Machine Learning on AWS - Technical
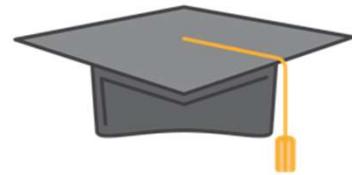
AWS Solutions Training for Partners

Vijay
**AWS Partner Trainer**

Machine Learning
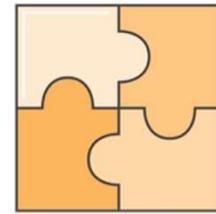
Prediction is the process of filling in missing information; it uses data you have to generate data you don't have.

Learning   Language   Perception   Problem Solving   Insight
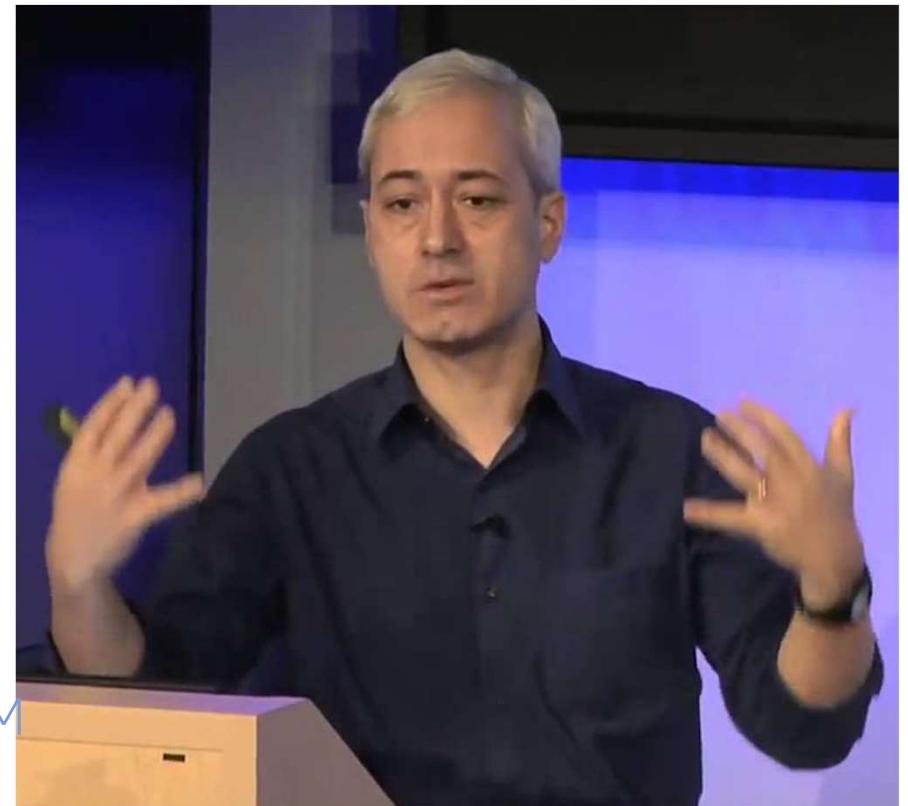
The Artificial Intelligence Landscape
Five Tribes / Two Breakthroughs

Tribe	Origins	Algorithm
Bayesians	Statistics	Probabilistic
Analogizers	Psychology	Kernel
Symbolists	Logic	Inverse Deduction
Evolutionaries	Biology	Genetic
Connectionists	Neuroscience	Back Propagation

- Computer Vision : CNNs
  - Static / Unstructured
- Natural Language Processing : RNNs / LSTM
  - Sequential / Structured
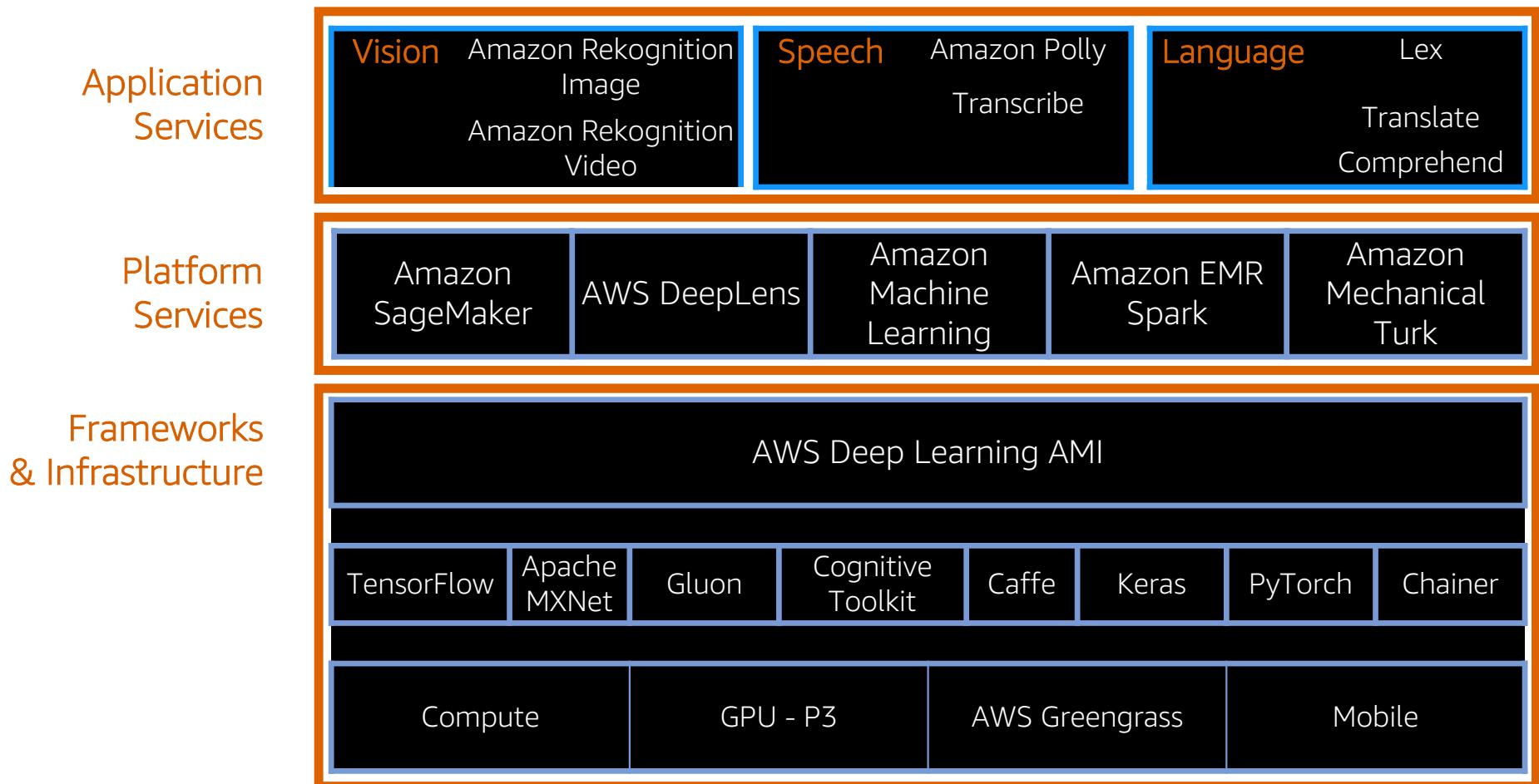
AWS Mission

Put machine learning in the hands of every developer, data scientist and architect

Customers Running ML on AWS

The AWS Machine Learning Stack

aws training and certification

Application Services

| Vision | Amazon Rekognition Image
Amazon Rekognition Video | Speech | Amazon Polly
Transcribe | Language | Lex
Translate
Comprehend |

Platform Services

| Amazon SageMaker | AWS DeepLens | Amazon Machine Learning | Amazon EMR Spark | Amazon Mechanical Turk |

Frameworks & Infrastructure

| AWS Deep Learning AMI |

| TensorFlow | Apache MXNet | Gluon | Cognitive Toolkit | Caffe | Keras | PyTorch | Chainer |

| Compute | GPU - P3 | AWS Greengrass | Mobile |

The AWS Machine Learning Stack

aws training and certification

Application Services

Vision	Amazon Rekognition Image Amazon Rekognition Video	Speech	Amazon Polly Transcribe	Language	Lex Translate Comprehend

Platform Services

Amazon SageMaker	AWS DeepLens	Amazon Machine Learning	Amazon EMR Spark	Amazon Mechanical Turk

Frameworks & Infrastructure

AWS Deep Learning AMI

TensorFlow	Apache MXNet	Gluon	Cognitive Toolkit	Caffe	Keras	PyTorch	Chainer

Compute	GPU - P3	AWS Greengrass	Mobile

Demo 1: Vision Services

The AWS Machine Learning Stack

aws training and certification

Application Services

Vision	Amazon Rekognition Image Amazon Rekognition Video	Speech	Amazon Polly Transcribe	Language	Lex Translate Comprehend

Platform Services

Amazon SageMaker	AWS DeepLens	Amazon Machine Learning	Amazon EMR Spark	Amazon Mechanical Turk

Frameworks & Infrastructure

AWS Deep Learning AMI

TensorFlow	Apache MXNet	Gluon	Cognitive Toolkit	Caffe	Keras	PyTorch	Chainer
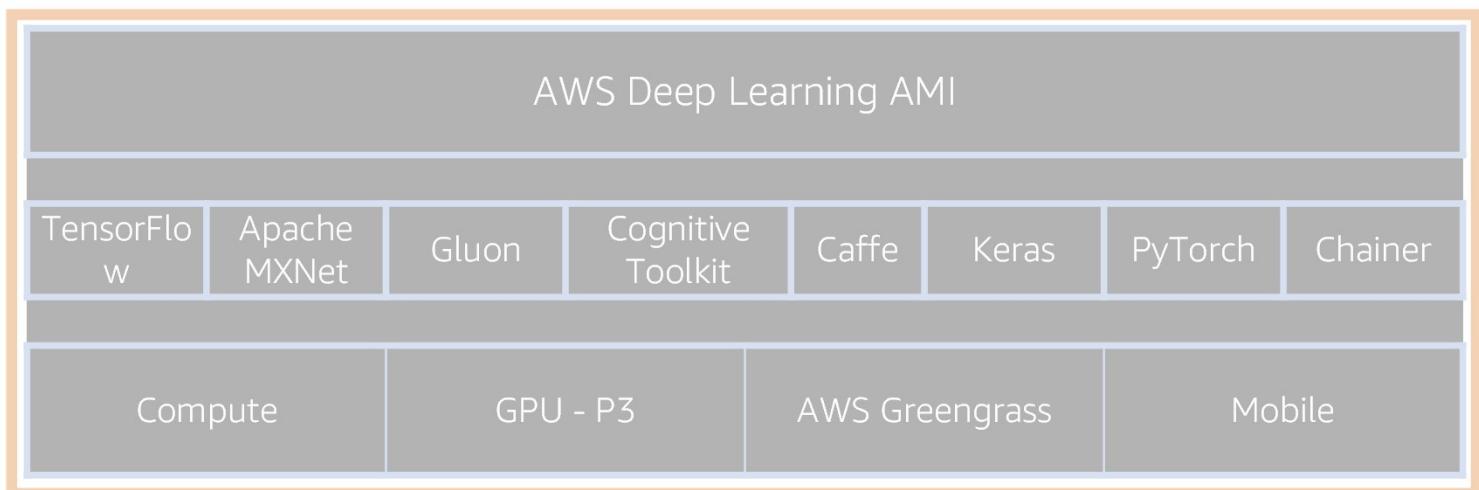
Compute	GPU - P3	AWS Greengrass	Mobile

Demo 2: Language Services

The AWS Machine Learning Stack

aws training and certification

Application Services

Vision	Amazon Rekognition Image	Speech	Polly	Language	Lex
	Amazon Rekognition Video		Transcribe		Translate
					Comprehend

Platform Services

Amazon SageMaker	AWS DeepLens	Amazon Machine Learning	Amazon EMR Spark	Amazon Mechanical Turk

Frameworks & Infrastructure

AWS Deep Learning AMI							
TensorFlow	Apache MXNet	Gluon	Cognitive Toolkit	Caffe	Keras	PyTorch	Chainer

Compute	GPU - P3	AWS Greengrass	Mobile

The AWS Machine Learning Stack

aws training and certification

Application Services

| Vision | Amazon Rekognition Image / Amazon Rekognition Video | Speech | Polly / Transcribe | Language | Lex / Translate / Comprehend |

Platform Services

| Amazon SageMaker | AWS DeepLens | Amazon Machine Learning | Amazon EMR Spark | Amazon Mechanical Turk |

Frameworks & Infrastructure

AWS Deep Learning AMI

| TensorFlow | Apache MXNet | Gluon | Cognitive Toolkit | Caffe | Keras | PyTorch | Chainer |

| Compute | GPU - P3 | AWS Greengrass | Mobile |

Demo 3: Deep Learning AMI

The AWS Machine Learning Stack

aws training and certification

Application Services

Vision	Amazon Rekognition Image	Speech	Amazon Polly	Language	Lex
	NEW Amazon Rekognition Video		**NEW** Transcribe		**NEW** Translate
					NEW Comprehend

Platform Services

NEW Amazon SageMaker	**NEW** AWS DeepLens	Amazon Machine Learning	Amazon EMR Spark	Amazon Mechanical Turk

Frameworks & Infrastructure

AWS Deep Learning AMI

TensorFlow	Apache MXNet	**NEW** Gluon	Cognitive Toolkit	Caffe	Keras	PyTorch	Chainer

Compute	**NEW** GPU - P3	**NEW** AWS Greengrass	Mobile

Amazon EC2 P3 Instances

The fastest, most powerful GPU instances in the cloud

- Up to 8 NVIDIA Tesla V100 GPUs
  - 16GB GPU memory with 900 GB/sec peak bandwidth
- 1 PetaFLOPs of computational performance
  - 14x better than P2
- 300 GB/s GPU-to-GPU communication (NVLink)
  - 9X better than P2

Airbnb
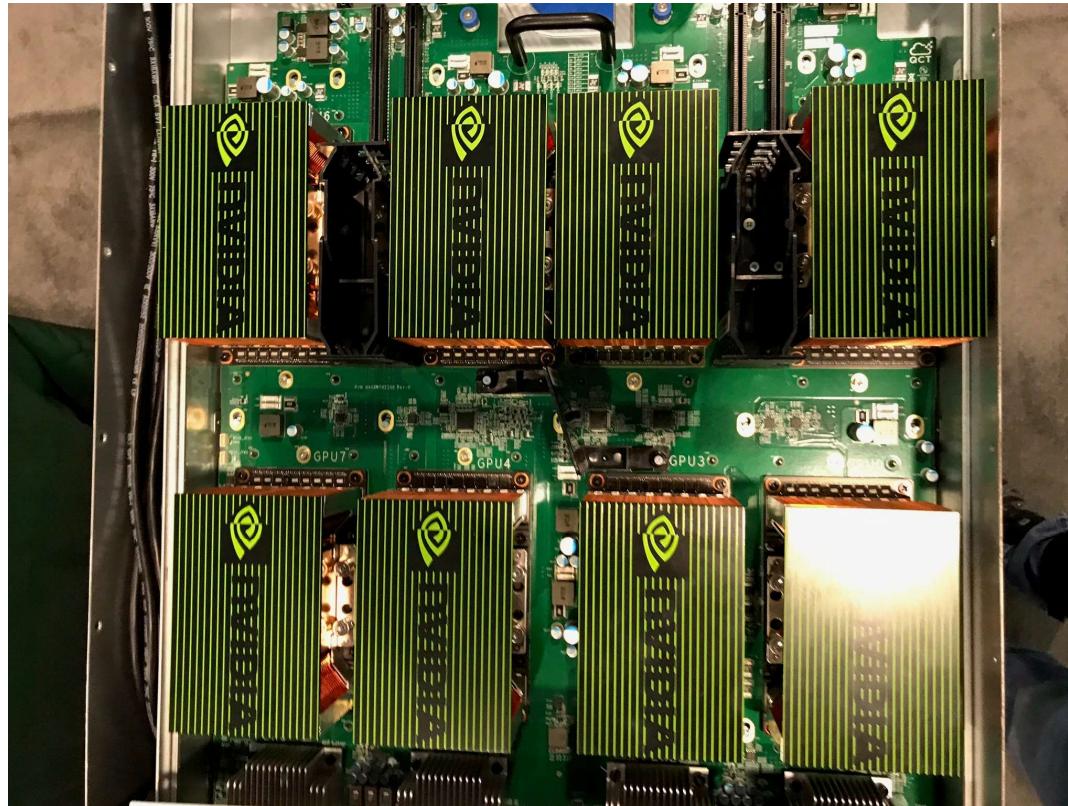
Toyota Research Institute

OpenAI

Amazon EC2 P3 Instances

Amazon EC2 P3 Instances

AWS Deep Learning Amazon Machine Image (AMI)

- Get started quickly with easy-to-launch tutorials
- Hassle-free setup and configuration
- Pay only for what you use – no additional charge for the AMI
- Accelerate your model training and deployment
- Support for popular deep learning frameworks

TensorFlow, MXNet, Gluon, Keras, Caffe2, PyTorch, Zendesk, Matric Analytics, SCDM, etc.

Amazon ML Solutions Lab

Lots of companies doing Machine Learning

Lack ML expertise

Unable to unlock business potential

Amazon ML Solutions Lab

Lots of companies doing Machine Learning

Lack ML expertise

Unable to unlock business potential

Amazon ML Solutions Lab provides the missing ML expertise

Brainstorming

Modeling

Education

Amazon ML Lab Customers

Johnson & Johnson
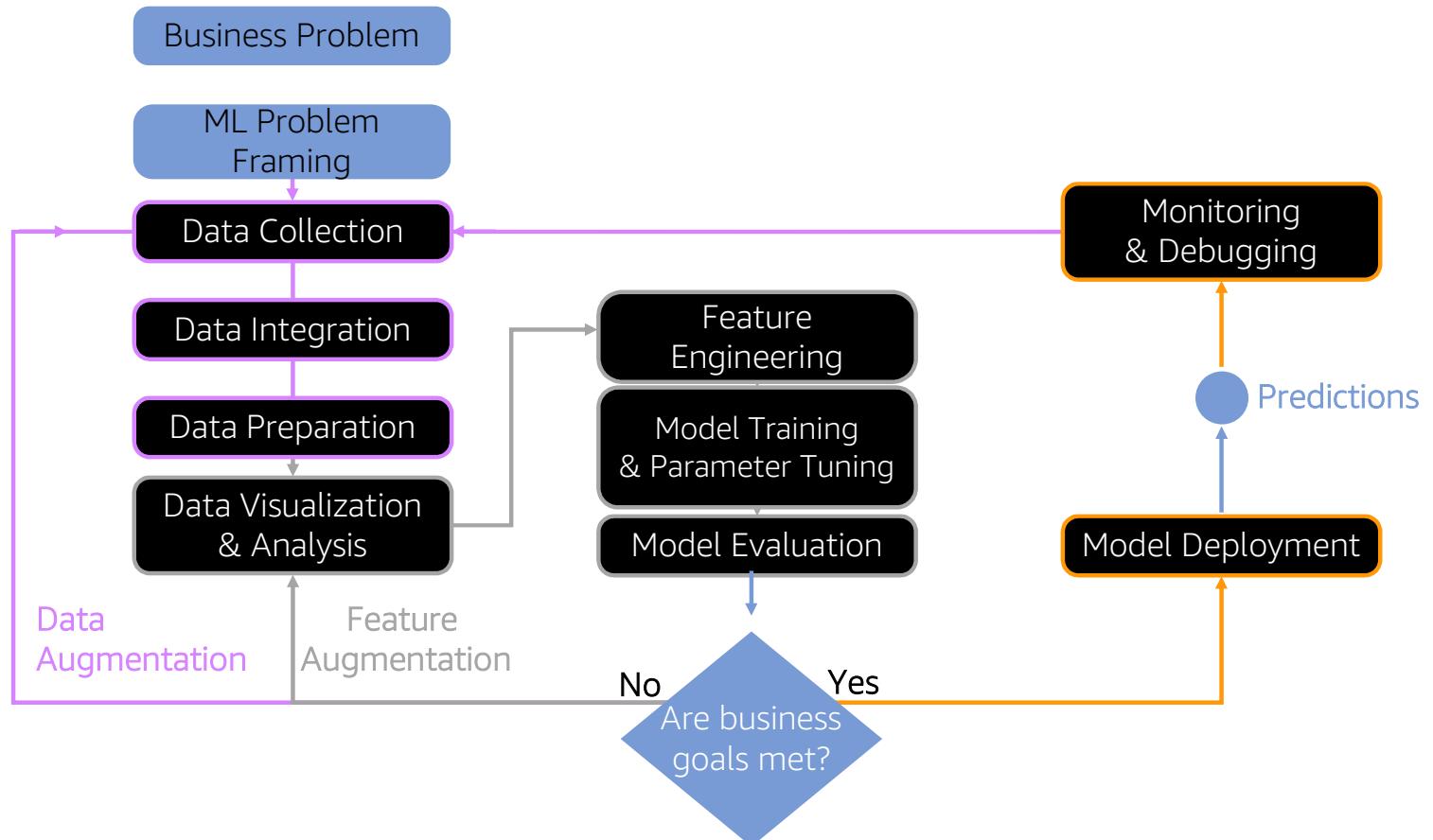
Toyota Research Institute

Washington Post

The Machine Learning Process

The Machine Learning Process
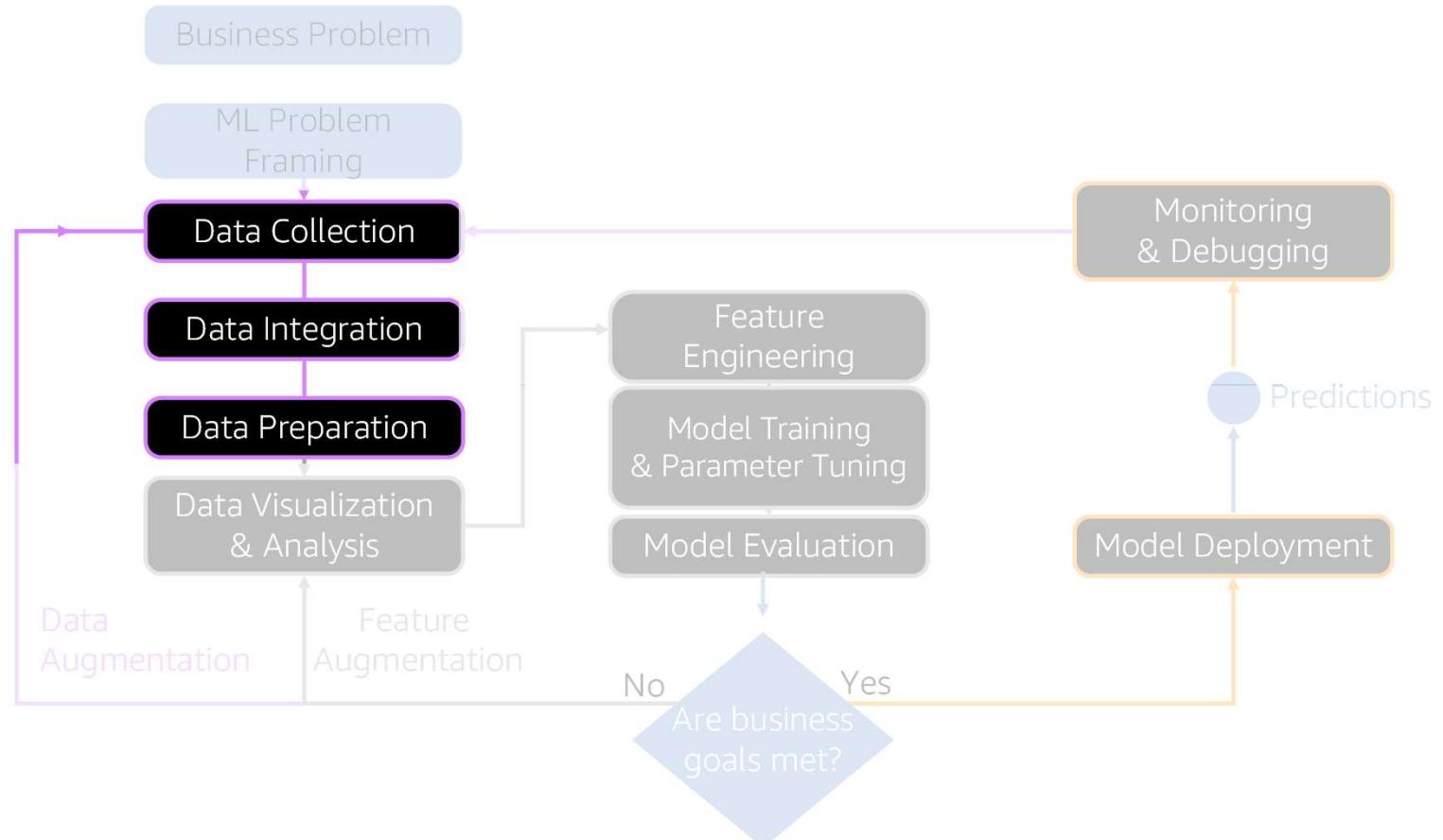
The ML Process
Integration: The Data Architecture

aws training and certification

Build the data platform:

- Amazon Simple Storage Service (Amazon S3)
- Amazon Athena
- Amazon EMR
- Amazon Redshift
- AWS Glue

Business Problem

ML Problem Framing

Data Collection

Data Integration

Data Preparation

Data Visualization & Analysis

Feature Engineering

Model Training & Parameter Tuning

Model Evaluation

Monitoring & Debugging

Predictions

Model Deployment

Data Augmentation

Feature Augmentation

No   Yes

Are business goals met?

The ML Process

aws training and certification

- Setup and Manage
  - Notebook Environments
  - Training Clusters
- Write Data Connectors
- Scale ML algorithms to large datasets
- Distribute ML training algorithm to multiple machines
- Secure model artifacts

Business Problem

ML Problem Framing

Data Collection

Data Integration

Data Preparation

Data Visualization & Analysis

Feature Engineering

Model Training & Parameter Tuning

Model Evaluation

Monitoring & Debugging

Predictions

Model Deployment

Data Augmentation

Feature Augmentation

No   Yes

Are business goals met?
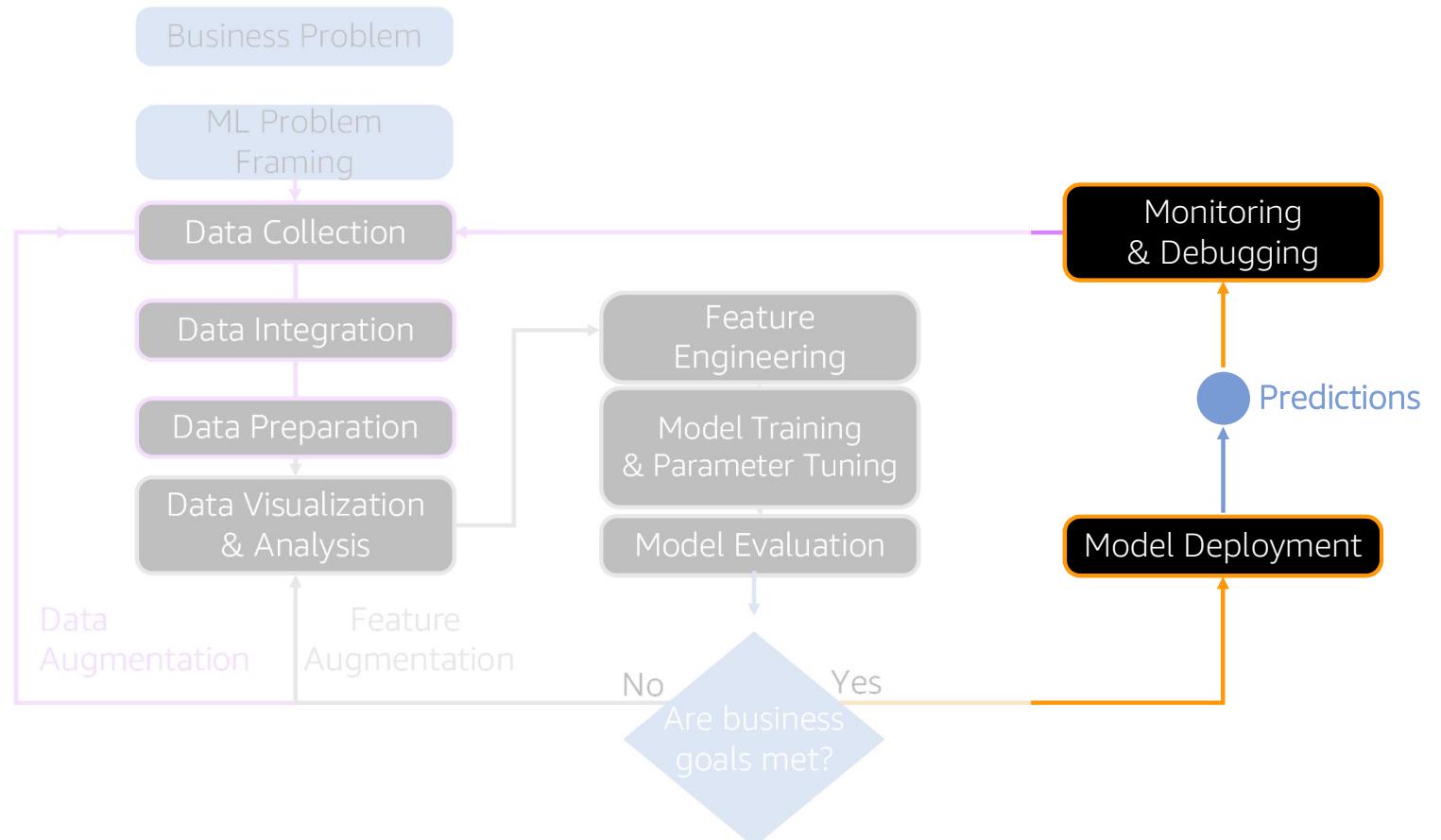
The ML Process
DevOps: Undifferentiated Heavy Lifting

aws training and certification
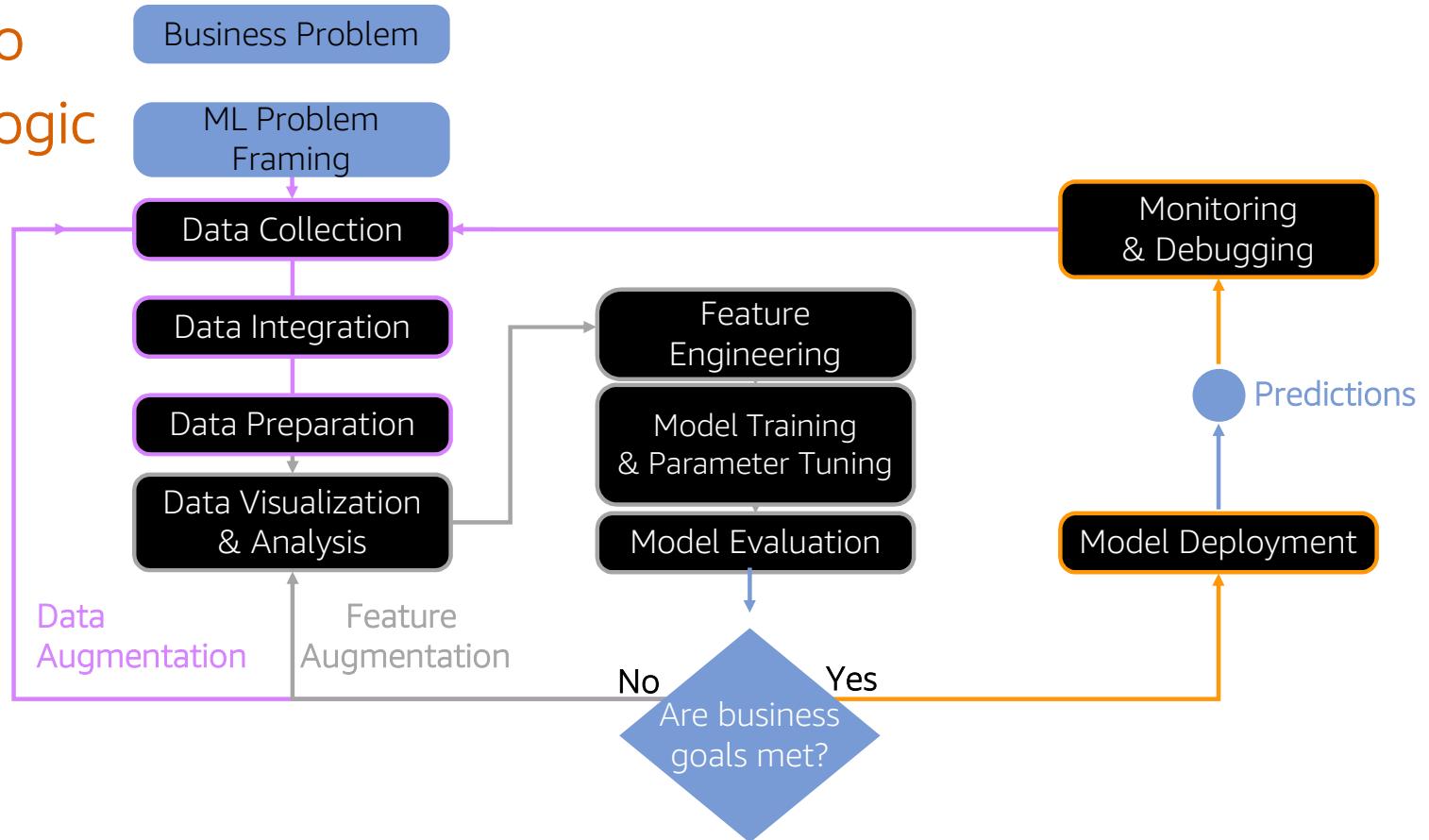
- Setup and Manage Inference Clusters

- Manage and Scale Model Inference APIs

- Monitor and Debug Model Predictions

- Models versioning and performance tracking

- Automate New Model version promotion to production (A/B testing)

Business Problem

ML Problem Framing

Data Collection

Data Integration

Data Preparation

Data Visualization & Analysis

Feature Engineering

Model Training & Parameter Tuning

Model Evaluation

Monitoring & Debugging

Predictions

Model Deployment

Data Augmentation

Feature Augmentation

No   Yes

Are business goals met?

Why Amazon SageMaker?

You Only Have to Write Business Logic

Flowchart:

- Business Problem
- ML Problem Framing
- Data Collection
- Data Integration
- Data Preparation
- Data Visualization & Analysis
- Feature Engineering
- Model Training & Parameter Tuning
- Model Evaluation
- Are business goals met?
  - No → Data Augmentation / Feature Augmentation
  - Yes → Model Deployment
- Predictions
- Monitoring & Debugging

Data Augmentation

Feature Augmentation

Amazon SageMaker

A Fully-Dockerized Lifecycle
From Discovery to Development and Deployment

aws training and certification
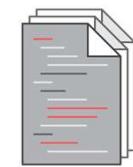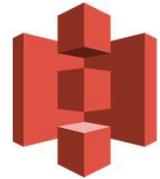
Data Scientists
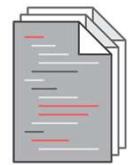
Amazon SageMaker
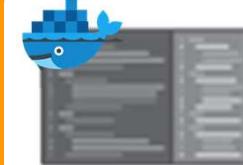
Model Artifacts

Amazon S3

Training Data

Training Algorithm
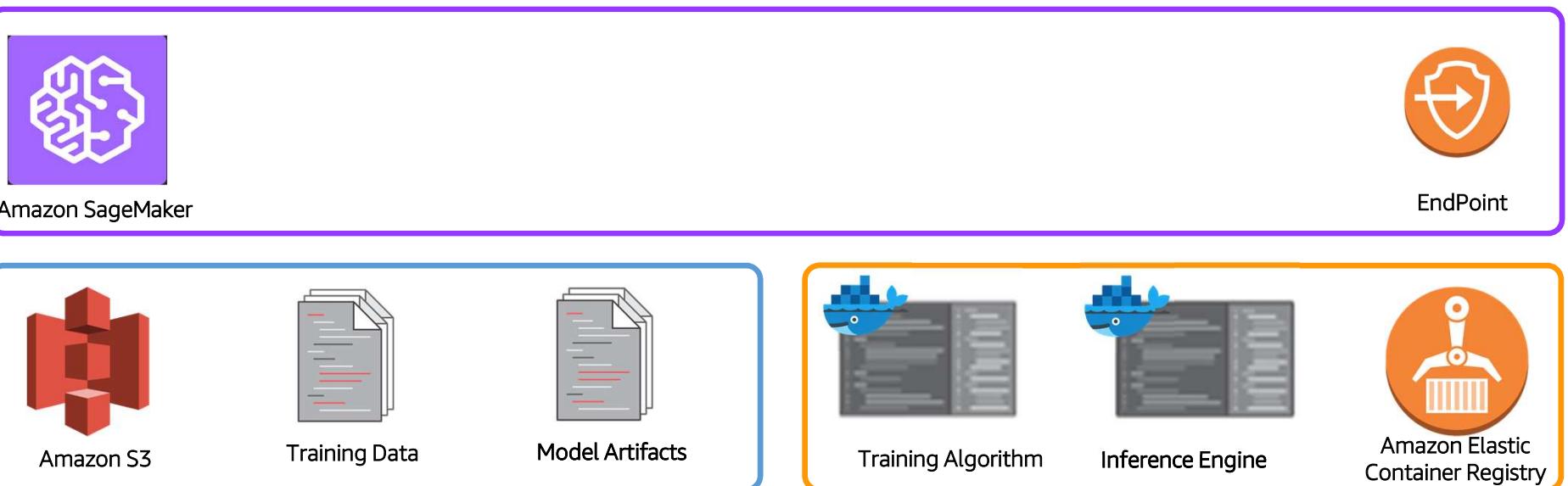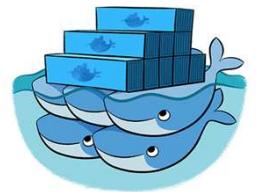
Inference Engine

Amazon Elastic Container Registry

A Fully-Dockerized Lifecycle
From discovery to development and deployment

Developers and Operations

Amazon SageMaker

EndPoint

Amazon S3

Training Data

Model Artifacts

Training Algorithm

Inference Engine

Amazon Elastic
Container Registry

A Fully-Dockerized Lifecycle
From discovery to development and deployment

Delighted Customers

API Gateway

Identification
Authorization
Logging
Analytics

Predictive Model

Amazon SageMaker
Launch Customers

Intuit

Digital Globe

ZipRecruiter

Hotels.com

Thomson Reuters

aws training and certification

INTUIT

"With Amazon SageMaker, we can accelerate our Artificial Intelligence initiatives at scale by building and deploying our algorithms on the platform. We will create novel large-scale machine learning and AI algorithms and deploy them on this platform to solve complex problems that can power prosperity for our customers."

Ashok Srivastava, Chief Data Officer, Intuit

Key Benefits of Amazon SageMaker at Intuit

From
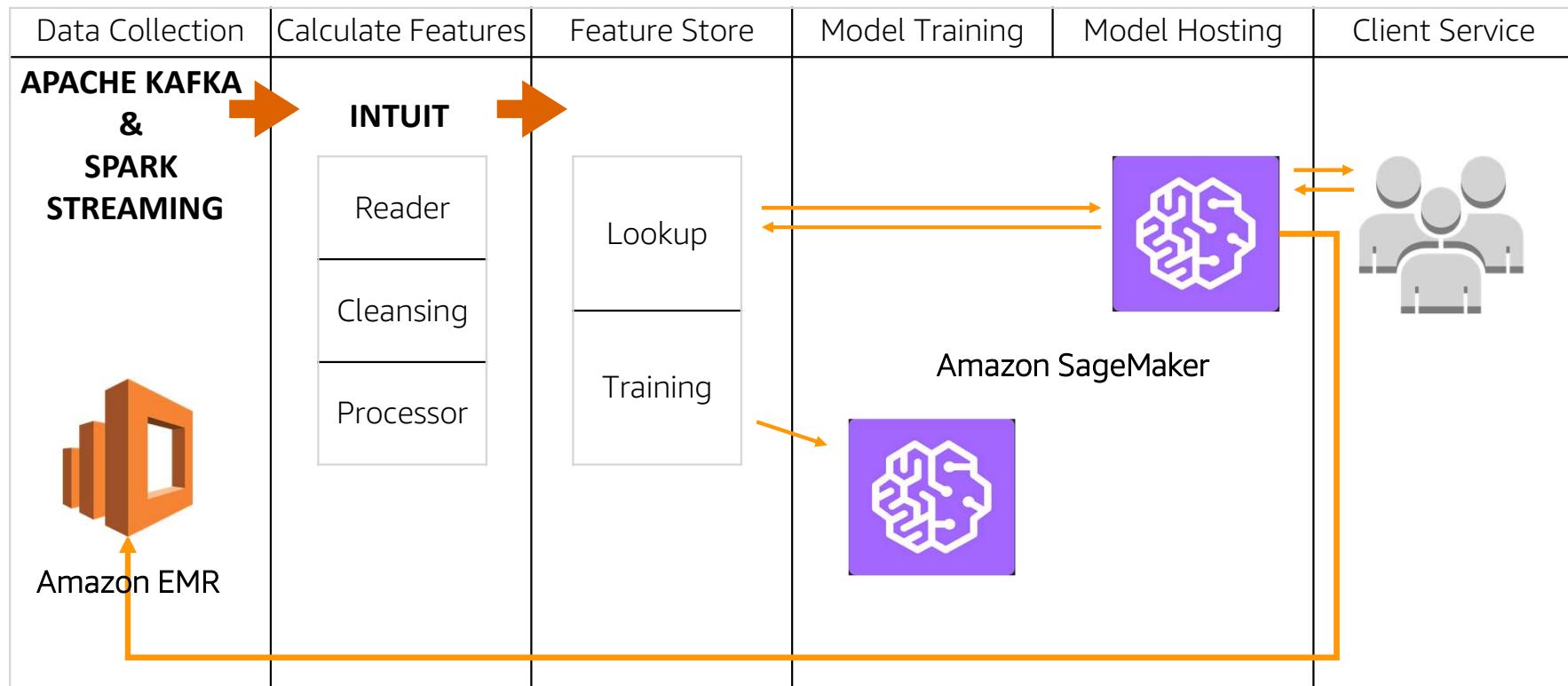
Ad-hoc setup and management of notebook environments

Limited choices for model deployment

Competing for compute resources across teams

To

Fraud Detection using SageMaker

Data Collection	Calculate Features	Feature Store	Model Training	Model Hosting	Client Service

APACHE KAFKA
&
SPARK
STREAMING

INTUIT

Reader

Cleansing

Processor

Lookup

Training

Amazon EMR

Amazon SageMaker

DigitalGlobe

"As the world's leading provider of high-resolution Earth imagery, data and analysis, DigitalGlobe works with enormous amounts of data every day. DigitalGlobe is making it easier for people to find, access, and run compute against our entire 100PB image library, which is stored in AWS's cloud, to apply deep learning to satellite imagery. We plan to use Amazon SageMaker to train models against petabytes of Earth observation imagery datasets using hosted Jupyter notebooks, so DigitalGlobe's Geospatial Big Data Platform (GBDX) users can just push a button, create a model, and deploy it all within one scalable distributed environment at scale."

Dr. Walter Scott, CTO of Maxar Technologies and founder of DigitalGlobe

ZipRecruiter

"*We're focused on making it faster and easier than ever to hire and get hired, training our machine learning algorithms against hundreds of millions of historical transactional activities in order to deliver highly relevant job matches as quickly as possible. Amazon SageMaker provided us with an answer to problems we had with ML workflow management, allowing us to train, evaluate and deploy models in a flexible way. In addition, Amazon SageMaker's modularity provides the ability to build and create models independently, which is a compelling feature for ZipRecruiter.*"
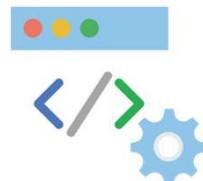
Avi Golan, VP of Engineering, ZipRecruiter

Amazon SageMaker's Components

Amazon SageMaker
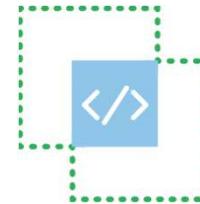
1 Notebook Instances

2 Algorithms

3 ML Training Service
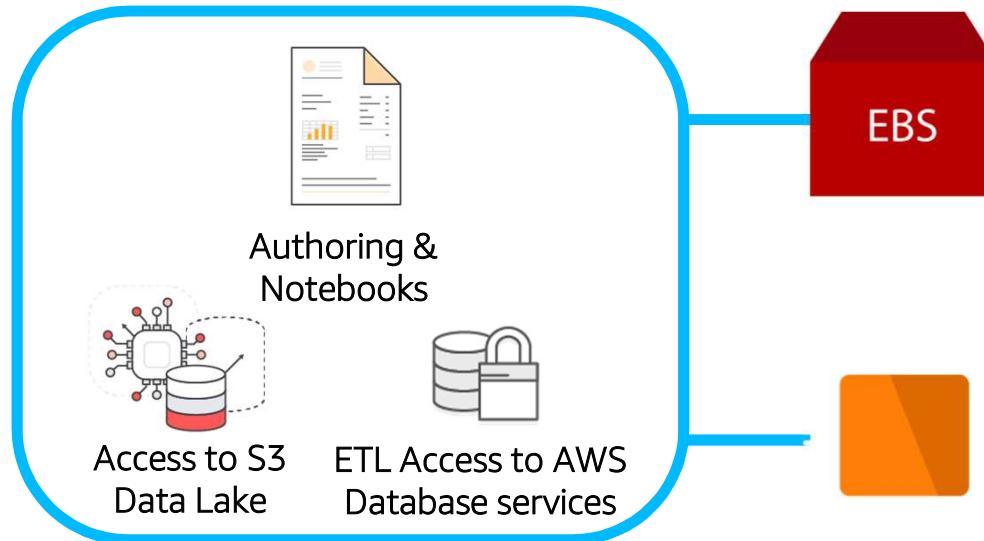
4 ML Hosting Service

SageMaker Notebook Instances
Zero Setup for Exploratory Data Analysis

Authoring & Notebooks

Access to S3 Data Lake

ETL Access to AWS Database services

EBS

Just add data!

- Recommendations/Personalization
- Fraud Detection
- Forecasting
- Image Classification
- Churn Prediction
- Marketing Email/Campaign Targeting
- Log processing and anomaly detection
- Speech to Text
- More…

Demo 4: A simple Jupyter Notebook

Pythagorean Theorem

aws training and certification
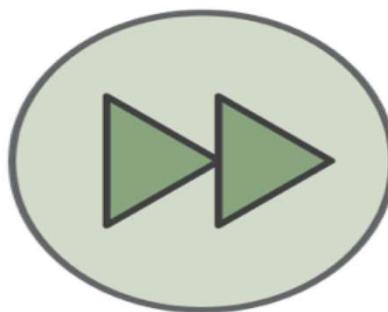
Demo 5: Predicting AWS Spot Pricing
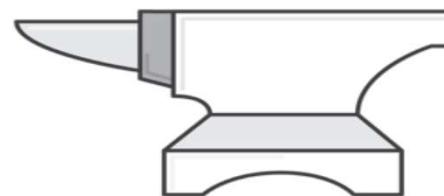
SageMaker Built-in Algorithms
10x Faster

Streaming datasets, for cheaper training

Train faster, in a single pass
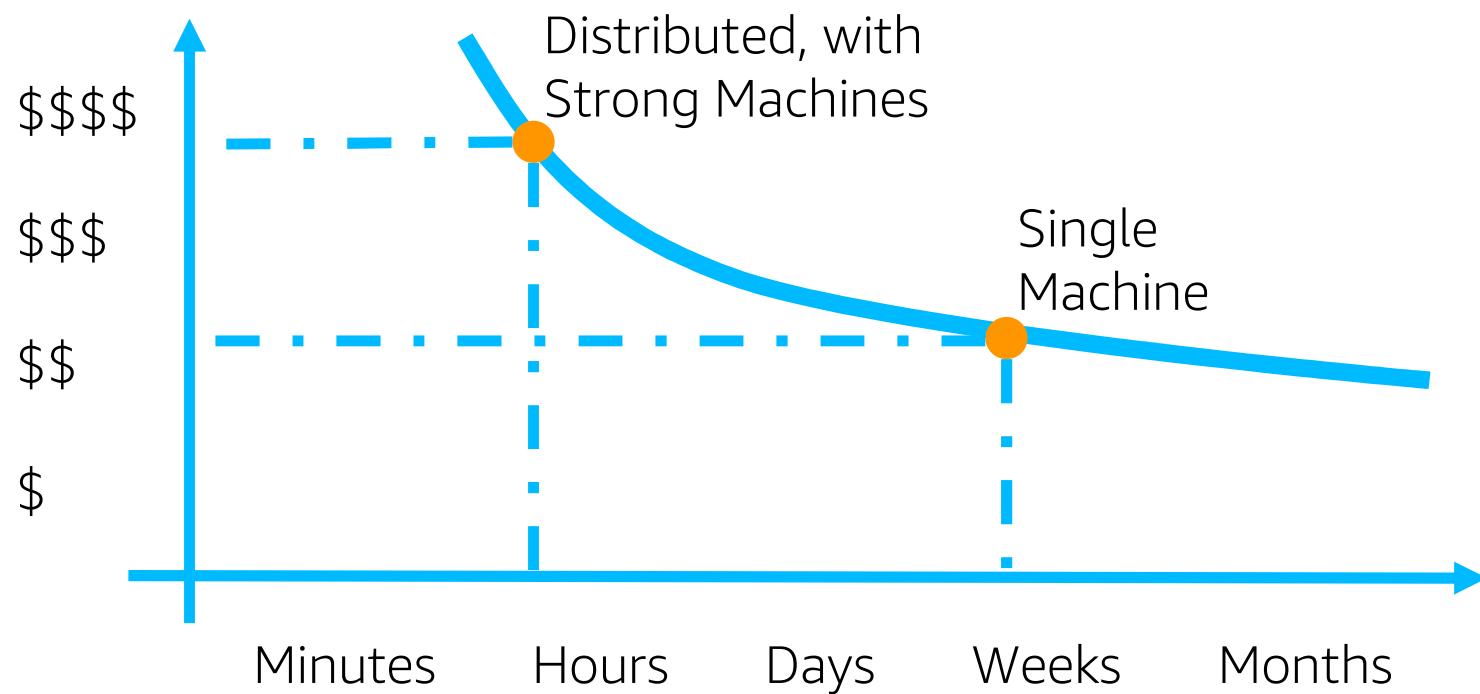
Greater reliability on extremely large datasets

Choice of several ML algorithms
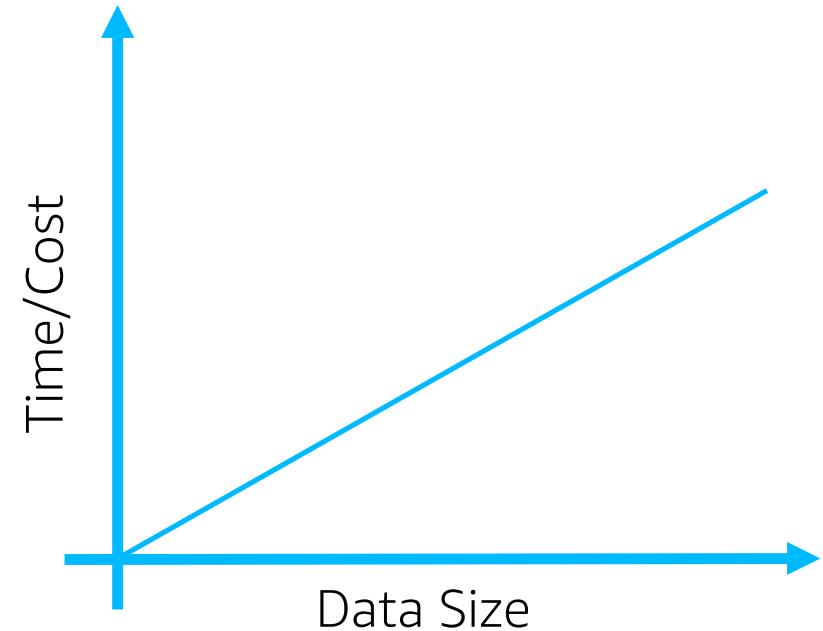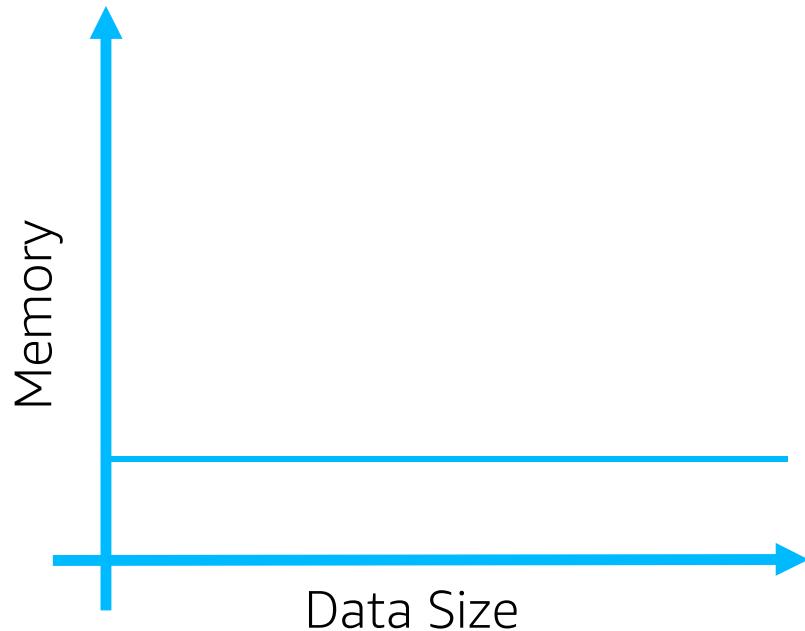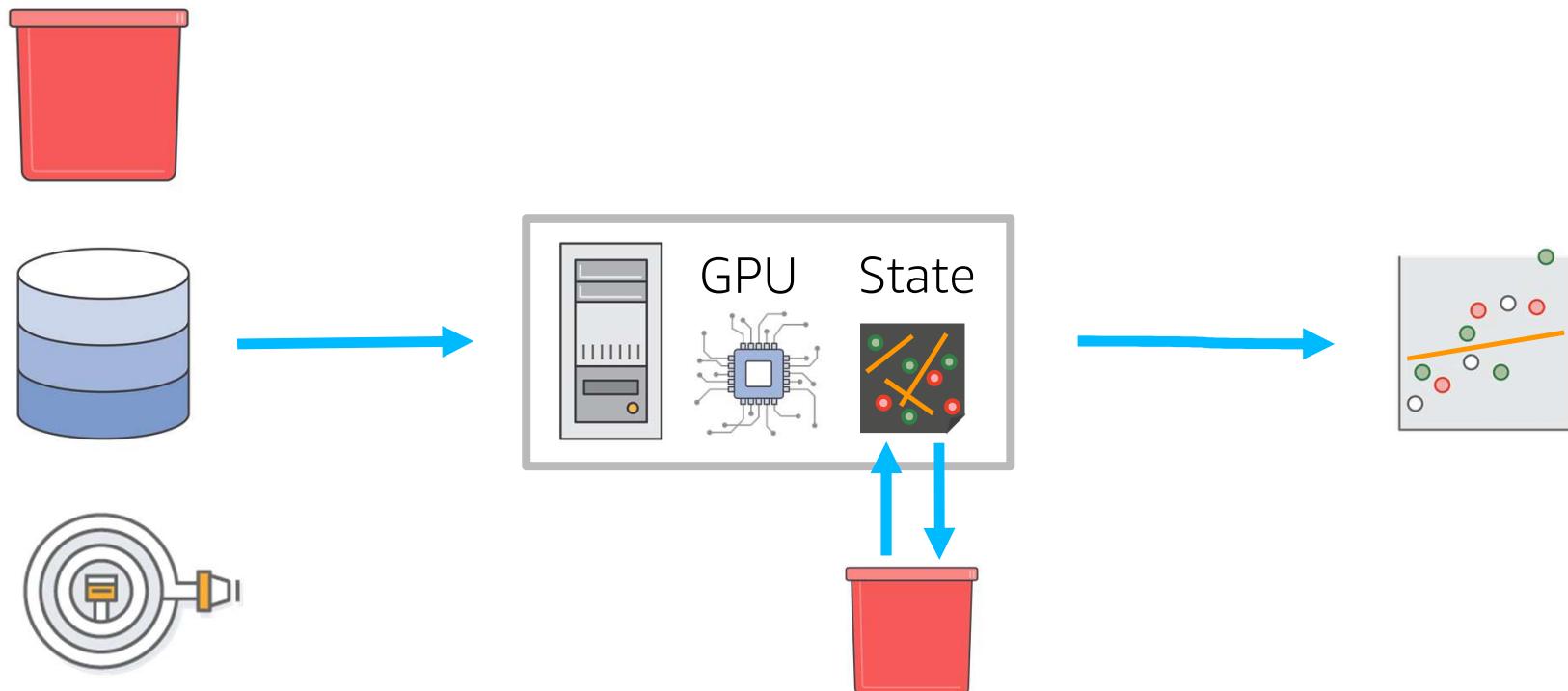
SageMaker Built-in Algorithms
Streaming
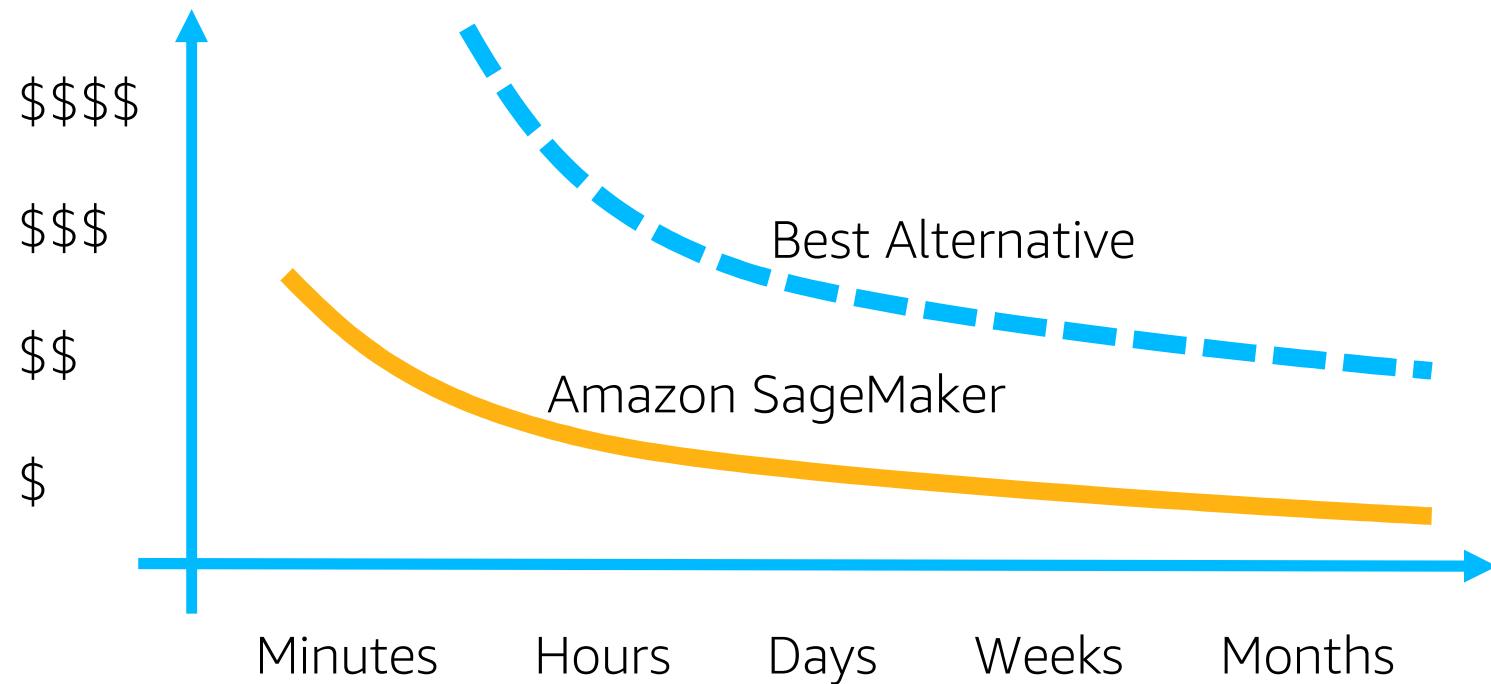
SageMaker Built-In Algorithms
Distributed Shared State

Infinitely Scalable ML Algorithms

Linear Learner
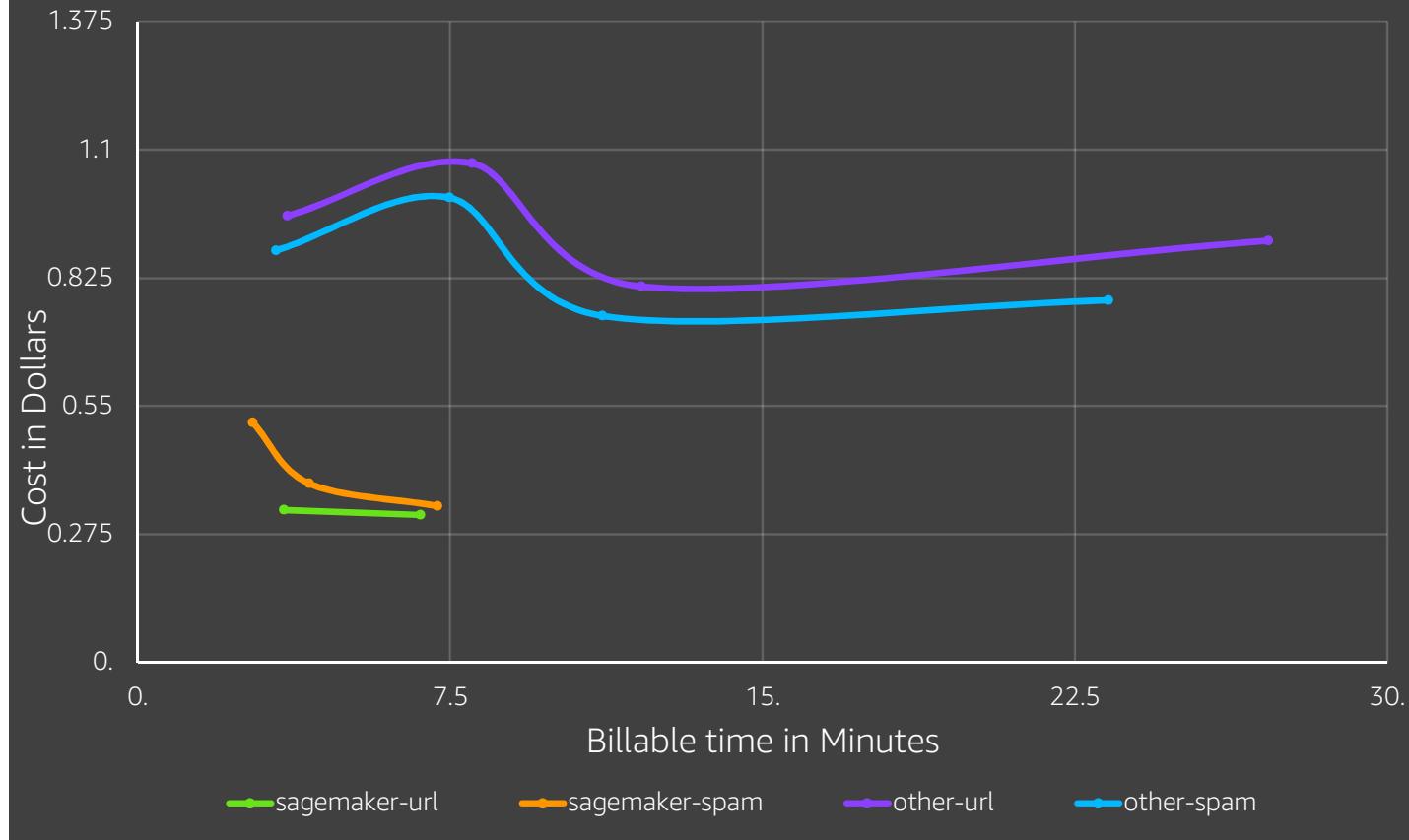
Regression (mean squared error)	
SageMaker	Other
1.02	1.06
1.09	1.02
0.332	0.183
0.086	0.129
83.3	84.5

Classification (F1 Score)	
SageMaker	Other
0.980	0.981
0.870	0.930
0.997	0.997
0.978	0.964
0.914	0.859
0.470	0.472
0.903	0.908
0.508	0.508

30 GB datasets for web-spam and web-url classification

Legend: sagemaker-url, sagemaker-spam, other-url, other-spam

Axes: Cost in Dollars (y-axis), Billable time in Minutes (x-axis)
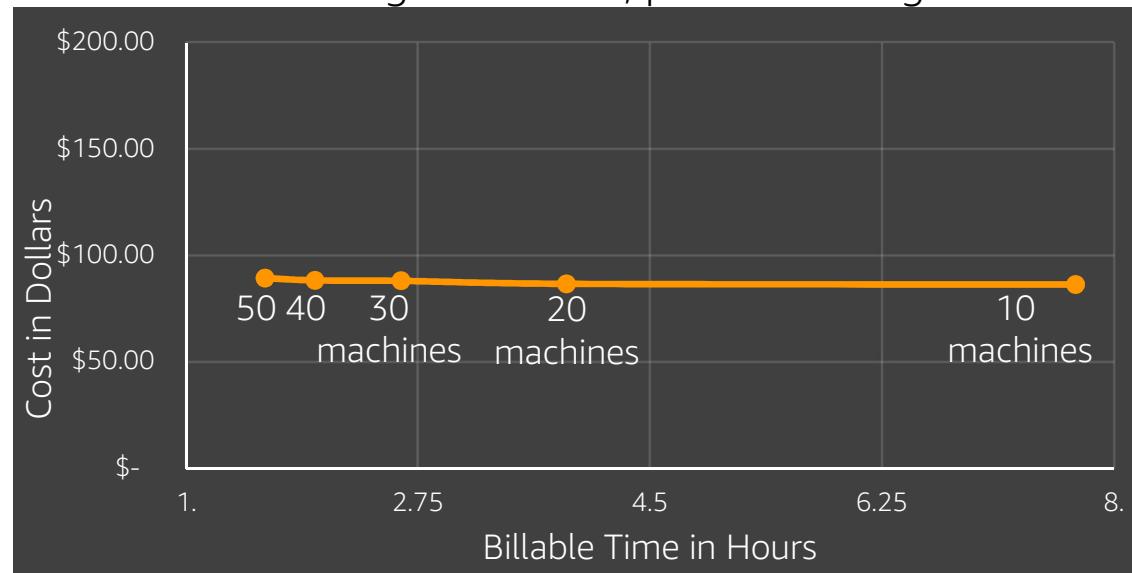
Factorization Machines

$$\tilde{y} = w_0 + \langle w_1, x \rangle + \sum_{i,j>i} x_i x_j \cdot \langle v_i, v_j \rangle$$

	Log_loss	F1 Score	Seconds
SageMaker	0.494	0.277	820
Other (10 Iter)	0.516	0.190	650
Other (20 Iter)	0.507	0.254	1300
Other (50 Iter)	0.481	0.313	3250

Click Prediction 1 TB advertising dataset, m4.4xlarge machines, perfect scaling.

K-Means Clustering
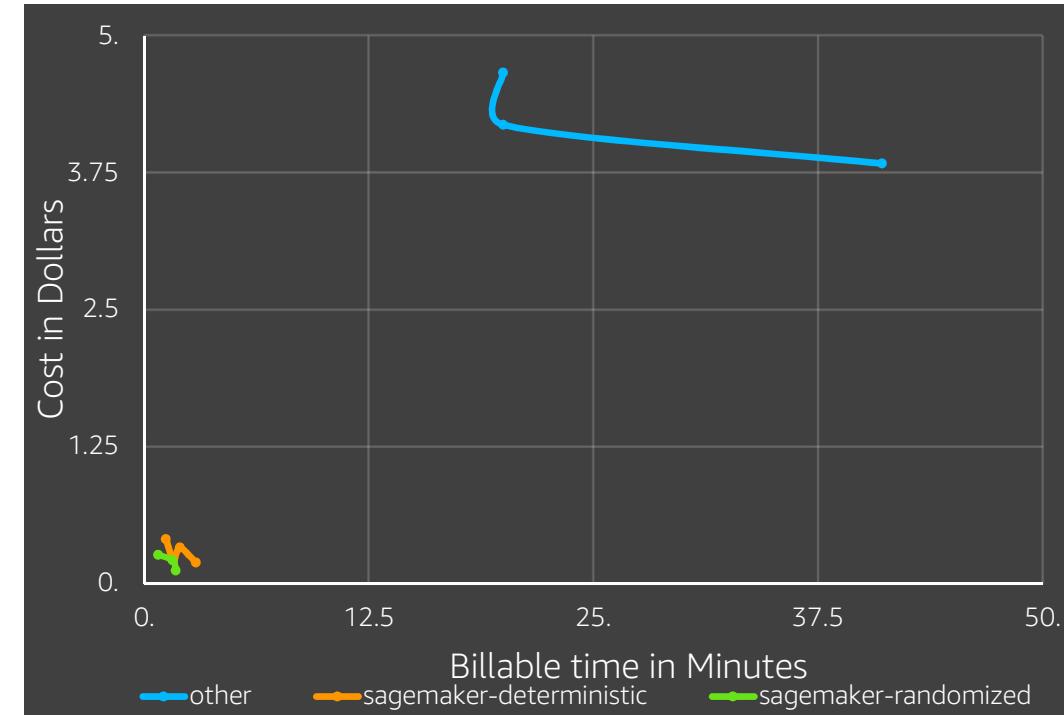
	k	SageMaker	Other
Text 1.2GB	10	1.18E3	1.18E3
	100	1.00E3	9.77E2
	500	9.18.E2	9.03E2
Images 9GB	10	3.29E2	3.28E2
	100	2.72E2	2.71E2
	500	2.17E2	Failed
Videos 27GB	10	2.19E2	2.18E2
	100	2.03E2	2.02E2
	500	1.86E2	1.85E2
Advertising 127GB	10	1.72E7	Failed
	100	1.30E7	Failed
	500	1.03E7	Failed
Synthetic 1100GB	10	3.81E7	Failed
	100	3.51E7	Failed
	500	2.81E7	Failed

Running Time vs. Number of Clusters

~10x Faster!

Billable Time in Minutes

Number of Clusters

sagemaker   other
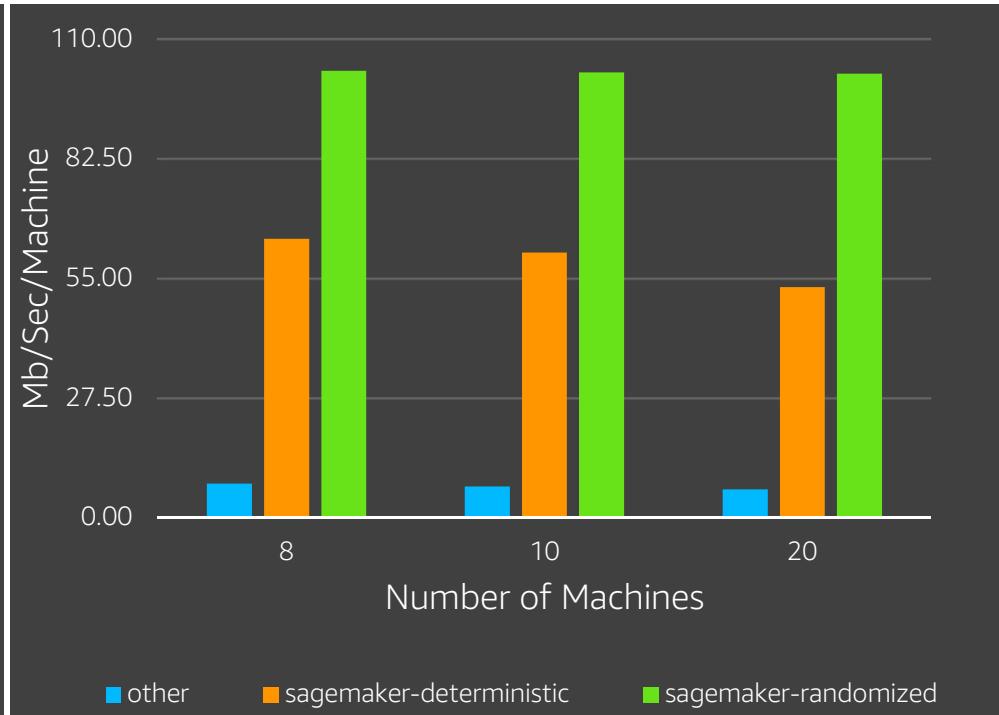
Principal Component Analysis (PCA)
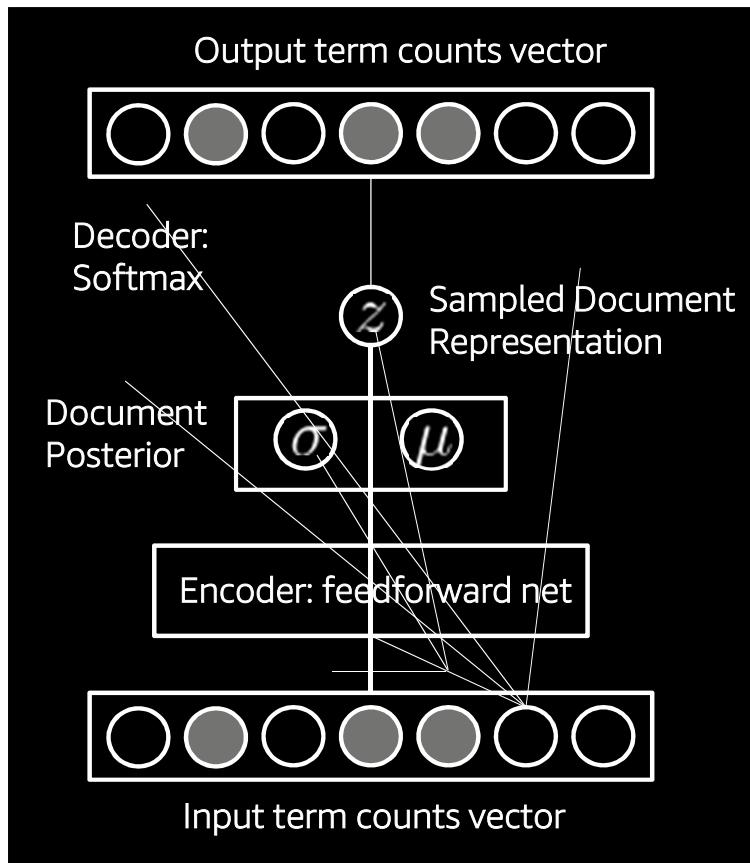
## Cost vs. Time
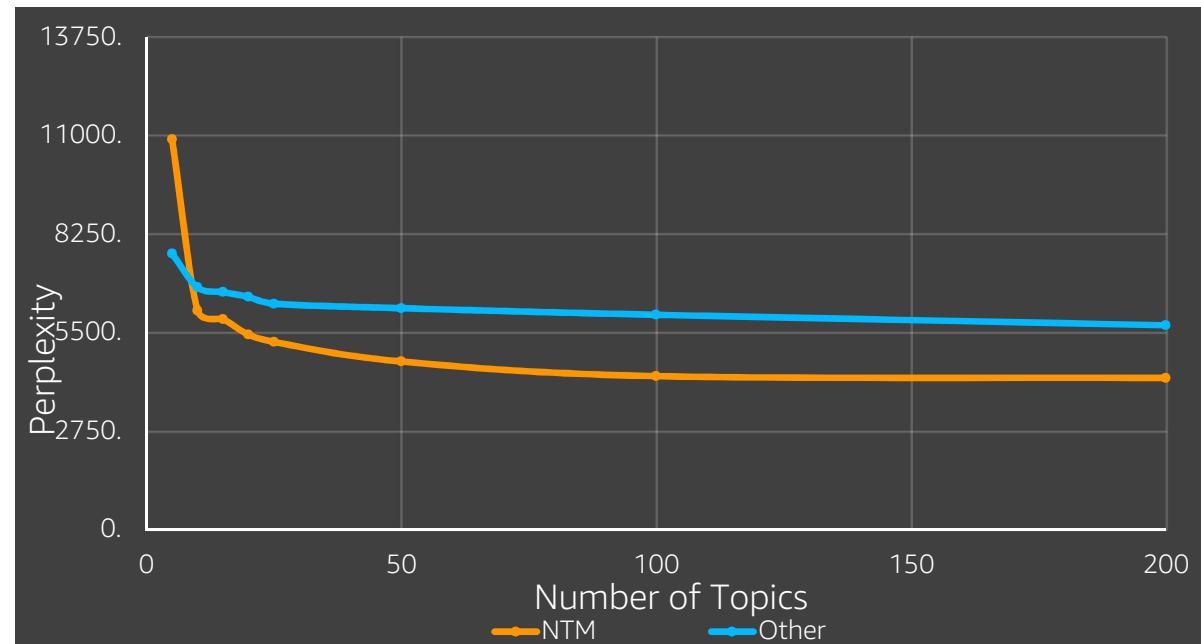
Throughput and Scalability

Neural Topic Modeling

Output term counts vector

Decoder: Softmax

Sampled Document Representation

Document Posterior

Encoder: feedforward net

Input term counts vector

z

σ μ
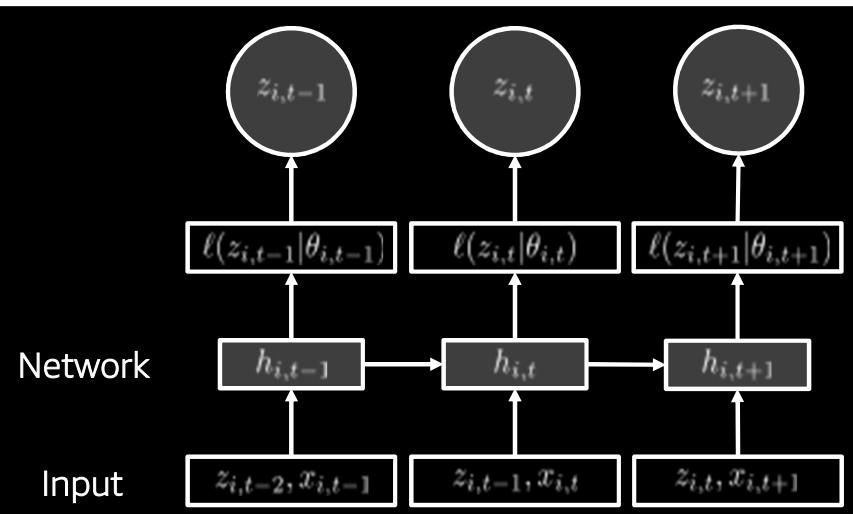
Perplexity vs. Number of Topic
(~200K documents, ~100K vocabulary)

Perplexity vs. Number of Topics

NTM Other

DeepAR

		Mean absolute percentage error		P90 Loss	
		DeepAR	R	DeepAR	R
Traffic Hourly occupancy rate of 963 bay area freeways		0.14	0.27	0.13	0.24
Electricity Electricity use of 370 homes over time		0.07	0.11	0.08	0.09
Page views Page view hits of websites	10k	0.32	0.32	0.44	0.31
	180k	0.32	0.34	0.29	NA

One hour on p2.xlarge, $1

More Great ML Algorithms

Spectral LDA

Training Time vs. Number of Topics

Boosted Decision Trees
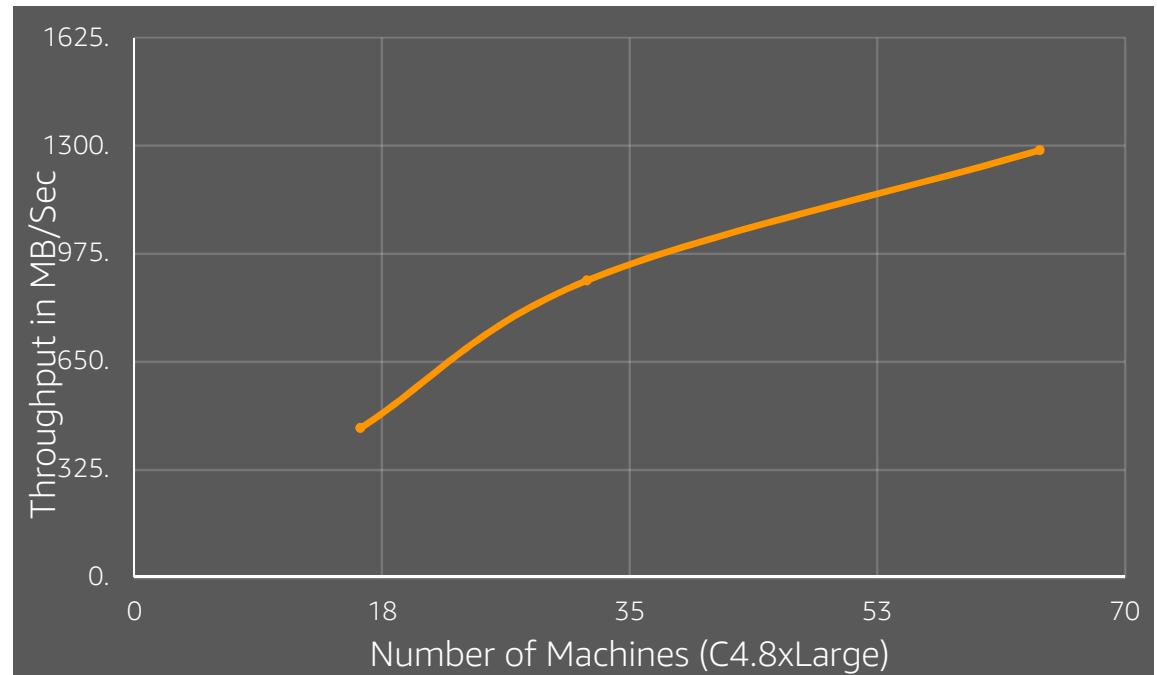
XGBoost is one of the most commonly used implementations of boosted decision trees in the world.

It is now available in Amazon SageMaker!

XGBoost

Throughput vs. Number of Machines

Sequence to Sequence

Based on Sockeye and Apache incubated MxNet, Multi-GPU, and can be used for Neural Machine Translation.

Supports both RNN/CNN as encoder/decoder

### English-German Translation

Best known result!
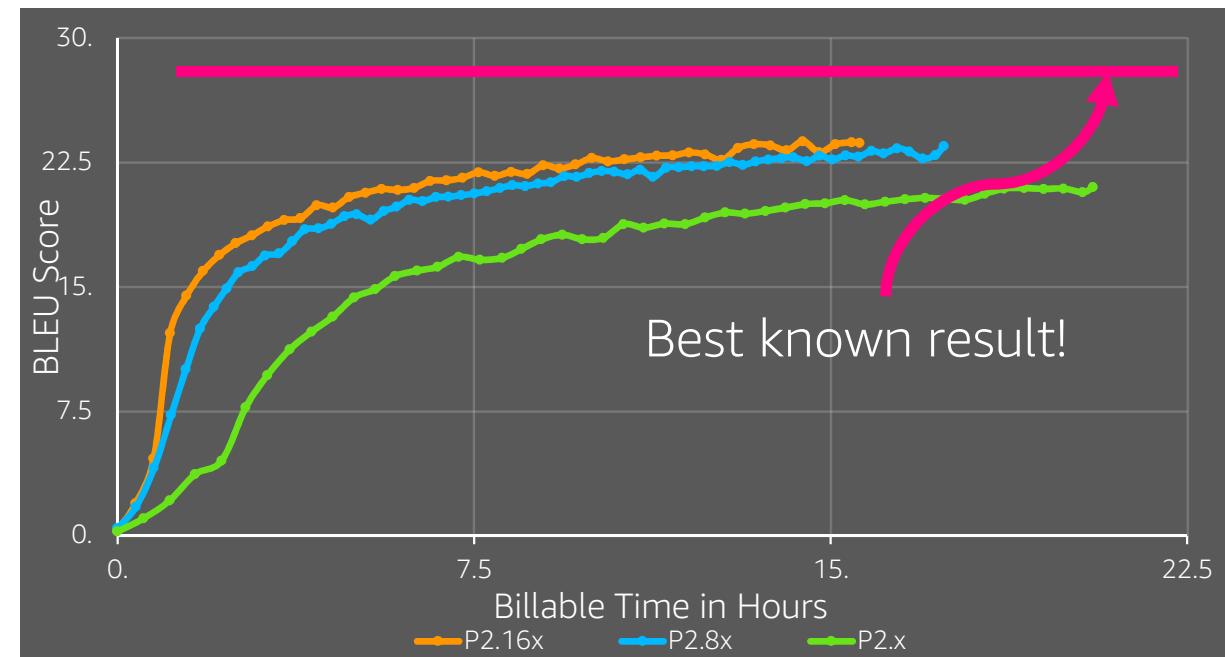
Image Classification
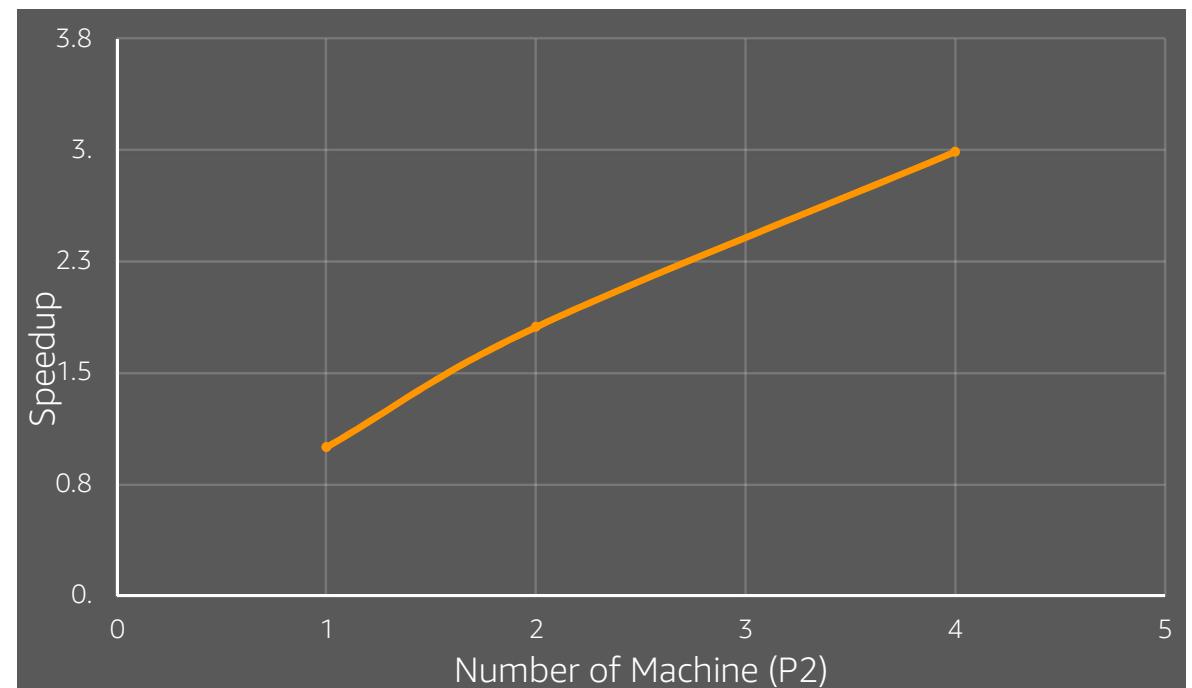
Implementation in MxNet of ResNet.
Other networks such as DenseNet and Inception will be added in the future.
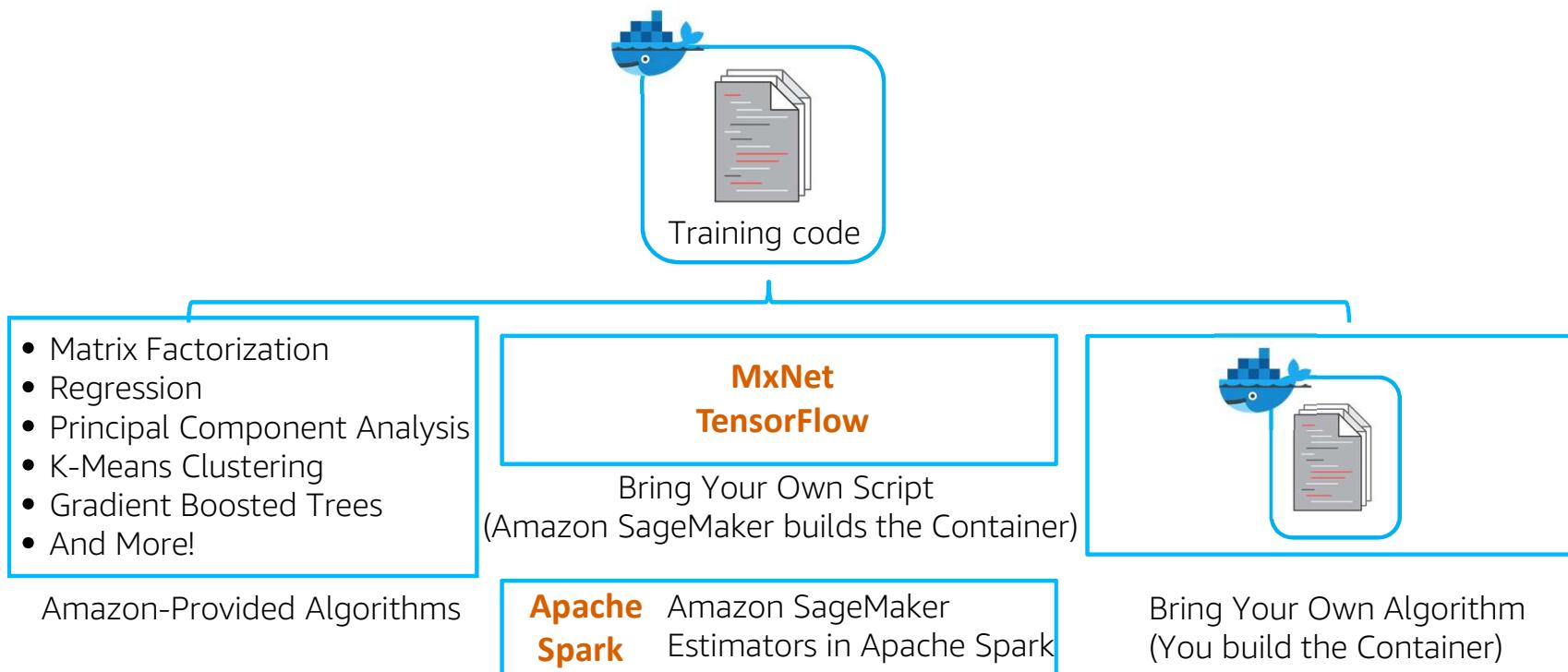
Transfer learning: begin with a model already trained on ImageNet!

Speedup with Horizontal Scaling
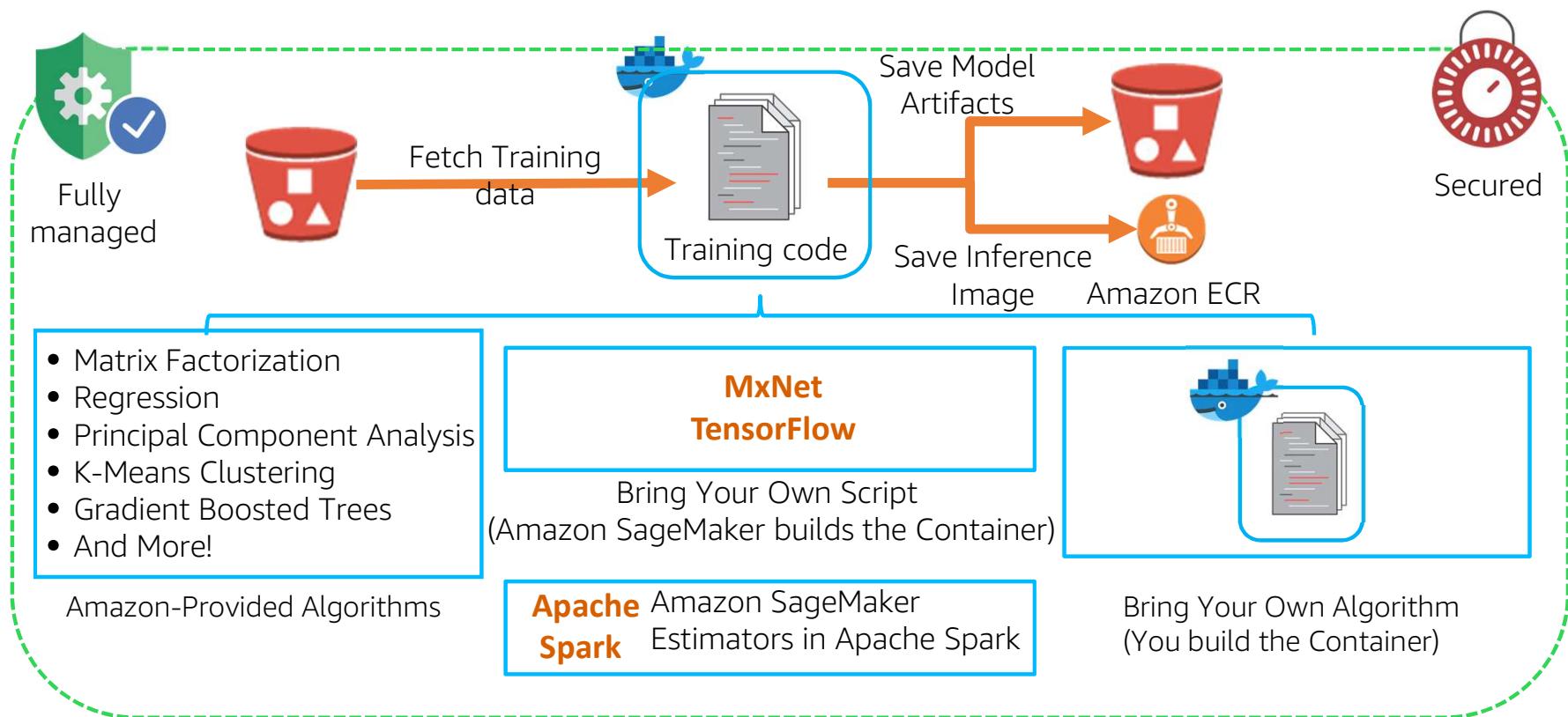
Amazon SageMaker Built-In Algorithms *10x Better*

Training code

- Matrix Factorization
- Regression
- Principal Component Analysis
- K-Means Clustering
- Gradient Boosted Trees
- And More!

Amazon-Provided Algorithms

MxNet
TensorFlow

Bring Your Own Script
(Amazon SageMaker builds the Container)

Apache Spark Amazon SageMaker Estimators in Apache Spark

Bring Your Own Algorithm
(You build the Container)

Amazon SageMaker Built-In Algorithms
Managed Distributed Training with Flexibility

Fully managed

Fetch Training data

Training code

Save Model Artifacts

Save Inference Image

Amazon ECR

Secured

- Matrix Factorization
- Regression
- Principal Component Analysis
- K-Means Clustering
- Gradient Boosted Trees
- And More!

Amazon-Provided Algorithms

MxNet
TensorFlow

Bring Your Own Script
(Amazon SageMaker builds the Container)

Apache Spark Amazon SageMaker Estimators in Apache Spark

Bring Your Own Algorithm
(You build the Container)

Demo 6: Using Amazon SageMaker Built-in Algorithms

Amazon SageMaker Hosting Service
Easy Model Deployment to Amazon SageMaker

aws training and certification
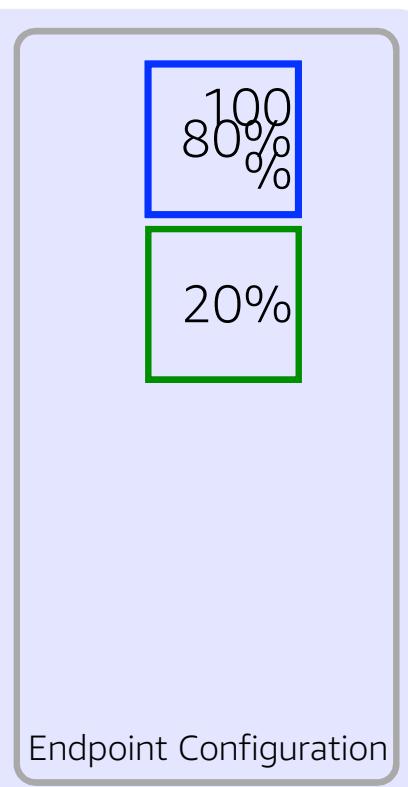
- InstanceType: c3.4xlarge
- InitialInstanceCount: 3
- ModelName: prod
- VariantName: primary
- InitialVariantWeight: 100

Inference EndPoint

Model Versions

Endpoint Configuration

100
80%

20%

Ground Truth   Amazon S3

Amazon ECR

SageMaker Hosting Service
## Easy Model Deployment to Amazon SageMaker

- Auto-Scaling Inference APIs
- A/B Testing (more to come)
- Low Latency & High Throughput
- Bring Your Own Model
- Python SDK
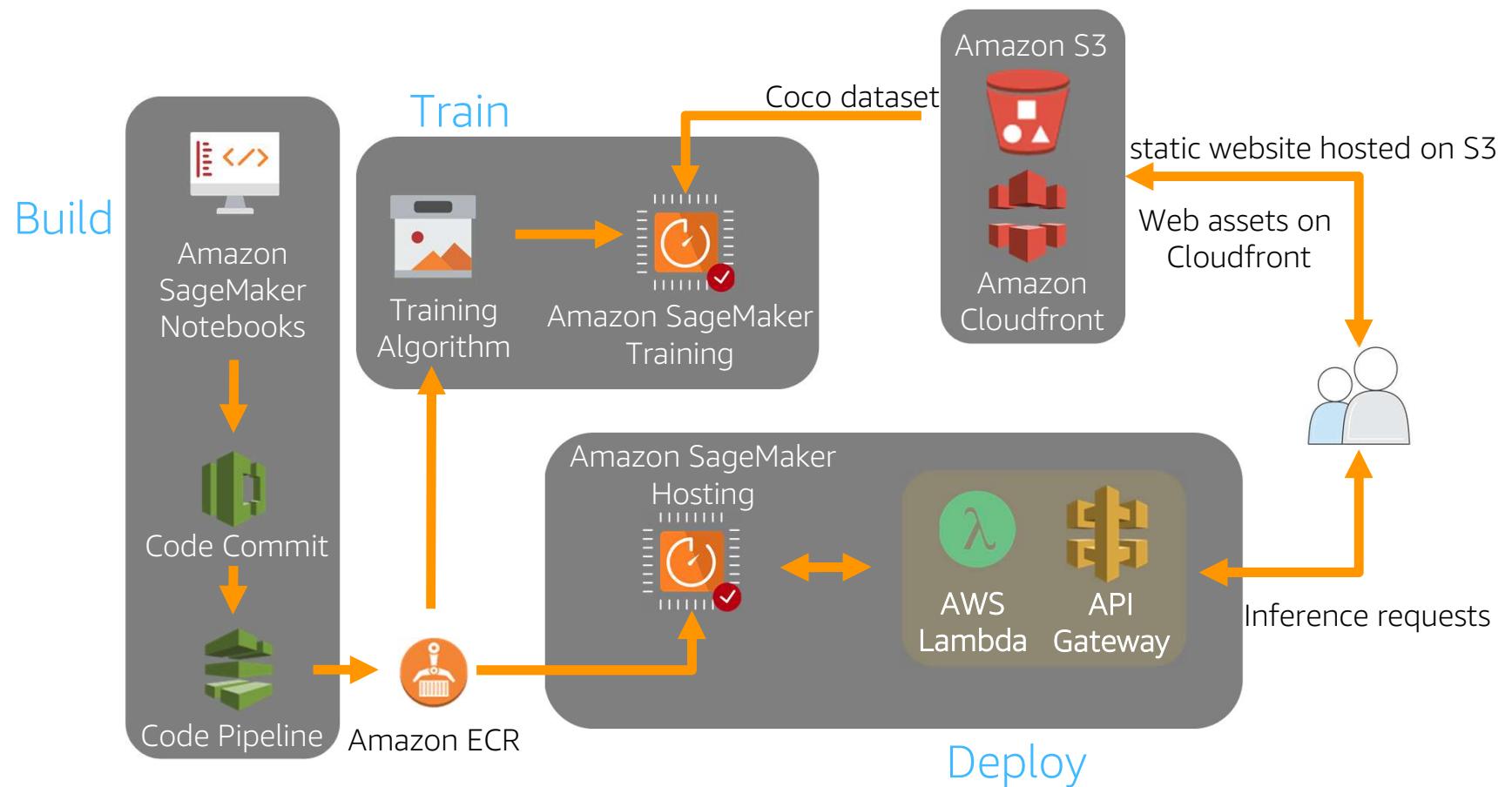
Demo 7: Analyzing Breast Cancer Datasets

Demo 8: Using Containers with Amazon SageMaker

Amazon SageMaker
Reference Architecture

aws training and certification

Build

Amazon SageMaker Notebooks

Code Commit

Code Pipeline

Amazon ECR

Train

Training Algorithm

Amazon SageMaker Training

Coco dataset

Amazon S3

Amazon Cloudfront

static website hosted on S3

Web assets on Cloudfront

Amazon SageMaker Hosting

AWS Lambda

API Gateway

Inference requests

Deploy

Amazon SageMaker
Technology Competency Partners

Data Services	Platform Solutions		SaaS and API Solutions	
Alteryx	Bonsai	DataRobot	Anodot	SigOpt
CrowdFlower	C3 IoT	DOMINO DATA LAB	Luminoso	Veritone
Paxata	Databricks	H2O.ai	Narrative Science	x.ai
TRIFACTA	Data Iku			

Call To Action

- Getting started with Amazon SageMaker: https://aws.amazon.com/sagemaker/

- Use the Amazon SageMaker SDK:
  - For Python: https://github.com/aws/sagemaker-python-sdk
  - For Spark: https://github.com/aws/sagemaker-spark

- SageMaker Examples: https://github.com/awslabs/amazon-sagemaker-examples

Thank You

aws training and certification