End-to-end Machine Learning using Kubeflow

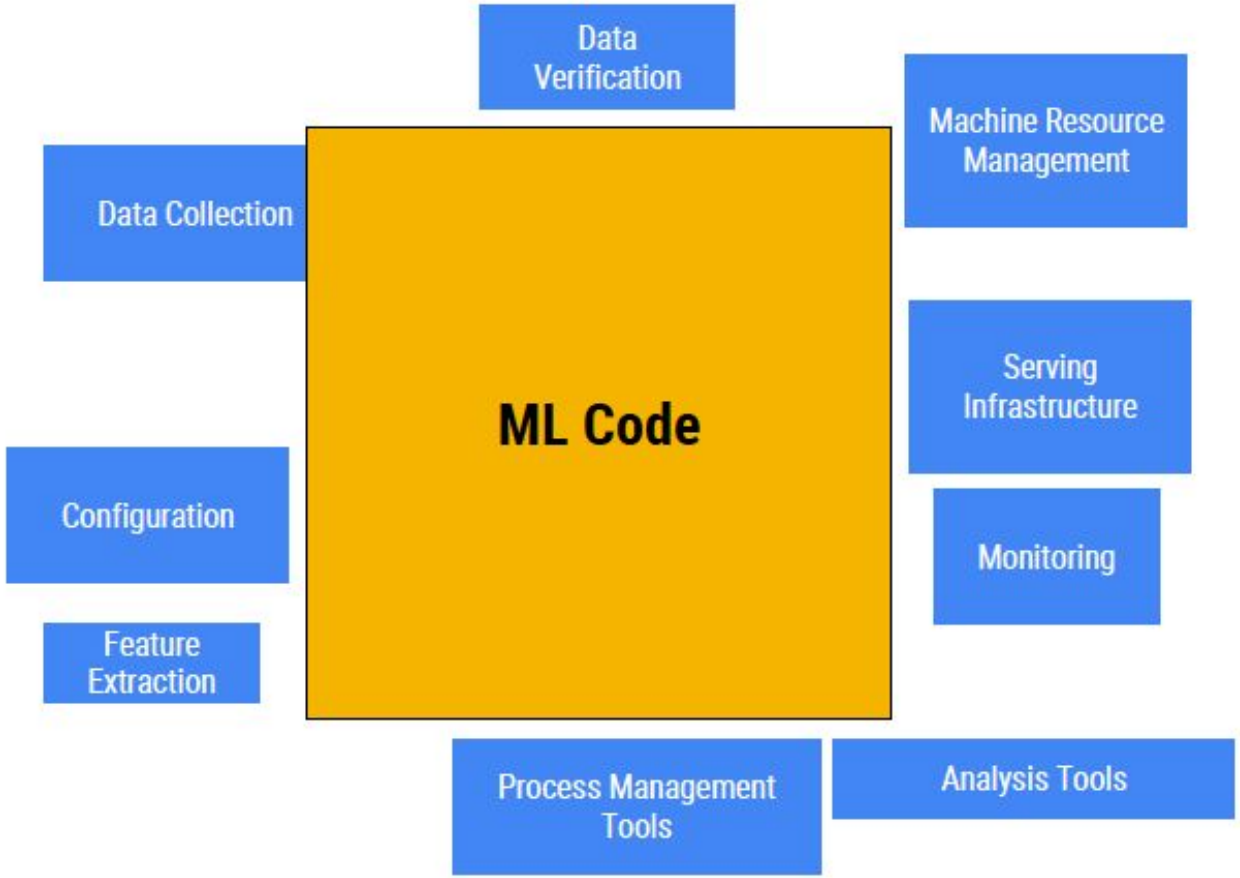
Johnu George, Amit Saha

07th Jan 2022, CODS-COMAD

ML science projects are cute…

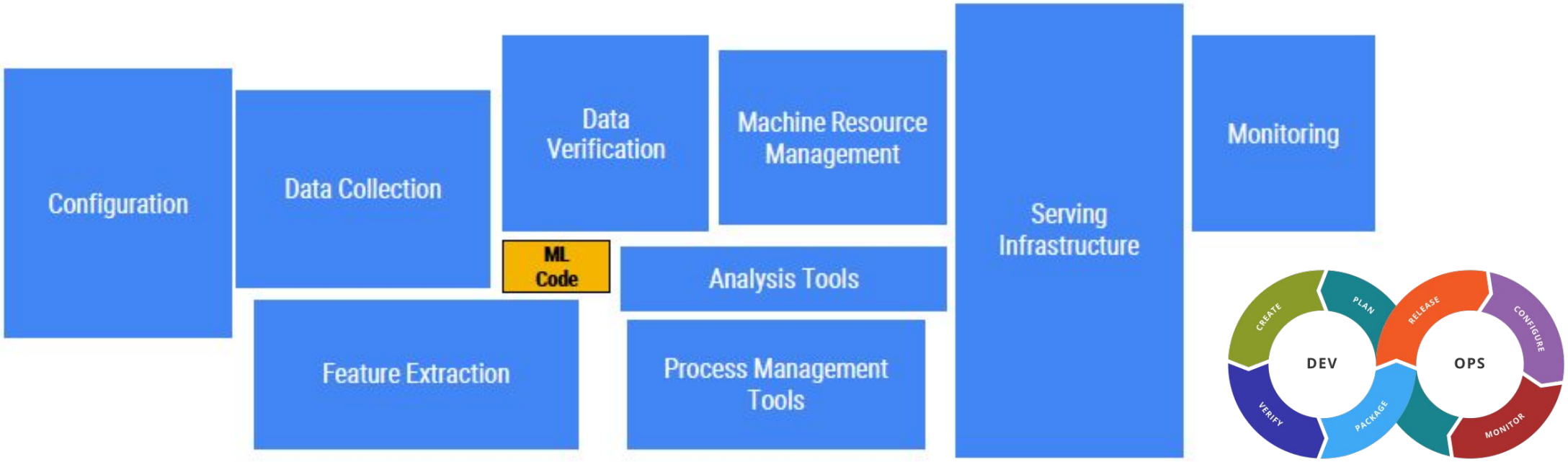
…but the calf won’t survive in the wild without the herd

Perception



SOURCE : Kubecon 2018 Talk: Building ML Products with Kubeflow

Reality



SOURCE : Kubecon 2018 Talk: Building ML Products with Kubeflow

ginablaber
@ginablaber

Follow

The story of enterprise Machine Learning: “It took me 3 weeks to develop the model. It’s been >11 months, and it’s still not deployed.” @DineshNirmalIBM #StrataData #strataconf

10:19 AM - 7 Mar 2018

Data Scientists can’t code!

https://insidebigdata.com/2019/08/13/help-my-data-scientists-cant-write-production-code/

Help! My Data Scientists Can’t Write (Production) Code!

August 13, 2019 by Editorial Team Leave a Comment

tweet share share email

Businesses across the world are hiring data scientists to beef up their efficiency and competitiveness via artificial intelligence (AI). Startup companies (dubbed AI-First companies) are disrupting traditional industries like banking, insurance, real estate and healthcare using AI technologies.

The demand for data scientists far exceeds supply. And, the problem is exacerbated by the fact that the data scientist profession is itself splitting into multiple sub-disciplines. Some examples of this divide include:

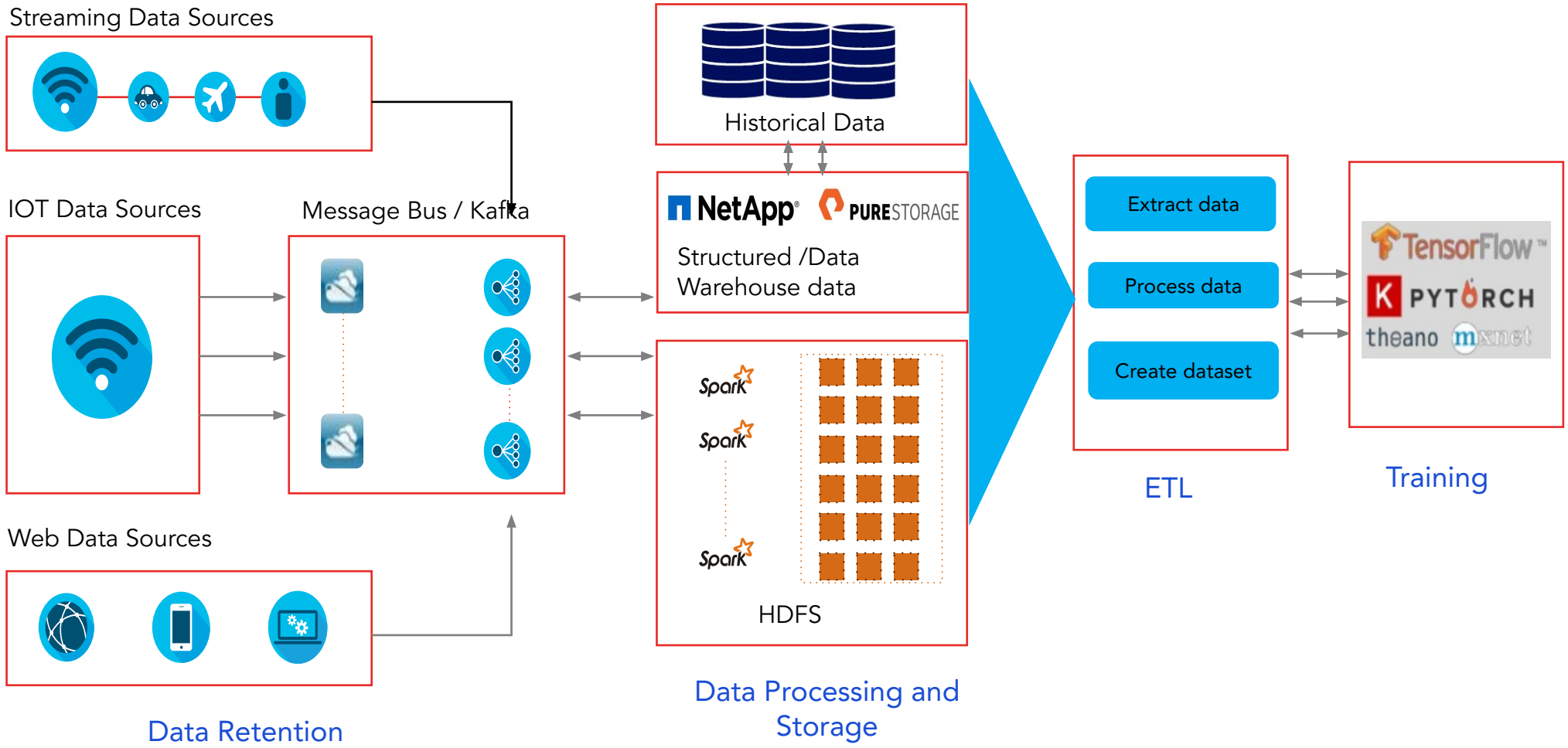
- Decision scientists have domain expertise and specialize in linking the domain knowledge and the algorithm to solve a problem.
- Data scientists have expertise with machine learning (ML) and related algorithmic fields at the application level, i.e., they know how to apply algorithms to data sets to generate successful experimental insights.
- ML research scientists can create new algorithms to solve more custom problems or adapt/exploit recent research advances.

Regardless of which of these skill-sets are needed, businesses face a common problem when

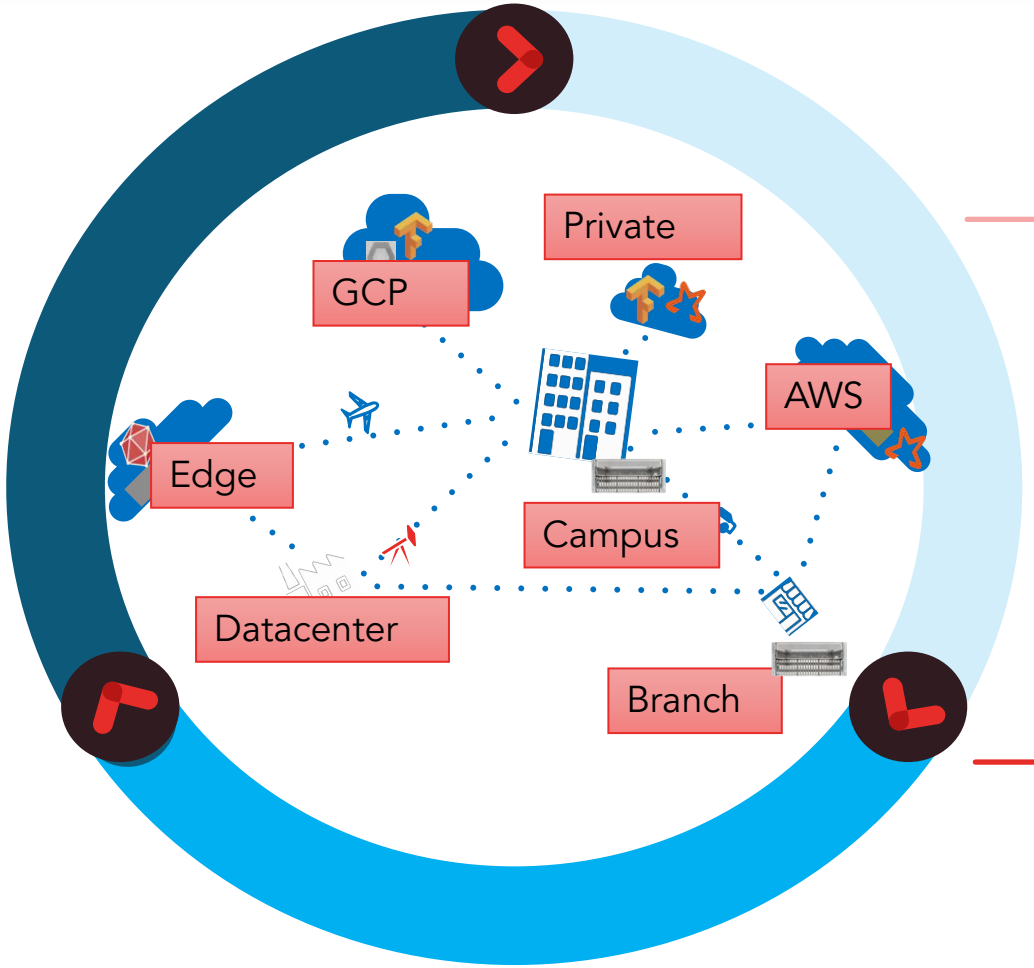
None of the above skill-sets are required to include a strong competency in production software development

application (which itself is a program, frequently in Python, R or Java) running for a business purpose. Not only that – the level of programming interest and capability varies widely from role to role and individual to individual.

The typical mess



… and it gets messier



Data

Ingest
Analyze
Transform
Validate
Split

Training

Model Dev
Model Validation
Training at Scale

Deployment/ Inference

Roll-out
Serving
Monitoring
Logging

Operationalizing ML is not easy

Kubeflow takes you from ML Science to ML Engineering in a multi-cloud world

Kubeflow

cisco

Agenda

01 ML in Production

02 Kubeflow Basics

03 Katib

04 Conclusion

www.kubeflow.org

Kubeflow is an open-source project dedicated to making deployments of machine learning (ML) workflows on Kubernetes simple, portable, and scalable

Kubeflow Features


Model Training


Model Serving


Data Science Services

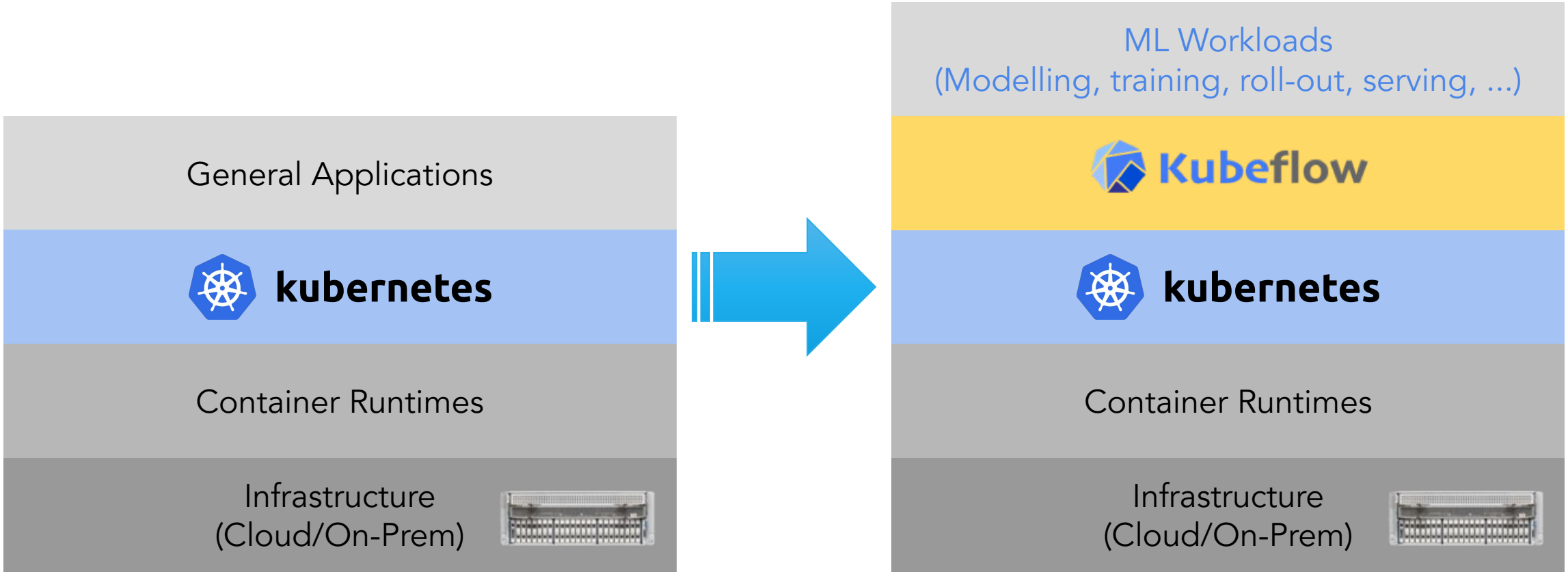

Experiment Tracking

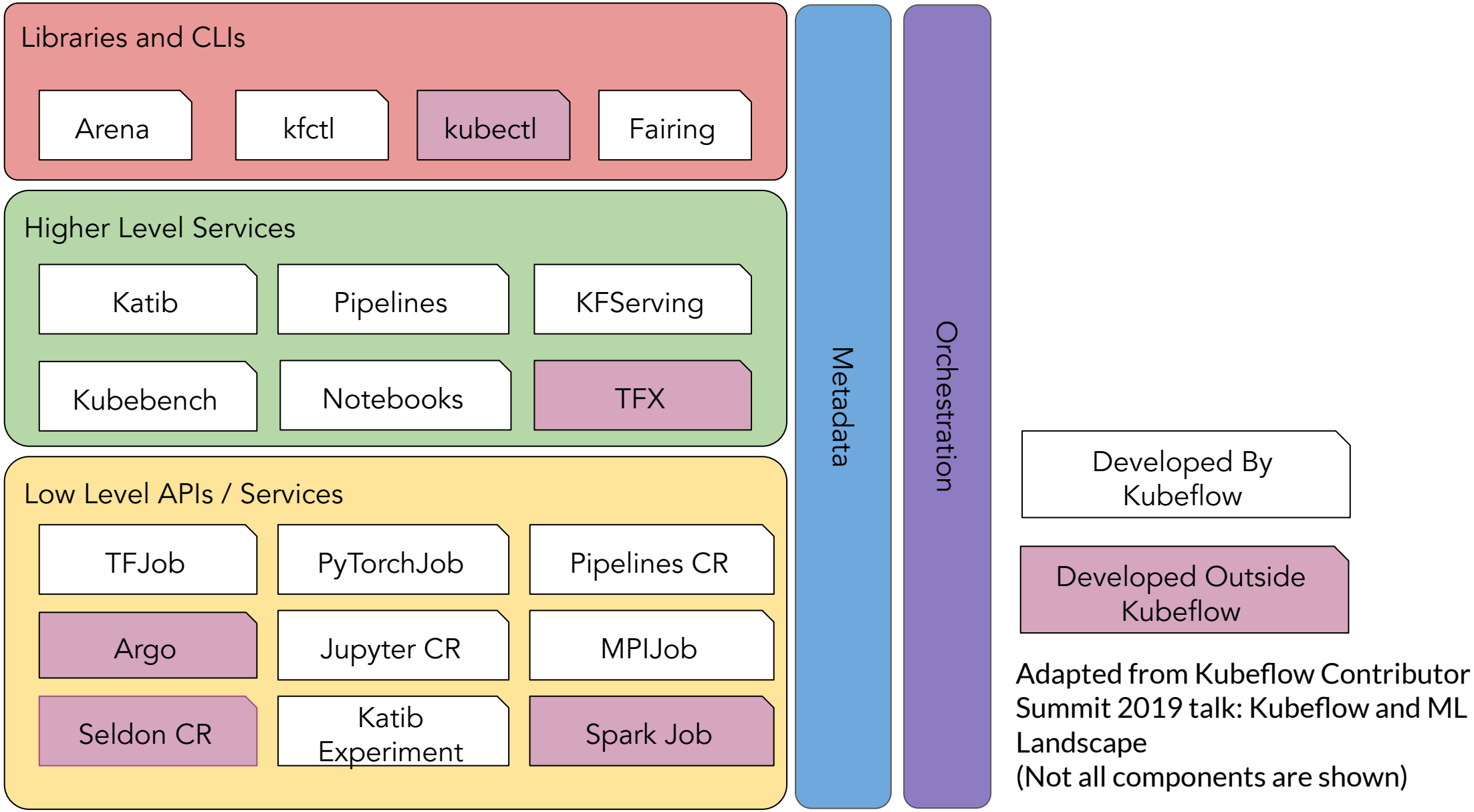

Model Management


Hyperparameter Tuning


Workflow Automation

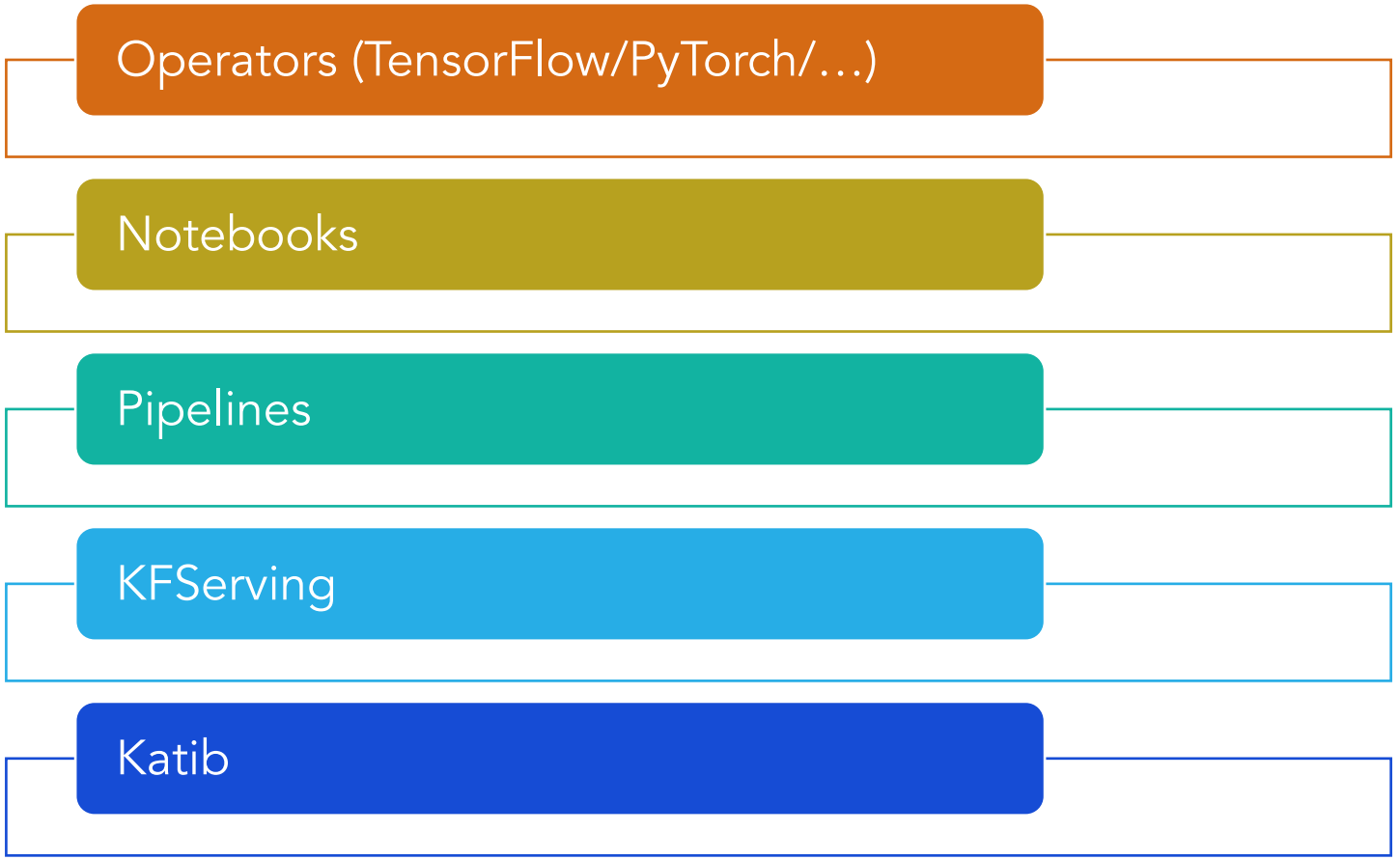
Kubeflow Architecture





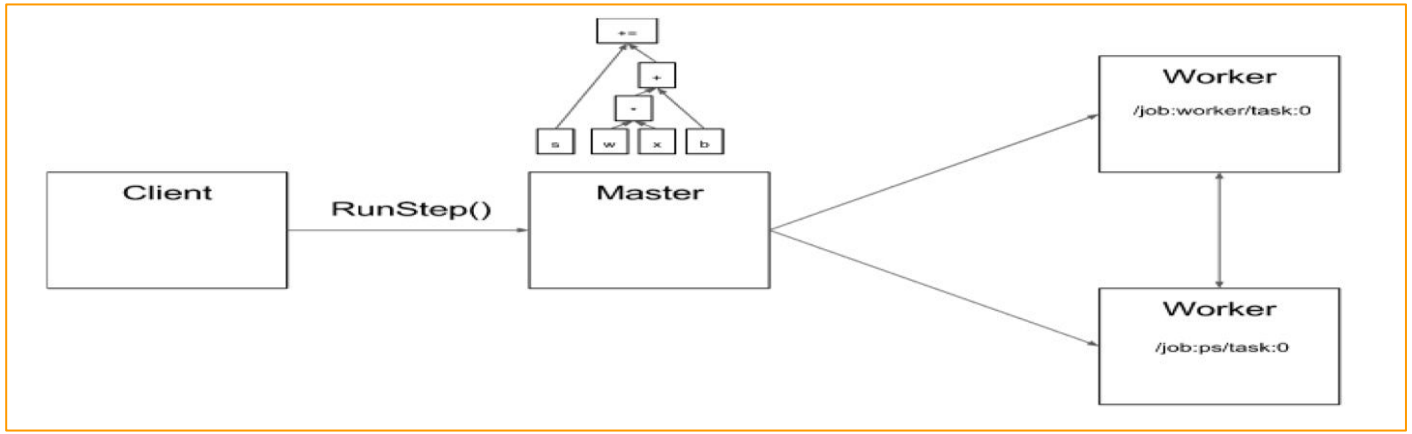
Adapted from Kubeflow Contributor Summit 2019 talk: Kubeflow and ML Landscape
(Not all components are shown)

Kubeflow Constructs

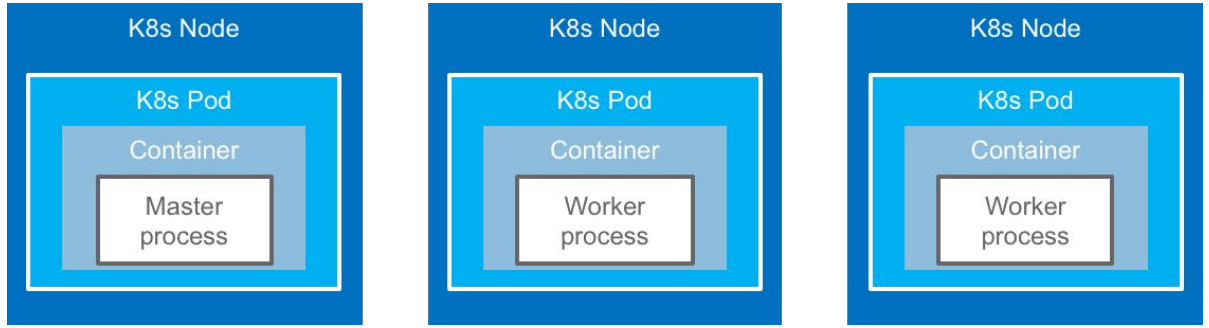


Operators

- Provides a set of K8s Custom Resources that turns distributed concepts in ML frameworks into K8s resources
- Makes it easy to configure and run local/distributed training jobs of various ML frameworks on K8s



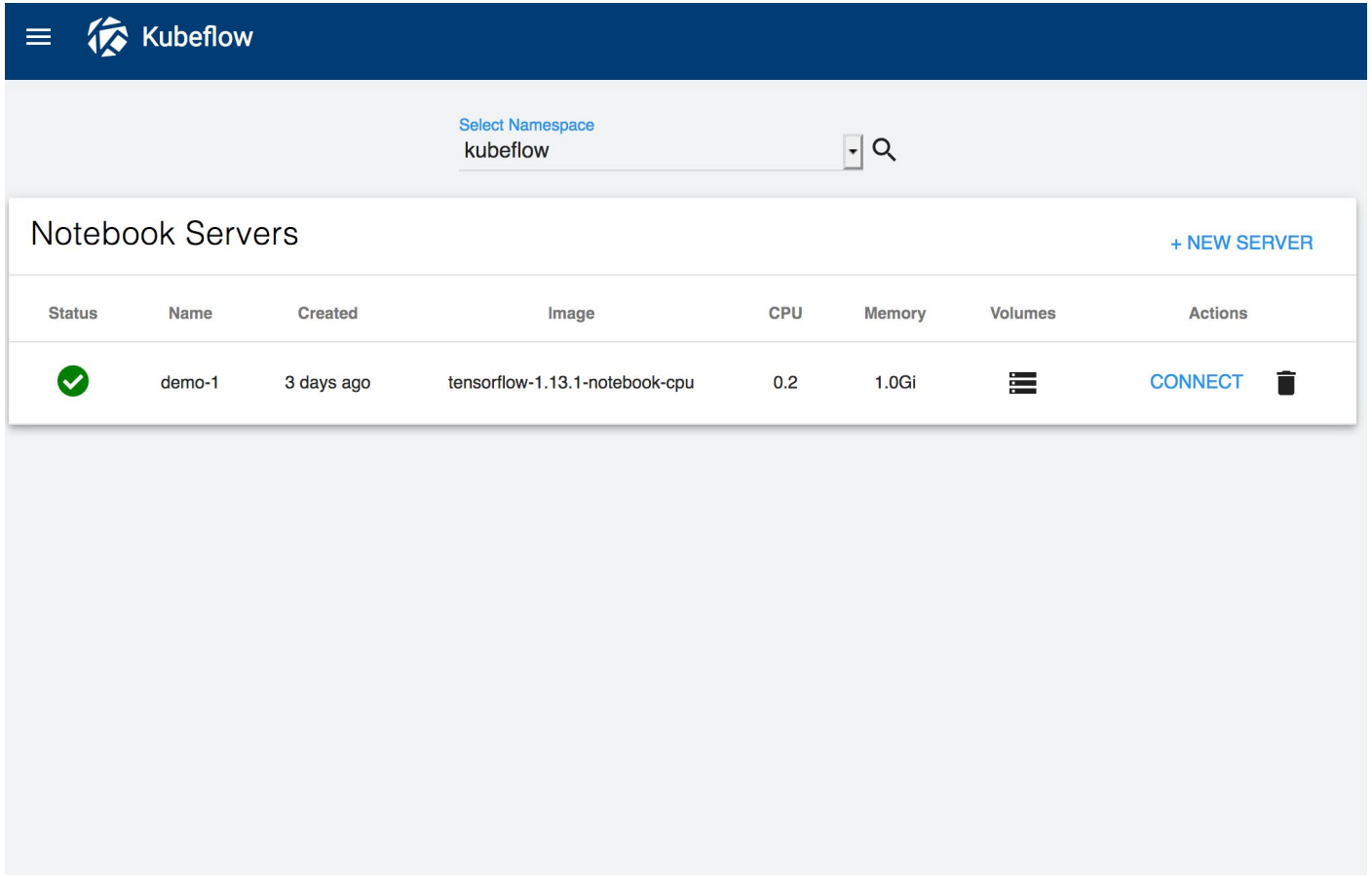
TFJob Operator



kubernetes

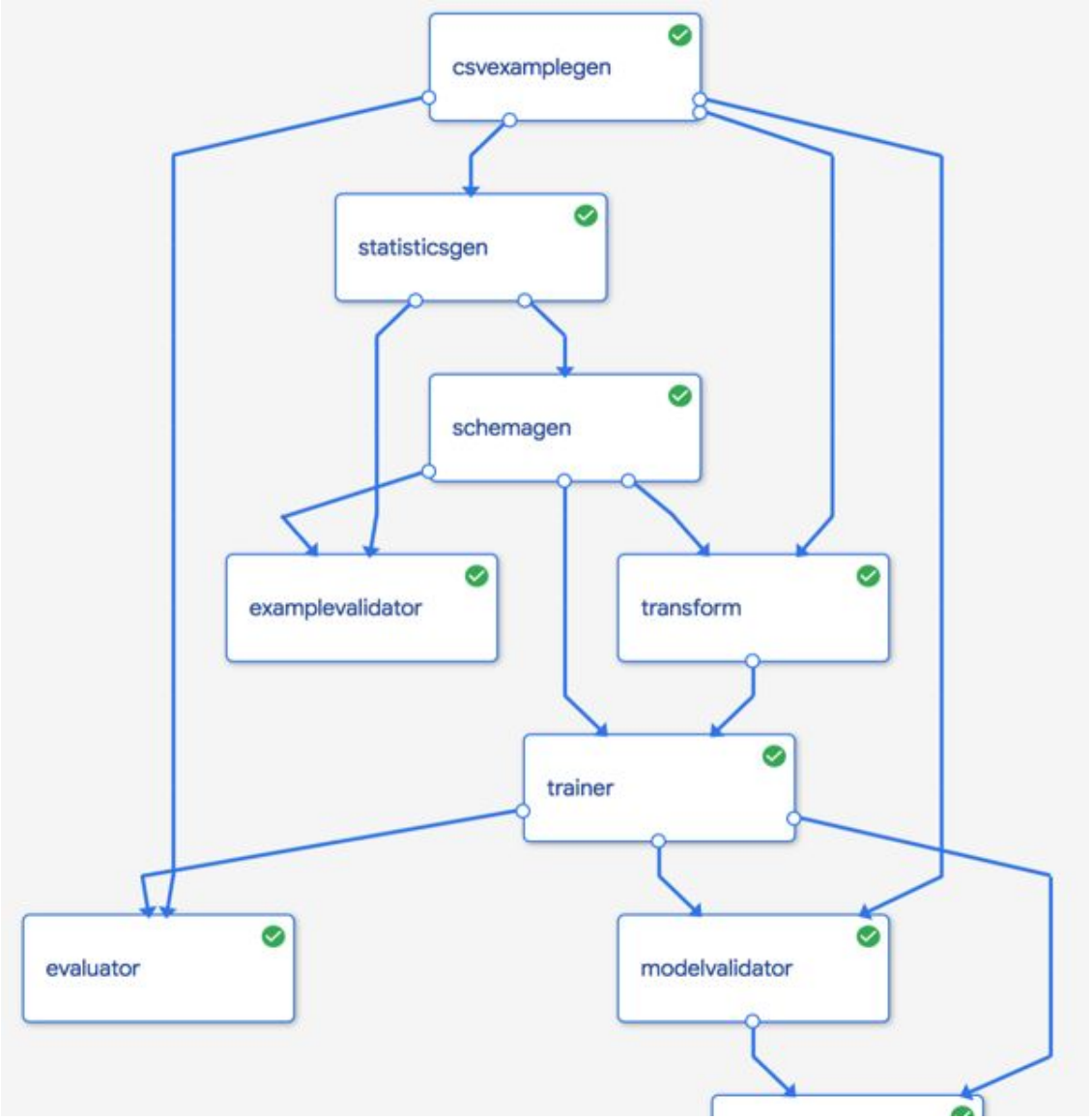
Notebooks

- Orchestration for Notebooks
  - Create and manage multiple Notebook servers in one place
- Integration with advanced DSLs
  - Kubeflow Fairing: build, train and serve, all from Notebooks
  - Kubeflow Pipeline SDK: create and deploy workflows from Notebooks



Pipelines

- Combine individual tasks into end-to-end workflows
- Provide orchestration and service integration
- Enable components & sharing
- Help with job tracking, experimentation, monitoring



Agenda

01
ML in Production

02
Kubeflow Basics

03
Katib

04
Conclusion

Katib – Hyperparameter Tuning

What is HP-tuning

Hyperparameters are parameters external to the model and are fixed before the training process

e.g., Learning rate, batch size, number of epochs, momentum

Hyperparameter Tuning finds values for hyperparameters that optimizes an objective function

"Meta" learning task

e.g., Finding the optimal batch size and learning rate to maximize prediction accuracy

Why is HP-tuning hard?

- Manual tuning is inefficient and error-prone
- More hyperparameters → exponential search space growth
- Need to tracking metrics across multiple jobs
- Managing resources and infrastructure for jobs
- Variety of frameworks and algorithms to support

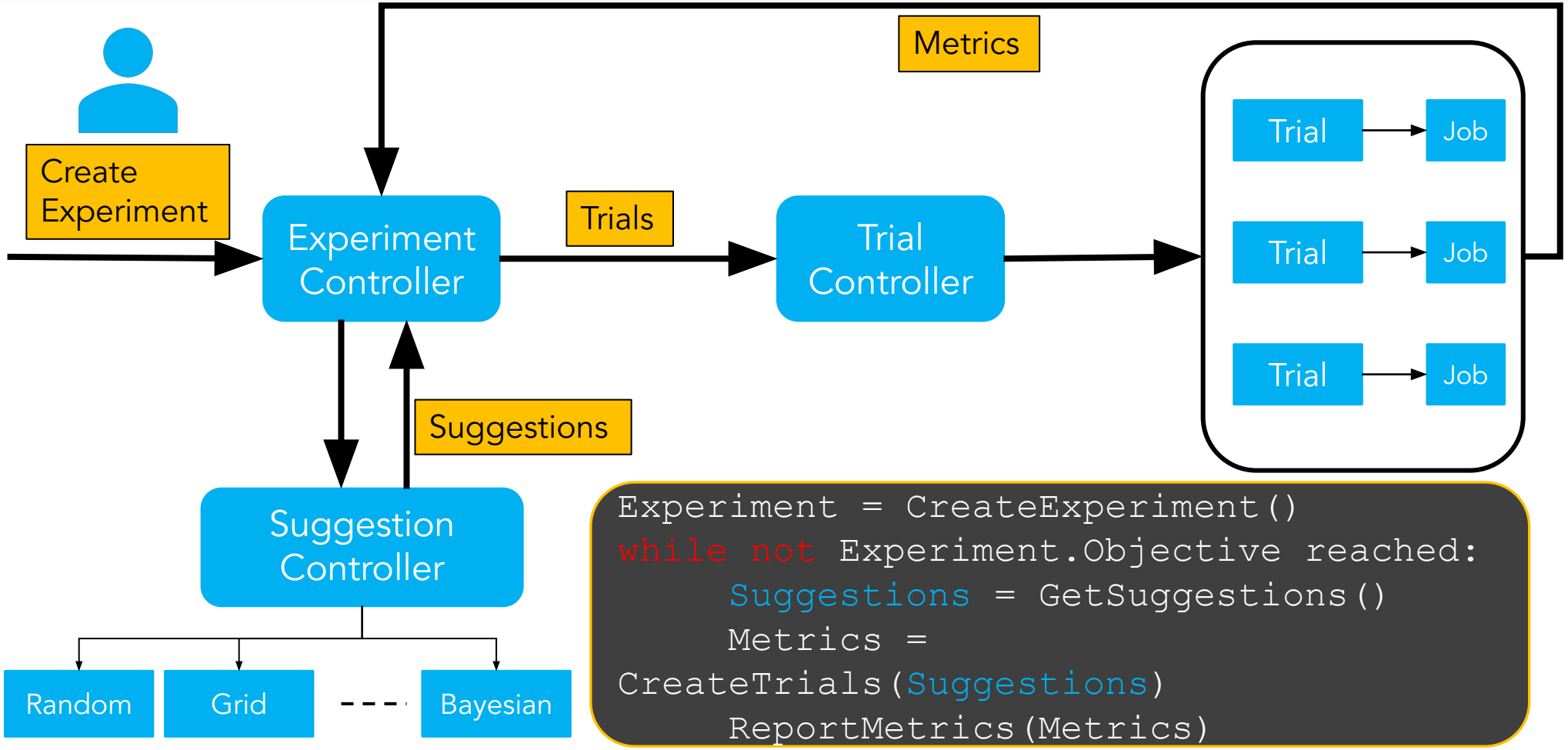
Katib – HP-tuner in Kubeflow

Framework agnostic – TensorFlow, PyTorch, MxNet, …

Customizable algorithm backend

- Random search
- Grid search
- Bayesian optimization
- Hyperband

Katib – System Architecture




```
Experiment = CreateExperiment()
while not Experiment.Objective reached:
    Suggestions = GetSuggestions()
    Metrics =
CreateTrials(Suggestions)
    ReportMetrics(Metrics)
```

Agenda

01
ML in Production

02
Kubeflow Basics

03
Katib

04
Conclusion

Conclusion

Production ML is non-trivial

Kubeflow for ML lifecycle management

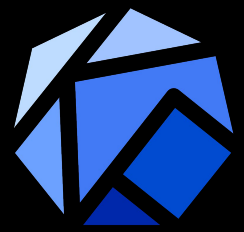
Enables modelling, training, tuning, serving, monitoring, etc.

Simplifies ML dev and ops

Runs on on-prem servers/software and on public cloud

Action Items (for you)

www.kubeflow.org
https://github.com/johnugeorge/cods-comad-2022-tutorial



Kubeflow