Machine Learning as Code: A Year of Democratizing ML with Kubernetes and Kubeflow

David Aronchick - Co-founder, Kubeflow
@aronchick

Jason "Jay" Smith - Customer Engineer, Google
@thejaysmith

One Year Ago...

What is Machine Learning?

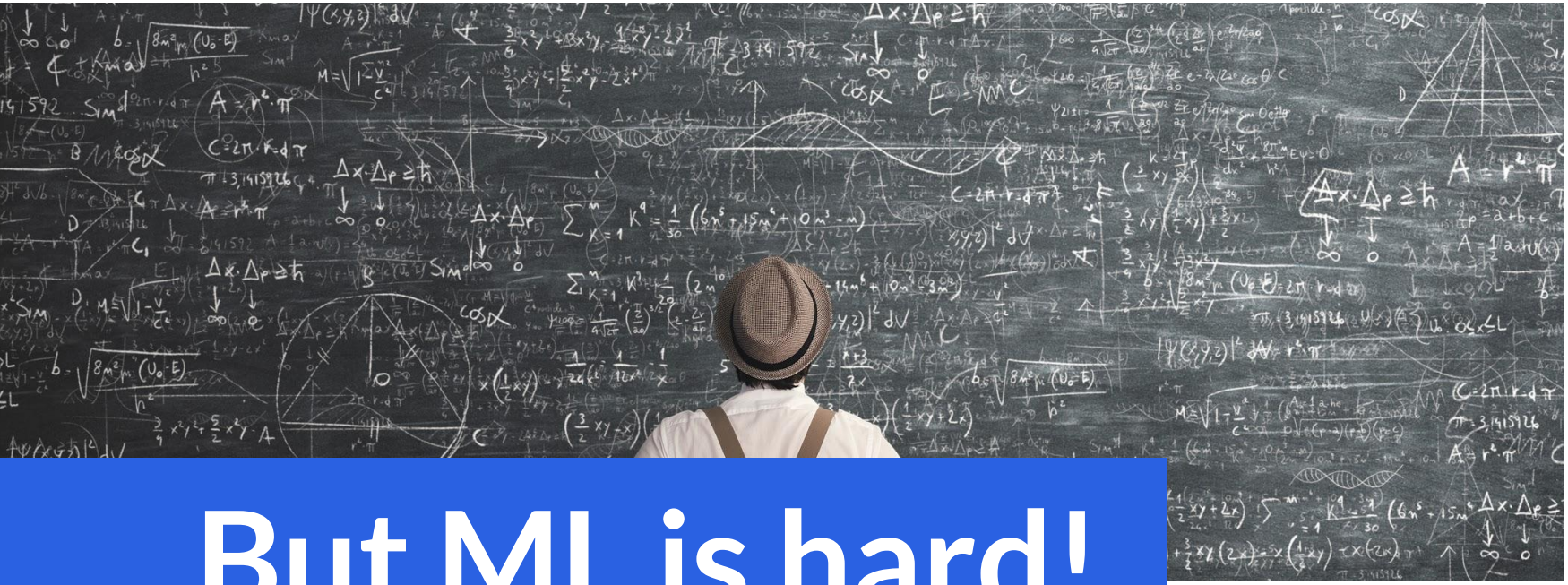
Machine Learning is a way of solving problems without explicitly knowing how to create the solution.

Google DC Ops

High PUE

Low PUE

PUE == Power Usage Effectiveness



But ML is hard!

Containers & Kubernetes

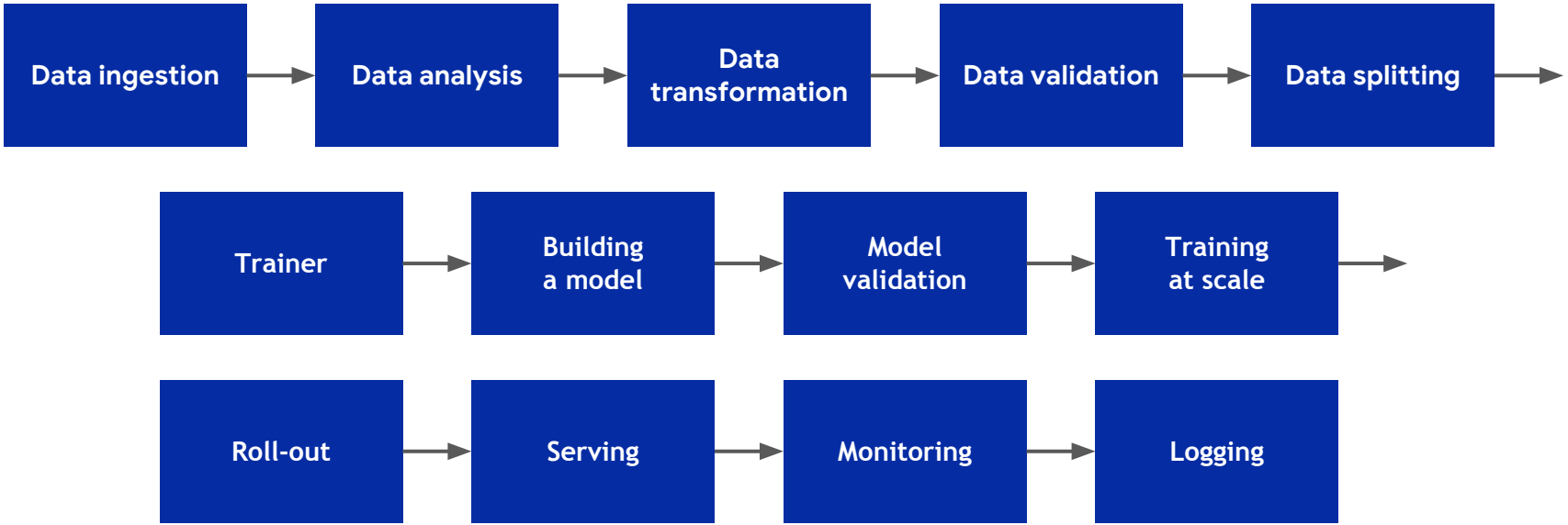
Cloud Native Apps

Cloud Native ML?

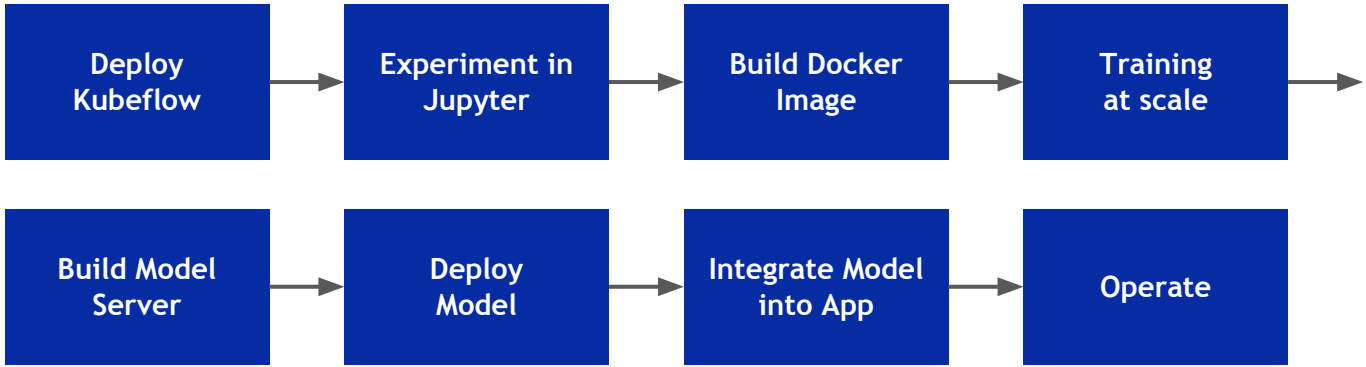
Platform

Building a model

Platform


Data ingestion
Data analysis
Data transformation
Data validation
Data splitting
Trainer
Building a model
Model validation
Training at scale
Roll-out
Serving
Monitoring
Logging

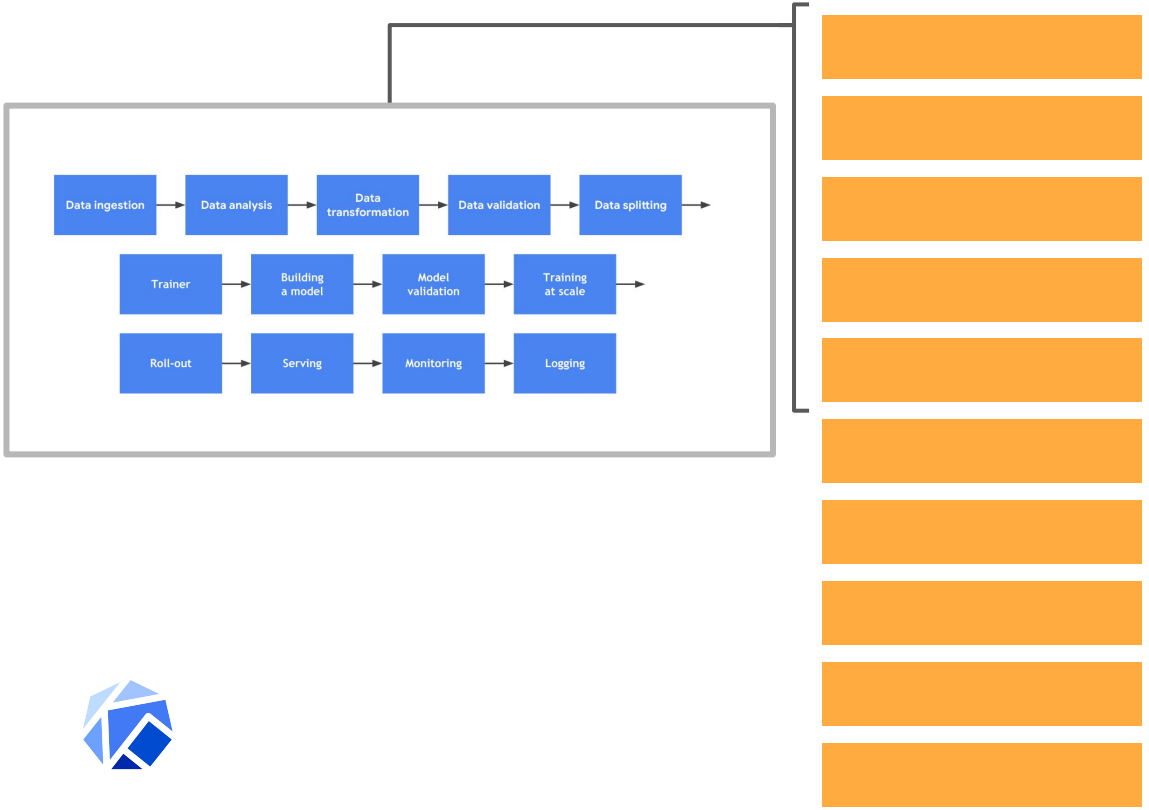
User Experience



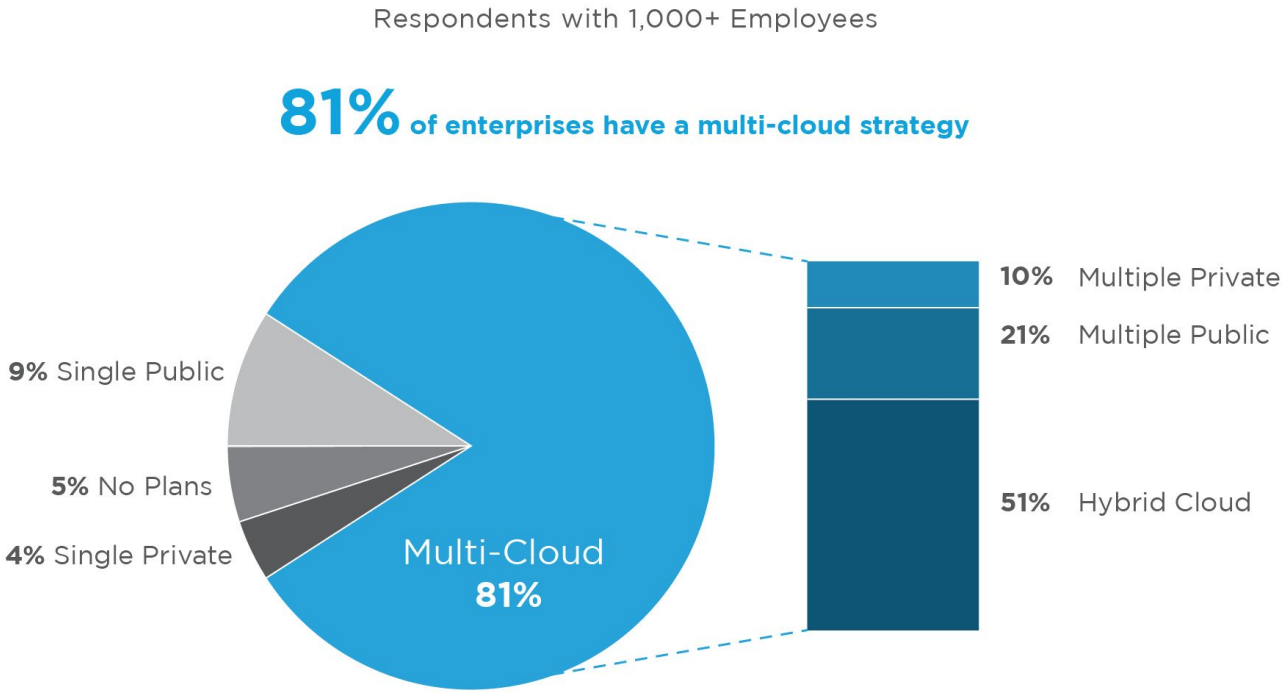
Experimentation

- Model
- UX
- Tooling
- Framework
- Storage
- Runtime
- Drivers
- OS
- Accelerator
- HW

Experimentation



Multi-Cloud is the Reality



Source: RightScale 2018 State of the Cloud Report

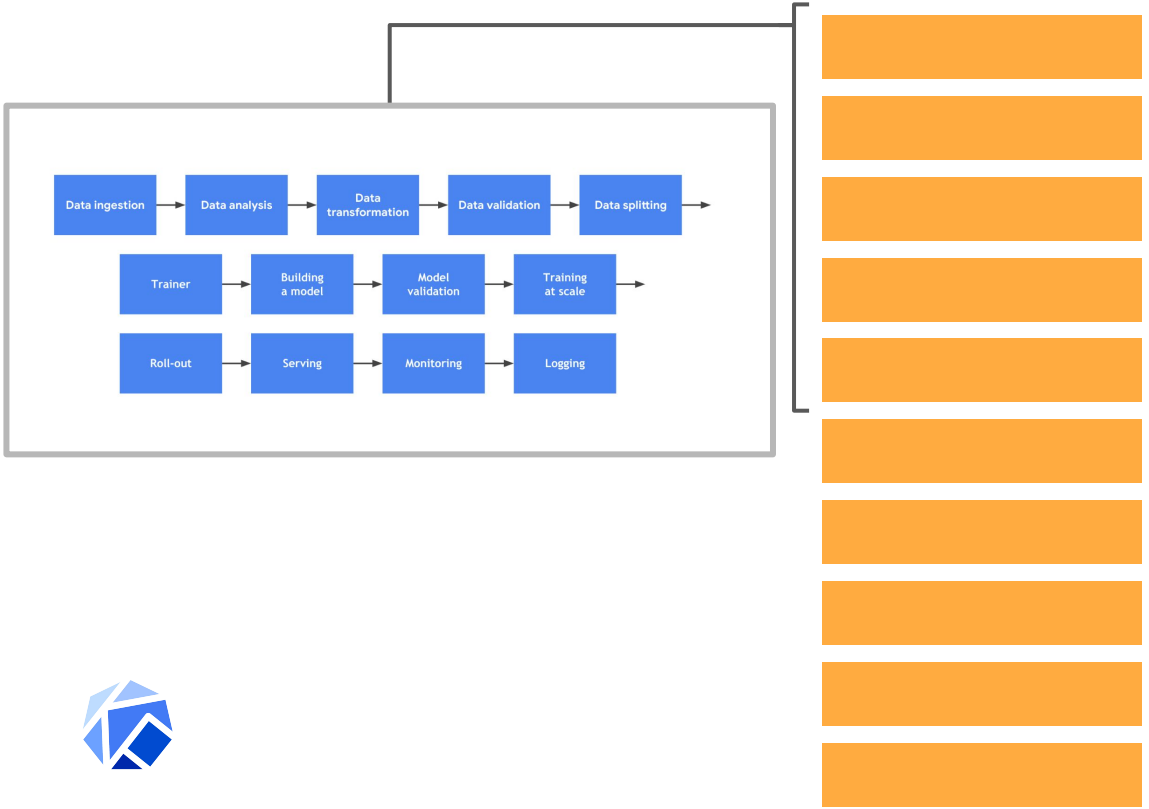
And Not Just One Cloud!

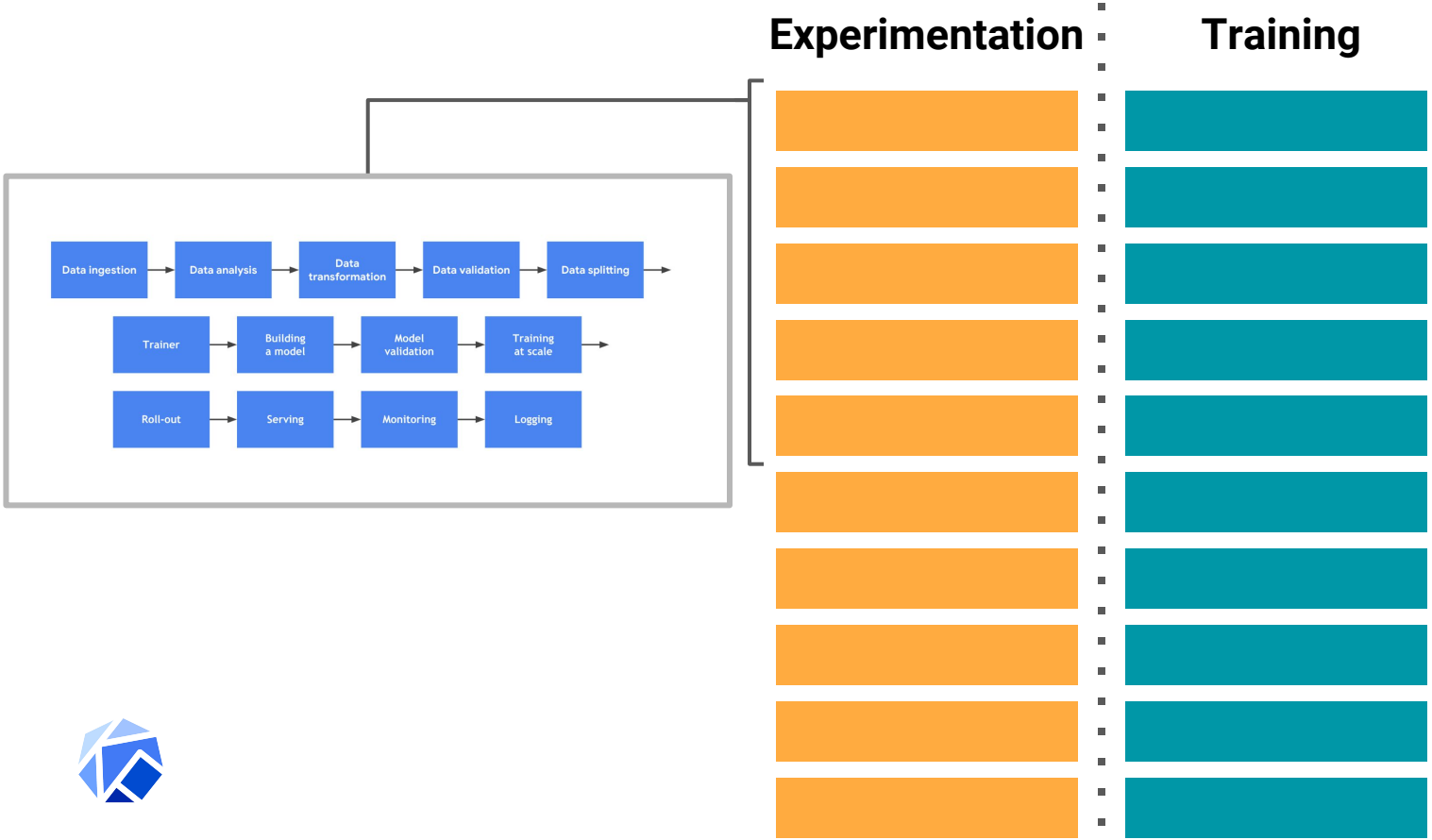
Companies using almost 5 public and private clouds on average

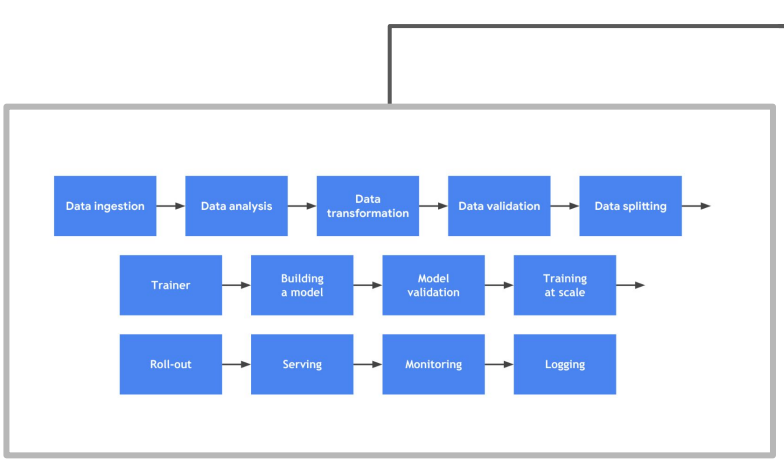
Public + Private Clouds Used	Average *All respondents*	Median *All respondents*
Running Applications	3.1	3.0
Experimenting	1.7	1.0
Total	**4.8**	**4.0**

Source: RightScale 2018 State of the Cloud Report

Experimentation




Experimentation
Training
Data ingestion
Data analysis
Data transformation
Data validation
Data splitting
Trainer
Building a model
Model validation
Training at scale
Roll-out
Serving
Monitoring
Logging



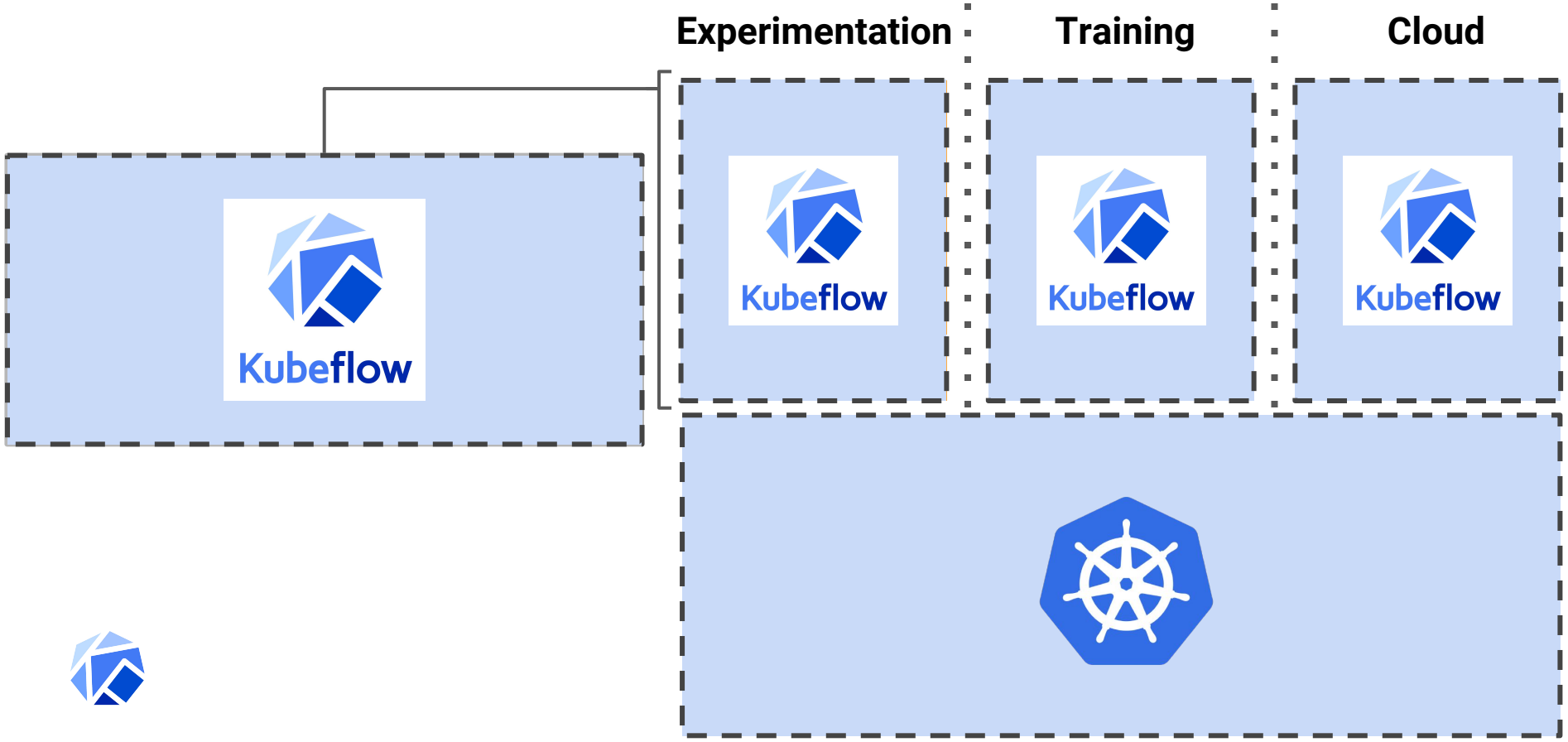
Experimentation

Training

Cloud

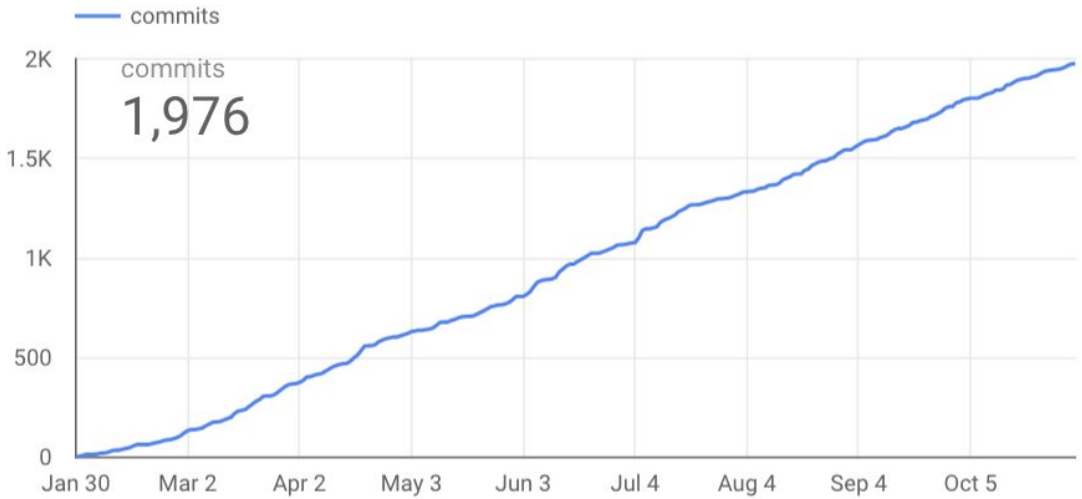
Kubecon 2017

Make it Easy for Everyone to Develop, Deploy and Manage Portable, Distributed ML on Kubernetes


Experimentation
Training
Cloud
Kubeflow
Kubeflow
Kubeflow
Kubeflow

Cloud Native ML!

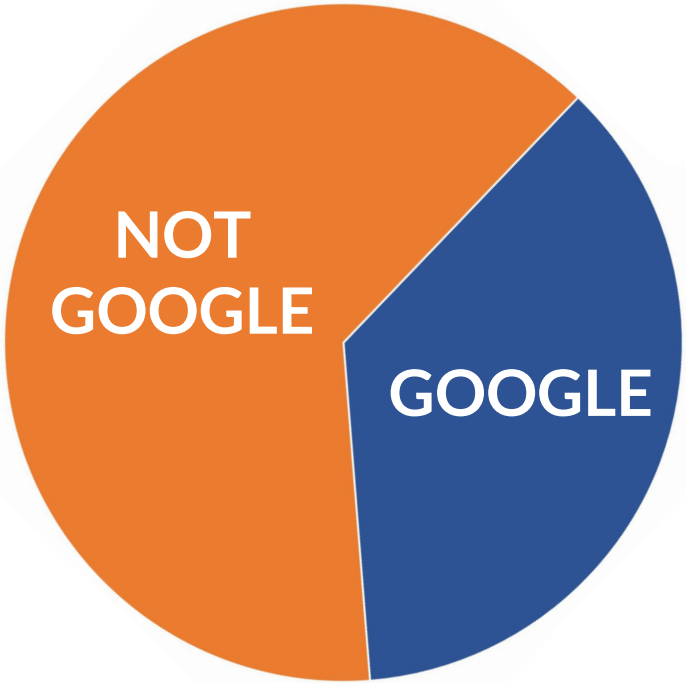
Momentum!



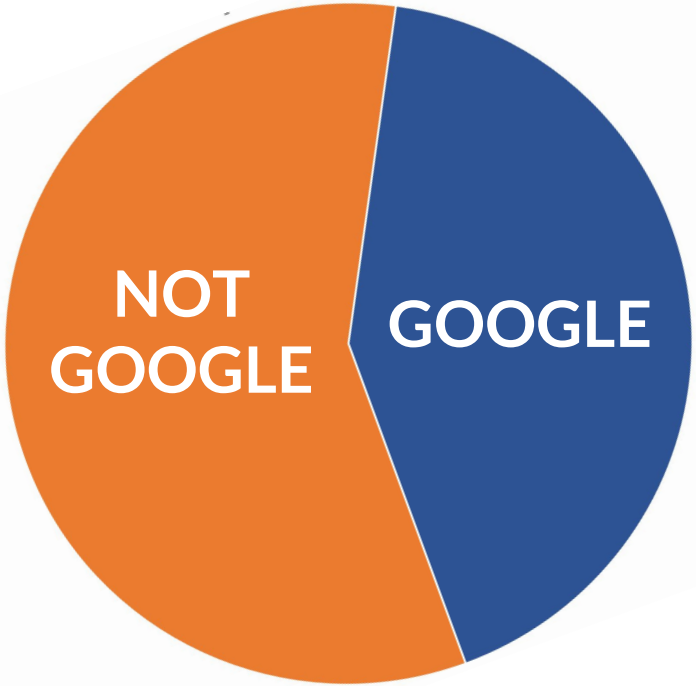
- 1900+ commits
- 100+ Community contributors
- 30+ Companies contributing, including:

Alibaba Cloud · 蚂蚁金服 ANT FINANCIAL · Caicloud · CANONICAL · cisco · vmware · DELL · GitHub · Google · heptio · HUAWEI · intel · Microsoft · MOMENTA · Pachyderm · jupyter · redhat · Seldon Technologies · UBER · weaveworks · Arrikto · lyft · IBM · aws · NVIDIA · intuit · DATAWIRE · NTT

Community Contributions



Kubernetes

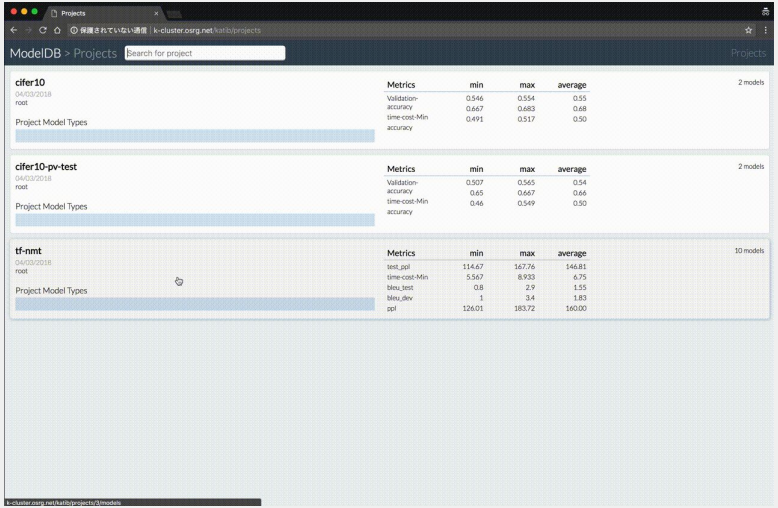


Kubeflow

Community Contribution

- Pluggable microservice architecture for HP tuning
  - Different optimization algorithms
  - Different frameworks
- StudyJob (K8s CRD)
  - Hides complexity from user
  - No code needed to do HP tuning

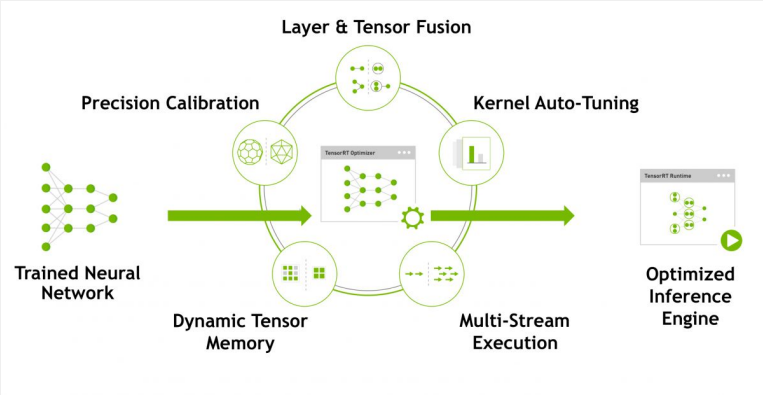
Katib from NTT



Community Contribution

- Production datacenter inferencing server
- Maximize real-time inference performance of GPUs
  - Multiple models per GPU per node
  - Supports heterogeneous GPUs & multi GPU nodes
- Integrates with orchestration systems and auto scalers via latency and health metrics

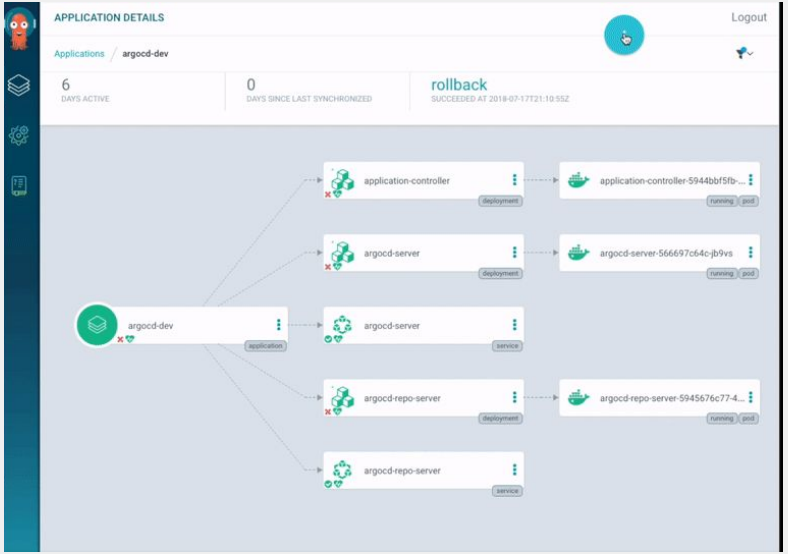
TensorRT from NVidia



Community Contribution

- Argo CRD for workflows
- Argo CRD is engine for Pipelines (more on that later)
- Argo CD for GitOps

Argo from Intuit



Community Contribution

- Jupyter Spawner
  - Simplifies starting a new notebook with all dependencies on KF
  - Contributions by Arrikto, Red Hat and Intel
- Seldon
  - Rich serving solution for multiple model types
  - Both commercial and OSS offering
- Kubebench
  - Run benchmark jobs on Kubeflow with various system and model settings
  - Leverages TFJobs & Argo
  - Major contributions from Cisco, others

Spawner

Kubebench

Introducing Kubeflow 0.4

(almost) Introducing Kubeflow 0.4

What's new in 0.4?

- Deploy
  - Application CRD
  - Simplified Setup
- Develop
  - Kubeflow Pipelines
  - TFJob/PyTorch beta

Click to Deploy

Click to Deploy

- **Problem**: It's too hard to install Kubeflow!
- **Solution**: A one-click installation tool, available via a clean web interface
- **How**:
  - Click to deploy uses a bootstrap container and kfctl.sh with all the necessary dependencies included
  - Also enables use of declarative infrastructure deployment (e.g. Deployment Manager on GCP)
  - **NO TEMPLATING TOOL NEEDED**

Demo

Kubeflow GitOps

GitOps

- **Problem:** Maintaining a cluster application is hard
- **Solution:** Implement a GitOps (coined by WeaveWorks) driven solution to manage the infrastructure and cluster code
- **How**:
  - ArgoCD runs the GitOps
  - Synchronize Kubeflow deployment with Git repository
  - https://www.kubeflow.org/docs/guides/gitops-for-kubeflow/

Demo

Kubeflow Pipelines

Pipelines

- **Problem:** ML solutions are often multi-stage
- **Solution:** Microservices platform designed to enable reusable components and workflow orchestration
- **How**:
  - Kubeflow Pipelines = a Python SDK for describing and containerizing ML tasks
  - Runs on Argo (already in the box) and offers experiment logging and analytics
  - Containerized steps lets you extend to your needs

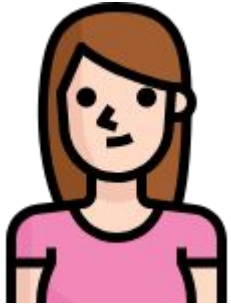
Demo

Auto-scaling

Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist

kubernetes



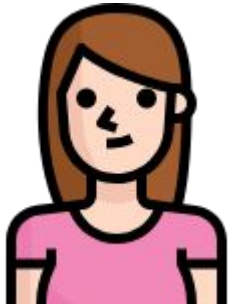
IT Ops

Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist

Model works great! But I need six nodes.

kubernetes



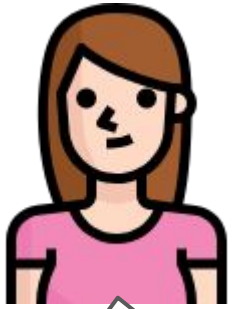
IT Ops

Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist

Model works great! But I need six nodes.

kubernetes



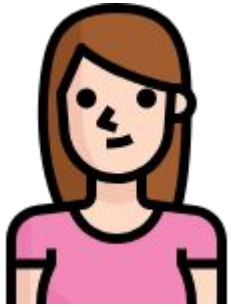
IT Ops

Sure thing, can I get to it after O(large number of things to do)?

Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist

kubernetes

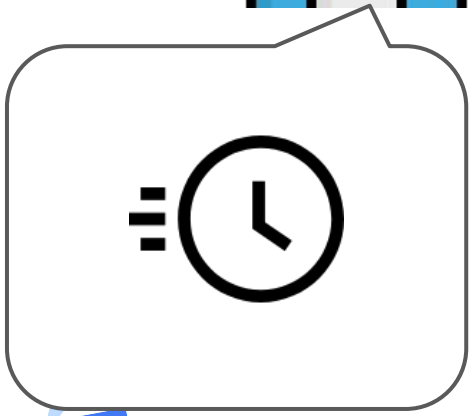


IT Ops

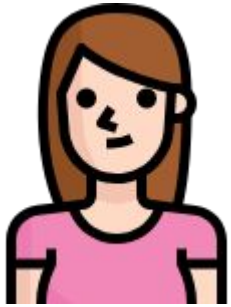
Rats. Ok, when you have the time.

Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist



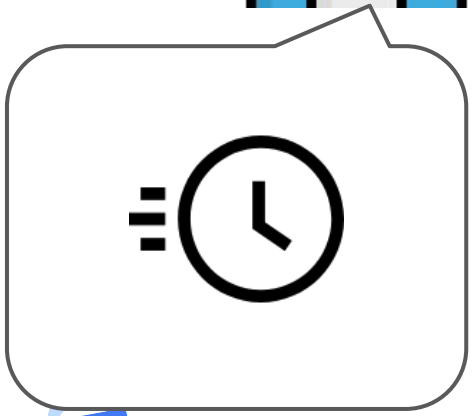
kubernetes

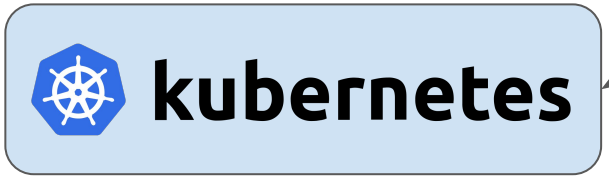


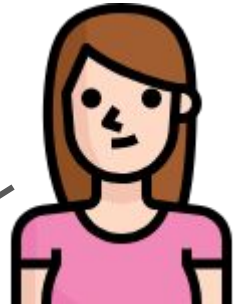
IT Ops

Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist



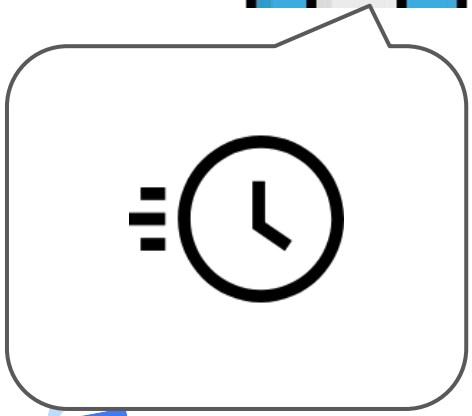




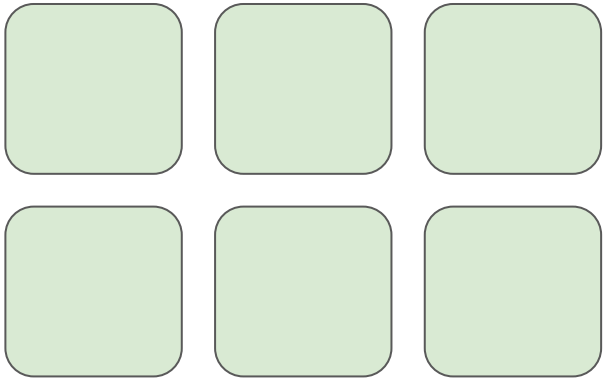
IT Ops

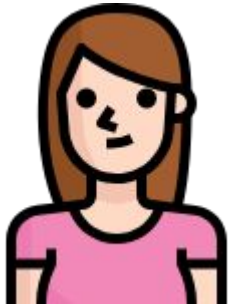
Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist



kubernetes



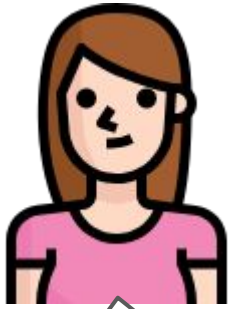


IT Ops

Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist

kubernetes



IT Ops

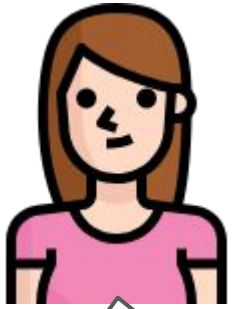
Whew... that took a while. Here you go!

Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist

Thanks!

kubernetes



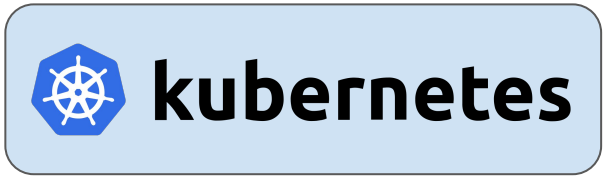
IT Ops

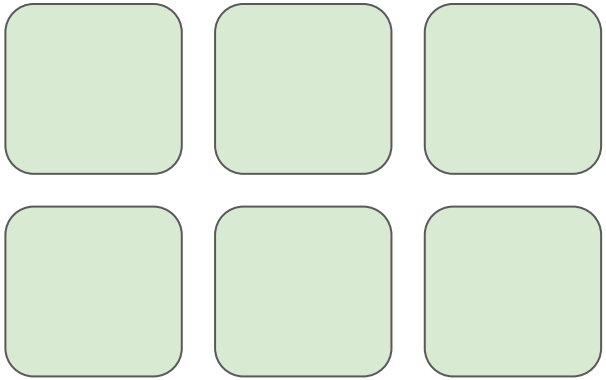
Whew... that took a while. Here you go!

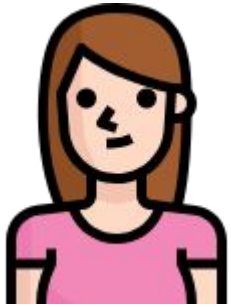
Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist

(Lots of Work)







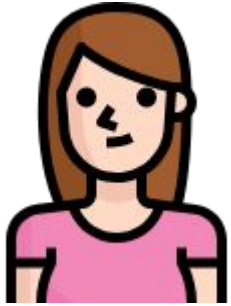
IT Ops

Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist

OK, I'm all done! Hope I'm not forgetting anything.

kubernetes

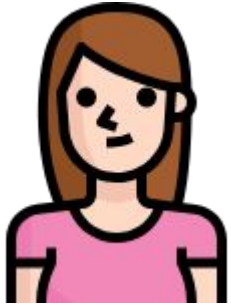


IT Ops

Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist

kubernetes

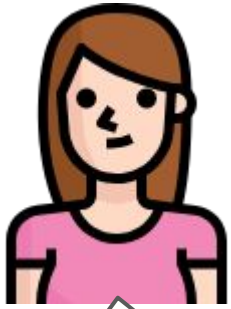


IT Ops

Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist

Oh noes! We forgot to turn it off!



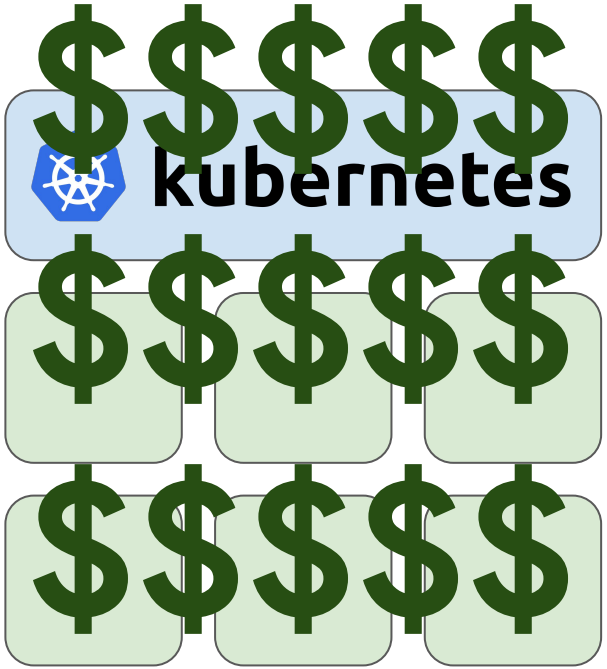
IT Ops

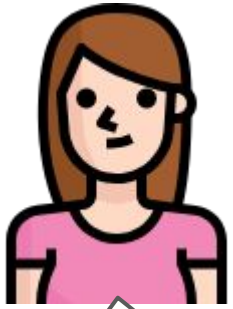
Oh noes! We forgot to turn it off!

Today, IT Ops Has a Lot of Stuff To Do...

Data Scientist

Oh noes! We forgot to turn it off!





IT Ops

Oh noes! We forgot to turn it off!

Autoscaling Jobs

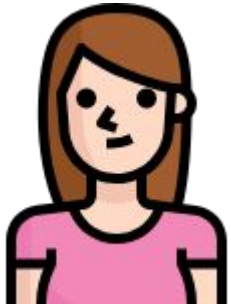
- Describe the job, let Kubernetes take care of the rest
  - CPU
  - RAM
  - Accelerators
- TF Jobs delete themselves when finished, node pool will auto scale back down (**PROTIP:** Save your logs elsewhere)
- Can be capped based on maximum scale parameters (your data scientists won't bankrupt you)

Let’s Give IT Ops the Day Off!

Data Scientist

Model works great! But I need six nodes.

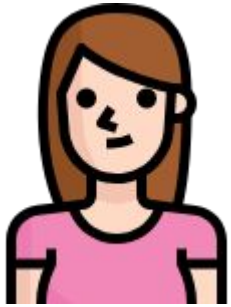
kubernetes



IT Ops

Let’s Give IT Ops the Day Off!

Data Scientist

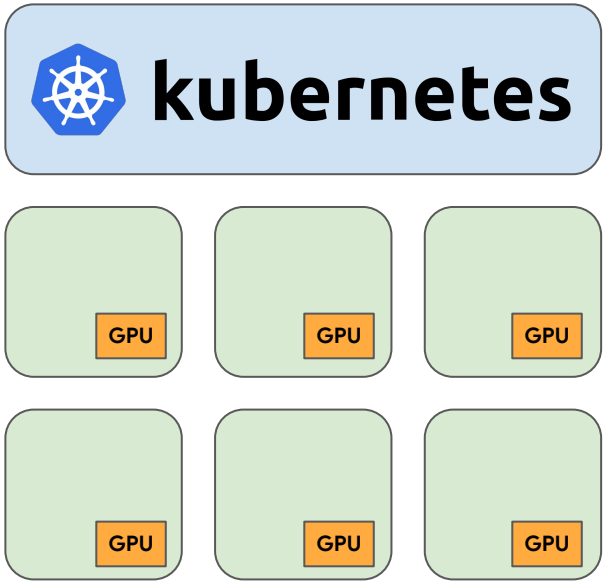


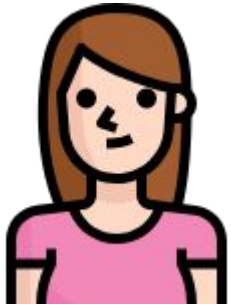
IT Ops

```
apiVersion: "kubeflow.org/v1alpha1"
kind: "TFJob"
spec:
  replicaSpecs:
    replicas: 6
    CPU: 1
    GPU: 1
    containers: gcr.io/myco/myjob:1.0
```

Let’s Give IT Ops the Day Off!

Data Scientist





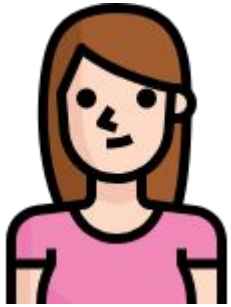
IT Ops

Let’s Give IT Ops the Day Off!

Data Scientist

Job’s Done!

kubernetes

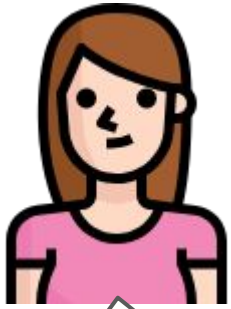


IT Ops

Let’s Give IT Ops the Day Off!

Data Scientist

kubernetes



IT Ops

Did you know that Youtube has 1 hour of cat videos uploaded every second?

Demo

Kubeflow Roadmap

We’re just getting started!

Our roadmap:

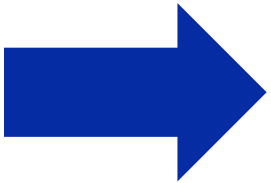
- Enterprise readiness (1.0, IAM/RBAC, clean upgrades)
- Better Jupyter Notebook Integration
- Pipeline Experiment Comparison & Model Management
- **You tell us!** (Or better yet, help!)

It’s a whole new world

- Data science will touch **EVERY** industry.
- We can’t ask people to become a PhD in statistics though.
- How do **WE** help <u>everyone</u> take advantage of this transformation?

Enabling ML EVERYWHERE

Let's give the people not in this room* the tools to change the world!



Nurses, Civil Engineers, Professors, Social Workers, Statisticians, Politicians, Teachers, Lawyers, Environmental Researchers, Housing Advocates, Scientists, Historians, ...

* Or watching this video

Kubeflow is open!

Open comm-unity

Open design

Open source

Open to ideas

Come Help!

- website: https://kubeflow.org
- github: https://github.com/kubeflow/kubeflow
- slack: kubeflow (http://kubeflow.slack.com)
- twitter: @kubeflow

David Aronchick @aronchick (aronchick@gmail.com)
Jason “Jay” Smith (jaysmith@google.com)