Machine Learning using Kubernetes

Arun Gupta, @arungupta

Centerpiece for digital transformation

Customer experience

Business operations

Decision making

Innovation

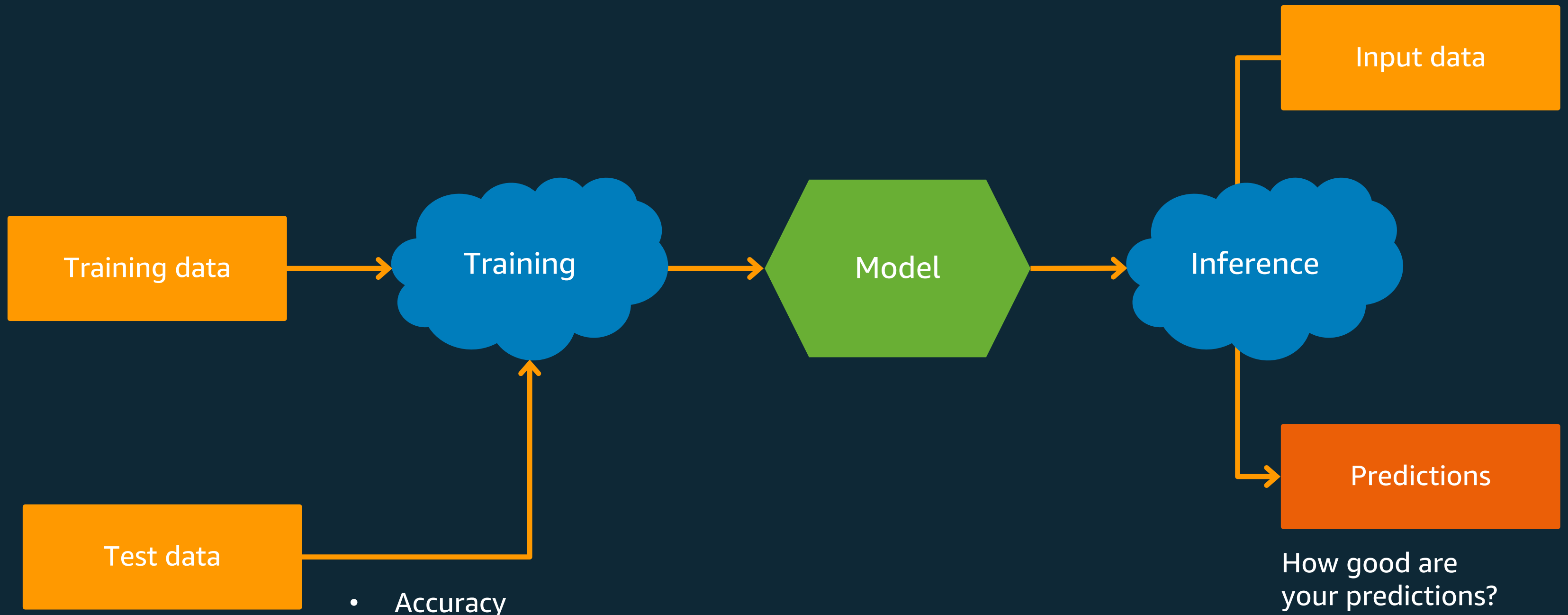
Competitive advantage

40% of digital transformation initiatives supported by AI in 2019 *—IDC 2018*

Our mission at AWS

Put machine learning in the hands of every developer

Machine Learning 101



- Accuracy
- Optimization

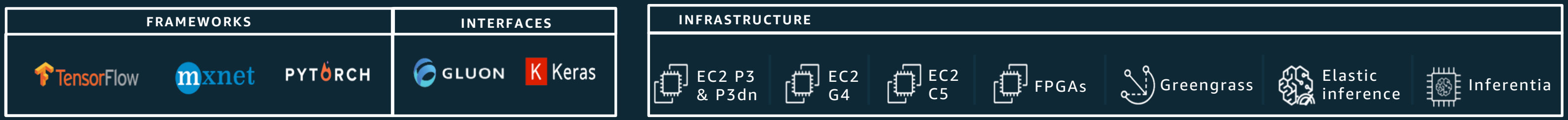
How good are
your predictions?

**A little less conversation,
a little more action, please**

—Elvis Presley

The Amazon ML stack:
Broadest & deepest set of capabilities

ML Frameworks + Infrastructure



FRAMEWORKS

TensorFlow

mxnet

PYTORCH

INTERFACES

GLUON

Keras

INFRASTRUCTURE

EC2 P3 & P3dn

EC2 G4

EC2 C5

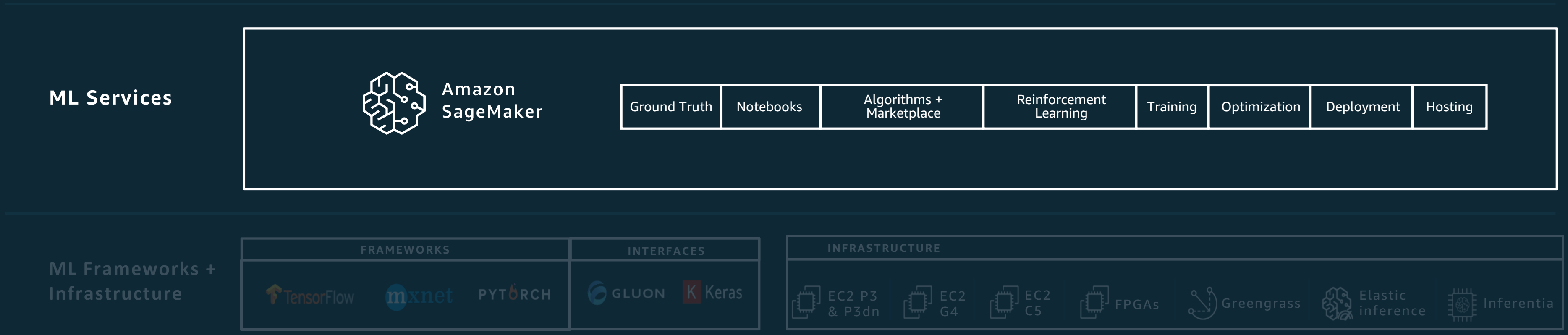
FPGAs

Greengrass

Elastic inference

Inferentia

The Amazon ML stack:
Broadest & deepest set of capabilities



ML Services

Amazon SageMaker

Ground Truth | Notebooks | Algorithms + Marketplace | Reinforcement Learning | Training | Optimization | Deployment | Hosting

ML Frameworks + Infrastructure

FRAMEWORKS

TensorFlow mxnet PYTORCH

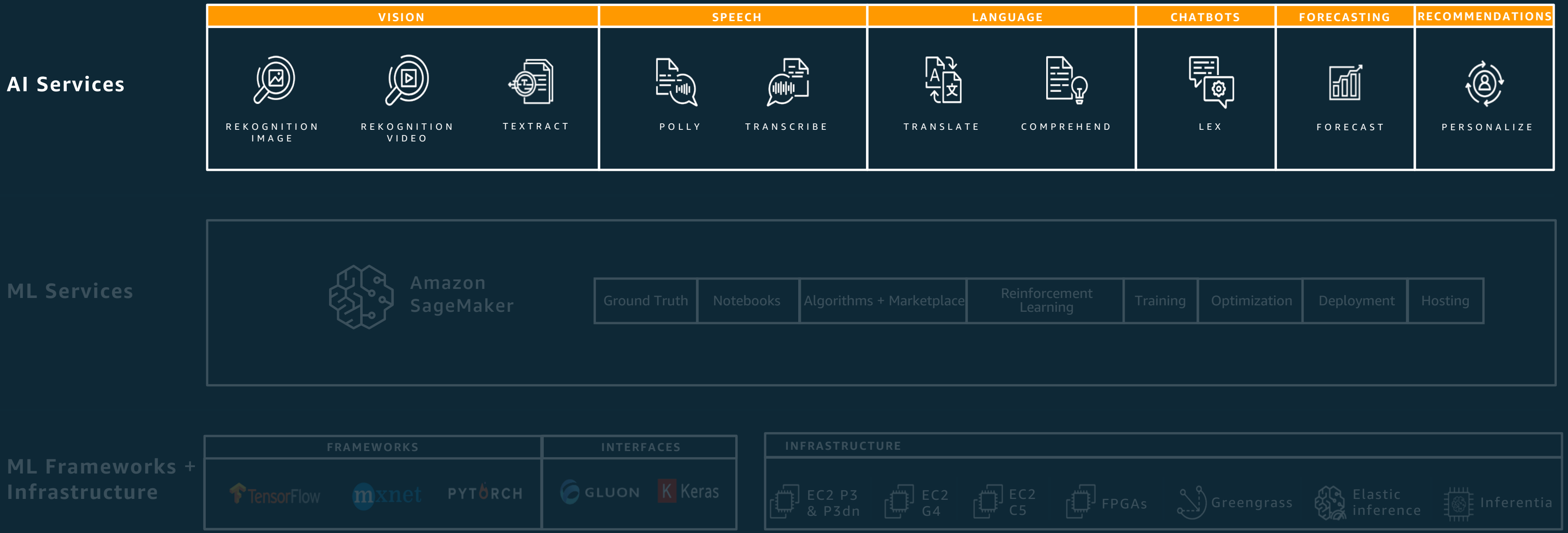
INTERFACES

GLUON Keras

INFRASTRUCTURE

EC2 P3 & P3dn | EC2 G4 | EC2 C5 | FPGAs | Greengrass | Elastic inference | Inferentia

The Amazon ML stack: Broadest & deepest set of capabilities



AI Services

VISION: REKOGNITION IMAGE, REKOGNITION VIDEO, TEXTRACT

SPEECH: POLLY, TRANSCRIBE

LANGUAGE: TRANSLATE, COMPREHEND

CHATBOTS: LEX

FORECASTING: FORECAST

RECOMMENDATIONS: PERSONALIZE

ML Services

Amazon SageMaker: Ground Truth, Notebooks, Algorithms + Marketplace, Reinforcement Learning, Training, Optimization, Deployment, Hosting

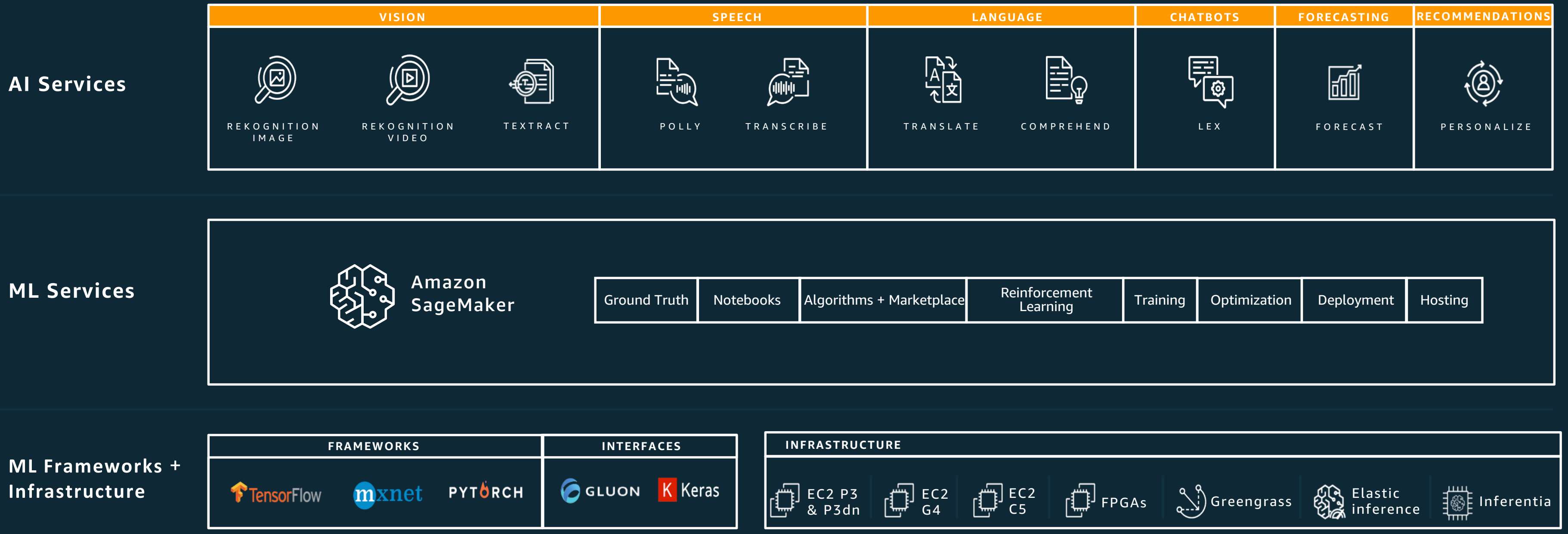
ML Frameworks + Infrastructure

FRAMEWORKS: TensorFlow, mxnet, PYTORCH

INTERFACES: GLUON, Keras

INFRASTRUCTURE: EC2 P3 & P3dn, EC2 G4, EC2 C5, FPGAs, Greengrass, Elastic inference, Inferentia

The Amazon ML stack:
Broadest & deepest set of capabilities



AI Services

- VISION: REKOGNITION IMAGE, REKOGNITION VIDEO, TEXTRACT
- SPEECH: POLLY, TRANSCRIBE
- LANGUAGE: TRANSLATE, COMPREHEND
- CHATBOTS: LEX
- FORECASTING: FORECAST
- RECOMMENDATIONS: PERSONALIZE

ML Services

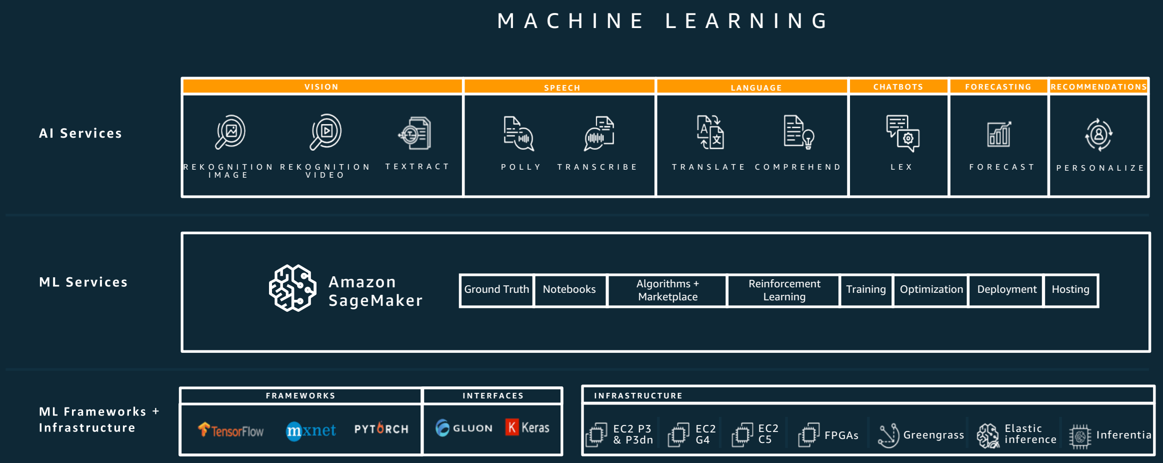
Amazon SageMaker

Ground Truth | Notebooks | Algorithms + Marketplace | Reinforcement Learning | Training | Optimization | Deployment | Hosting

ML Frameworks + Infrastructure

- FRAMEWORKS: TensorFlow, mxnet, PYTORCH
- INTERFACES: GLUON, Keras
- INFRASTRUCTURE: EC2 P3 & P3dn, EC2 G4, EC2 C5, FPGAs, Greengrass, Elastic inference, Inferentia

"A little less conversation, a little more action, please"



MACHINE LEARNING

AI Services

VISION			SPEECH		LANGUAGE		CHATBOTS	FORECASTING	RECOMMENDATIONS
REKOGNITION IMAGE	REKOGNITION VIDEO	TEXTRACT	POLLY	TRANSCRIBE	TRANSLATE	COMPREHEND	LEX	FORECAST	PERSONALIZE

ML Services

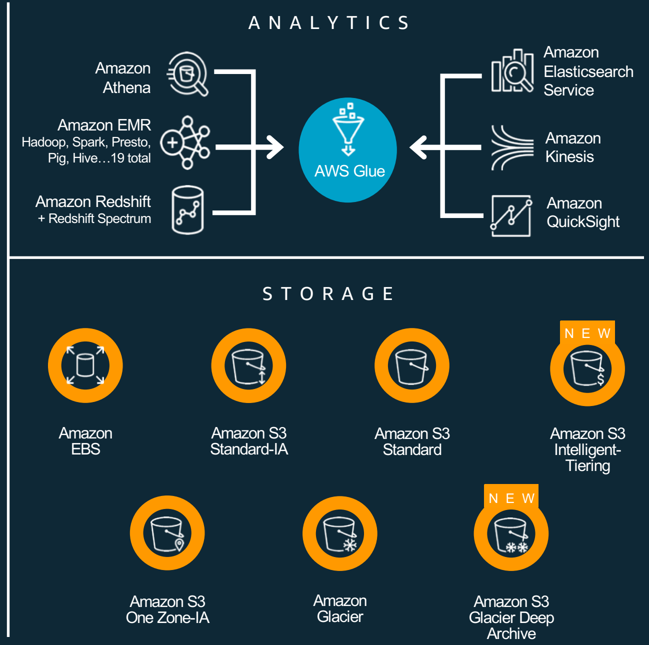
Amazon SageMaker

Ground Truth	Notebooks	Algorithms + Marketplace	Reinforcement Learning	Training	Optimization	Deployment	Hosting

ML Frameworks + Infrastructure

FRAMEWORKS			INTERFACES	
TensorFlow	mxnet	PYTORCH	GLUON	Keras

INFRASTRUCTURE						
EC2 P3 & P3dn	EC2 G4	EC2 C5	FPGAs	Greengrass	Elastic inference	Inferentia



ANALYTICS

Amazon Athena

Amazon EMR
Hadoop, Spark, Presto, Pig, Hive…19 total

Amazon Redshift
+ Redshift Spectrum

AWS Glue

Amazon Elasticsearch Service

Amazon Kinesis

Amazon QuickSight

STORAGE

Amazon EBS

Amazon S3 Standard-IA

Amazon S3 Standard

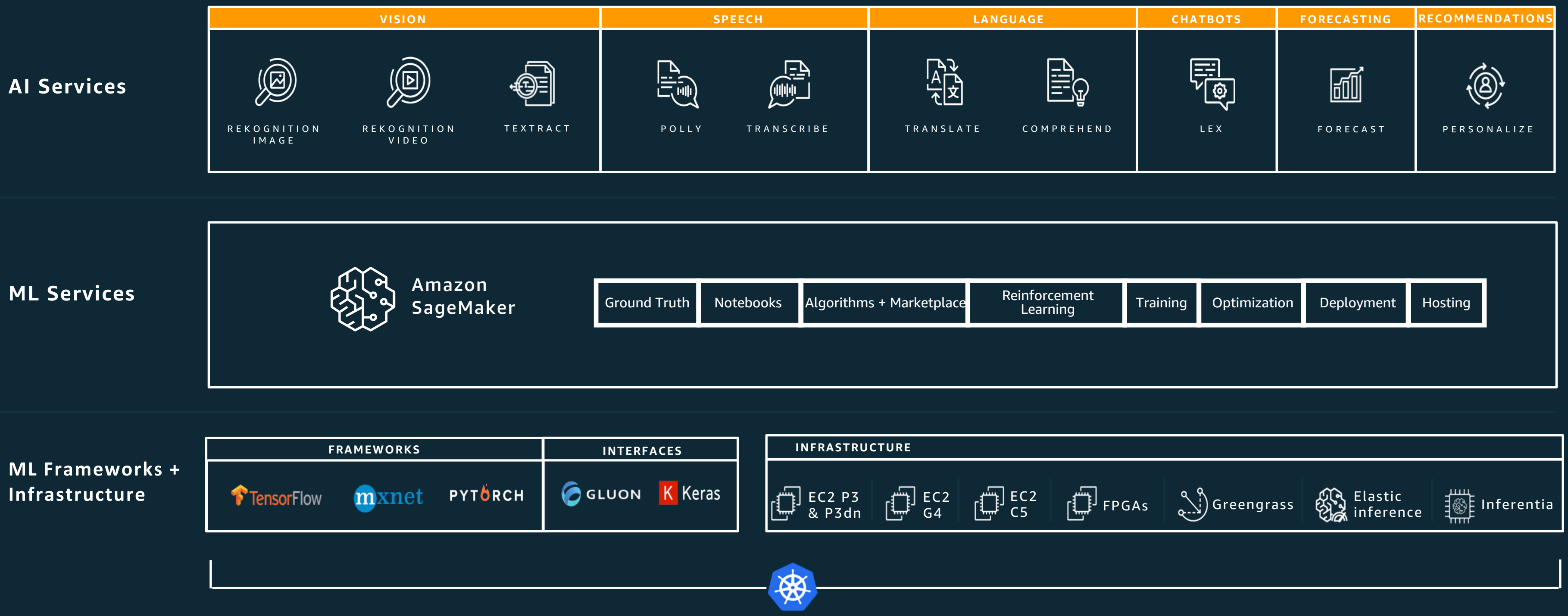
NEW
Amazon S3 Intelligent-Tiering

Amazon S3 One Zone-IA

Amazon Glacier

NEW
Amazon S3 Glacier Deep Archive

Machine Learning using Kubernetes



AI Services

VISION: REKOGNITION IMAGE, REKOGNITION VIDEO, TEXTRACT

SPEECH: POLLY, TRANSCRIBE

LANGUAGE: TRANSLATE, COMPREHEND

CHATBOTS: LEX

FORECASTING: FORECAST

RECOMMENDATIONS: PERSONALIZE

ML Services

Amazon SageMaker

Ground Truth | Notebooks | Algorithms + Marketplace | Reinforcement Learning | Training | Optimization | Deployment | Hosting

ML Frameworks + Infrastructure

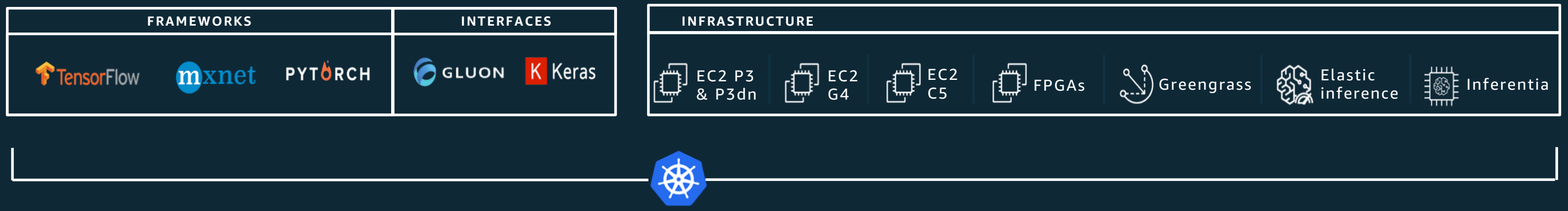
FRAMEWORKS: TensorFlow, mxnet, PYTORCH

INTERFACES: GLUON, Keras

INFRASTRUCTURE: EC2 P3 & P3dn, EC2 G4, EC2 C5, FPGAs, Greengrass, Elastic inference, Inferentia

Machine Learning using Kubernetes

ML Frameworks + Infrastructure



FRAMEWORKS

TensorFlow

mxnet

PYTORCH

INTERFACES

GLUON

Keras

INFRASTRUCTURE

EC2 P3 & P3dn

EC2 G4

EC2 C5

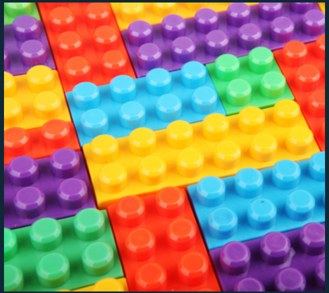
FPGAs

Greengrass

Elastic inference

Inferentia

Why Machine Learning on Kubernetes?



Composability

Portability

Scalability

http://www.shutterstock.com/gallery-635827p1.html

Amazon EKS—run Kubernetes in cloud

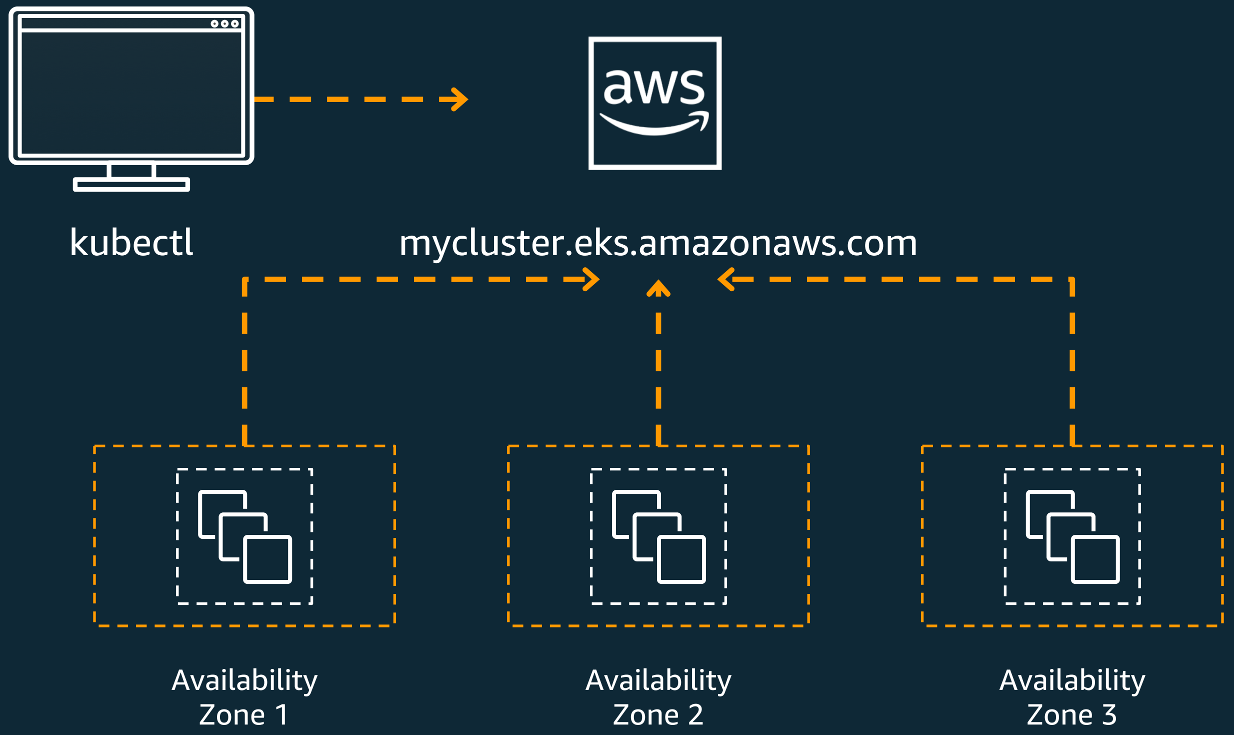
Managed Kubernetes control plane, attach data plane

Native upstream Kubernetes experience

Platform for enterprises to run production-grade workloads

Integrates with additional AWS services

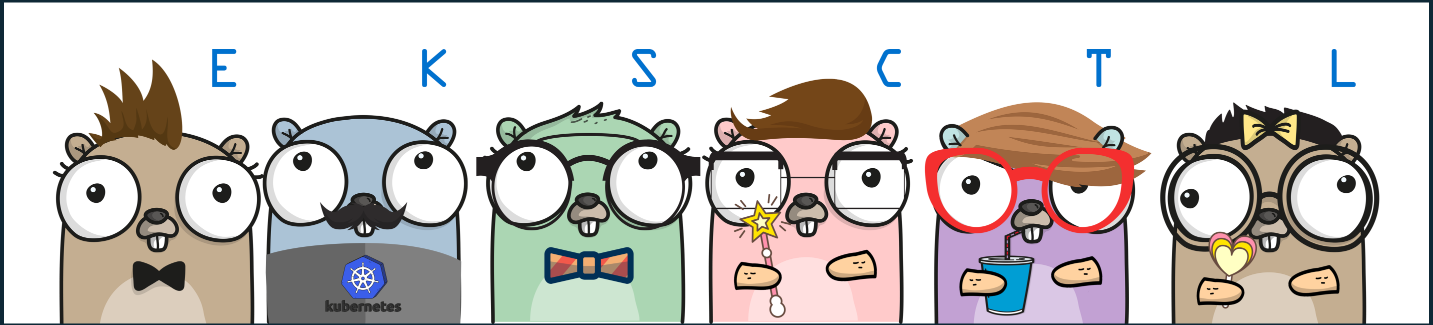
Amazon EKS deployment



Getting started with Amazon EKS

eksctl CLI—create Amazon EKS clusters (eksctl.io)

Creates all resources needed for the cluster



Creating an EKS cluster using eksctl

`brew tap weaveworks/tap`
`brew install weaveworks/tap/eksctl`
	Install

`eksctl create cluster`	Auto generated cluster name 2x `m5.large` nodes Uses AWS EKS AMI `us-west-2` region Dedicated VPCs Static AMI resolver
`eksctl create cluster --node-type=p2.xlarge`	GPU-powered cluster

GPUs for Machine Learning training

- Training maps to matrix multiplications
- Coupled with extremely high memory bandwidth

Matrix A

a_{11}	a_{12}	a_{13}
a_{21}	a_{22}	a_{23}
a_{31}	a_{32}	a_{33}

✖

Matrix B

b_{11}	b_{12}	b_{13}
b_{21}	b_{22}	b_{23}
b_{31}	b_{32}	b_{33}

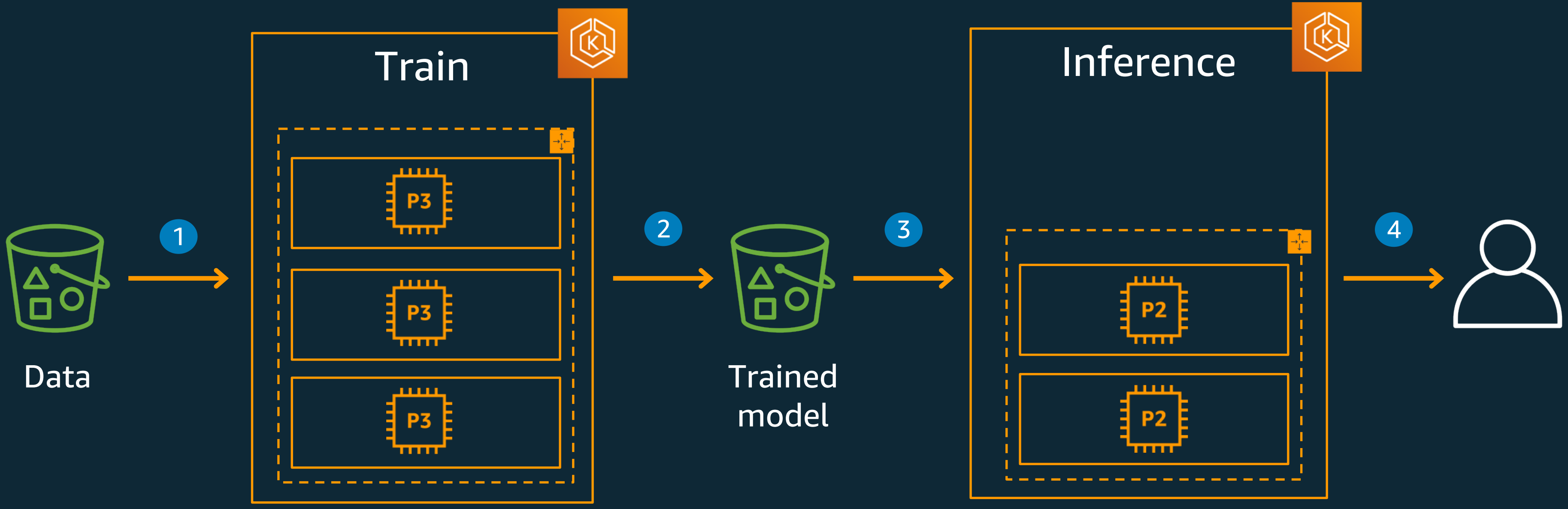
=

Matrix C

$a_{11}.b_{11} + a_{12}.b_{21} + a_{13}.b_{31}$	$a_{11}.b_{12} + a_{12}.b_{22} + a_{13}.b_{32}$	$a_{11}.b_{13} + a_{12}.b_{23} + a_{13}.b_{33}$
$a_{21}.b_{11} + a_{22}.b_{21} + a_{23}.b_{31}$	$a_{21}.b_{12} + a_{22}.b_{22} + a_{23}.b_{32}$	$a_{21}.b_{13} + a_{22}.b_{23} + a_{23}.b_{33}$
$a_{31}.b_{11} + a_{32}.b_{21} + a_{33}.b_{31}$	$a_{31}.b_{12} + a_{32}.b_{22} + a_{33}.b_{32}$	$a_{31}.b_{13} + a_{32}.b_{23} + a_{33}.b_{33}$

Operations can be parallelized across 1,000s of cores

Set up K8s for ML—option 1



Dedicated K8s cluster

Create K8s cluster for ML—option 1

```
eksctl create cluster \
  --name training \
  --nodes=4 \
  --node-type=p3.8xlarge
```

Create training cluster

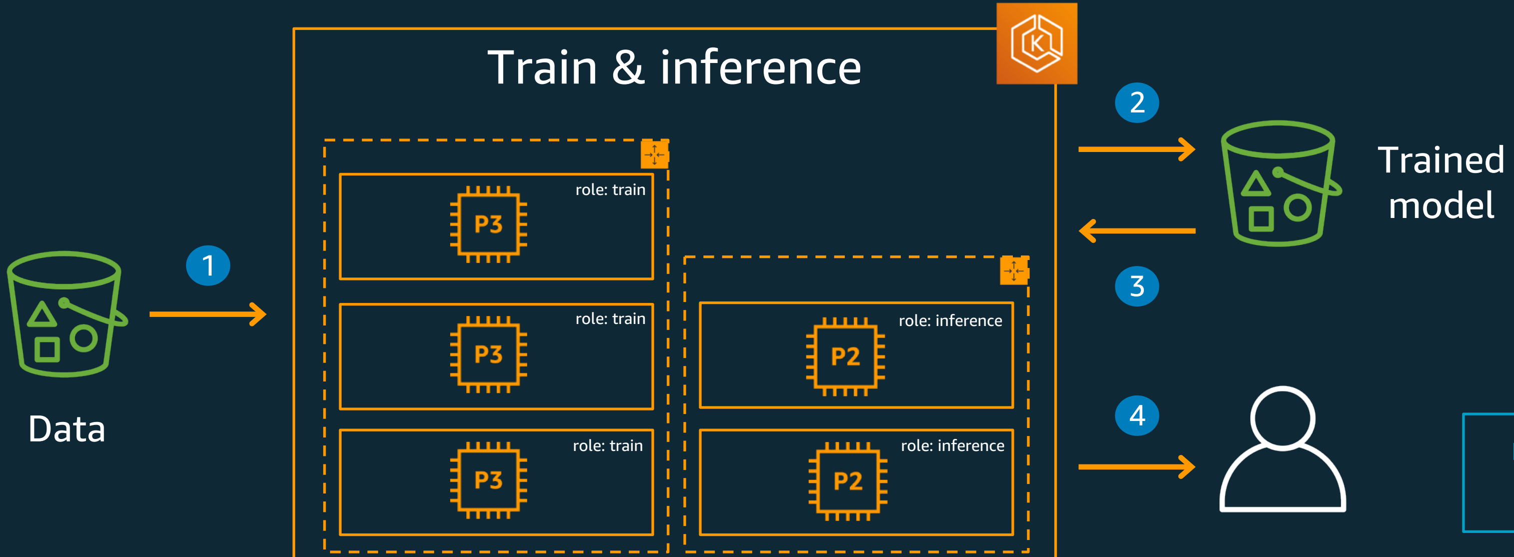
```
eksctl create cluster \
  --name inference \
  --nodes=2 \
  --node-type=p2.xlarge
```

Create inference cluster

Scaling the cluster

CLUSTER AUTOSCALER	ESCALATOR
Burst-able workloads	Batch or job-based workloads
Aggressively move pods for utilization, can be configured for completion	Wait for the jobs to be completed
Scale up based upon metrics	Aggressively scale up to reduce wait-time for pods
Takes over desired instance knob of auto-scaling group	
Run them in same cluster with different node groups	

Set up K8s for ML—option 2a



```
nodeSelector:
  role: train
```

Unified K8s cluster

Create K8s cluster for ML—option 2

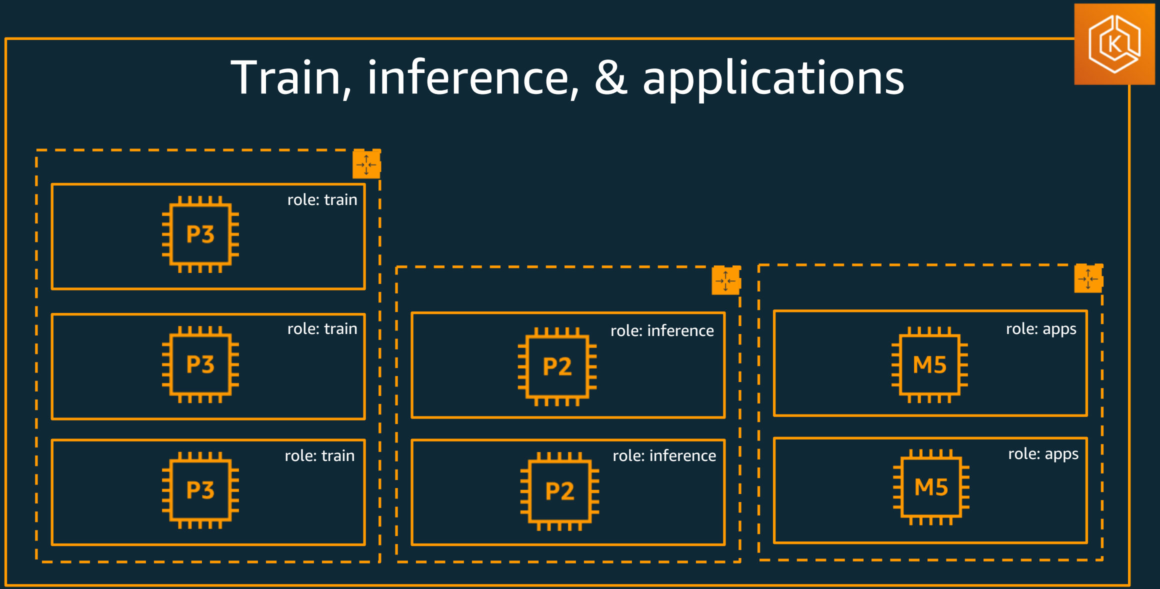
```
apiVersion: eksctl.io/v1alpha4
kind: ClusterConfig

metadata:
  name: gpu-cpu-cluster
  region: us-west-2

nodeGroups:
  - name: ng-train
    labels: {role: train}
    instanceType: p3.8xlarge
    desiredCapacity: 4
  - name: ng-inference
    labels: {role: inference}
    instanceType: m5. 2xlarge
    desiredCapacity: 4
```

Eksctl cluster configuration with two node groups

```
eksctl create cluster -f config.yaml
```

Create cluster

Set up K8s for ML—option 2b



```
nodeSelector:
    role: train
```

Unified K8s cluster

Challenges in setting up containers for ML

Takes days to test and configure

Must optimize for performance & scale

Rebuild and re-optimize new framework versions

AWS deep learning containers

Optimized and customizable containers for deep learning environments

Pre-packaged Docker container images fully configured and validated

Best performance and scalability without tuning

Works with Amazon EKS, Amazon ECS, and Amazon EC2

KEY FEATURES

Customizable container images

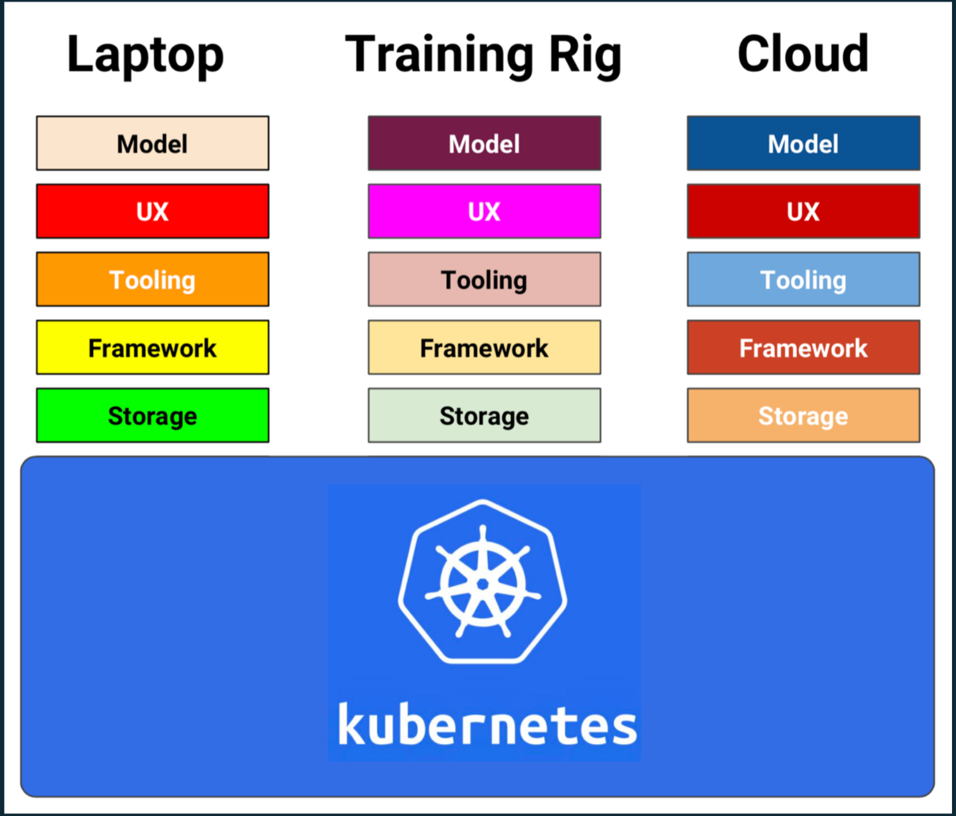
Support for TensorFlow, Apache MXNet

Single and multi-node training and inference

16 container images

TensorFlow	mxnet
Training	Inference
GPU	CPU
Python 2.7	Python 3.6

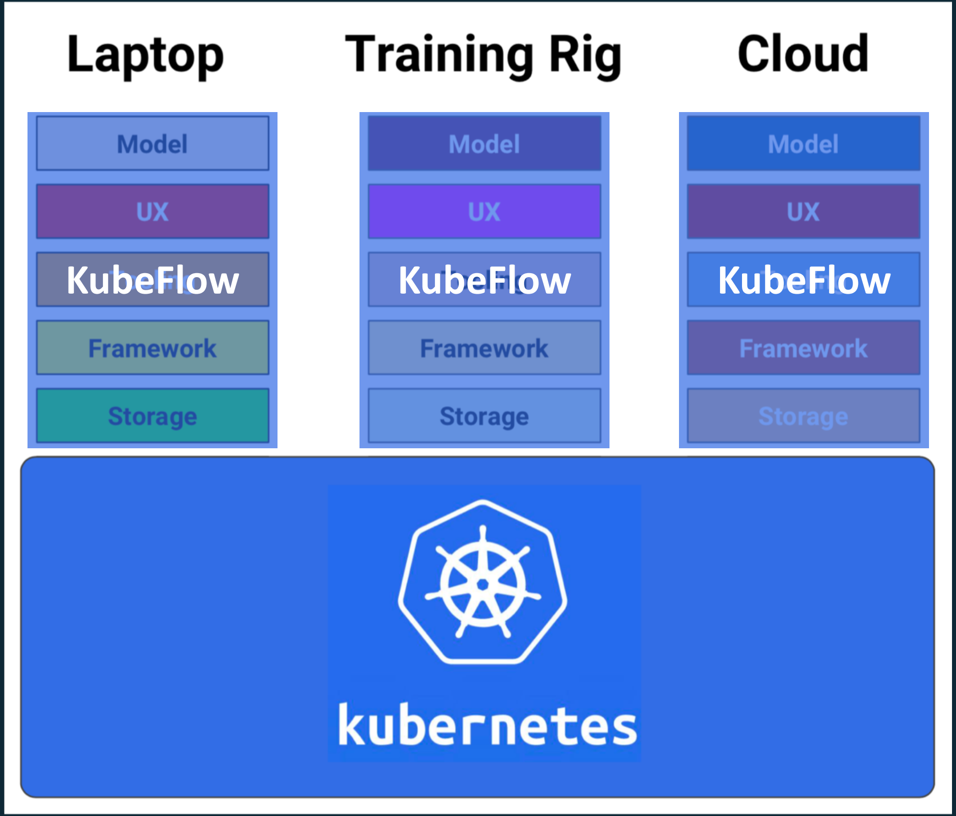
ML on K8s—without KubeFlow



Laptop	Training Rig	Cloud
Model	Model	Model
UX	UX	UX
Tooling	Tooling	Tooling
Framework	Framework	Framework
Storage	Storage	Storage

Credits: @aronchik

ML on K8s—with KubeFlow



Credits: @aronchik

What's in KubeFlow?

jupyter

Notebook for collaborative & interactive training

Serving deployment & training controller

argo

For workflows

SELDON

For complex inference and non TF models

Framework operators

ReverseProxy (ambassador)

Wiring to make it work on any K8s anywhere

MNIST database

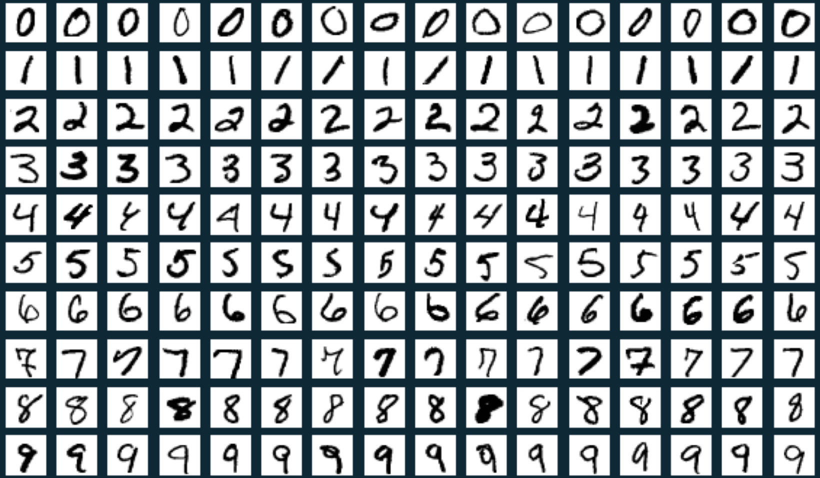
Database of gray-scaled handwritten digits

Training set of 60k

Test set of 10k

Size-normalized (28x28 pixels)

Centered in a fixed-size image



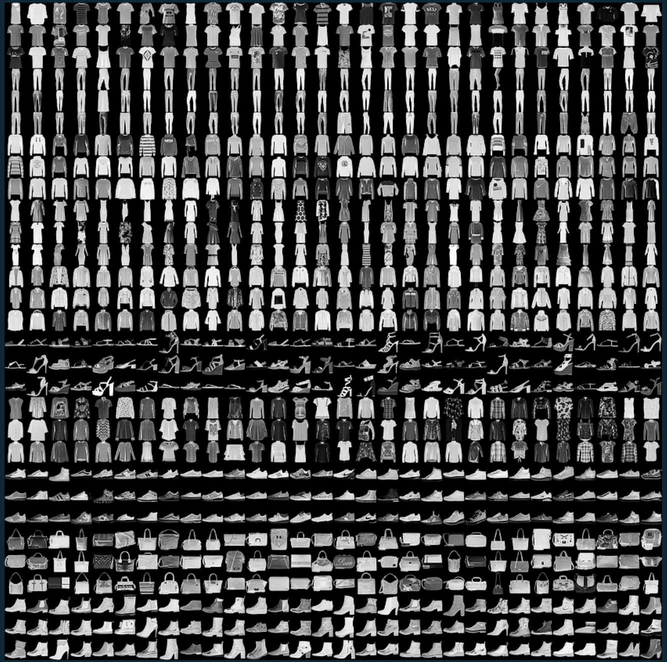
http://yann.lecun.com/exdb/mnist/

Fashion MNIST

Database of Zalando's article images

Labels assigned to 10 items

Drop-in replacement for MNIST



https://github.com/zalandoresearch/fashion-mnist

TensorFlow

Open source library to develop and train ML models

Created by Google Brain team

Can run on desktop, servers, mobiles, edge devices

AWS is the platform of choice to run TensorFlow

Browsi
BTGPactual digital
COX AUTOMOTIVE
THOMSON REUTERS
edmunds.com
GoodData
grammarly
Expedia
hudl
IDEXX
intuit
Tradelegs
MATRIX ANALYTICS
Pinterest
Relay42
SCDM
DEV FACTORY
Rakuten Intelligence
jayway.
digitizing ideas
zendesk
Zocdoc

85% of all TensorFlow workloads in the cloud runs on AWS

Source: Nucleus Research, November 2018

Train twice as fast with TensorFlow

65%

Scaling efficiency with 256 GPUs

STOCK TENSORFLOW

90%

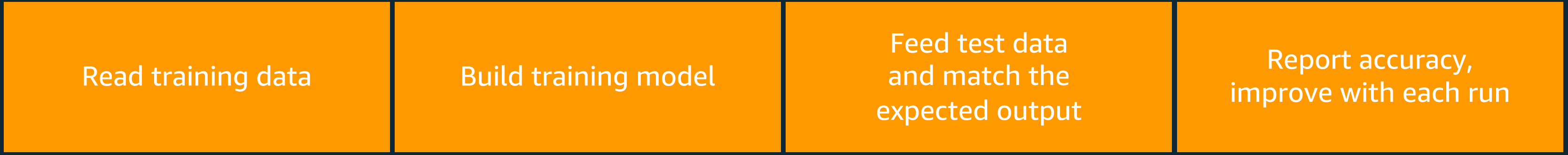
Scaling efficiency with 256 GPUS

AWS-OPTIMZED TENSORFLOW

Machine Learning using TensorFlow on K8s

Download Keras-consumable Fashion-MNIST training and test data

Run 40 epochs on the model



Read training data

Build training model

Feed test data and match the expected output

Report accuracy, improve with each run

Export the model to S3 bucket

Apache MXNet

Programmable

Simple syntax,
multiple languages

Portable

Highly efficient models
for mobile and IoT

High performance

Near linear scaling across
hundreds of GPUs

Most open

Accepted into the
Apache Incubator

Best on AWS

Optimized for deep
learning on AWS

Advantages of KubeFlow on AWS

EKS cluster provision with eksctl

External traffic with AWS ALB Ingress Controller

Amazon FSx CSI driver to manage Lustre file system

Centralized and unified K8s logs in CloudWatch

TLS and Auth with AWS Certificate Manager and AWS Cognito

Private access for your K8s API server endpoint

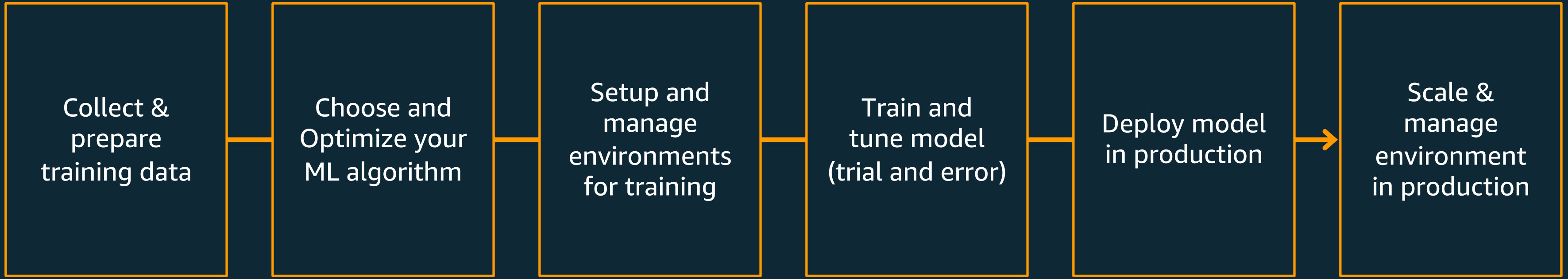
Detect GPU instance and install Nvidia device plugin

Distributed training using Horovod

Distributed Training framework for TensorFlow, Keras, PyTorch, and MXNet

Traditional Russian dance where participants dance in a circle with linked hands

Machine Learning pipeline



Collect & prepare training data

Choose and Optimize your ML algorithm

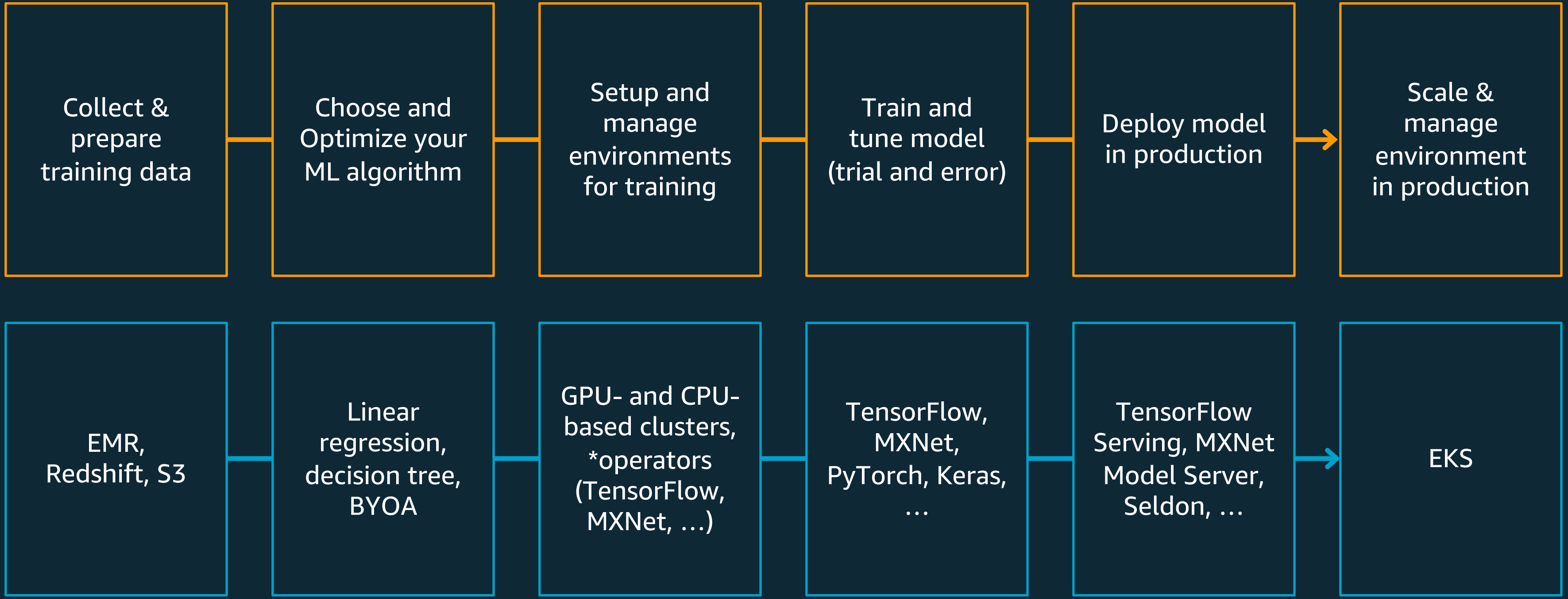
Setup and manage environments for training

Train and tune model (trial and error)

Deploy model in production

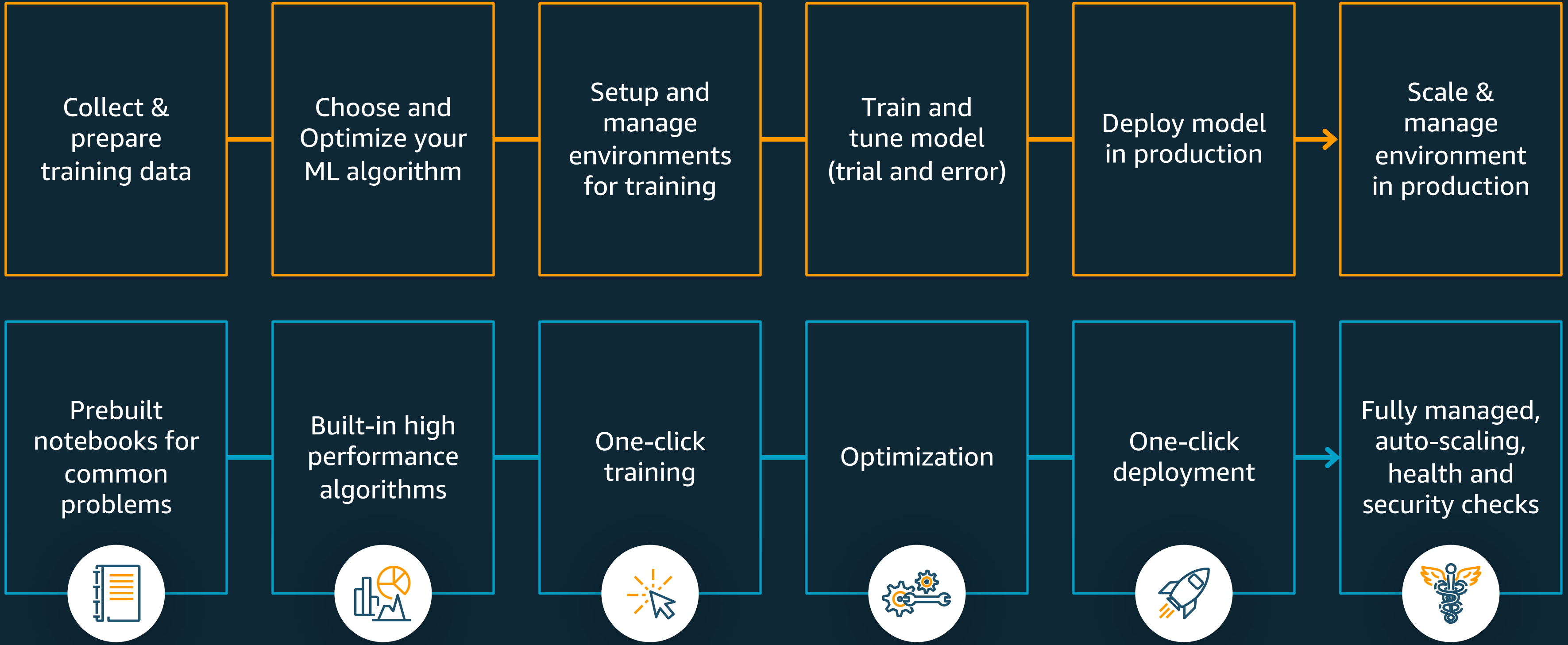
Scale & manage environment in production

Machine Learning pipeline for K8s



Collect & prepare training data	Choose and Optimize your ML algorithm	Setup and manage environments for training	Train and tune model (trial and error)	Deploy model in production	Scale & manage environment in production
EMR, Redshift, S3	Linear regression, decision tree, BYOA	GPU- and CPU-based clusters, *operators (TensorFlow, MXNet, …)	TensorFlow, MXNet, PyTorch, Keras, …	TensorFlow Serving, MXNet Model Server, Seldon, …	EKS

Machine Learning pipeline using SageMaker



Collect & prepare training data

Choose and Optimize your ML algorithm

Setup and manage environments for training

Train and tune model (trial and error)

Deploy model in production

Scale & manage environment in production

Prebuilt notebooks for common problems

Built-in high performance algorithms

One-click training

Optimization

One-click deployment

Fully managed, auto-scaling, health and security checks

References

https://github.com/aws-samples/machine-learning-using-k8s/