Case Study: Machine Learning as a Service in Production

November 20th, 2019
San Diego

Itay Gabbay
Machine Learning Lead, MOD (Israel)

Tushar Katarki
Product Manager, ML on OpenShift at Red Hat

משרד הביטחון
MINISTRY OF DEFENCE

Digital Transformation

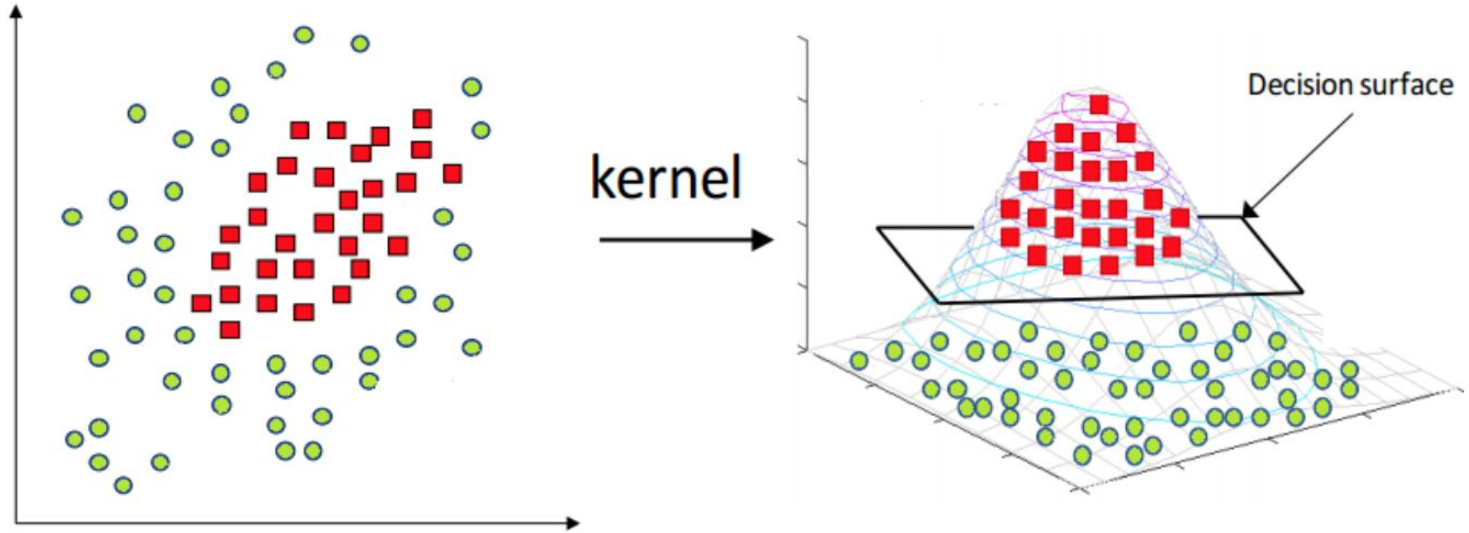
Private Cloud

Accelerate R&D Projects

Machine Learning

Why Machine Learning?

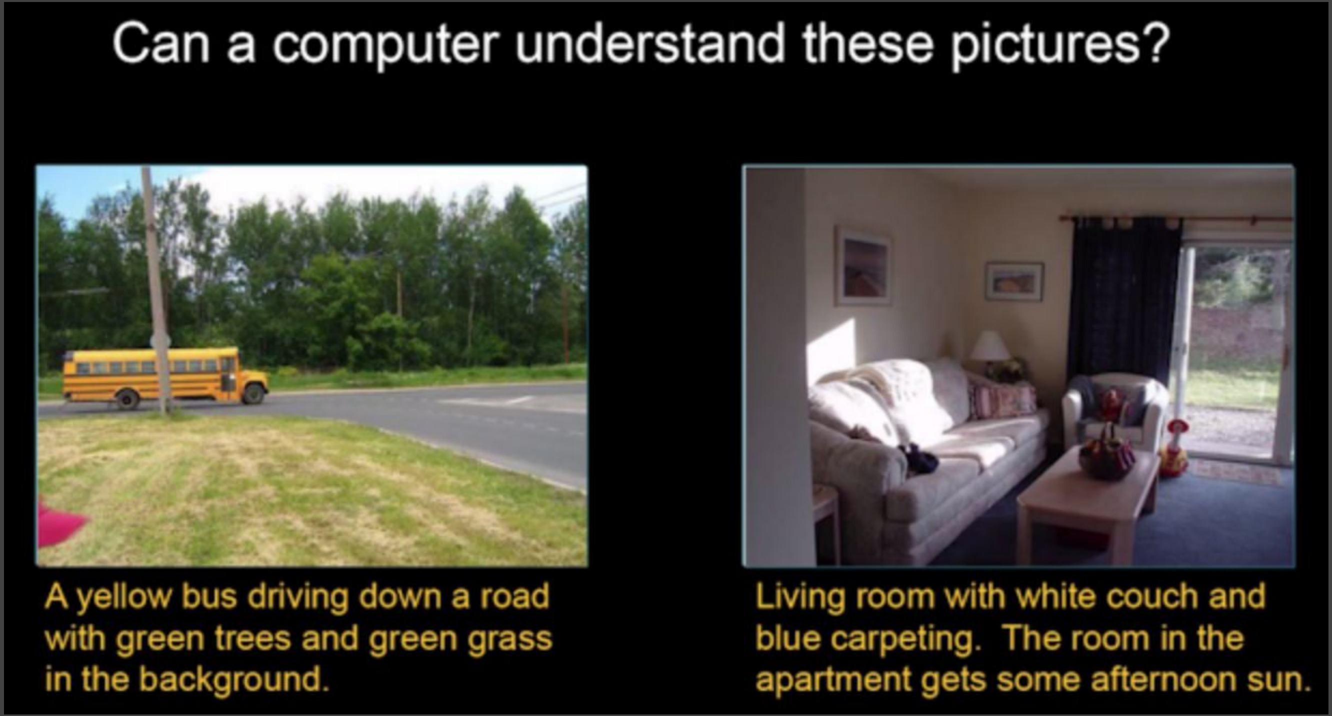
Machine Learning vs. Traditional Methods



SVM algorithm maps the points into 3D where they are separable by a plane (linear hyperplane in 3D)

Source: https://towardsdatascience.com/why-use-machine-learning-instead-of-traditional-statistics-334c2213700a

Deep Learning



Can a computer understand these pictures?

A yellow bus driving down a road with green trees and green grass in the background.

Living room with white couch and blue carpeting. The room in the apartment gets some afternoon sun.

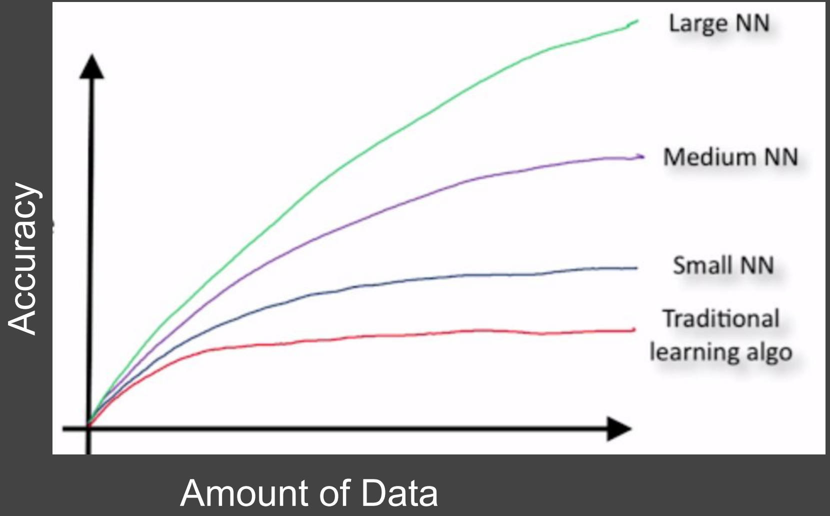
Source: https://www.nextbigfuture.com/2016/03/how-scale-is-enabling-deep-learning.html

AI vs Humans !

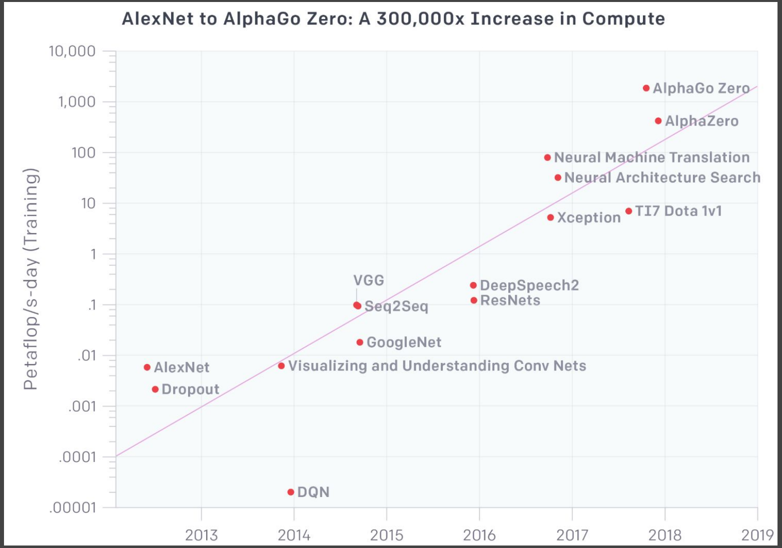
Go champion Lee Sedol, on the right, concedes the second of possible five games vs. Google's AlphaGo AI.

Characteristics of Machine Learning

Data and Compute Intensive

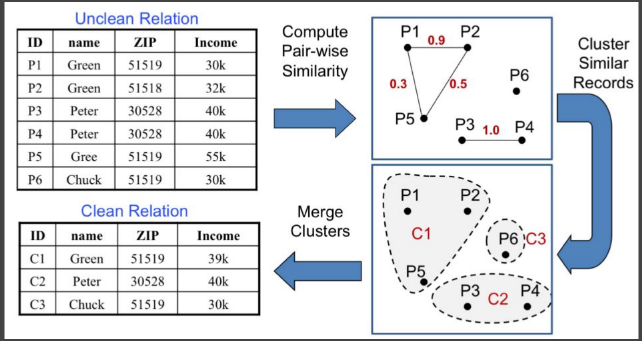


Source: Andrew Yan-Tak Ng, Chief Scientist at Baidu Research



Source: OpenAI

ML is Iterative !

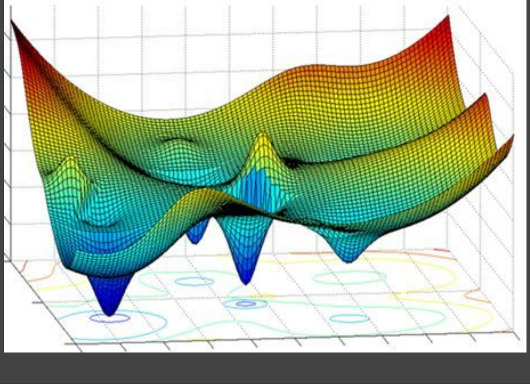


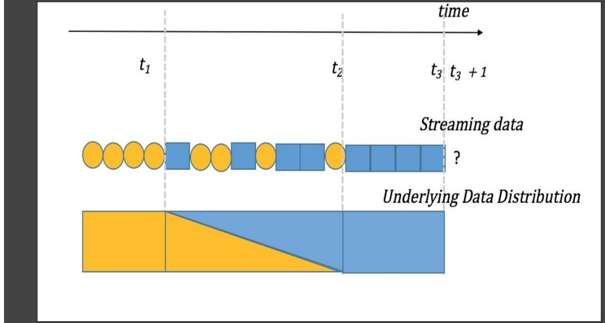
Unclean Relation

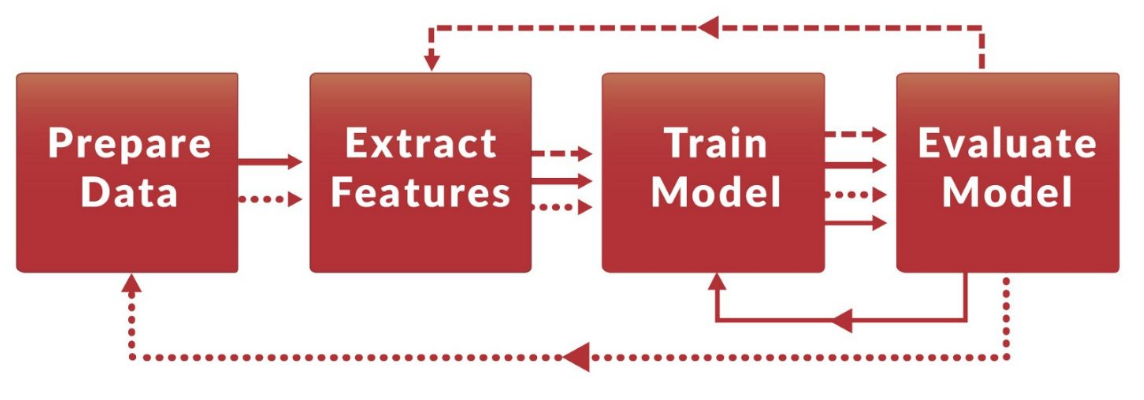
ID	name	ZIP	Income
P1	Green	51519	30k
P2	Green	51518	32k
P3	Peter	30528	40k
P4	Peter	30528	40k
P5	Gree	51519	55k
P6	Chuck	51519	30k

Clean Relation

ID	name	ZIP	Income
C1	Green	51519	39k
C2	Peter	30528	40k
C3	Chuck	51519	30k







Collaboration and Sharing

The Journey @MOD

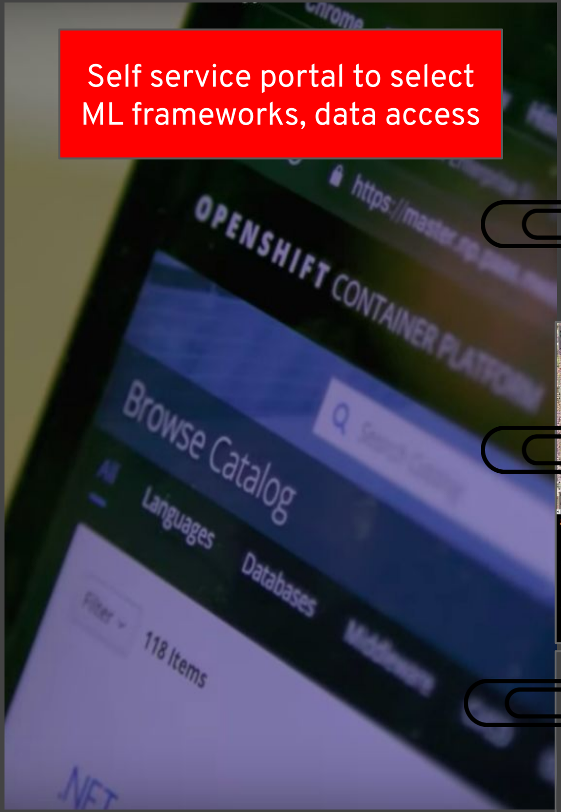
Top 3 Requirements

Access to elastic compute

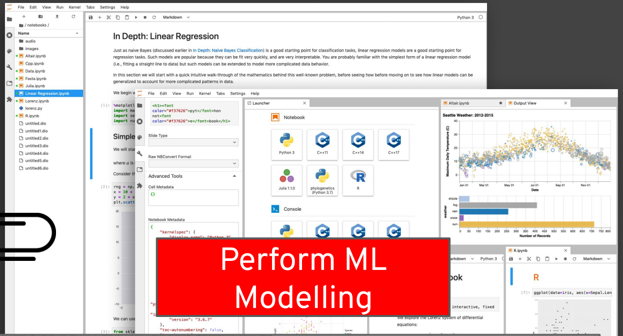
Deploy models into production

Optimize and Scale Machine Learning

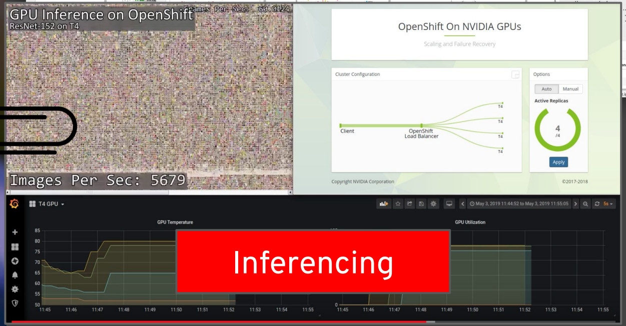
As a Data Scientist, I want a ***"self-service cloud like"*** *experience for my Machine Learning projects, where I can access a* ***rich set of modelling frameworks, data, and computational resources,*** ***share and collaborate*** *with colleagues, and deliver my work into* ***production*** *with* ***speed,*** *agility and repeatability to* ***drive organizational value!***



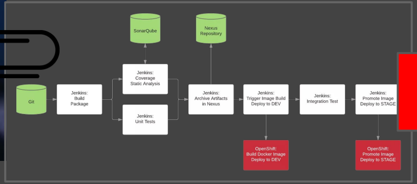
Self service portal to select ML frameworks, data access



Perform ML Modelling

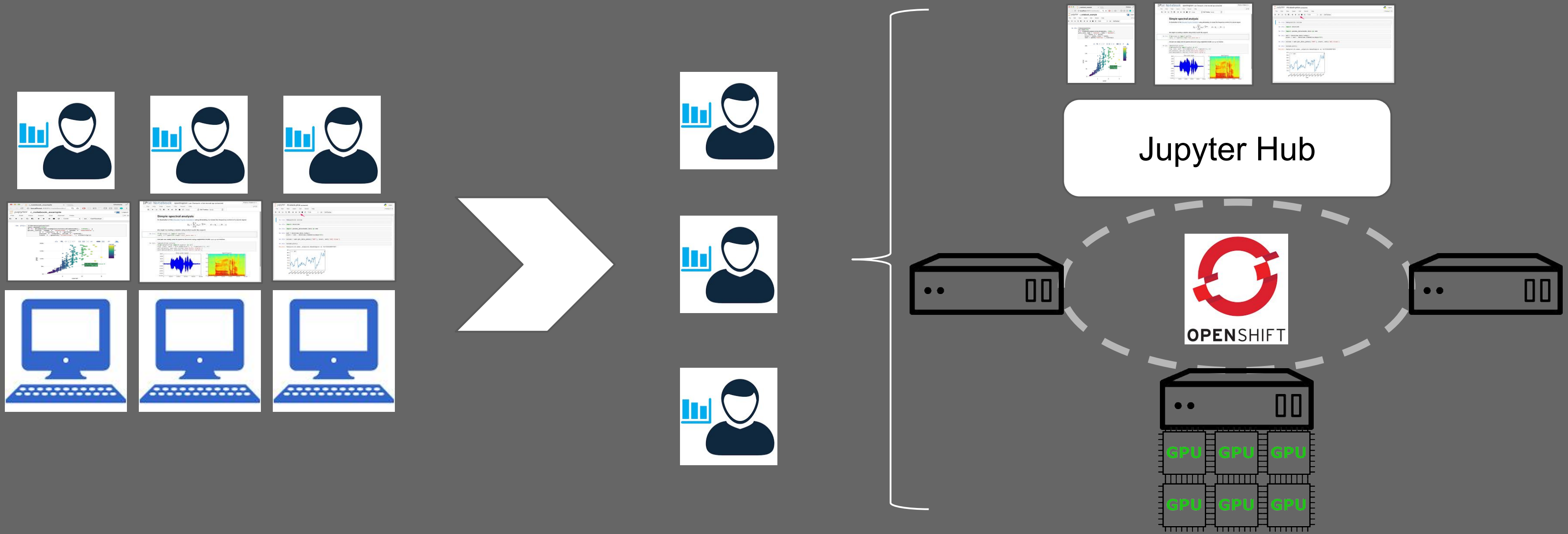


Inferencing


SonarQube
Nexus Repository
Git
Jenkins: Build Package
Jenkins: Coverage Static Analysis
Jenkins: Unit Tests
Jenkins: Archive Artifacts in Nexus
Jenkins: Trigger Image Build Deploy to DEV
Jenkins: Integration Test
Jenkins: Promote Image Deploy to STAGE
OpenShift: Build Docker Image Deploy to DEV
OpenShift: Promote Image Deploy to STAGE

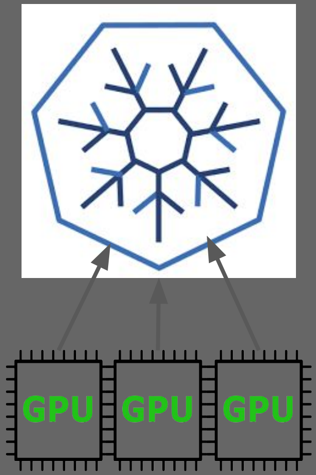
Deployment in production

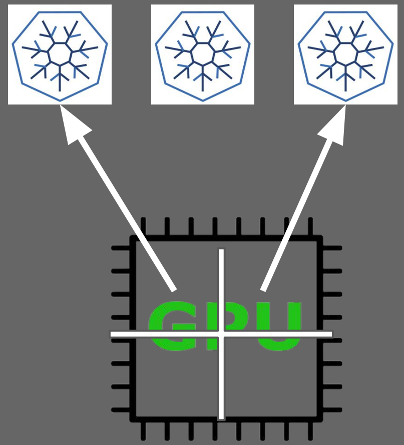
Step 1: A Self Service Cloud



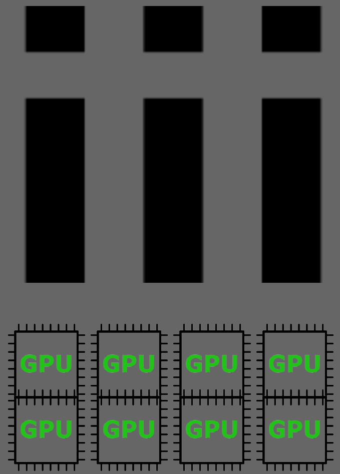
Jupyter Hub

Multiple GPUs and Sharing GPUs

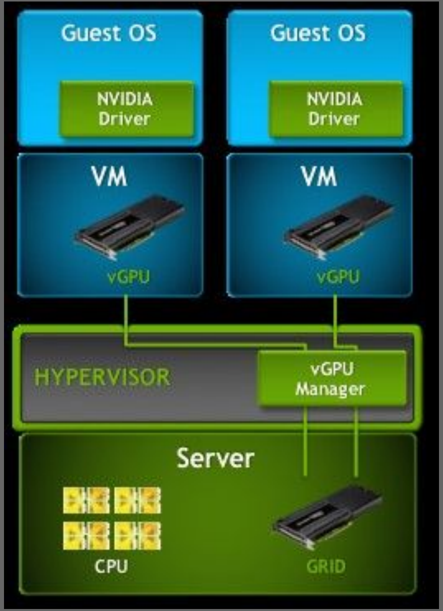


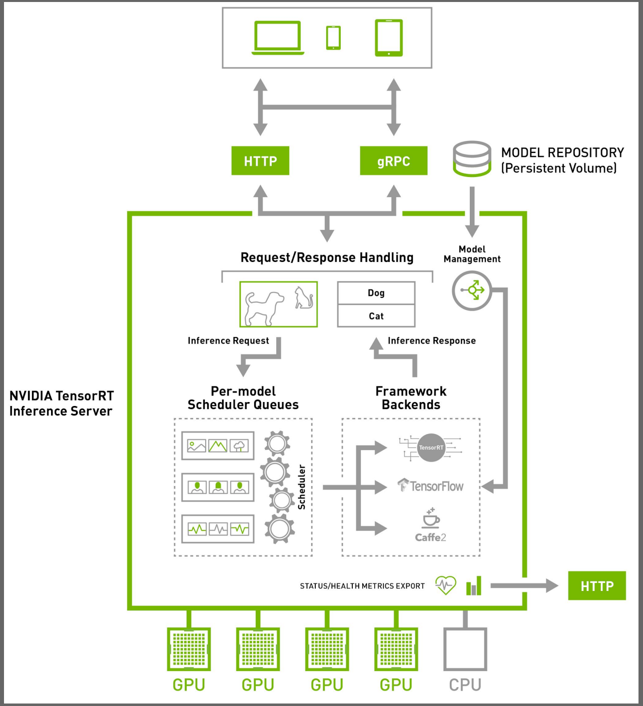


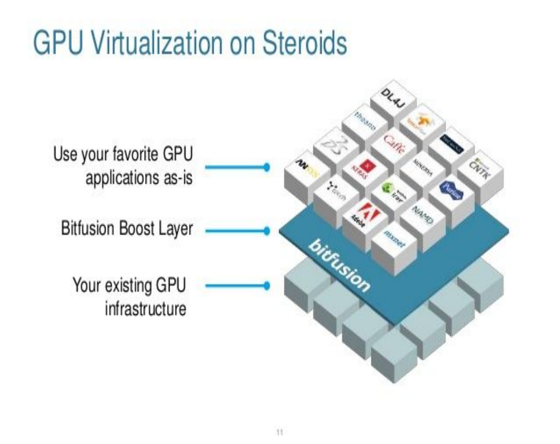
Kubernetes Job



FYI - Other GPU Sharing techniques







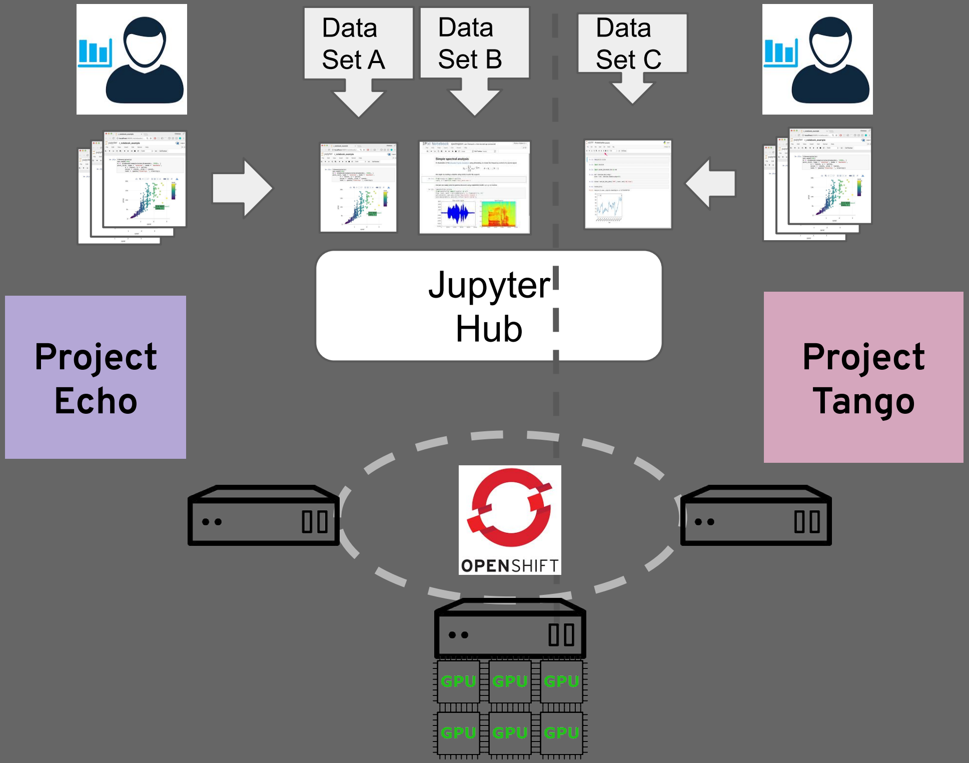
GPU Virtualization on Steroids

Use your favorite GPU applications as-is

Bitfusion Boost Layer

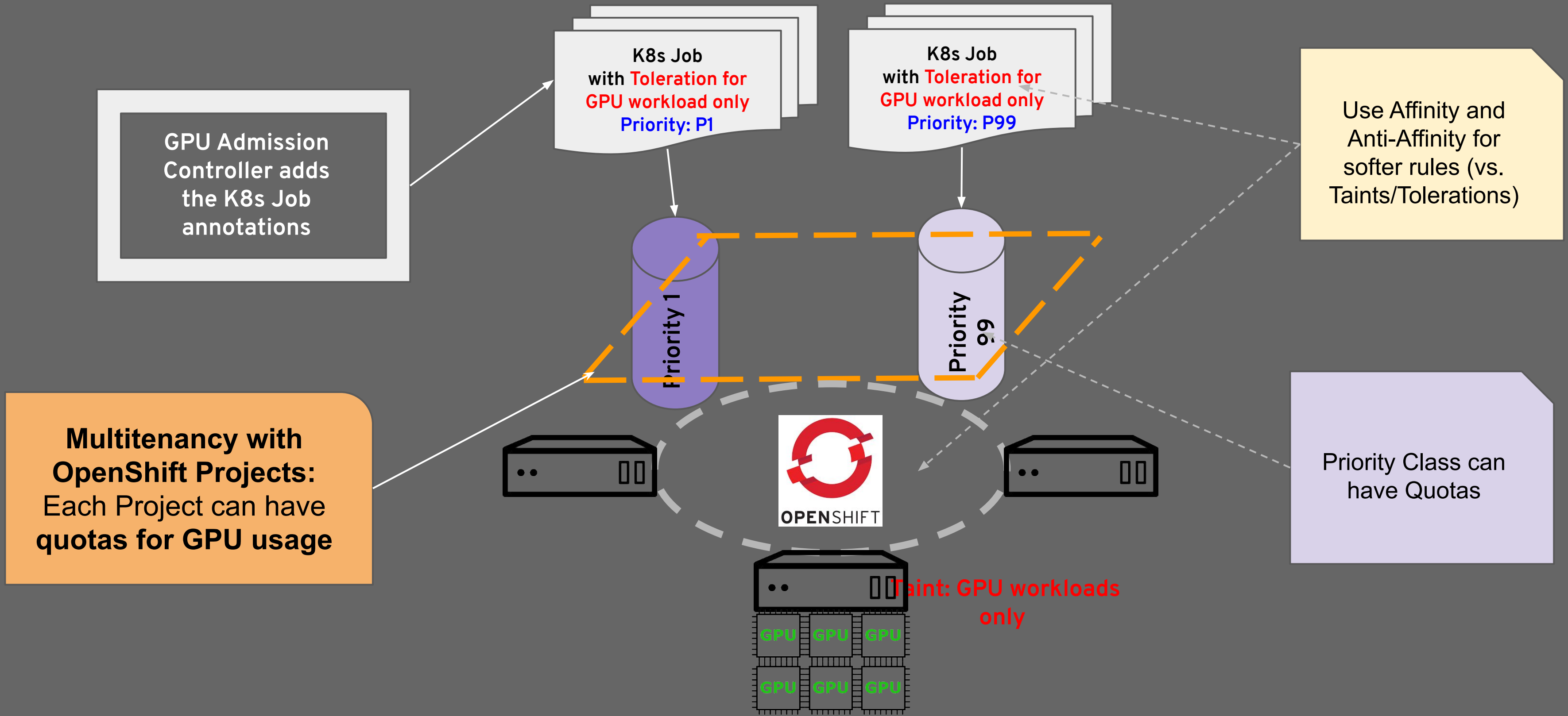
Your existing GPU infrastructure

Multi Tenancy



- ➔ Identity and Role Based Access
- ➔ Network Policies
- ➔ Storage Classes
- ➔ Resource Isolation
- ➔ Resource Quotas

Being good neighbours !



GPU Admission Controller adds the K8s Job annotations

K8s Job with Toleration for GPU workload only Priority: P1

K8s Job with Toleration for GPU workload only Priority: P99

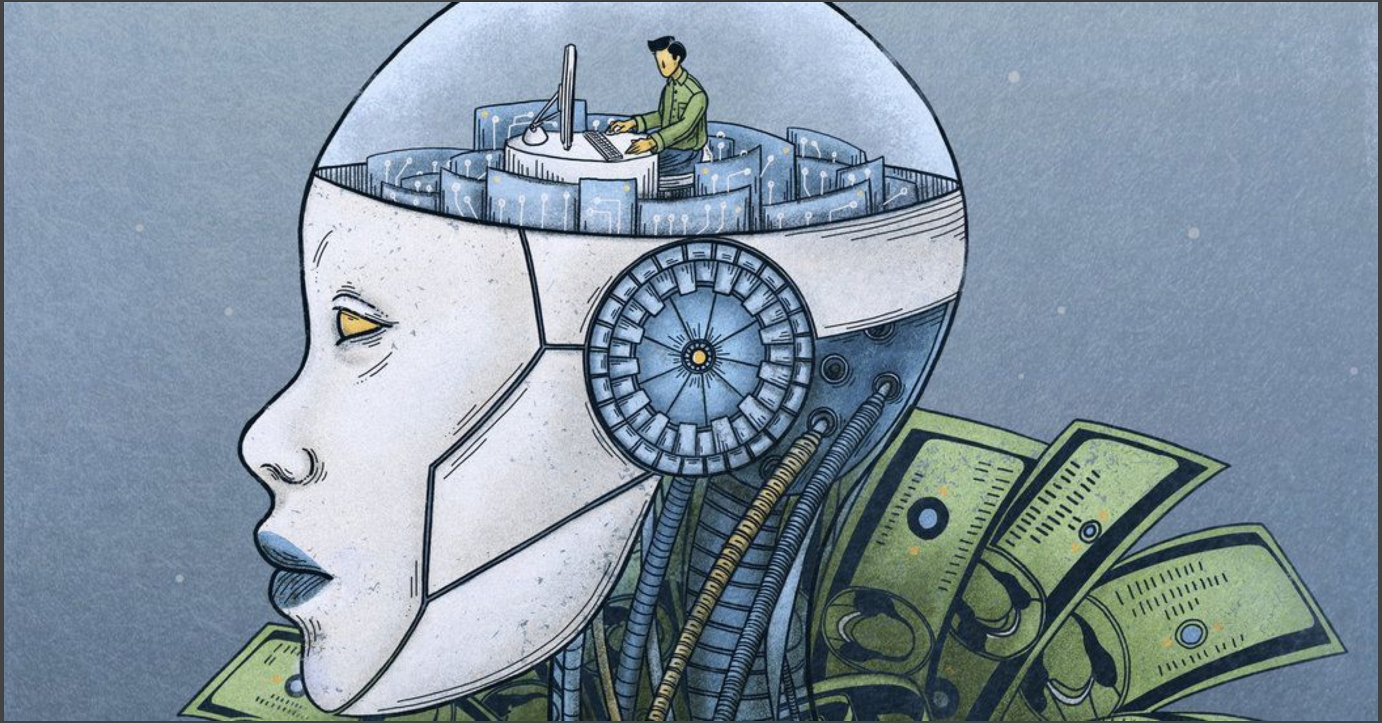
Use Affinity and Anti-Affinity for softer rules (vs. Taints/Tolerations)

Multitenancy with OpenShift Projects: Each Project can have **quotas for GPU usage**

Priority Class can have Quotas

Taint: GPU workloads only

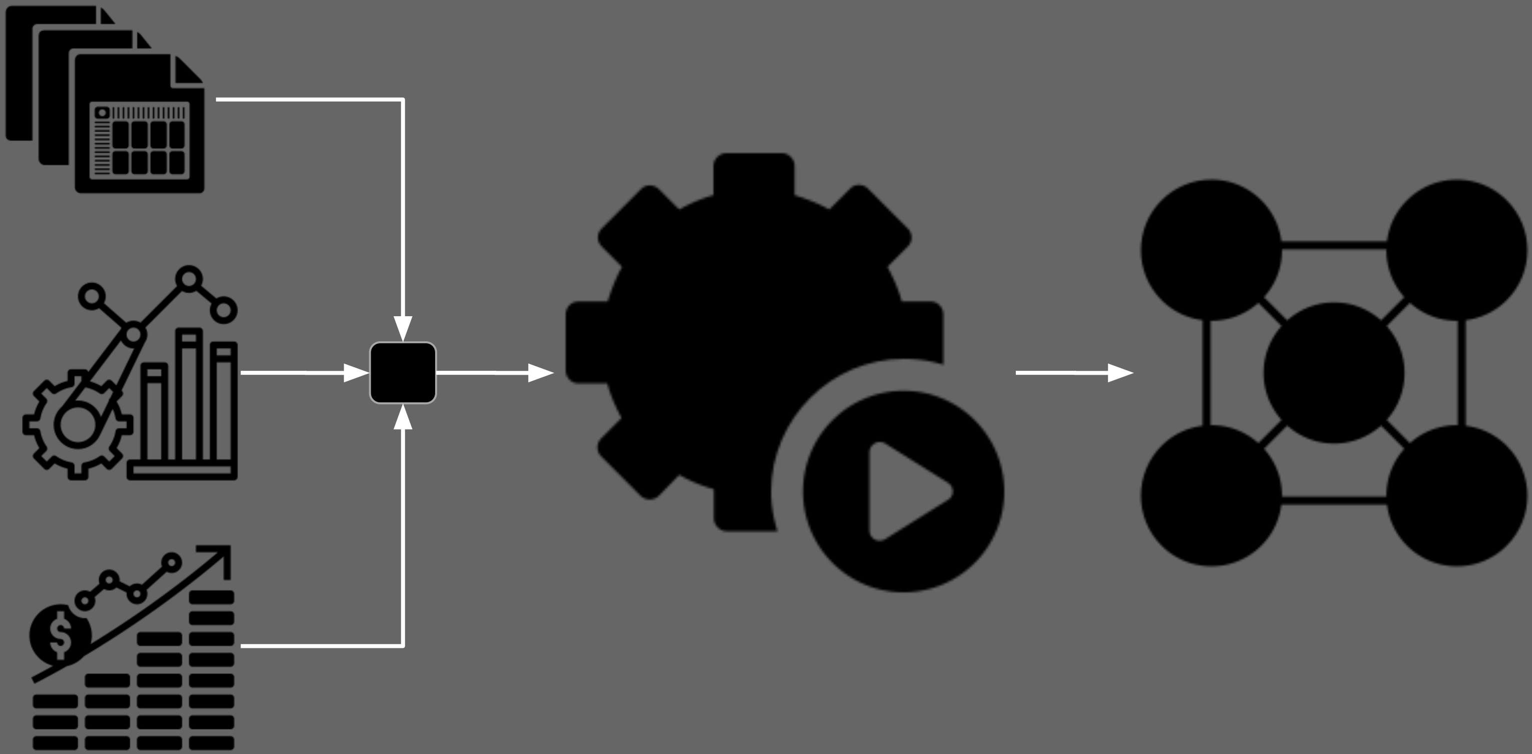
Step 2: Scaling ML Talent

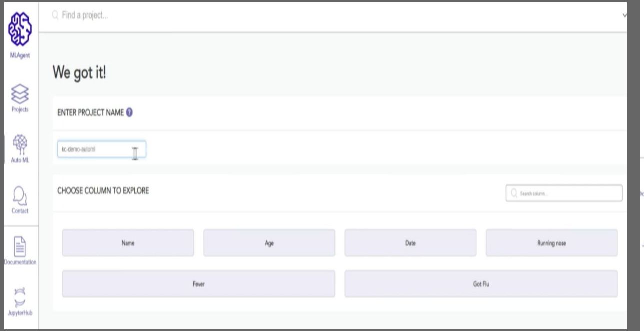


WHAT IF I TOLD YOU

YOU CAN BE A ML EXPERT

Introducing AutoML



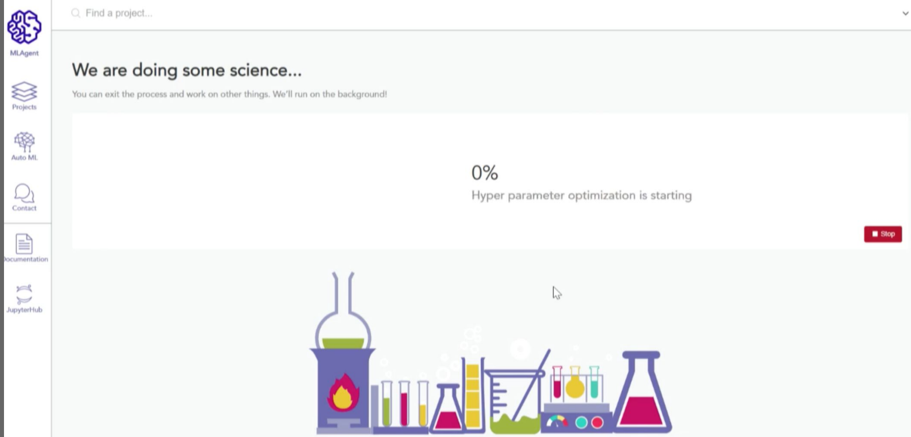


We got it!

ENTER PROJECT NAME

kc-demo-automl

CHOOSE COLUMN TO EXPLORE

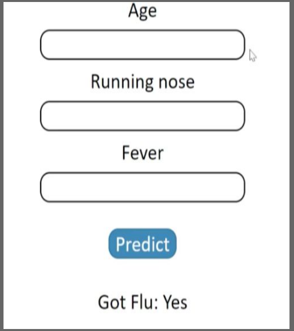


We are doing some science...

You can exit the process and work on other things. We'll run on the background!

0%

Hyper parameter optimization is starting

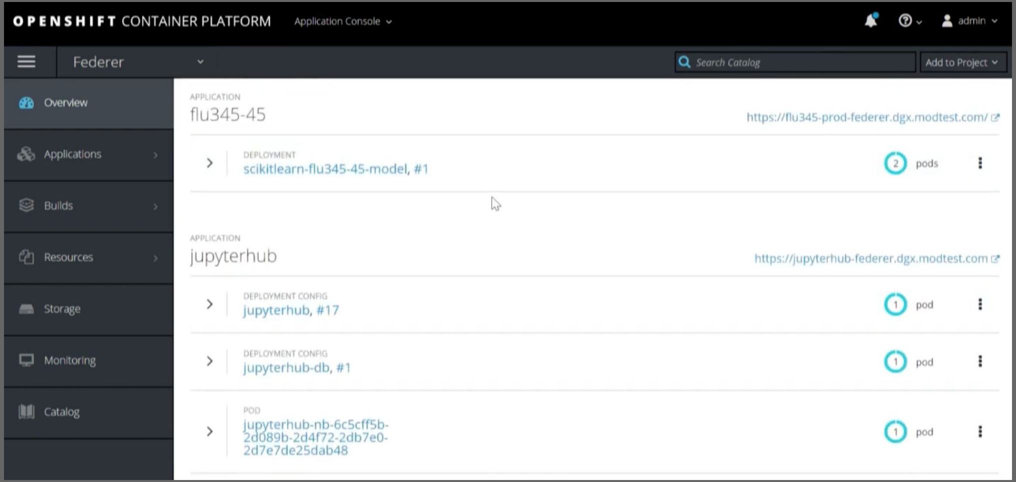


Age

Running nose

Fever

Got Flu: Yes



APPLICATION
flu345-45
https://flu345-prod-federer.dgx.modtest.com/

DEPLOYMENT
scikitlearn-flu345-45-model, #1 — 2 pods

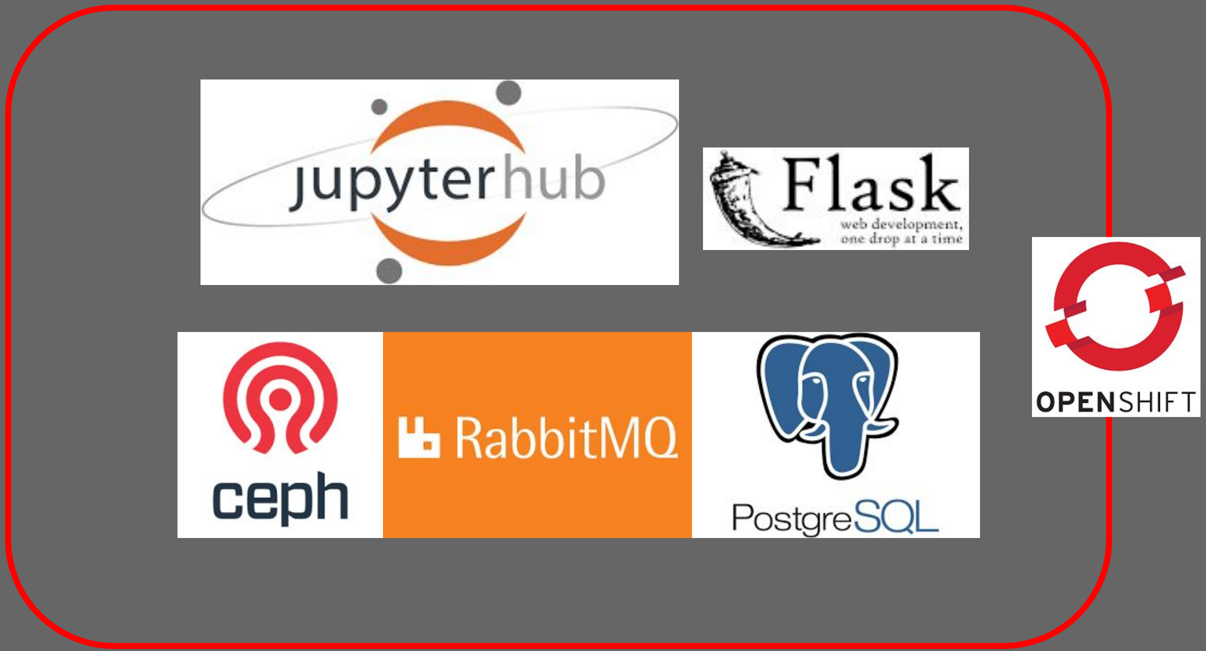
APPLICATION
jupyterhub
https://jupyterhub-federer.dgx.modtest.com

DEPLOYMENT CONFIG
jupyterhub, #17 — 1 pod

DEPLOYMENT CONFIG
jupyterhub-db, #1 — 1 pod

POD
jupyterhub-nb-6c5cff5b-2d089b-2d4f72-2db7e0-2d7e7de25dab48 — 1 pod

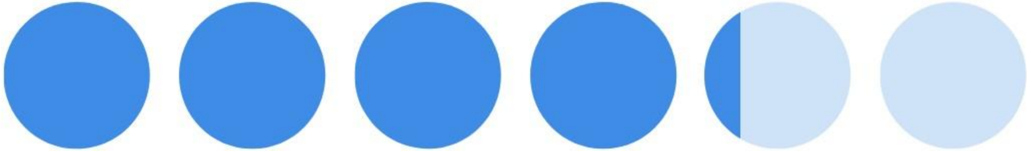
AutoML Design Elements



30%

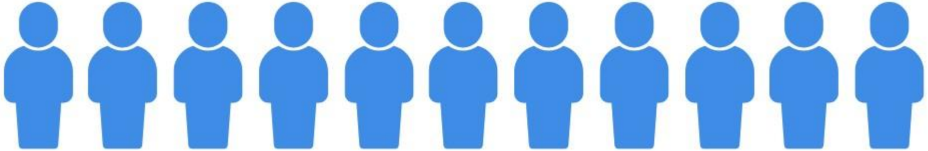
Average improvement in machine learning model performance

70%



Increase in machine learning models number

600%



Growth of users in the previous 6 months

Challenges and Next Steps

➔ **Automate** development, debug and deployment of **notebooks**

➔ Better way to **save and catalog** experiments

➔ AutoML for unstructured data - **images, audio**

➔ Supported way for **GPU sharing**

➔ Multi-cluster

What’s Next

Community First

NVIDIA NGC

Kubeflow

Tensorflow

PYTORCH

Spark

jupyter

APACHE kafka

cassandra

OPEN DATA HUB

AI Platform powered by Open Source

ML-as-a-service reference architecture on OpenShift and open source and ISV content

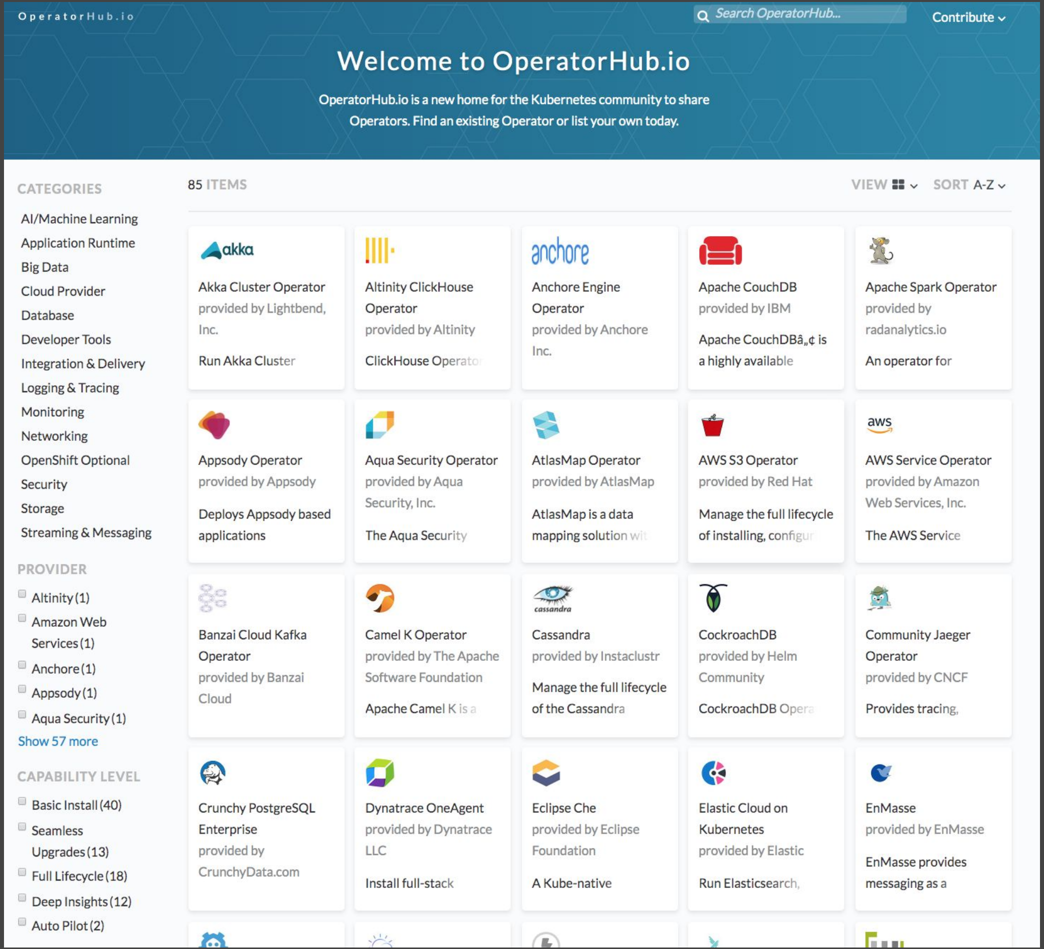
Integration & Delivery

Logging & Trac

The registry for Kubernetes Operators

OperatorHub.io

Home for k8s community to share operators for various apps/tools



OperatorHub.io

Search OperatorHub...

Contribute

Welcome to OperatorHub.io

OperatorHub.io is a new home for the Kubernetes community to share Operators. Find an existing Operator or list your own today.

CATEGORIES

- AI/Machine Learning
- Application Runtime
- Big Data
- Cloud Provider
- Database
- Developer Tools
- Integration & Delivery
- Logging & Tracing
- Monitoring
- Networking
- OpenShift Optional
- Security
- Storage
- Streaming & Messaging

PROVIDER

- □ Altinity (1)
- □ Amazon Web Services (1)
- □ Anchore (1)
- □ Appsody (1)
- □ Aqua Security (1)

Show 57 more

CAPABILITY LEVEL

- □ Basic Install (40)
- □ Seamless Upgrades (13)
- □ Full Lifecycle (18)
- □ Deep Insights (12)
- □ Auto Pilot (2)

85 ITEMS

VIEW SORT A-Z

Akka Cluster Operator
provided by Lightbend, Inc.
Run Akka Cluster

Altinity ClickHouse Operator
provided by Altinity
ClickHouse Operator

Anchore Engine Operator
provided by Anchore Inc.

Apache CouchDB
provided by IBM
Apache CouchDBâ„¢ is a highly available

Apache Spark Operator
provided by radanalytics.io
An operator for

Appsody Operator
provided by Appsody
Deploys Appsody based applications

Aqua Security Operator
provided by Aqua Security, Inc.
The Aqua Security

AtlasMap Operator
provided by AtlasMap
AtlasMap is a data mapping solution wit

AWS S3 Operator
provided by Red Hat
Manage the full lifecycle of installing, configur

AWS Service Operator
provided by Amazon Web Services, Inc.
The AWS Service

Banzai Cloud Kafka Operator
provided by Banzai Cloud

Camel K Operator
provided by The Apache Software Foundation
Apache Camel K is a

Cassandra
provided by Instaclustr
Manage the full lifecycle of the Cassandra

CockroachDB
provided by Helm Community
CockroachDB Opera

Community Jaeger Operator
provided by CNCF
Provides tracing,

Crunchy PostgreSQL Enterprise
provided by CrunchyData.com

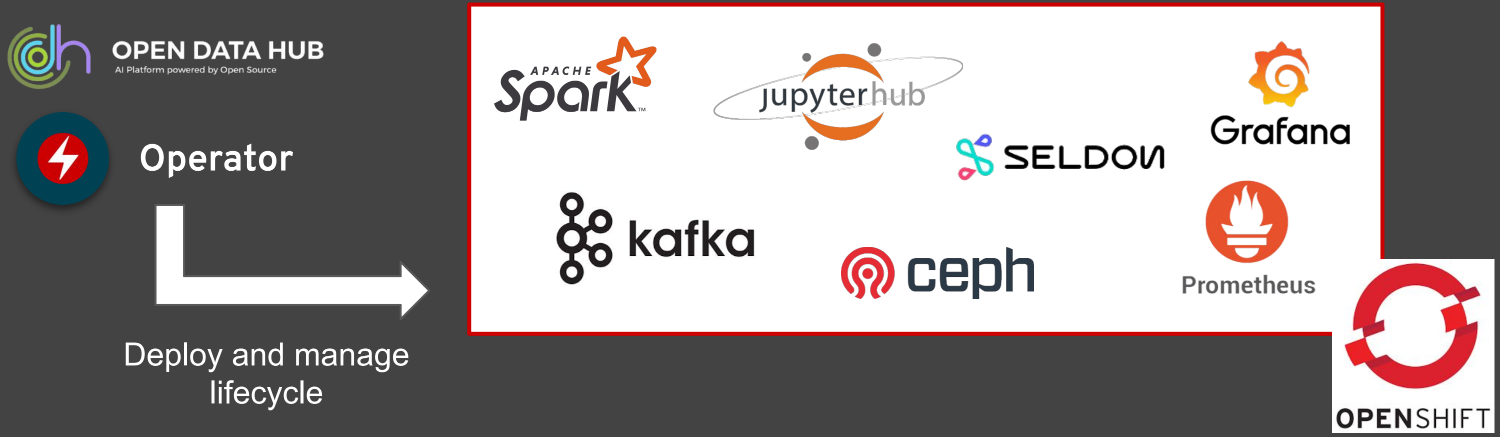
Dynatrace OneAgent
provided by Dynatrace LLC
Install full-stack

Eclipse Che
provided by Eclipse Foundation
A Kube-native

Elastic Cloud on Kubernetes
provided by Elastic
Run Elasticsearch,

EnMasse
provided by EnMasse
EnMasse provides messaging as a

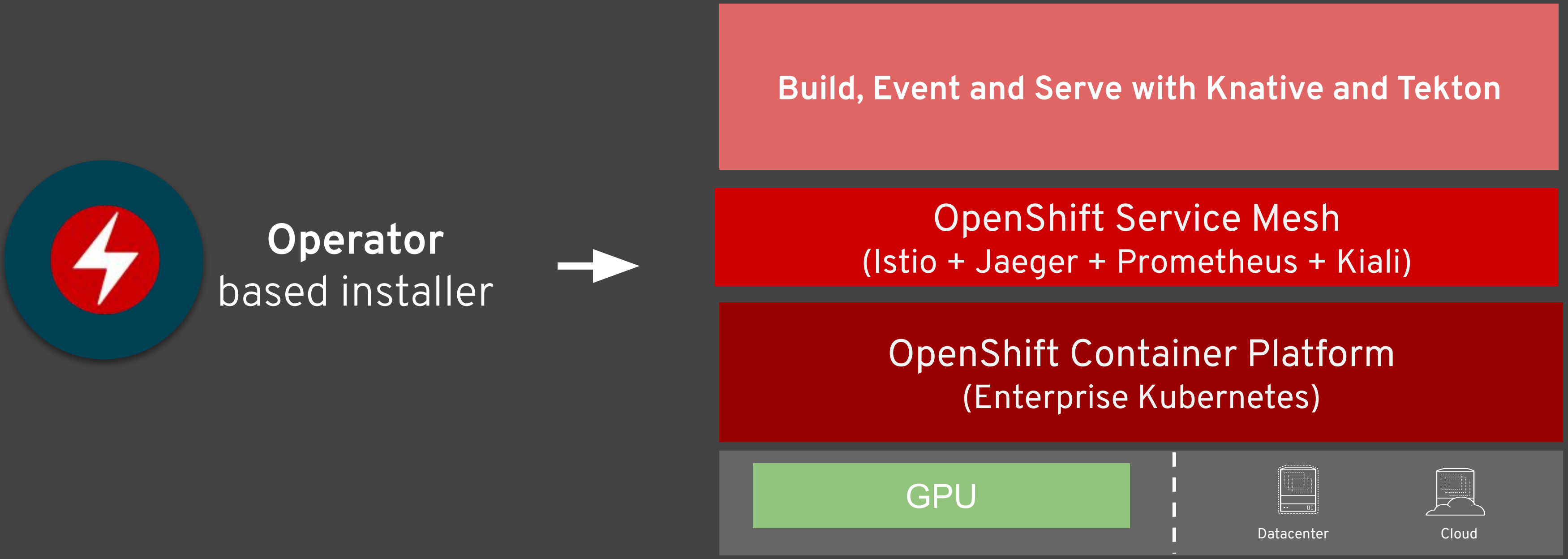
Open Data Hub



Operator

Deploy and manage lifecycle

OpenShift 4



Operator
based installer

→

Build, Event and Serve with Knative and Tekton

OpenShift Service Mesh
(Istio + Jaeger + Prometheus + Kiali)

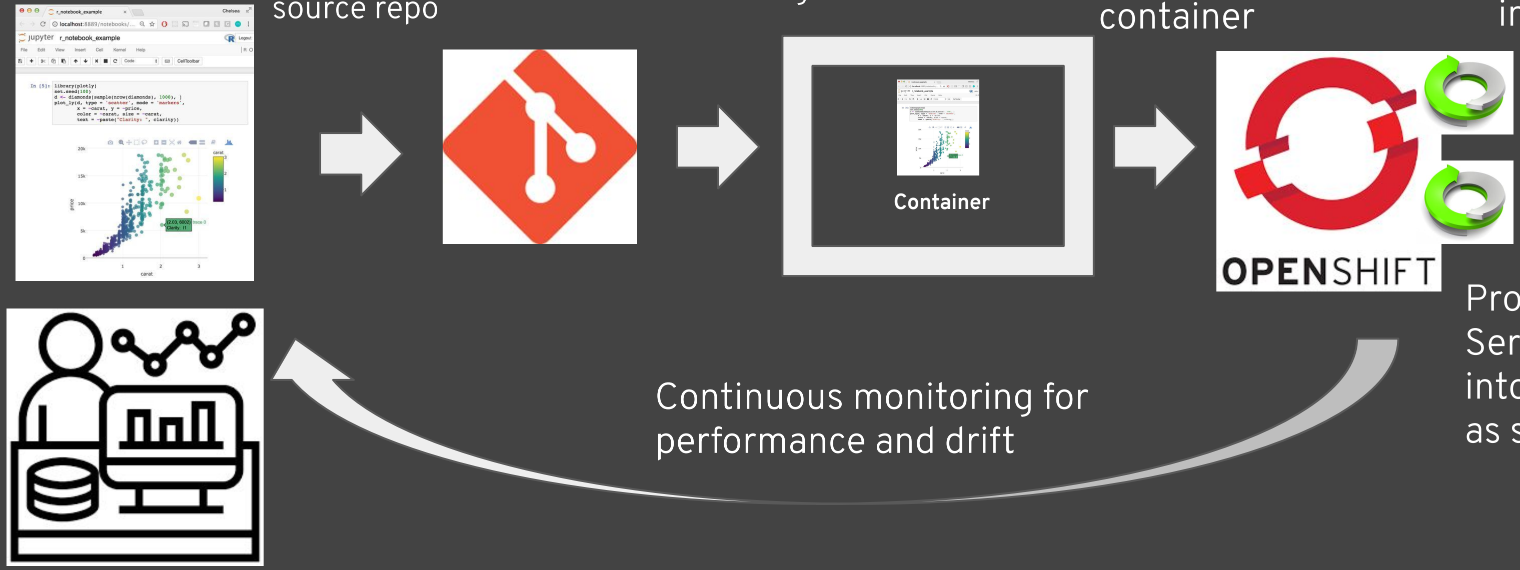
OpenShift Container Platform
(Enterprise Kubernetes)

GPU

Datacenter

Cloud

From experimentation to production with CI/CD



```
In [5]: library(plotly)
        set.seed(100)
        d <- diamonds[sample(nrow(diamonds), 1000), ]
        plot_ly(d, type = 'scatter', mode = 'markers',
                x = ~carat, y = ~price,
                color = ~carat, size = ~carat,
                text = ~paste("Clarity: ", clarity))
```

r_notebook_example
localhost:8889/notebooks/...
jupyter r_notebook_example
File Edit View Insert Cell Kernel Help
Code
CellToolbar
Logout
carat
price
Container
OPENSHIFT

Check-in to source repo

Source-2-image

Deloy notebook container

Model test & iteration and integration

Promote and Serve models into production as services

Continuous monitoring for performance and drift

DEMO - Self Service with Open Data Hub

SUMMARY

- MOD Case Study: Machine Learning-as-a-service platform
  - Why and how they built a **cloud-like** experience
  - AutoML
- Kubernetes and OpenShift and open source tools
- OperatorHub and OpenDataHub

THANK YOU !

- → Contact:
- → tkatarki@redhat.com
- → @tkatarki

- → Q/A