Microsoft

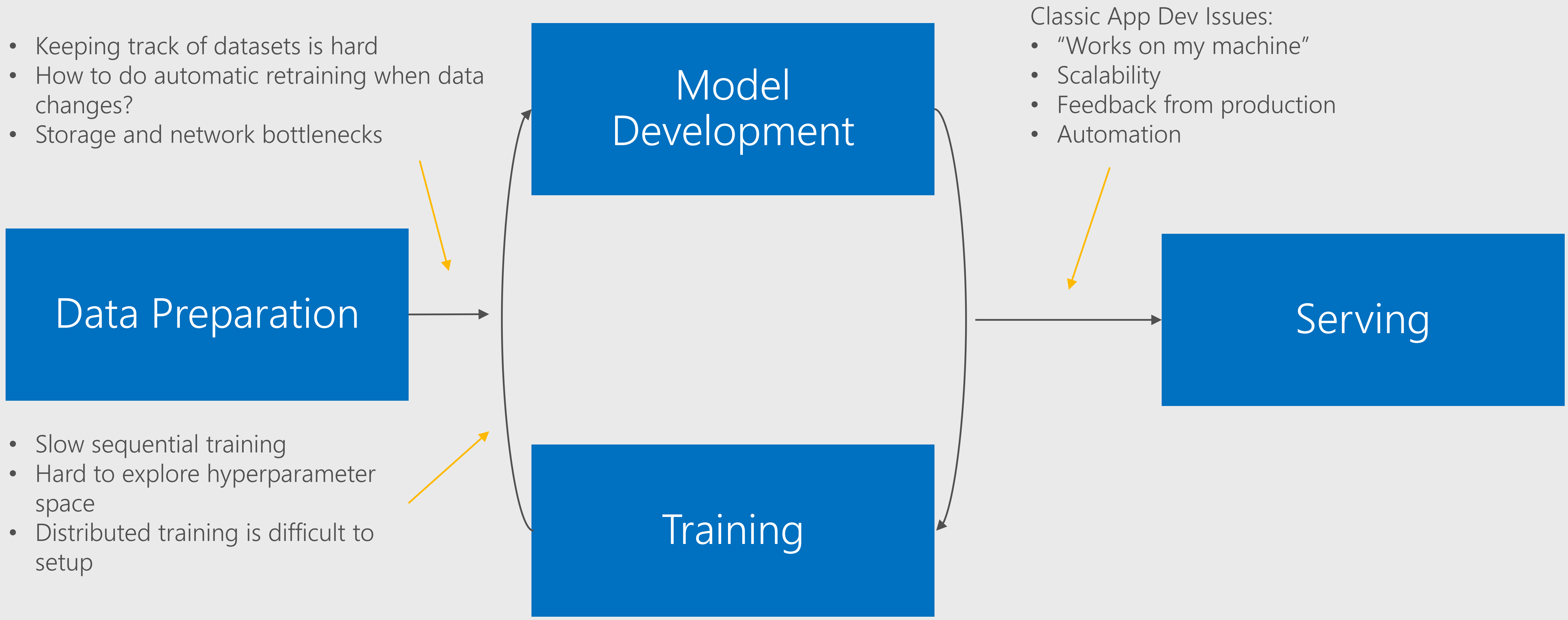
Machine Learning Lifecycle with Kubeflow on Azure Kubernetes Service (AKS)

SATO Naoki (Neo) - @satonaoki
Azure Technologist
Microsoft

Agenda

- What is the typical ML workflow and some of their shortcomings
- Why DevOps?
- Why Containers, Kubernetes, and Helm?
- Intro to Kubeflow, Helm, Argo
- Demos
  - Image classification with Inception v3 and transfer learning
  - Automate repeatable ML experiments with containers
  - Deploy ML components to Kubernetes with Kubeflow
  - Scale and test ML experiments with Helm
  - Manage training jobs and pipelines with Argo
  - Serve trained models for inference with TF Serving
  - Rapid prototyping with self-service Jupyter notebook from JupyterHub

Simplified ML Workflow/Pipeline



What is DevOps?

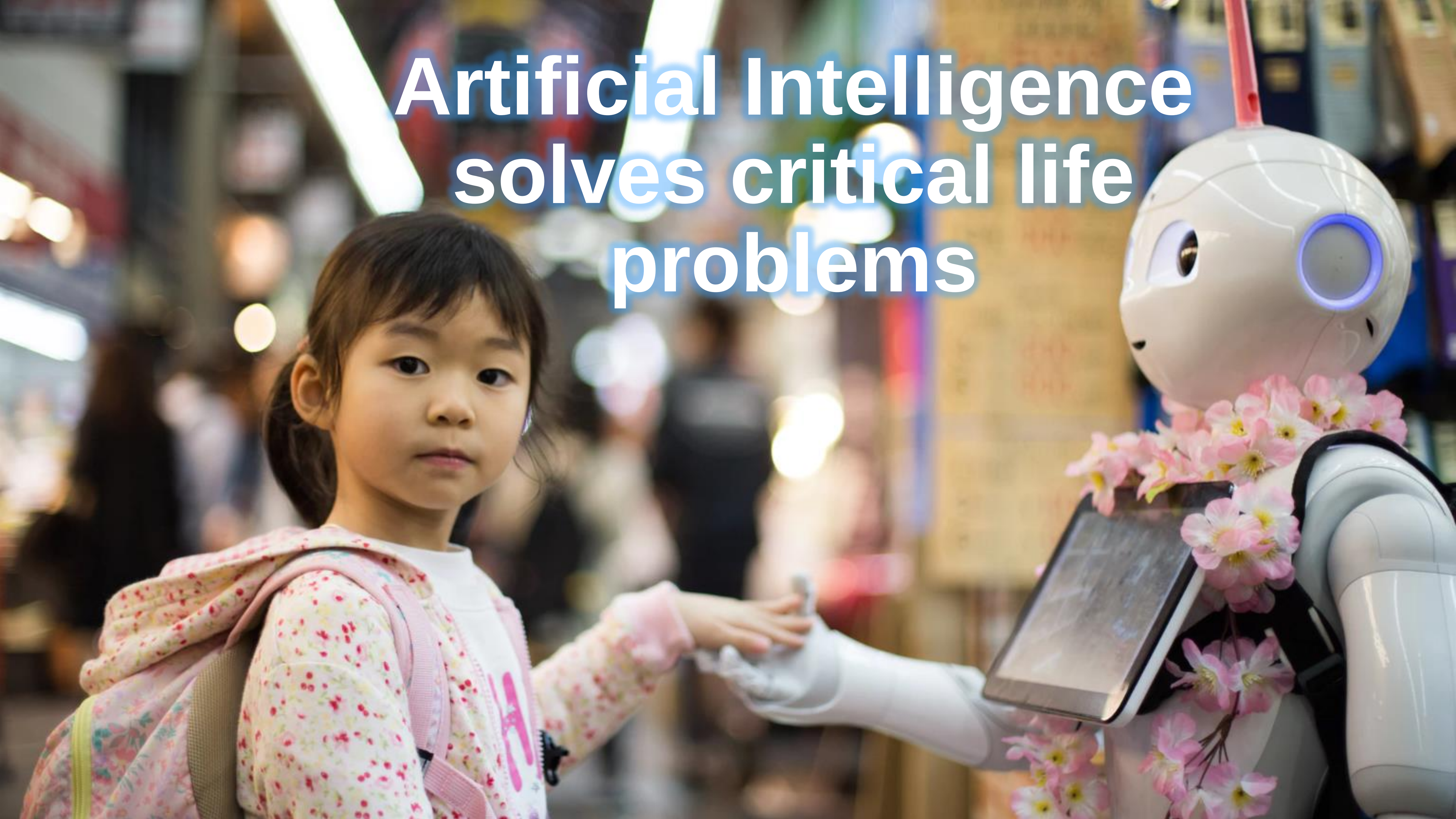
- "A cross-disciplinary community of practice dedicated to the study of building, evolving and operating rapidly-changing resilient systems at scale" (Jez Humble)
- Applying Agile practices to operations
  - Infrastructure as code
  - Ops teams embracing source control (git)
  - Automated testing
  - Repeatable/consistent
  - CI/CD
- This has worked well for App Dev. Now time for AI/ML
  - But, must ensure data scientist are not hindered by structure

Why Containers, Kubernetes & Helm?

- Container
  - Contains everything needed to run your application
  - Build once run anywhere
  - Starts in seconds: Great for scalability
  - Images are stored in a centralized place (Docker Hub, Azure Container Registry, gcr, ECR etc.)
- Container orchestration
  - Automating deployment, scaling, and management of containerized applications
  - Declarative
  - Can be a mix of GPU or CPU nodes
- Massive Scale
  - OpenAI dedicates up to 10k cores for a single experiment
- Autoscaling capabilities: Pay for what you use, scale down when idle
- Parallel training instead of sequential: huge time saver for large trainings

Kubeflow

- Machine Learning Toolkit for Kubernetes
  - To make ML workflows on Kubernetes simple, portable, and scalable
- Training controllers – simplify and manage the deployment of training jobs
  - TFJob – custom resource to handle drivers and config
  - Tensorflow, PyTorch, MXNet, Chainer, and more
- JupyterHub to create and manage interactive Jupyter notebooks
- Model serving – serve exported models with TF Serving or Seldon
- Additional components for storage, workflow, etc.


Artificial Intelligence solves critical life problems

BBC Sign in

NEWS

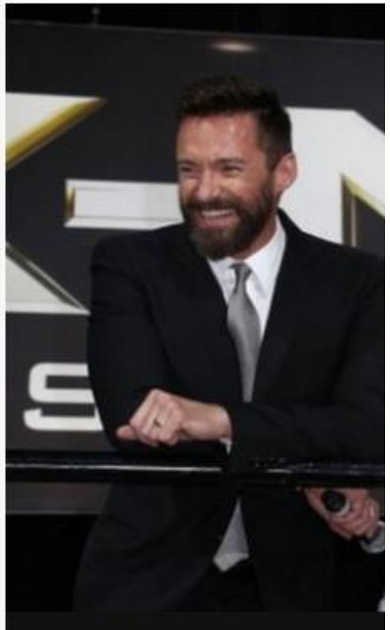
Home | Video | World | US & C

Asia | China | India

'Disappearance
Fan Bingbing c

By Kerry Allen
BBC Monitoring

1 August 2018



Fan Bingbing recently received glob

CNN World + | Live TV | U.S. Edition +

Chinese star Fan Bingbing seen i
disappearance

By Ben Westcott, CNN

Updated 6:28 AM ET, Wed October 17, 2018

Fan Bingbing seen after lengthy disappearance 00:35

ENTERTAINMENT

Fan Bingbing outside the airport in Beijing.

Home / Entertainment / Movies

Fan Bingbing spotted for first time in months, outside Beijing airport

OCTOBER 17, 2018 | ENTERTAINMENT, MOVIES, PEOPLE

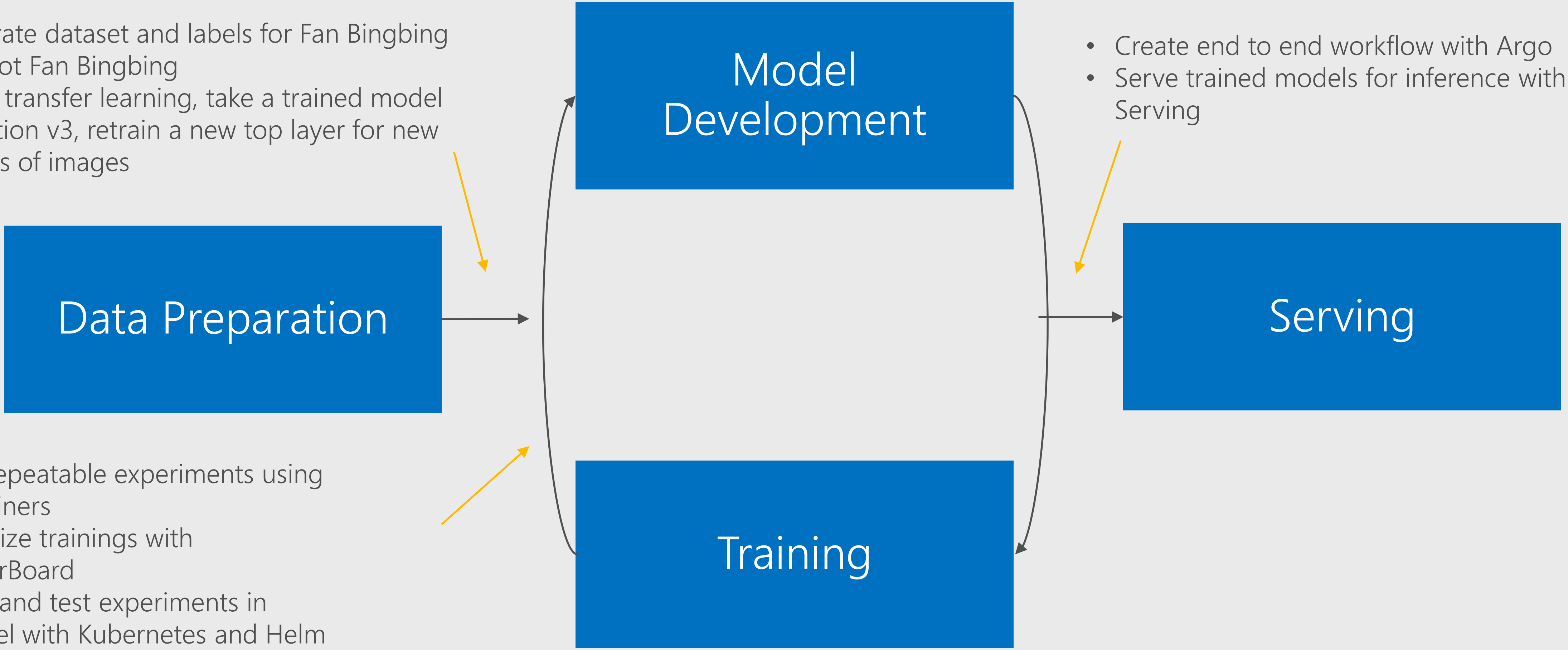
BY AGENCY

Demo: Find 范冰冰

Image classification with Inception v3 and transfer learning

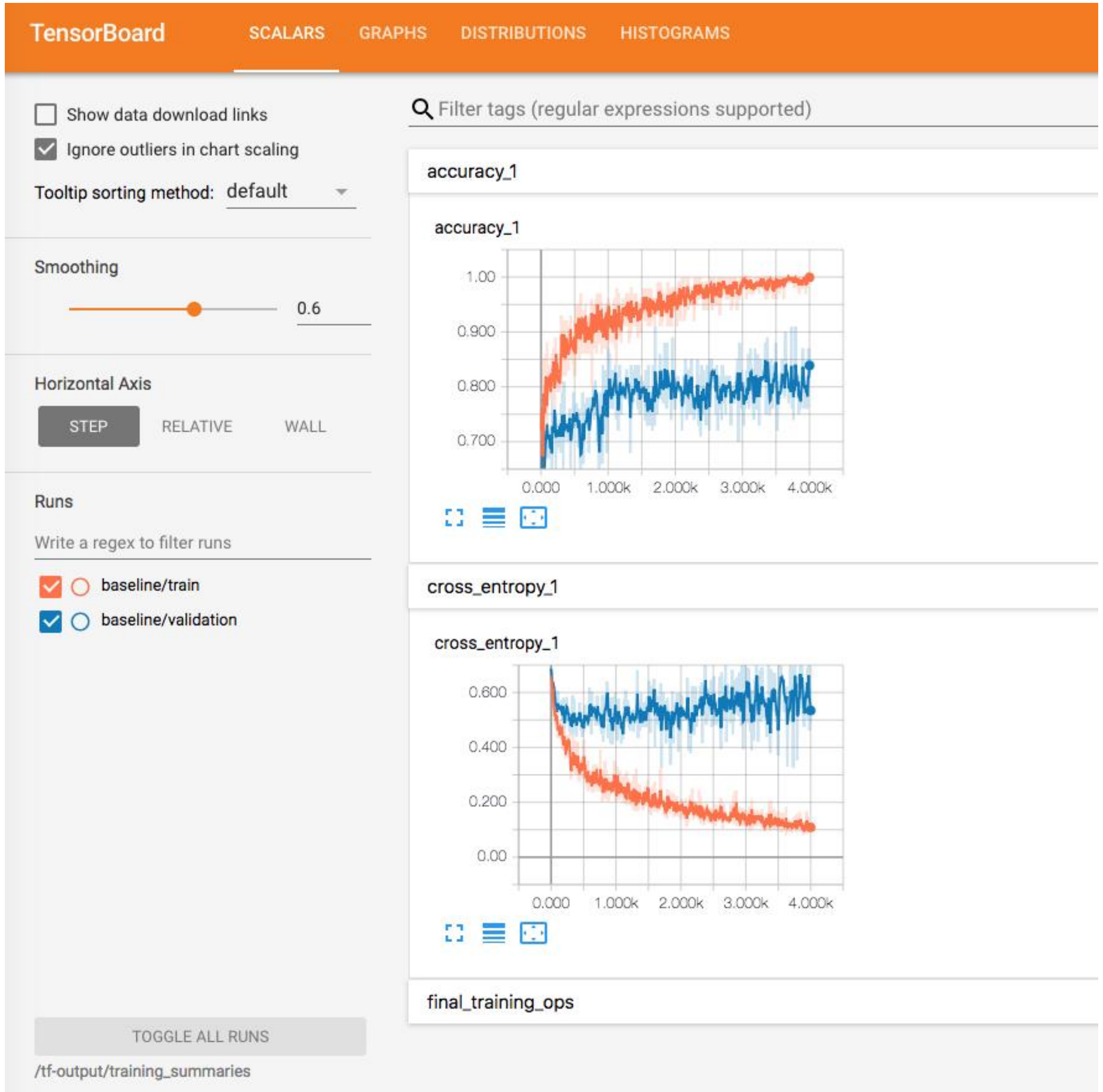
- Generate dataset and labels for Fan Bingbing and not Fan Bingbing
- Using transfer learning, take a trained model Inception v3, retrain a new top layer for new classes of images

- Create end to end workflow with Argo
- Serve trained models for inference with TF Serving



- Run repeatable experiments using containers
- Visualize trainings with TensorBoard
- Scale and test experiments in parallel with Kubernetes and Helm

Demo: Run TensorFlow Training with Containers



- https://www.youtube.com/watch?v=7Ndx3HKaS5s

Demo: Serving the Model with TF Serving

- Options for serving
  - Wrap model in a web framework (eg – Flask)
  - Tensorflow Serving
  - Seldon

- https://www.youtube.com/watch?v=t13F33I27TI

Demo:
Run TensorFlow Training with Kubeflow

- https://www.youtube.com/watch?v=lvH3ivDrocw
- https://www.youtube.com/watch?v=OQvO0pFaeEc

Demo: Scale and Test Experiments in Parallel using Kubernetes, TFJob, Helm, Virtual Kubelet, & ACI

- Spin up pods for each variation of hyperparameters
- One centralized TensorBoard instance
- Autoscaling will create / remove container instances as needed to save cost

- https://www.youtube.com/watch?v=EtOuo1dj56c
- https://www.youtube.com/watch?v=E1p9bTN-fYc

Demo:
Create End to End ML Pipelines with Argo

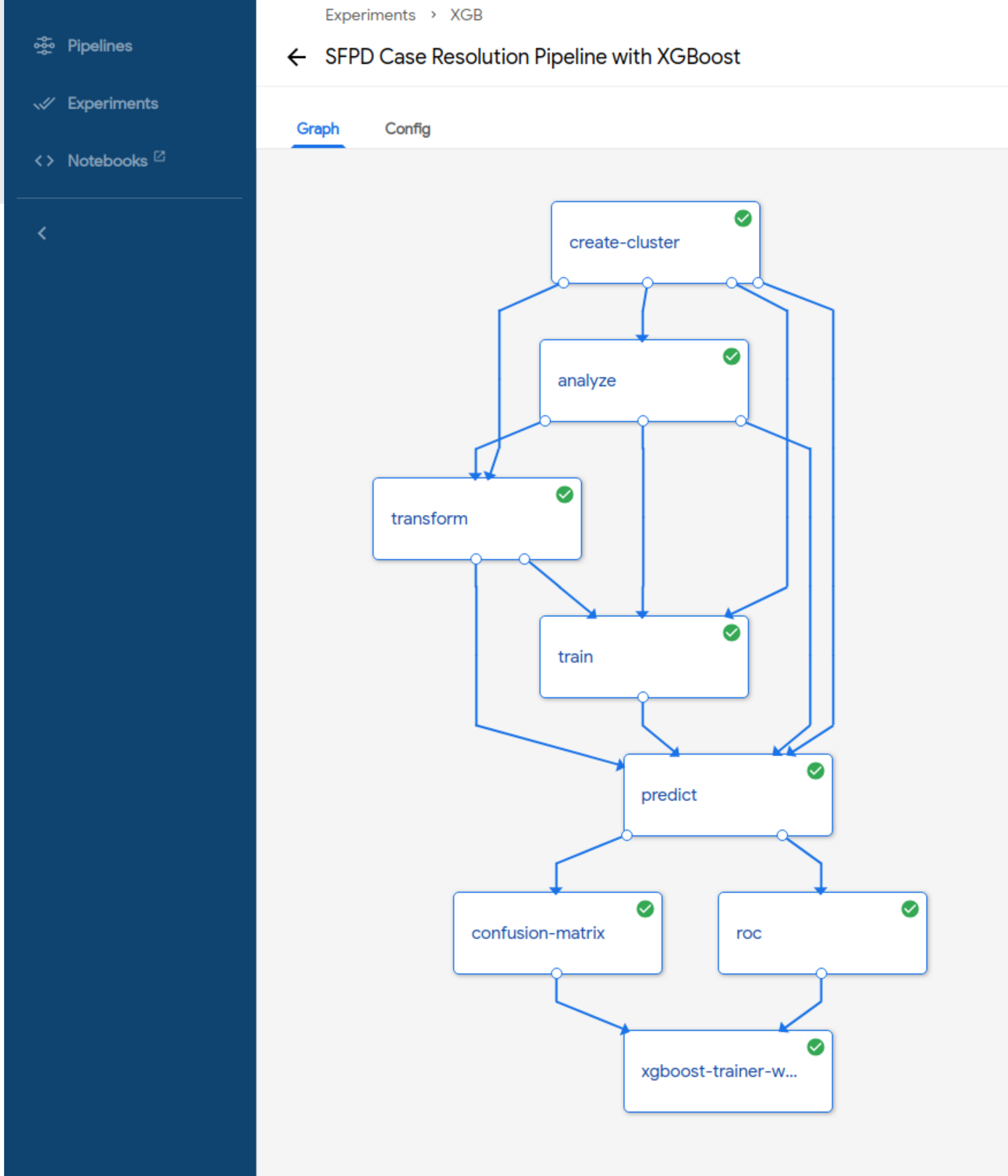
- https://www.youtube.com/watch?v=5zJrvWy9srs
- https://www.youtube.com/watch?v=2P50c-srlkA

Kubeflow Pipelines

```
...
create_cluster_op =
CreateClusterOp('create-cluster',
project, region, output)

analyze_op = AnalyzeOp('analyze',
project, region,
create_cluster_op.output, schema,
train_data,
'%s/{{workflow.name}}/analysis' % output)

transform_op = TransformOp('transform’,
project, region,
create_cluster_op.output, train_data,
eval_data, target, analyze_op.output,
'%s/{{workflow.name}}/transform' %
output)
...
```



Demo: Rapid prototyping with self-service Jupyter notebook from JupyterHub

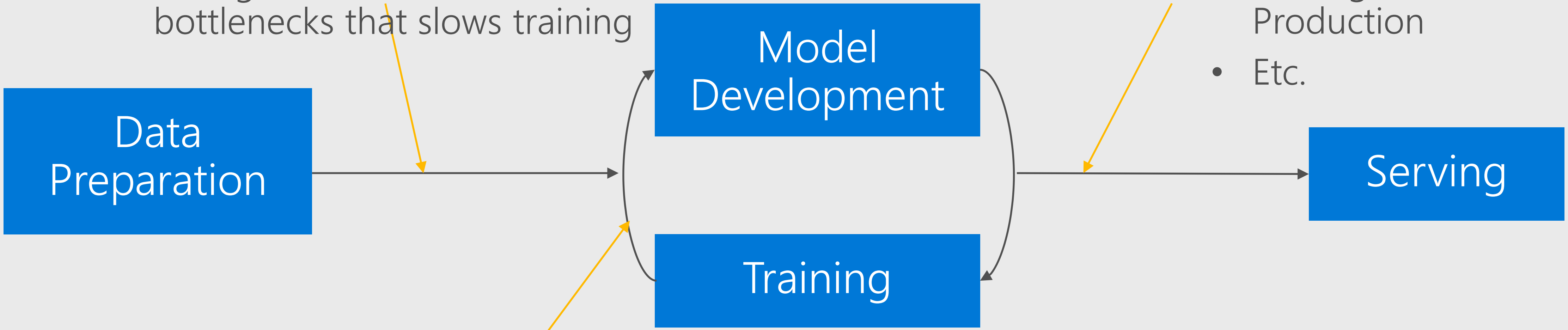
- https://www.youtube.com/watch?v=kGr6mTUEBhs
- https://www.youtube.com/watch?v=8MTGAT6qsXo

What's Next?

- Keeping track of datasets
- How to do automatic retraining when data changes ?
- Storage and Network bottlenecks that slows training

Classic App dev. issues:

- Reproducibility ("it works on my machine")
- Scalability
- Getting feedback from Production
- Etc.



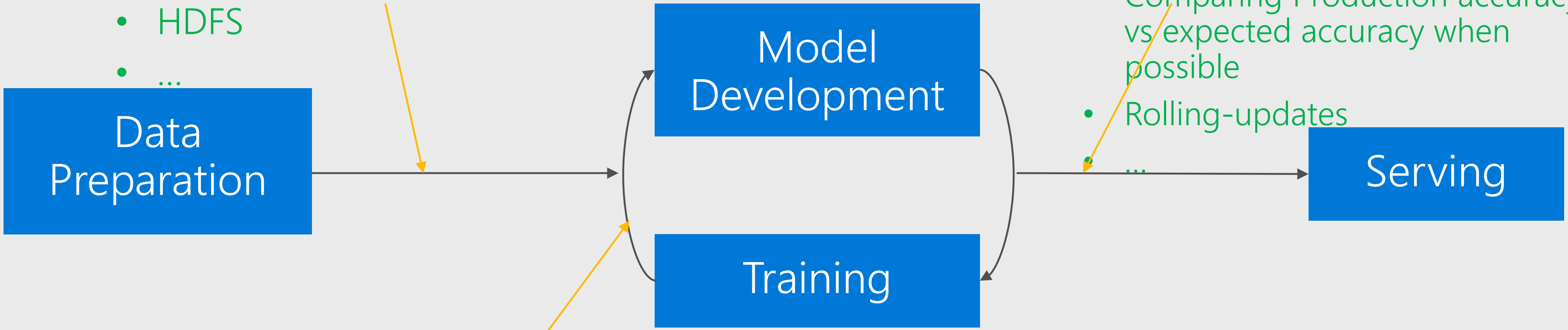
- Slow Sequential Training
- Hard to explore hyper-parameters space
- Distributed training is hard to set up

What’s Next?

- Pachyderm can version datasets and trigger
  new trainings when changes occur
- Distributed File Systems
  - NFS
  - HDFS
  - …

Classic DevOps solutions:

- Containers
- CI/CD
- Autoscaling
- A/B testing and canary release of Models
- Comparing Production accuracy vs expected accuracy when possible
- Rolling-updates
- …



(one) Solution is Kubernetes:

- Highly Scalable
- Easy to explore hyper-parameters space
- Easy to do distributed training

But really, Data Scientists shouldn’t have to care about containers, kubernetes and all that stuff

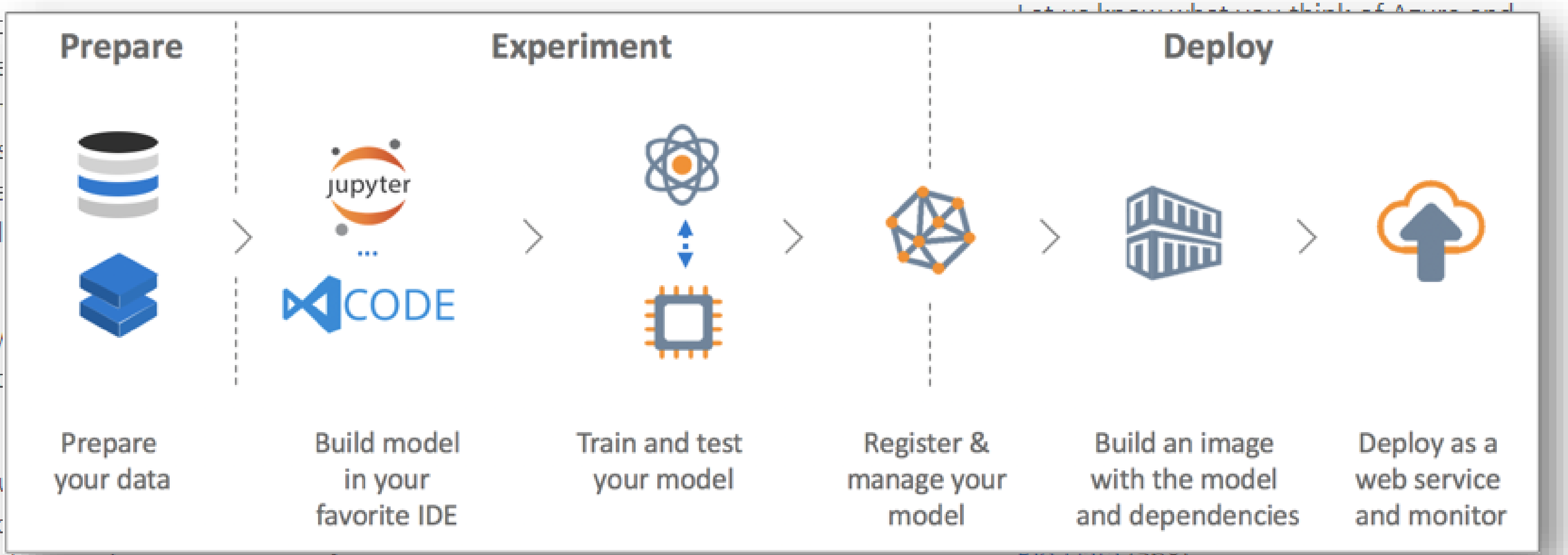
Announcing general availability of Azure Machine Learning service: A look under the hood

Posted on December 4, 2018

Venky Veeraraghavan, Group Program Manager, Microsoft Azure

Today, we are announcing the general availability of Azure Machine Learning service.

Azure Machine Learning service contains many advanced
building, training, and deploying machine learning mode
skill levels to identify suitable algorithms and hyperparar
such as PyTorch, TensorFlow, and scikit-learn allow data s
for machine learning further improve productivity by ena
deployed in the cloud and on the edge. All these capabil
anywhere, including data scientists' workstations.



We built Azure Machine Learning service working closely
to improve customer service, build better products and c
examples.

TAL, a 150-year-old leading life insurance company in Au
customer experience. Traditionally, TAL's quality assuranc
cases. Using Azure Machine Learning service, it is now able to review 100 percent of cases.

"Azure Machine Learning regularly lets TAL's data scie
delivering faster outcomes and the opportunity to roll out many more models than was previously possible. There is

https://azure.microsoft.com/blog/azure-machine-learning-service-a-look-under-the-hood/

Subscribe

Explore

See where we're heading. Check out upcoming changes to Azure products

Azure updates

Big Data (368)

Blockchain (72)

Business Intelligence (90)

Resources

- Source code for this talk:
  https://github.com/ritazh/kubecon-ml
- Kubeflow labs for AKS:
  https://github.com/Azure/kubeflow-labs
- Provision a Kubernetes cluster on Azure:
  https://github.com/Azure/kubeflow-labs/tree/master/2-kubernetes#provisioning-a-kubernetes-cluster-on-azure

Microsoft