A tale of ~~two~~ one world

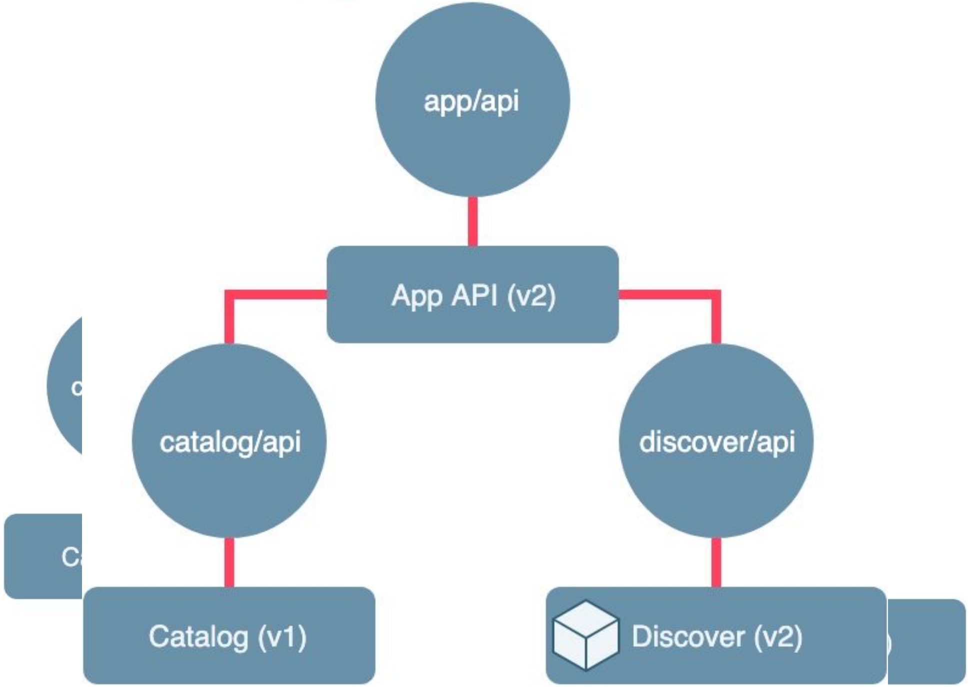
TensorFlow
Extended

Deployment of Services

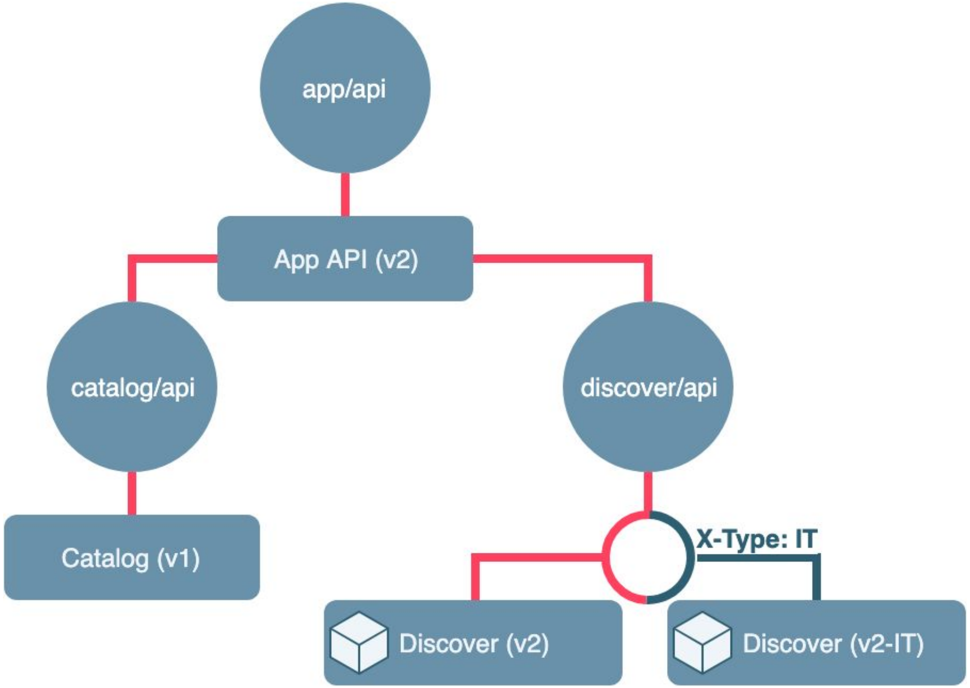
Demo: Canary Release Service



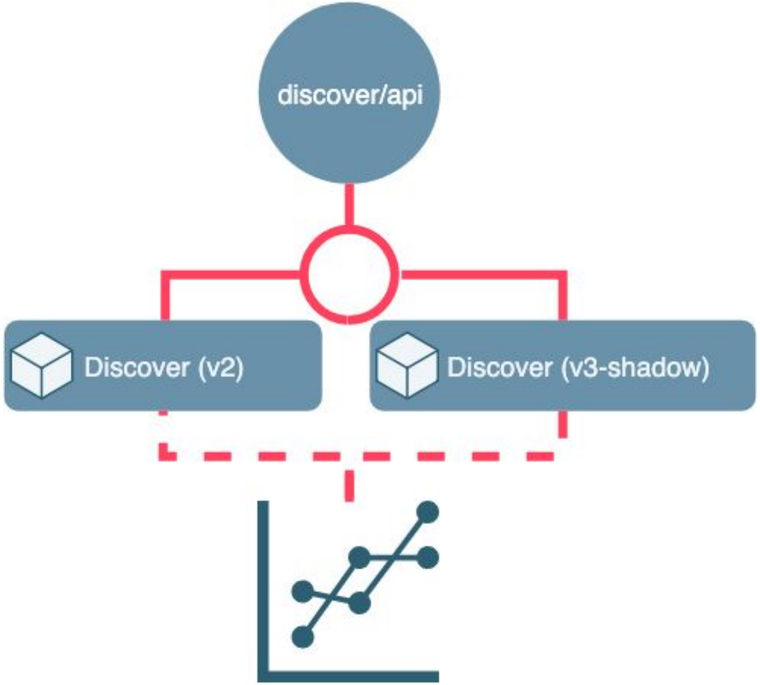
Demo: Canary Release ML Model



Scenario: Model Segmentation



Scenario: Shadow Traffic



Talk with us

Join use for a chat the at the Mesosphere stand ….

What’s next for Vamp?

- Integrating Istio for smart networking
- AI/ML based workflows & policies for self-learning and optimisation
- More CI and APM integrations
- Additional release, validation, experimentation and optimisation workflows
- Hybrid containers & serverless support
- Cloud-based version with new & improved core “engine”