ML Immersion Day

Mat Rowlands, Solutions Architect, AWS

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

Today's Agenda

09.00 – 09.30	Welcome & Registration
09.30 – 10.00	Intro to AI / ML on AWS
10.00 – 11.15	Hands on with AI services
11.15 – 11.30	Coffee Break
11.30 – 12.00	Intro to Amazon SageMaker
12.00 – 13.00	Lunch
13.00 – 14.00	Predictions with SageMaker workshop
14.00 – 15.00	Image recognition on SageMaker workshop
15.00 – 15.15	Tea Break!
15.15 – 16.15	TensorFlow on SageMaer workshop
16.15 – 16.30	Wrap Up, Q&A and survey

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

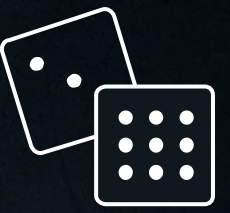
Centerpiece for digital transformation

Customer experience

Business operations

Decision making

Innovation



Competitive advantage

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

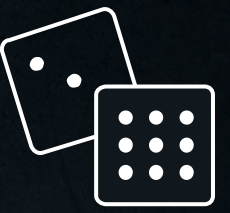
Centerpiece for digital transformation

Customer experience

Business operations

Decision making

Innovation



Competitive advantage

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

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Our mission at AWS

Put machine learning in the hands
of every developer

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

More machine learning happens on AWS than anywhere else

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Why AWS for AI?

Broadest and deepest set of AI and ML services

200 new features & services launched this last year alone

Unmatched flexibility

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Why AWS for AI?

Broadest and deepest set of AI and ML services

Accelerate your adoption of ML with SageMaker

200 new features & services launched this last year alone

Unmatched flexibility

70% cost reduction in data-labeling

10x faster performance

75% lower inference cost

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Why AWS for AI?

Broadest and deepest set of AI and ML services

200 new features & services launched this last year alone

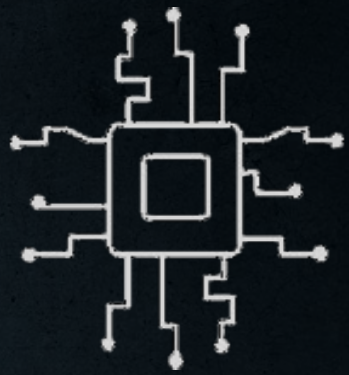
Unmatched flexibility

Accelerate your adoption of ML with SageMaker

70% cost reduction in data-labeling

10x faster performance

75% lower inference cost



Built on the most comprehensive cloud platform optimized for ML

AWS holds the top spots on Stanford's benchmark, for fastest training time, lowest cost, lowest inference latency

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

The Amazon ML stack: Broadest & deepest set of capabilities

AI SERVICES

Vision


REKOGNITION IMAGE


REKOGNITION VIDEO

Speech


TEXTRACT


POLLY


TRANSCRIBE

Language


TRANSLATE


COMPREHEND

Chatbots


LEX

Forecasting


FORECAST

Recommendations


PERSONALIZE

ML SERVICES


AMAZON SAGEMAKER

BUILD

Pre-built algorithms & notebooks

Data labelling (GROUND TRUTH)

Algorithms & models (AWS MARKETPLACE FOR MACHINE LEARNING)

TRAIN

One-click model training & tuning

Optimization (NEO)

Reinforcement learning

DEPLOY

One-click deployment & hosting

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

TensorFlow

mxnet

PYTORCH

Interfaces

GLUON

Keras

Infrastructure


EC2 P3 & P3N


EC2 C5


FPGAs


GREENGRASS


ELASTIC INFERENCE

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

The Amazon ML stack: Broadest & deepest set of capabilities

AI SERVICES

Vision


REKOGNITION IMAGE


REKOGNITION VIDEO

Speech


TEXTRACT


POLLY


TRANSCRIBE

Language


TRANSLATE


COMPREHEND

Chatbots


LEX

Forecasting


FORECAST

Recommendations


PERSONALIZE

ML SERVICES


AMAZON SAGEMAKER

BUILD

Pre-built algorithms & notebooks

Data labelling (GROUND TRUTH)

Algorithms & models (AWS MARKETPLACE FOR MACHINE LEARNING)

TRAIN

One-click model training & tuning

Optimization (NEO)

Reinforcement learning

DEPLOY

One-click deployment & hosting

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

TensorFlow

mxnet

PYTORCH

Interfaces

GLUON

Keras

Infrastructure


EC2 P3 & P3N


EC2 C5


FPGAs


GREENGRASS


ELASTIC INFERENCE

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

The Amazon ML stack: Broadest & deepest set of capabilities

AI SERVICES

Vision | Speech | Language | Chatbots | Forecasting | Recommendations


REKOGNITION IMAGE


REKOGNITION VIDEO


TEXTRACT


POLLY


TRANSCRIBE


TRANSLATE


COMPREHEND


LEX


FORECAST


PERSONALIZE

ML SERVICES


AMAZON SAGEMAKER

BUILD
- Pre-built algorithms & notebooks
- Data labelling (GROUND TRUTH)
- Algorithms & models (AWS MARKETPLACE FOR MACHINE LEARNING)

TRAIN
- One-click model training & tuning
- Optimization (NEO)
- Reinforcement learning

DEPLOY
- One-click deployment & hosting

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks | Interfaces | Infrastructure

TensorFlow
mxnet
PYTORCH
GLUON
Keras


EC2 P3 & P3N


EC2 C5


FPGAs


GREENGRASS


ELASTIC INFERENCE

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

So, what's in it for me?

	Data Scientist	Data Engineer	Developer	Decision Maker
AI SERVICES	Rapid Prototype Feature testing Focus on more complex tasks No infrastructure No deployment	Quickly iterate on data format and features Focus on more complex tasks No infrastructure No deployment	Quick integration Self serve No Data Science background No infrastructure No deployment	Solve business problems and add rich features quickly No Data Science background No infrastructure No deployment
ML SERVICES	Optimized Frameworks Unified environment Jupyter notebooks Full control Modular architecture	Unified deployment Integrates with IAM, VPC Elastic scaling Labeling Performance optimized	Easy integration Example notebooks A/B testing models	Optimize for a domain specific problem Optimize for performance
ML FRAMEWORKS & INFRASTRUCTURE	Choose optimal hardware Target IoT devices Full control	Choose optimal hardware	Endpoint data structures	Solving business differentiating problems

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

AI Services

AI SERVICES


REKOGNITION IMAGE

Vision


REKOGNITION VIDEO


TEXTRACT

Speech


POLLY


TRANSCRIBE

Language


TRANSLATE


COMPREHEND

Chatbots


LEX

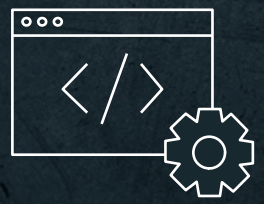
Forecasting


FORECAST

Recommendations


PERSONALIZE

Pre-trained AI services that require no ML skills or training



Easily add intelligence to your existing apps and workflows

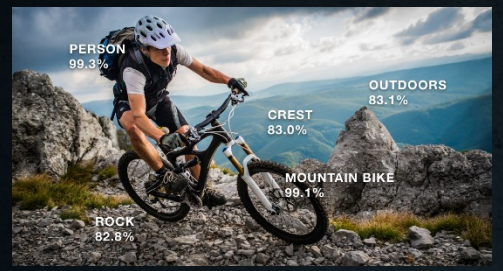
Quality and accuracy from continuously-learning APIs

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

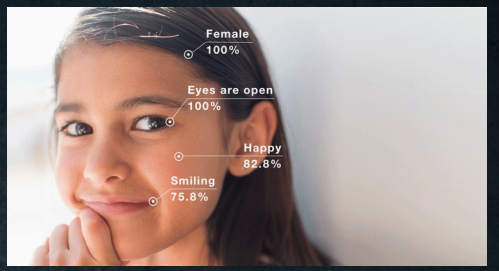
Amazon Rekognition

Deep learning-based image and video analysis



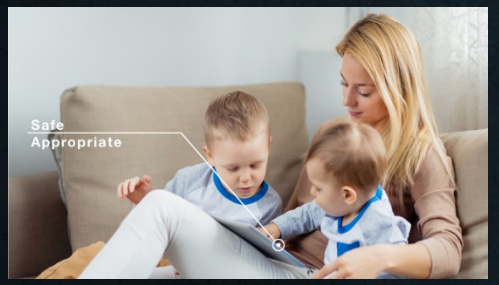
Object, Scene & Activity Recognition

Facial Recognition

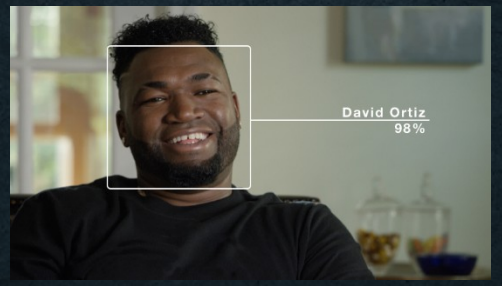


Facial Analysis

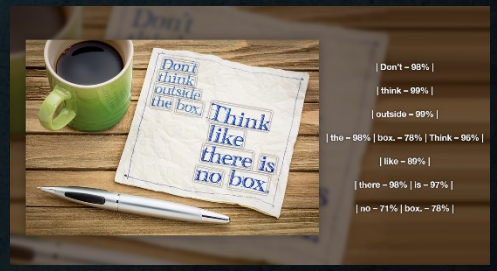
Person Tracking



Unsafe Content Detection



Celebrity Recognition

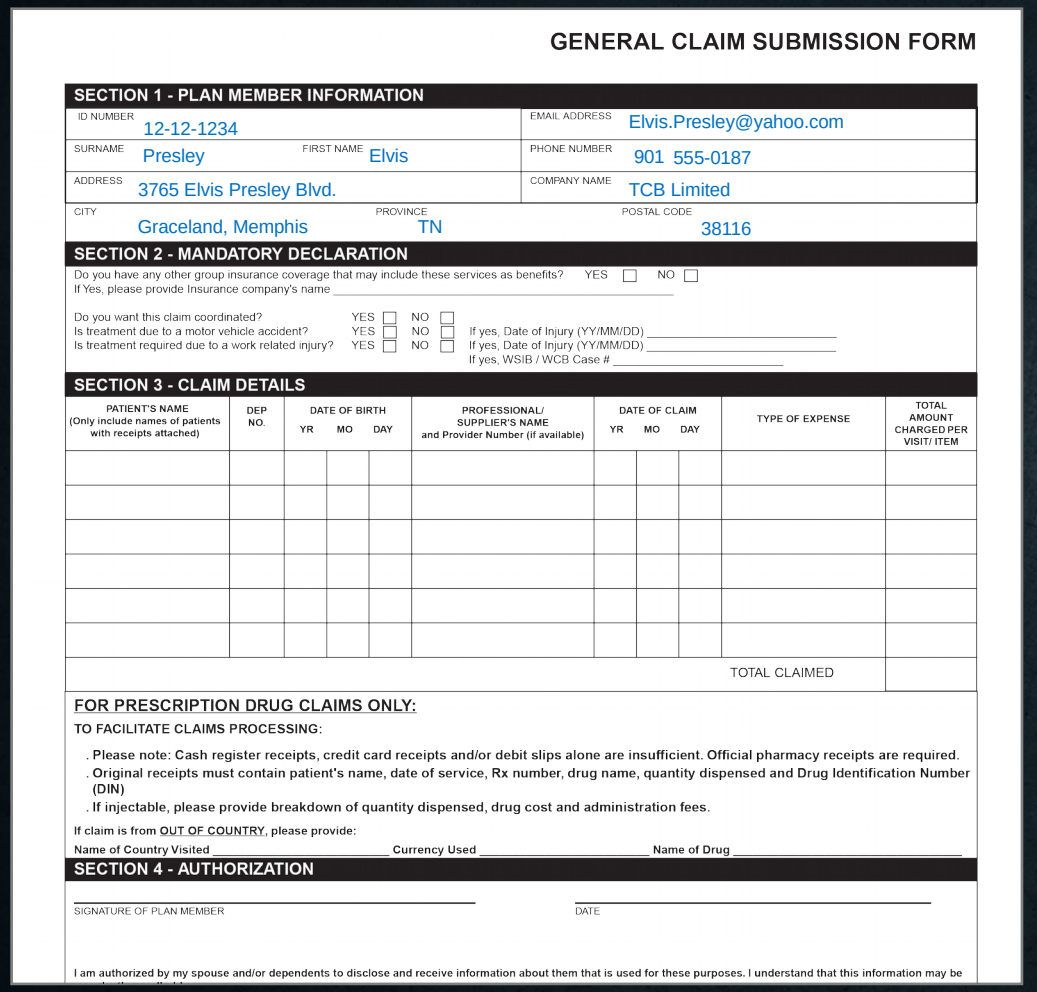


Text in Images

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Textract

Automatic document processing without data entry, or writing rules



GENERAL CLAIM SUBMISSION FORM

SECTION 1 - PLAN MEMBER INFORMATION

ID NUMBER	12-12-1234	EMAIL ADDRESS	Elvis.Presley@yahoo.com
SURNAME	Presley	FIRST NAME Elvis	PHONE NUMBER 901 555-0187
ADDRESS	3765 Elvis Presley Blvd.		COMPANY NAME TCB Limited
CITY	Graceland, Memphis	PROVINCE TN	POSTAL CODE 38116

SECTION 2 - MANDATORY DECLARATION

Do you have any other group insurance coverage that may include these services as benefits? YES ☐ NO ☐
If Yes, please provide Insurance company's name _____

Do you want this claim coordinated? YES ☐ NO ☐
Is treatment due to a motor vehicle accident? YES ☐ NO ☐ If yes, Date of Injury (YY/MM/DD) _____
Is treatment required due to a work related injury? YES ☐ NO ☐ If yes, Date of Injury (YY/MM/DD) _____
If yes, WSIB / WCB Case # _____

SECTION 3 - CLAIM DETAILS

PATIENT'S NAME (Only include names of patients with receipts attached)	DEP NO.	DATE OF BIRTH YR	MO	DAY	PROFESSIONAL/ SUPPLIER'S NAME and Provider Number (if available)	DATE OF CLAIM YR	MO	DAY	TYPE OF EXPENSE	TOTAL AMOUNT CHARGED PER VISIT/ ITEM
									TOTAL CLAIMED	

FOR PRESCRIPTION DRUG CLAIMS ONLY:

TO FACILITATE CLAIMS PROCESSING:

. Please note: Cash register receipts, credit card receipts and/or debit slips alone are insufficient. Official pharmacy receipts are required.
. Original receipts must contain patient's name, date of service, Rx number, drug name, quantity dispensed and Drug Identification Number (DIN)
. If injectable, please provide breakdown of quantity dispensed, drug cost and administration fees.

If claim is from OUT OF COUNTRY, please provide:
Name of Country Visited _____ Currency Used _____ Name of Drug _____

SECTION 4 - AUTHORIZATION

_____ SIGNATURE OF PLAN MEMBER

_____ DATE

I am authorized by my spouse and/or dependents to disclose and receive information about them that is used for these purposes. I understand that this information may be

OCR++

Name	Presley, Elvis Aaron
Address	Graceland, Memphis, TN
Id	12-12-1234
Company	TCB Limited

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Polly

Turn text into lifelike speech using deep learning

Wide Selection of Voices and Languages

Synchronize Speech

Fine-grained Control

Unlimited Replay

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

aws

Amazon Transcribe

Automatic conversion of speech into accurate, grammatically correct text

Multiple languages

Intelligent punctuation and formatting

Timestamp generation

Support for telephony audio

Recognize multiple speakers

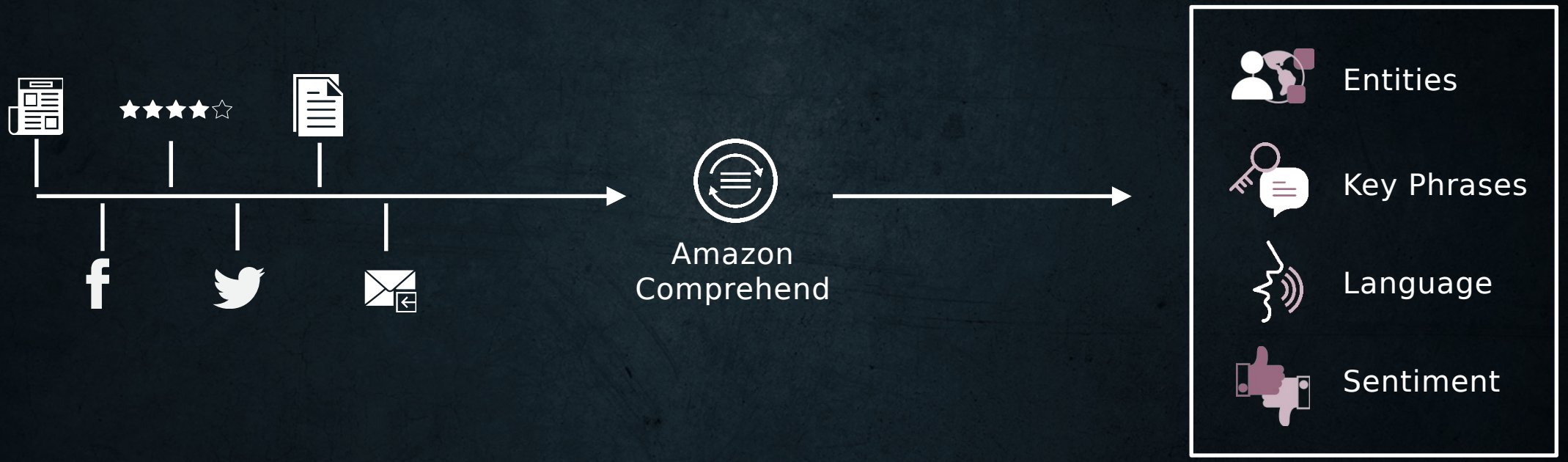
Custom vocabulary

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

aws

Amazon Comprehend

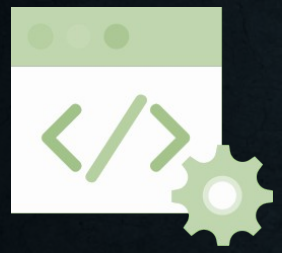
Discover insights and relationships in text



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Lex

Conversational interfaces for your applications powered by the same deep learning technologies as Alexa

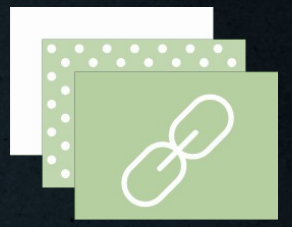


Integrated development in the AWS console

Trigger Lambda functions

Multi-step conversations

One-click deployment

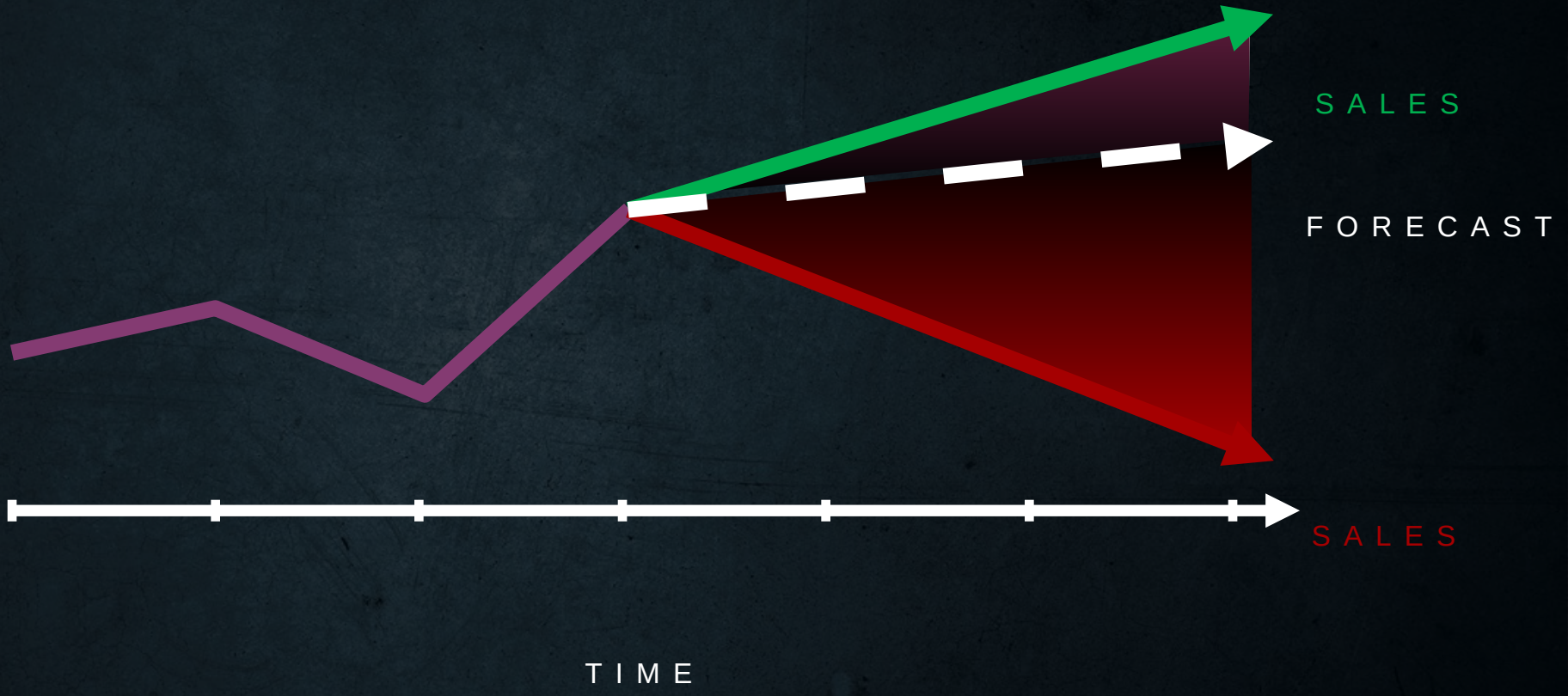


Enterprise connectors

Fully managed

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

The perils of poor predictions in forecasting



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Forecast

Accurate time-series forecasting service, based on the same technology used at Amazon.com

Any historical time-series

Integrates with SAP and Oracle Supply Chain

Integrates with Amazon Timestream

Custom forecasts with 3 clicks

50% more accurate

1/10th the cost

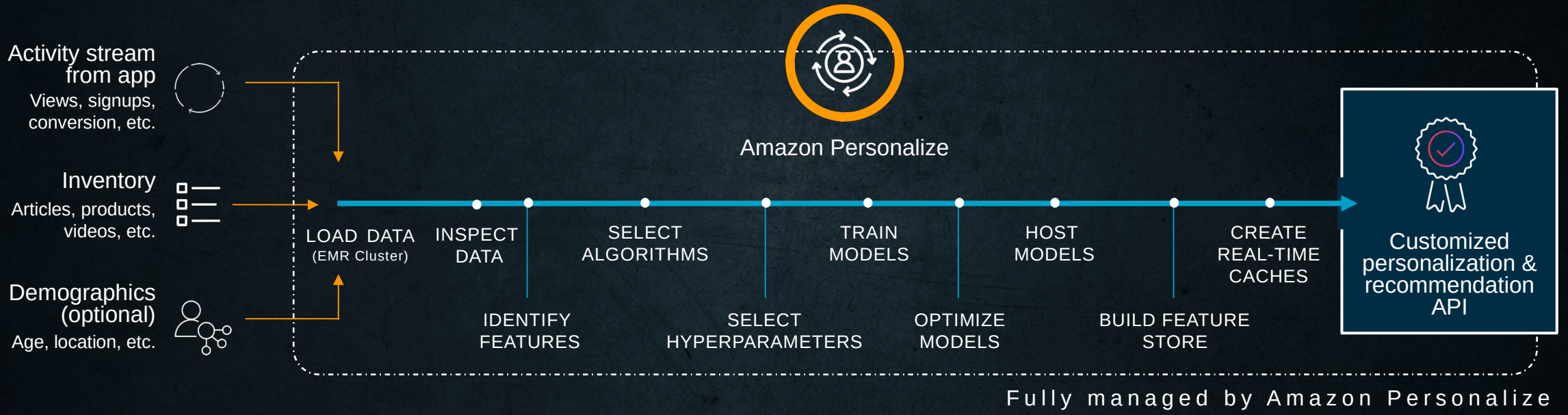
Generate forecasts for:

Retail demand	Travel demand	AWS usage
Revenue forecasts	Web traffic	Advertising demand

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Amazon Personalize

Real-time personalization and recommendation service, based on the same technology used at Amazon.com



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Enough of me chatting!

Let’s build a chat bot
https://bit.ly/30rxCeN

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

aws

The Amazon ML stack: Broadest & deepest set of capabilities

AI SERVICES

Vision


REKOGNITION IMAGE


REKOGNITION VIDEO

Speech


TEXTRACT


POLLY


TRANSCRIBE

Language


TRANSLATE


COMPREHEND

Chatbots


LEX

Forecasting


FORECAST

Recommendations


PERSONALIZE

ML SERVICES


AMAZON SAGEMAKER

BUILD

Pre-built algorithms & notebooks

Data labelling (GROUND TRUTH)

Algorithms & models (AWS MARKETPLACE FOR MACHINE LEARNING)

TRAIN

One-click model training & tuning

Optimization (NEO)

Reinforcement learning

DEPLOY

One-click deployment & hosting

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

TensorFlow

mxnet

PYTORCH

Interfaces

GLUON

Keras

Infrastructure


EC2 P3 & P3N


EC2 C5


FPGAs


GREENGRASS


ELASTIC INFERENCE

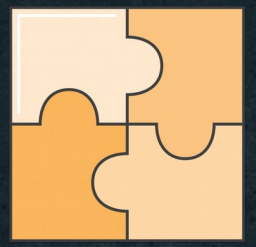
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Solving Some Of The Hardest Problems In Computer Science

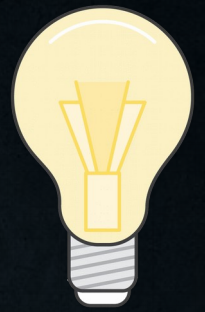
Learning

Language



Perception

Problem Solving



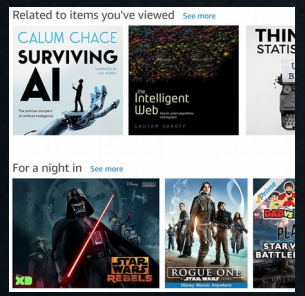
Reasoning

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Machine Learning at Amazon: A long heritage

Personalized Recommendations



Fulfillment automation & Inventory Management

Drones

Voice driven Interactions

Inventing entirely new Customer experiences

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

ML Services

Amazon SageMaker

1 BUILD

- Fast & accurate data labelling
- Built-in, high performance algorithms & notebooks

2 TRAIN

- One-click training and tuning
- Model optimization

3 DEPLOY

- One-click deployment
- Fully managed hosting with auto-scaling and elastic inference

TensorFlow · mxnet · PYTORCH · Chainer · HOROVOD · GLUON · Keras

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

ML Services

Amazon SageMaker

1

BUILD

Fast & accurate data labelling

Built-in, high performance algorithms & notebooks

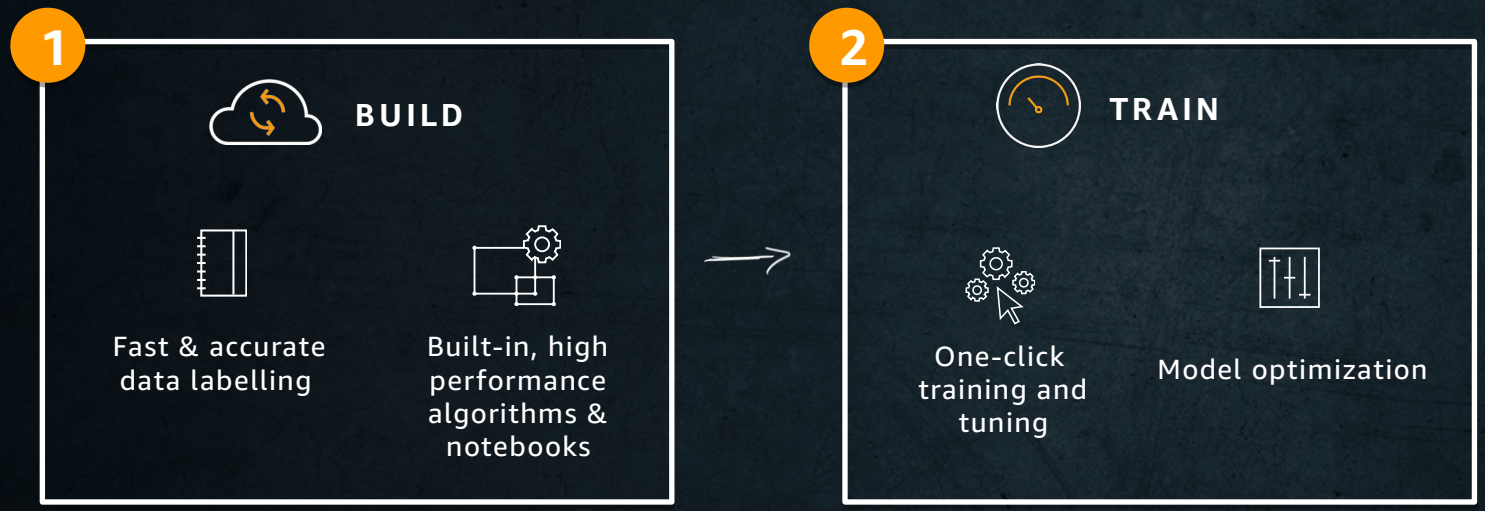
TensorFlow mxnet PYTORCH Chainer HOROVOD GLUON Keras

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

ML Services

Amazon SageMaker



TensorFlow mxnet PYTORCH Chainer HOROVOD GLUON Keras

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

ML Services

Amazon SageMaker

1 BUILD

- Fast & accurate data labelling
- Built-in, high performance algorithms & notebooks

2 TRAIN

- One-click training and tuning
- Model optimization

3 DEPLOY

- One-click deployment
- Fully managed hosting with auto-scaling and elastic inference

TensorFlow · mxnet · PYTORCH · Chainer · HOROVOD · GLUON · Keras

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Amazon SageMaker: Customer Story

"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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Key benefits of SageMaker at Intuit

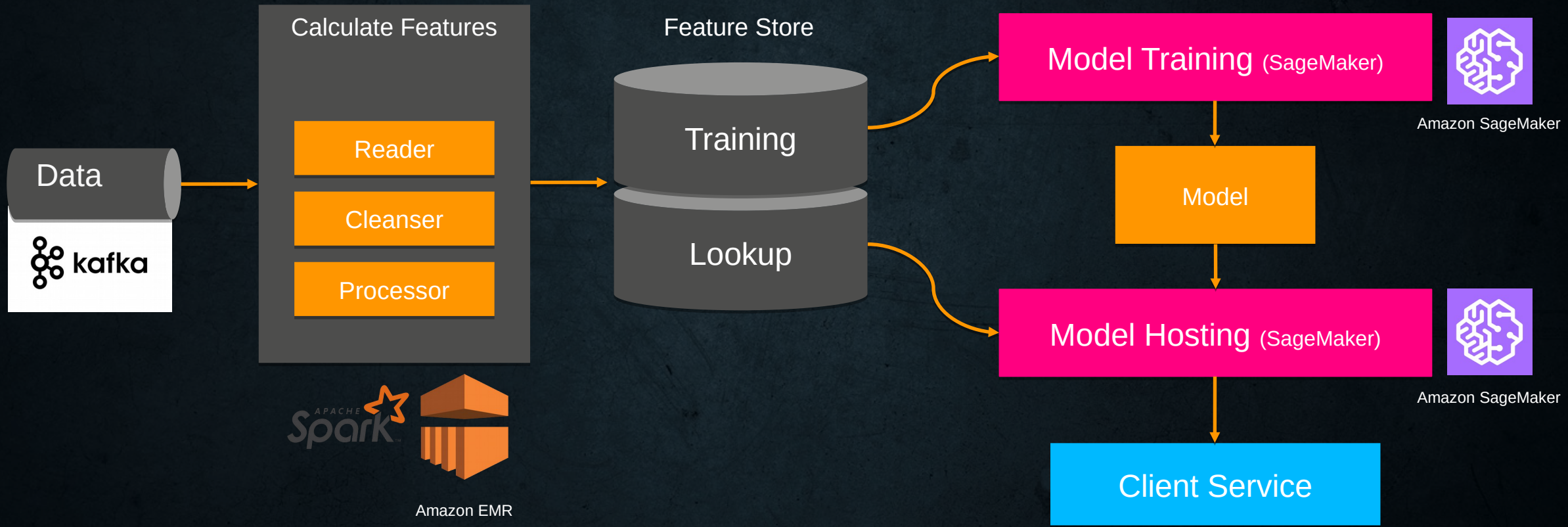
From	To
Ad-hoc setup and management of notebook environments	Easy data exploration in SageMaker notebooks
Limited choices for model deployment	Building around virtualization for flexibility
Competing for compute resources across teams	Auto-scalable model hosting environment

intuit

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Near real-time fraud detection in AWS using SageMaker



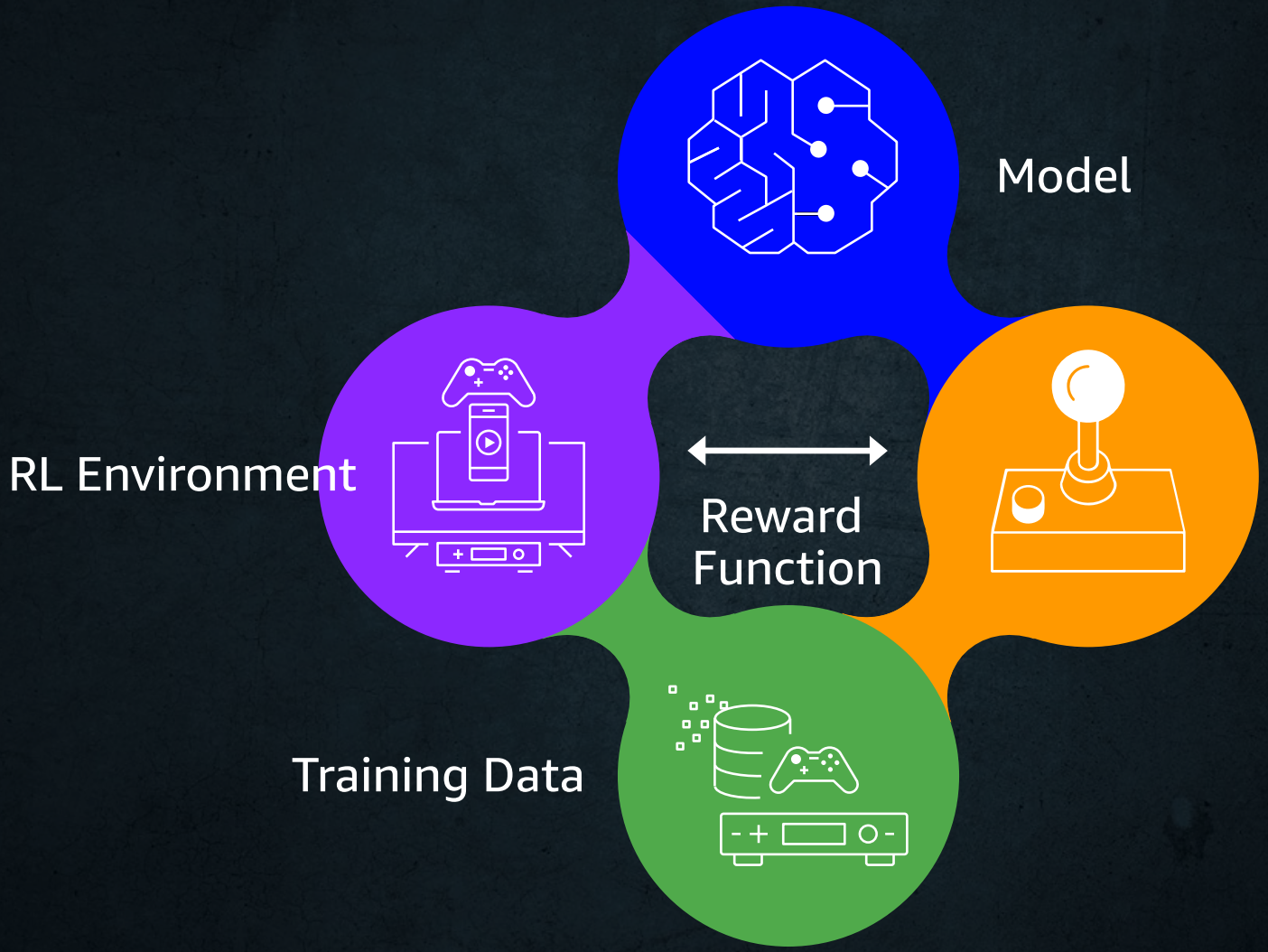
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

How do you teach machine learning models to make decisions when there is no training data?

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Reinforcement learning at work



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Reinforcement Learning

Learn by interacting with the real world

Model the real-world problem as a simulation environment

Trial and error

Observe results

Optimize learning strategy to maximize long-term reward

Model learns how to make complex decisions

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

What is an RL environment?

Representation of the real world

Programmed to represent real-world conditions

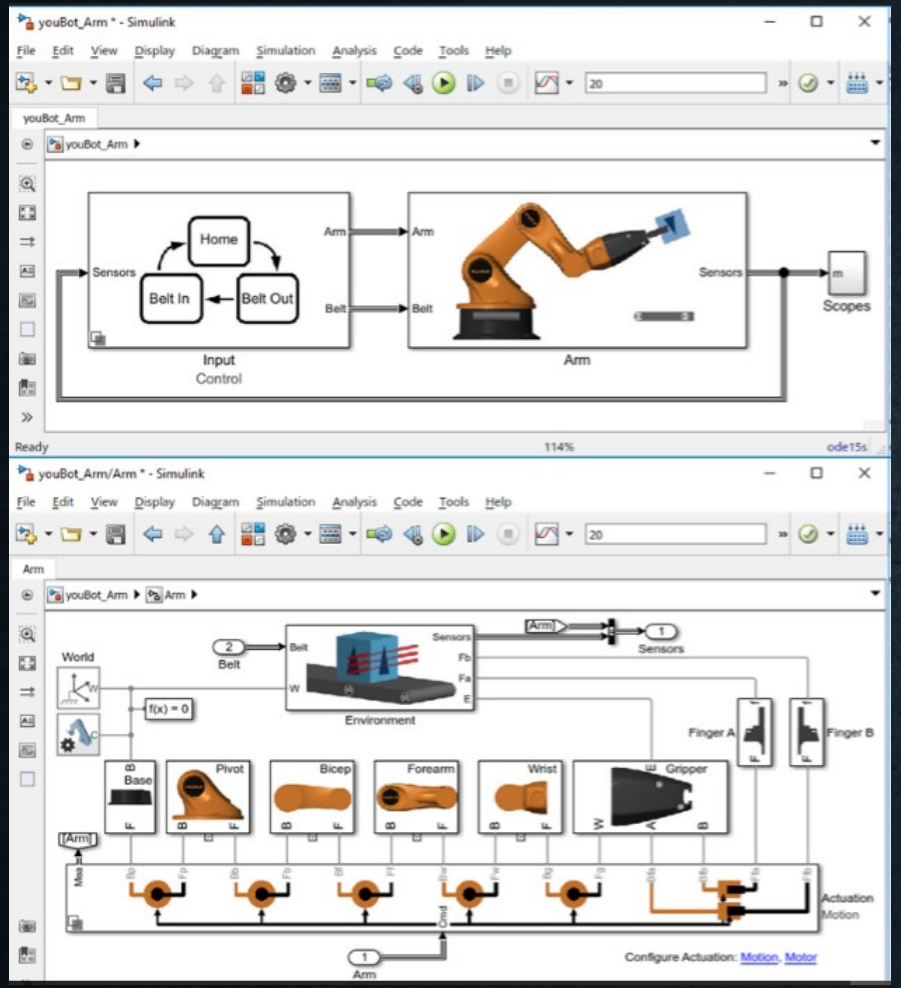
Enables interaction with user or a computer program

Dynamic and updates itself based on the interactions and programmed behaviour

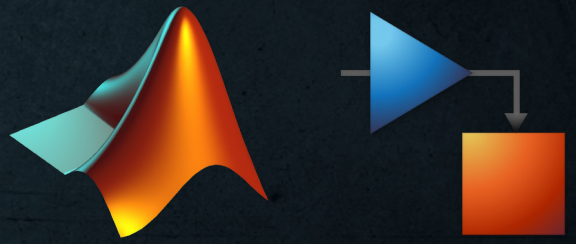
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Environments can be sophisticated



MATLAB and Simulink for modeling and simulation



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Or you can write your own RL environment

```
import gym
from gym.spaces import Discrete, Box

def __init__():
    #Initialize the RL environment

def reset(self):
    #Reset the RL environment

def step(self, action):
    #Take an action in the RL environment
    #Return the observation
    #return reward (positive or negative)

def render():
    #Render the RL environment
```

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

This makes RL applicable in many domains and not just gaming

Robotics

Industrial control

HVAC

Autonomous vehicles

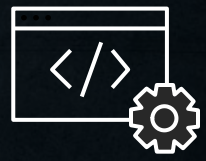
Advertising

NLP

Operations

Finance

Resource allocation



Online content delivery

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

ML Workshop featuring Amazon SageMaker

End-to-End Managed ML Platform

https://bit.ly/2VCfBqq

Module 1 – Predicting the future

- Launching Amazon SageMaker
- Binding to a Git repository
- Working with Jupyter Notebooks
- Using built in XG-Boost algorithm
- Training a model
- Deploying an endpoint
- Getting predictions

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Amazon SageMaker

1

Notebook Instances

2

Algorithms

3

ML Training Service

4

ML Hosting Service

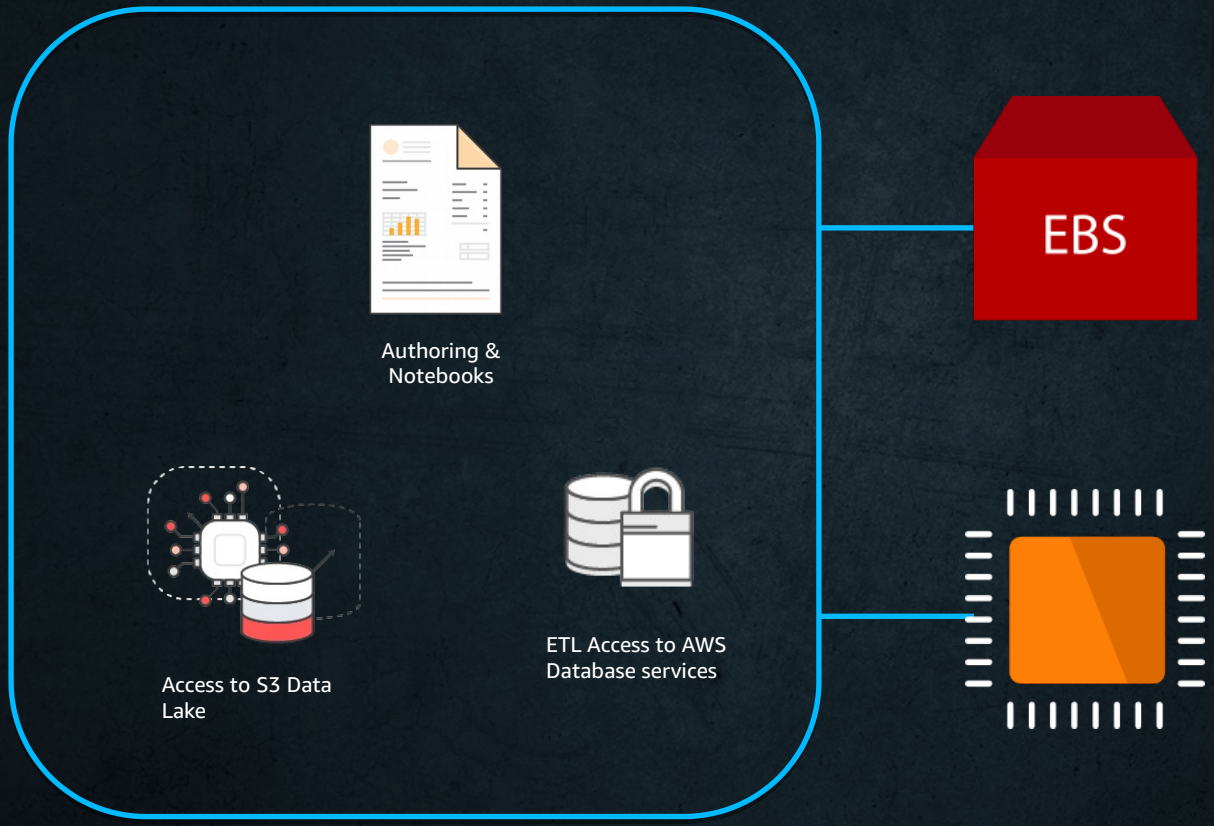
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

1

Zero Setup For Exploratory Data Analysis

Notebook Instances



"Just add data"

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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

2

Amazon SageMaker: 10x better algorithms

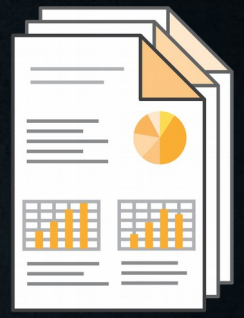
I

Algorithms

Streaming datasets, for cheaper training

Train faster, in a single pass

Greater reliability on extremely large datasets

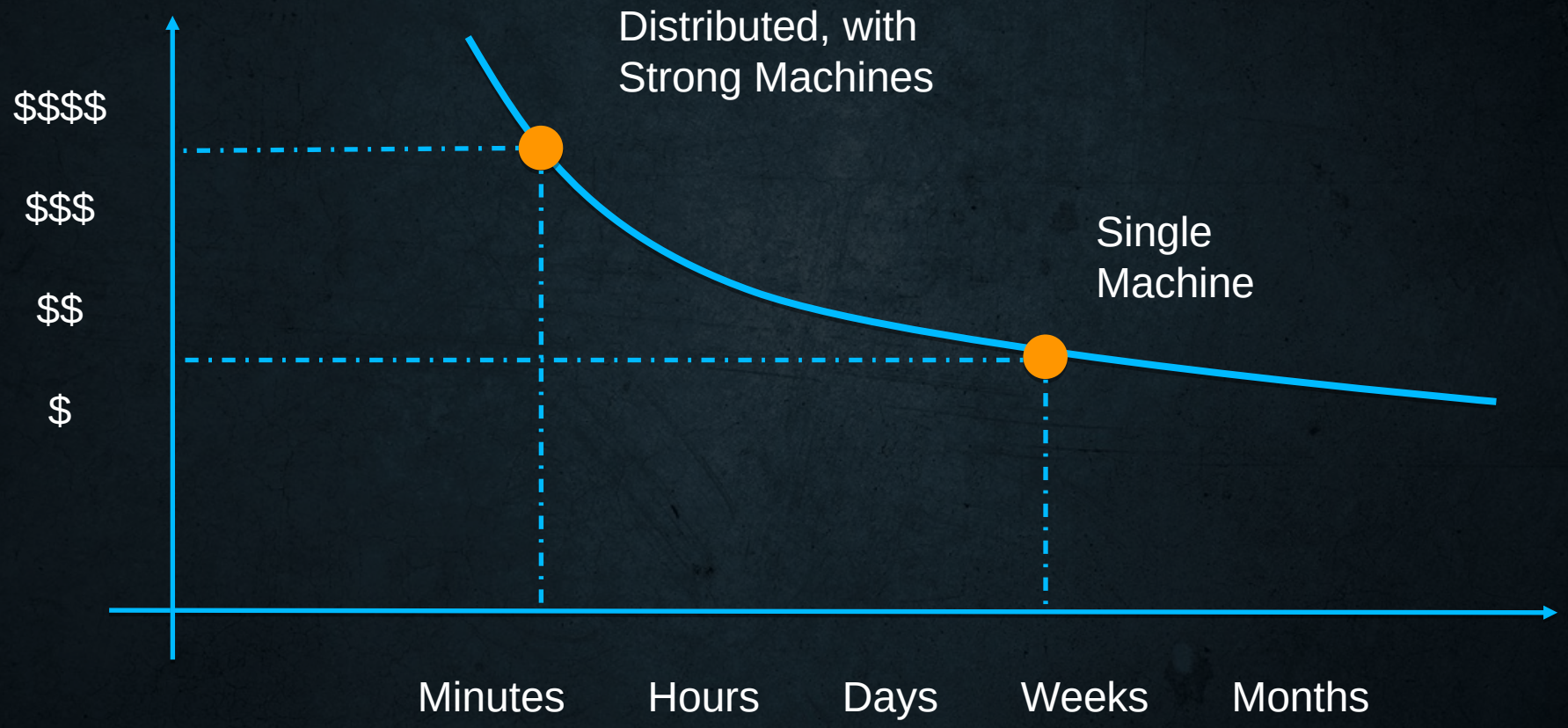


Choice of several ML algorithms

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

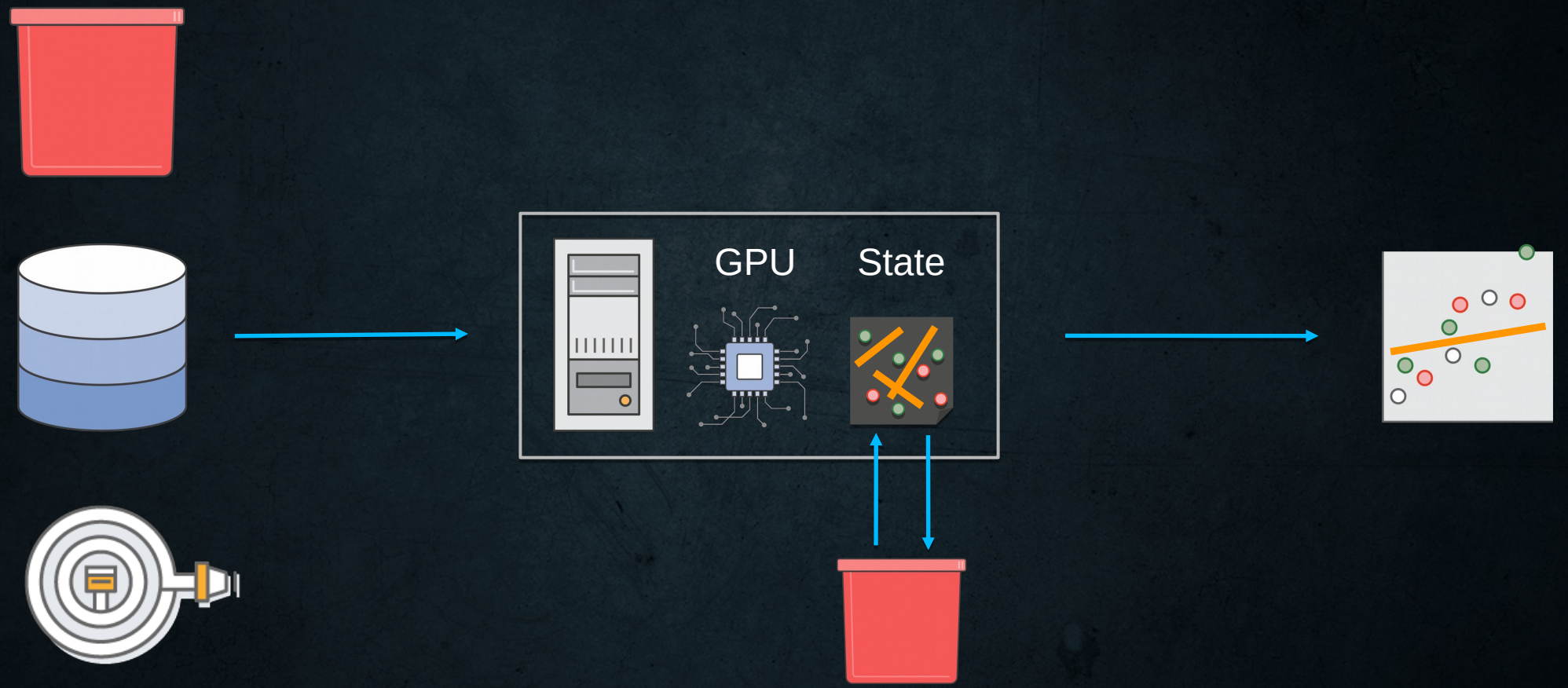
aws

Cost vs. Time



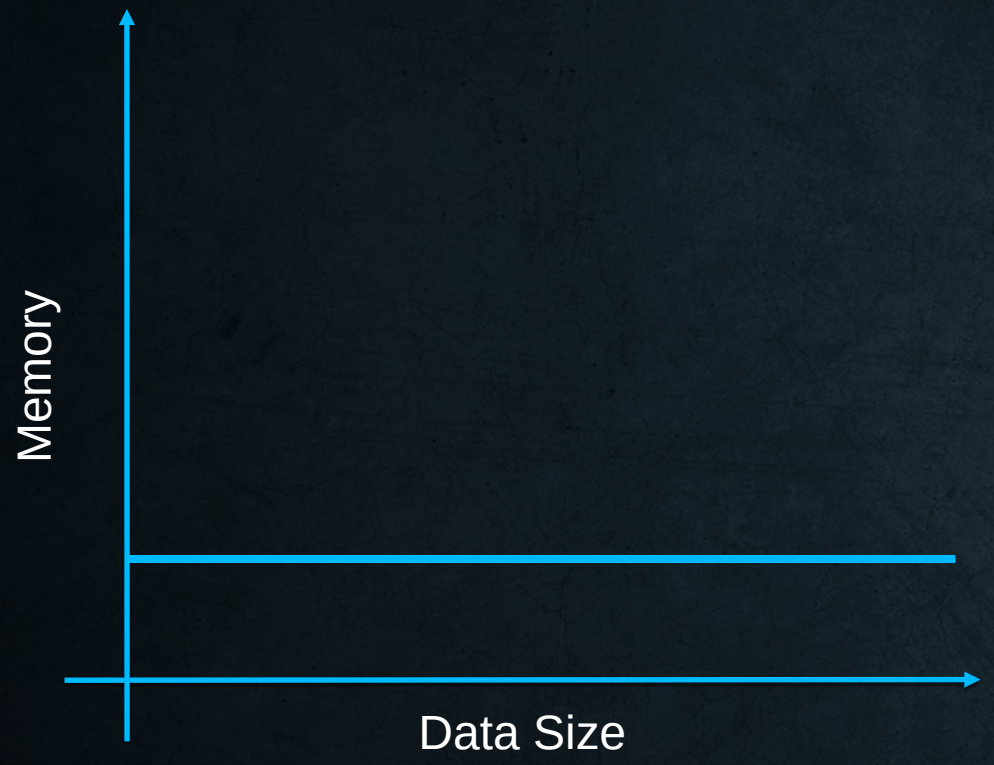
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

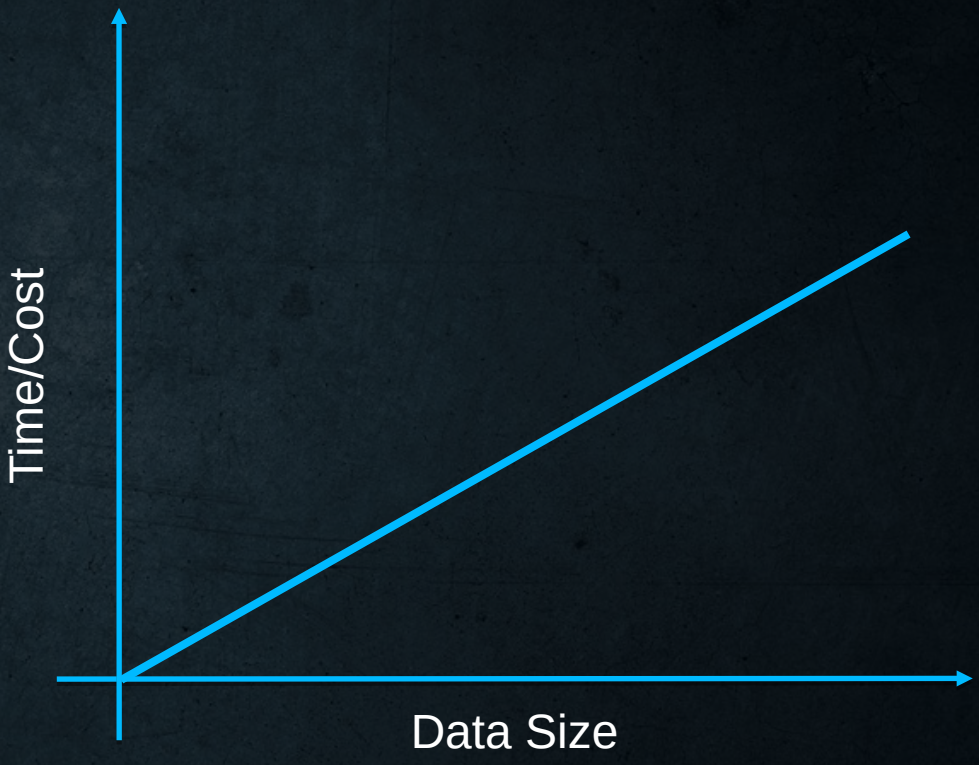
Streaming



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

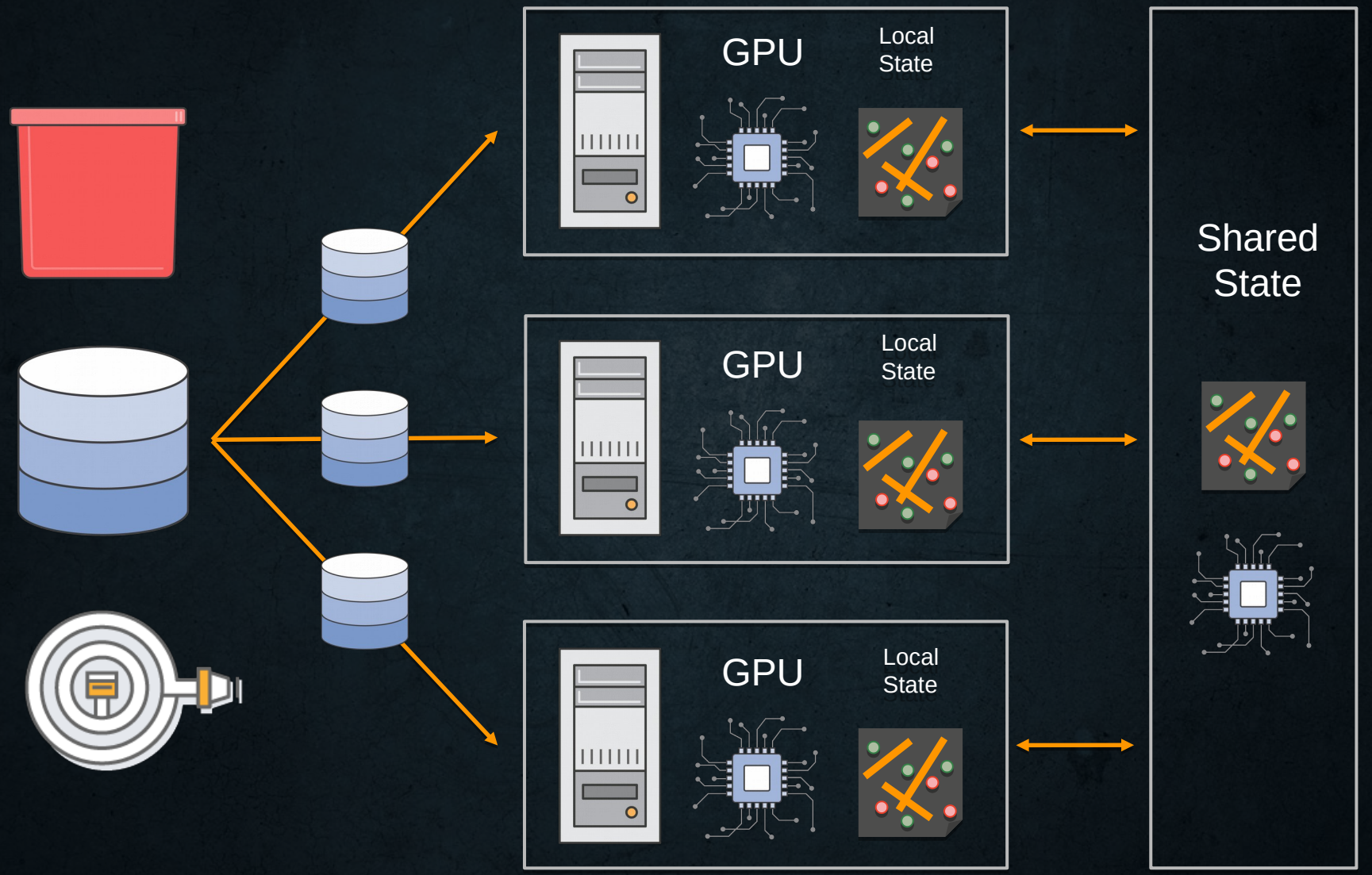
Streaming





© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

Shared State



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

Pipe Mode

- Data flows on-the-fly during training
- Shorter startup times
- Higher I/O throughput with optimized agent
- No dataset size limits
- Requires protobuf recordIO format
- Available algorithms:
  - Principal Component Analysis (PCA)
  - K-Means Clustering
  - Factorization Machines
  - Latent Dirichlet Allocation (LDA)
  - Linear Learner (Classification and Regression)
  - Neural Topic Modelling
  - Random Cut Forest
- Or use it with your own algorithms!

PCA + K-Means on NY Taxi Dataset

Pipe

File

https://aws.amazon.com/blogs/machine-learning/using-pipe-input-mode-for-amazon-sagemaker-algorithms/

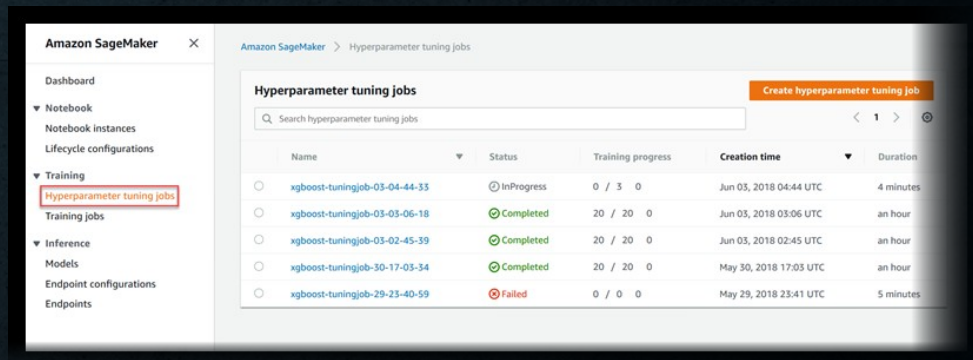
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

Automatic Model Tuning

1. Define objective metrics
   (Not needed for built-in)
2. Define hyper-parameter ranges
3. Let automatic tuning do the job!

- Based on Bayesian optimization
- Intelligently navigates parameter space to find optimal parameters
- Tries to balance explore/exploit trade-off

```
"ParameterRanges": {
      "CategoricalParameterRanges": [
        {
          "Name": "tree_method",
          "Values": ["auto", "exact", "approx", "hist"]
      ],
      "ContinuousParameterRanges": [
        {
          "Name": "eta",
          "MaxValue" : "0.5",
          "MinValue": "0"
        }
      ],
      "IntegerParameterRanges": [
        {
          "Name": "max_depth",
          "MaxValue": "10",
          "MinValue": "1",
        }
      ]
    }
```



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

3

Managed Distributed Training with Flexibility

ML Training Service

Secured

Fetch Training data

Training code

Save Model Artifacts

Save Inference Image

Amazon ECR

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

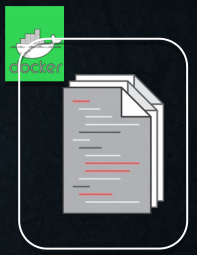
Amazon provided Algorithms

mxnet TensorFlow™ CHAINER

Bring Your Own Script
(SageMaker builds the Container)

APACHE Spark

SageMaker Estimators in Apache Spark



Bring Your Own Algorithm (You build the Container)

CPU GPU

HPO

Fully managed

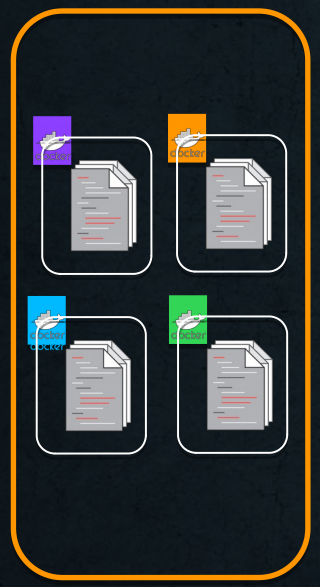
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

4

Easy Model Deployment to Amazon SageMaker

ML Hosting Service



Versions of the same inference code saved in inference containers. Prod is the primary one, 50% of the traffic must be served there!

Amazon ECR

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

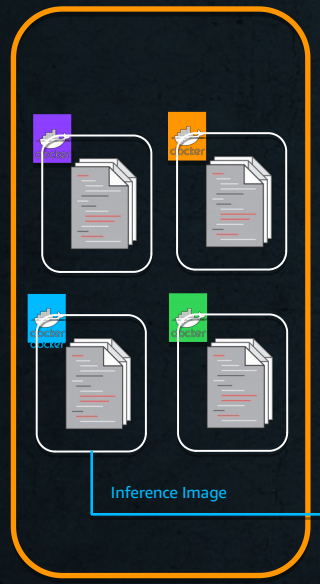
aws

4

Easy Model Deployment to Amazon SageMaker

ML Hosting Service

Model Artifacts



Inference Image

Amazon ECR

ModelName: prod

Create a Model

Versions of the same inference code saved in inference containers. Prod is the primary one, 50% of the traffic must be served there!

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

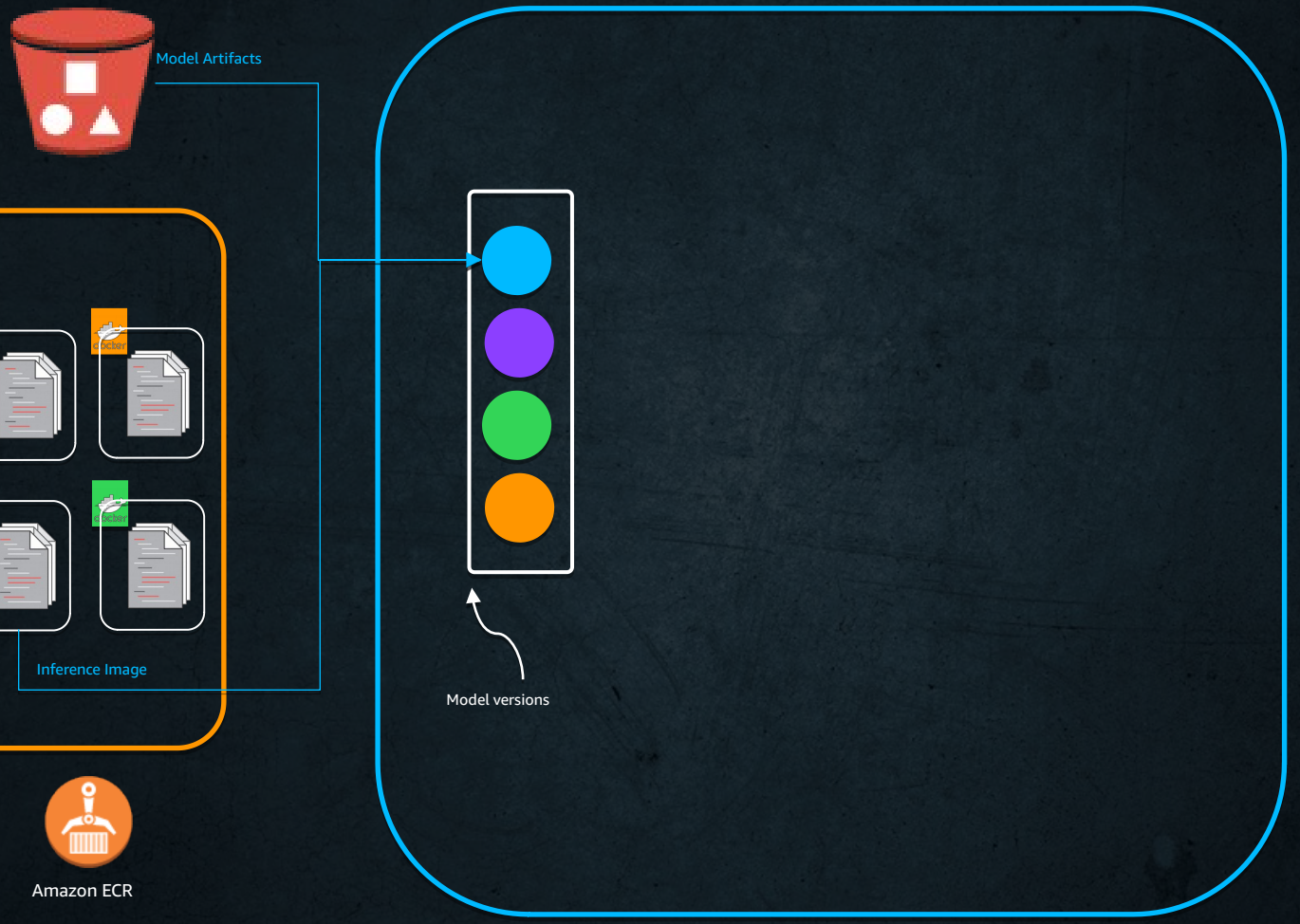
aws

4

Easy Model Deployment to Amazon SageMaker

ML Hosting Service

Versions of the same inference code saved in inference containers. Prod is the primary one, 50% of the traffic must be served there!



Create versions of a Model

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

4

Easy Model Deployment to Amazon SageMaker

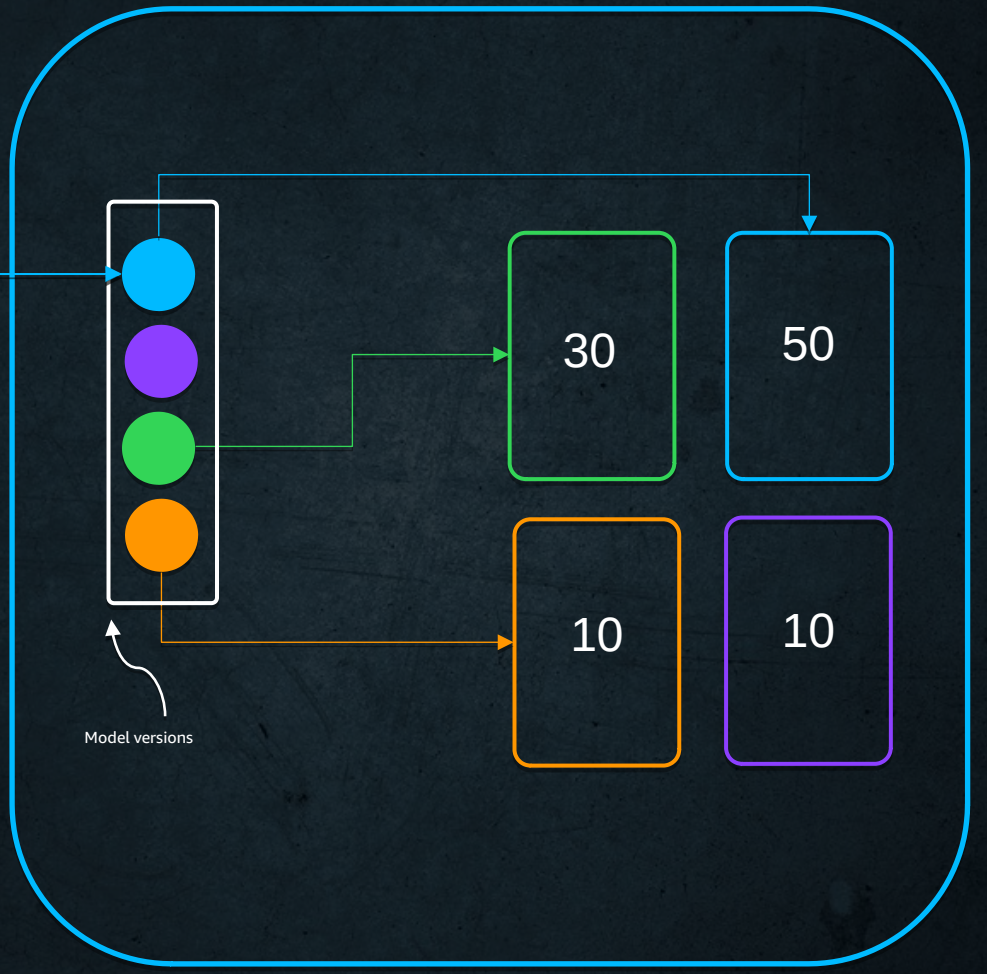
ML Hosting Service

Versions of the same inference code saved in inference containers. Prod is the primary one, 50% of the traffic must be served there!

Model Artifacts

Inference Image

Amazon ECR



InstanceType: c3.4xlarge

InitialInstanceCount: 3

ModelName: prod

VariantName: primary

InitialVariantWeight: 50

ProductionVariant

Create weighted ProductionVariants

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

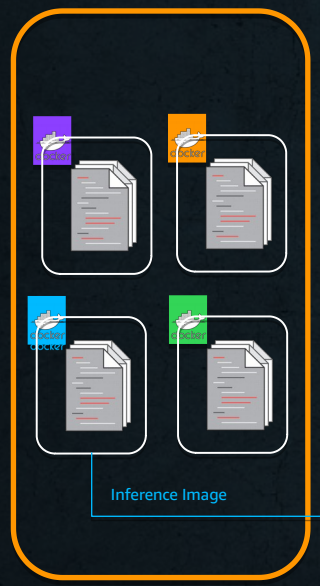
aws

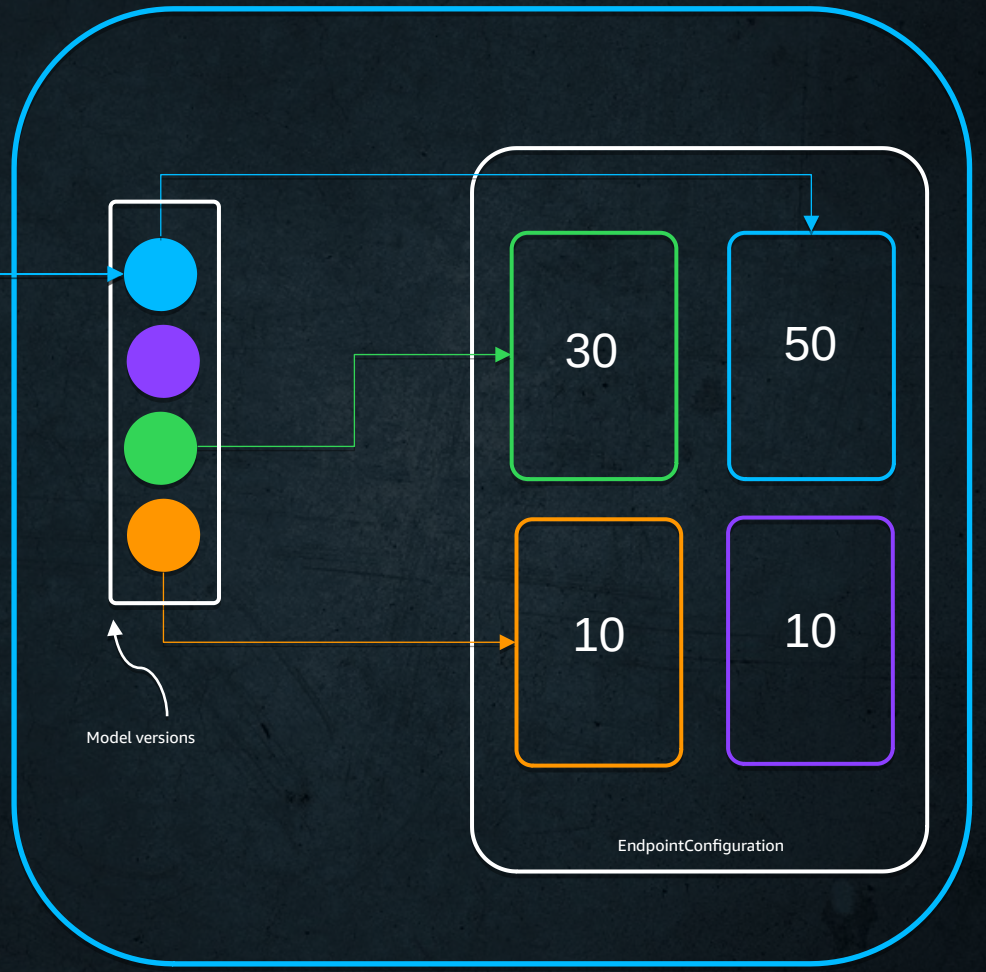
4

Easy Model Deployment to Amazon SageMaker

ML Hosting Service

Versions of the same inference code saved in inference containers. Prod is the primary one, 50% of the traffic must be served there!





InstanceType: c3.4xlarge

InitialInstanceCount: 3

ModelName: prod

VariantName: primary

InitialVariantWeight: 50

ProductionVariant

Create an EndpointConfiguration from one or many ProductionVariant(s)

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

4

Easy Model Deployment to Amazon SageMaker

ML Hosting Service

- ✓ Auto-Scaling Inference APIs
- ✓ A/B Testing
- ✓ Low Latency & High Throughput
- ✓ Bring Your Own Model
- ✓ Python SDK

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Amazon SageMaker workflows

Experiment Management

Organize, track, and evaluate model training experiments with SageMaker search

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Amazon SageMaker workflows

Experiment Management

Organize, track, and evaluate model training experiments with SageMaker search

Collaboration

Link GitHub, AWS CodeCommit, and self-hosted Git repositories to notebooks

Clone public and private repositories

Secure information with AWS Identity and Access Management (IAM), LDAP, and AWS Secrets Manager

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Amazon SageMaker workflows

Experiment Management

Organize, track, and evaluate model training experiments with SageMaker search

Collaboration

Link GitHub, AWS CodeCommit, and self-hosted Git repositories to notebooks

Clone public and private repositories

Secure information with AWS Identity and Access Management (IAM), LDAP, and AWS Secrets Manager

Automation

Use AWS Step Functions to automate end-to-end workflows

Integrate with Apache Airflow

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

The Amazon ML stack: Broadest & deepest set of capabilities

AI SERVICES

Vision


REKOGNITION IMAGE


REKOGNITION VIDEO

Speech


TEXTRACT


POLLY


TRANSCRIBE

Language


TRANSLATE


COMPREHEND

Chatbots


LEX

Forecasting


FORECAST

Recommendations


PERSONALIZE

ML SERVICES


AMAZON SAGEMAKER

BUILD

Pre-built algorithms & notebooks

Data labelling (GROUND TRUTH)

Algorithms & models (AWS MARKETPLACE FOR MACHINE LEARNING)

TRAIN

One-click model training & tuning

Optimization (NEO)

Reinforcement learning

DEPLOY

One-click deployment & hosting

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

TensorFlow

mxnet

PYTORCH

Interfaces

GLUON

Keras

Infrastructure


EC2 P3 & P3N


EC2 C5


FPGAs


GREENGRASS


ELASTIC INFERENCE

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Highest-performing infrastructure for your business

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

TensorFlow

mxnet

PYTORCH

Interfaces

GLUON

Keras

Infrastructure

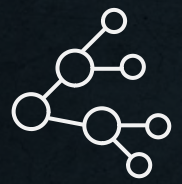
EC2 P3 & P3N

EC2 C5

FPGAs

GREENGRASS

ELASTIC INFERENCE



Build custom algorithms using the ML frameworks

Fastest and lowest-cost compute options for ML workloads

Elastic compute to provision just-right compute for your ML workloads

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

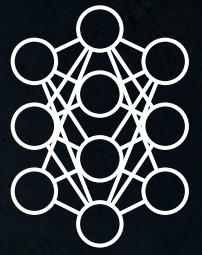
aws

Amazon EC2 P3dn instance

The largest P3 instance, optimized for distributed training

Reduce machine learning training time

Better GPU utilization



Support larger, more complex models

KEY FEATURES

100Gbps of networking bandwidth

8 NVIDIA Tesla V100 GPUs

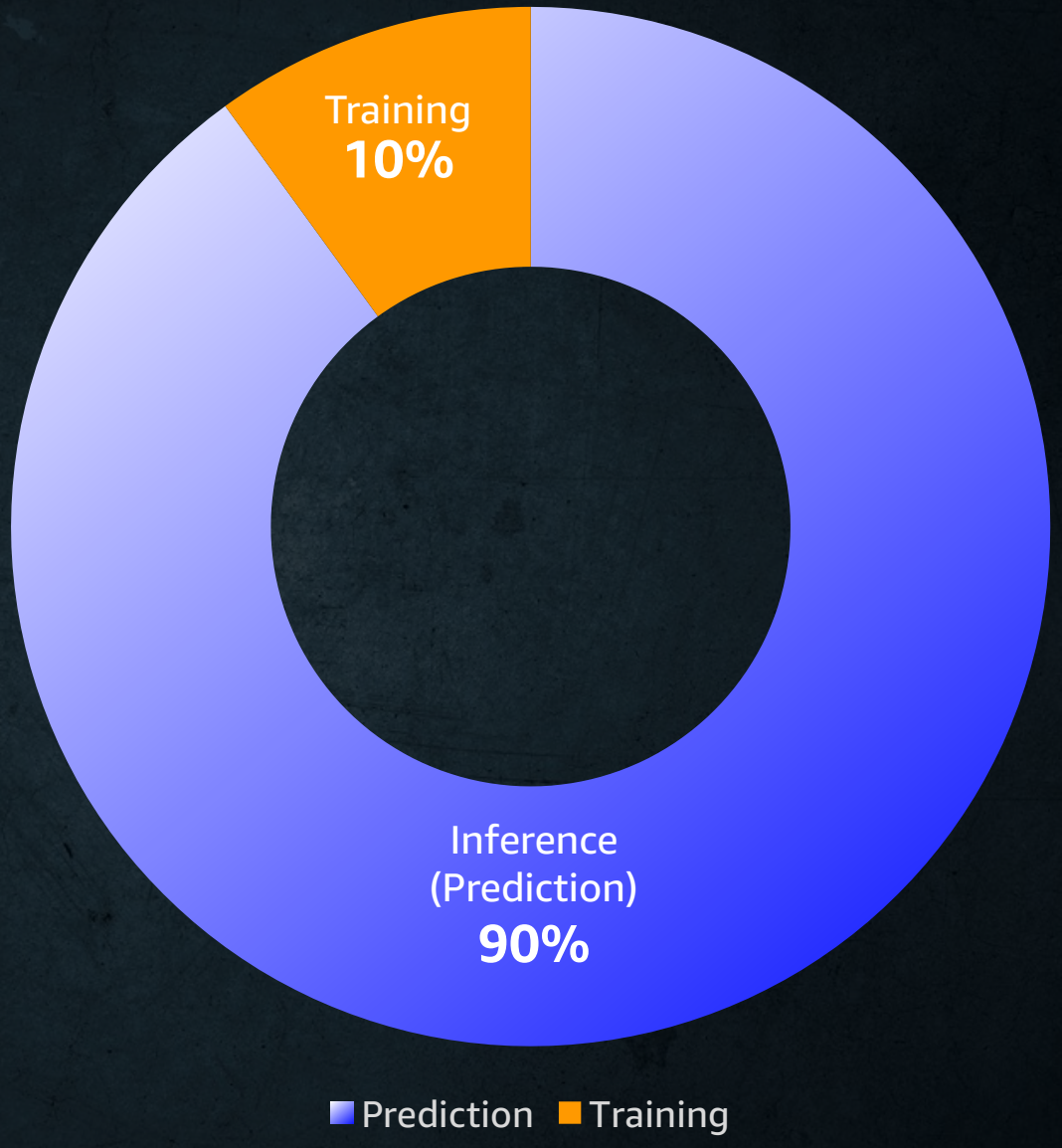
32GB of memory per GPU (2x more P3)

96 Intel Skylake vCPUs (50% more than P3) with AVX-512

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Predictions drive complexity and cost in production



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

Amazon Elastic Inference

Add GPU acceleration to any Amazon EC2 instance for faster inference at much lower cost (up to 75% savings)

Provision Elastic Inference capacity inside VPC

Starting at 1 TFLOPS	Up to 32 TFLOPS
Simple speech and language models	Recommendation engines or fraud detection models

Any instance family

mxnet

TensorFlow

PYTORCH

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

aws

Lowering inference costs with Amazon Elastic Inference

ResNet-50
Computer vision
deep learning model

360,000
images per hour,
inference

$0.22
per hour
on medium EI accelerator

LOWEST COST AVAILABLE

75%
lower cost

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

aws

ML Workshop featuring Amazon SageMaker

End-to-End Managed ML Platform

https://bit.ly/2VCfBqq

Module 2 – Image recognition

- Using built in Image Classification algorithm
- Training a model with Transfer Learning
- Using Elastic Inference
- Using Automatic Tuning
- Deploying an endpoint
- Getting predictions

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

AWS is framework agnostic

Choose from popular frameworks

TensorFlow mxnet PYTORCH Chainer Caffe2 ONNX Keras GLUON

Run them fully managed

Or run them yourself

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

AWS: The platform of choice to run TensorFlow

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

Source: Nucleus Research, November 2018

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

The best place to run TensorFlow

Stock
TensorFlow

65%

scaling efficiency
with 256 GPUs

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

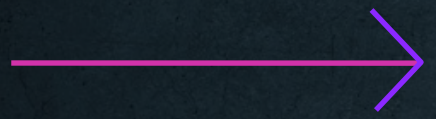
aws

The best place to run TensorFlow

Stock TensorFlow

65%

scaling efficiency with 256 GPUs



NEW

AWS-Optimized TensorFlow

90%

scaling efficiency with 256 GPUs

Available with Amazon SageMaker and the AWS Deep Learning AMIs

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

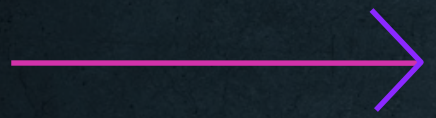
The best place to run TensorFlow

NEW

Stock TensorFlow

65%

scaling efficiency with 256 GPUs



AWS-Optimized TensorFlow

90%

scaling efficiency with 256 GPUs

Available with Amazon SageMaker and the AWS Deep Learning AMIs

30m

training time

14m

training time

Fastest time for TensorFlow

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

AWS is the best platform for TensorFlow

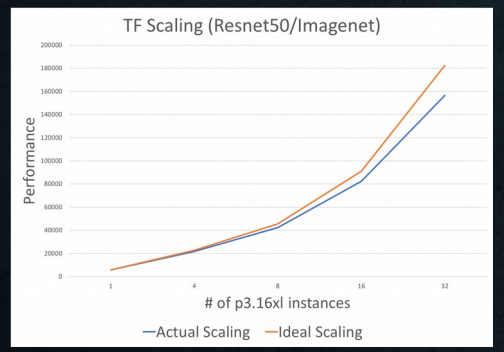
Fast

Versatile

Scalable

Time-to-train (CNN, ResNet50)

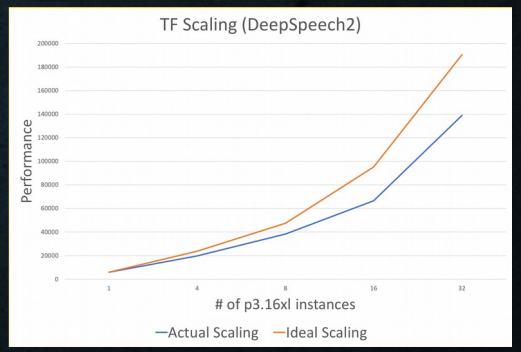
7h → 14m with 256 GPUs



Within 90% of ideal scaling

Time-to-train (LSTM, DeepSpeech2)

6d → 3.5h with 256 GPUs



Within 73% of ideal scaling

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

ML Workshop featuring Amazon SageMaker

End-to-End Managed ML Platform

https://bit.ly/2VCfBqq

Module 3 – Tensor Flow

- Bring your own training script for Tensor Flow
- Training locally
- Using Batch Transform for bulk predictions
- Making your own container
- Deploying an endpoint
- Getting predictions

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

The Amazon ML stack: Broadest & deepest set of capabilities

AI SERVICES

Vision


REKOGNITION IMAGE


REKOGNITION VIDEO

Speech


TEXTRACT


POLLY


TRANSCRIBE

Language


TRANSLATE


COMPREHEND

Chatbots


LEX

Forecasting


FORECAST

Recommendations


PERSONALIZE

ML SERVICES


AMAZON SAGEMAKER

BUILD

Pre-built algorithms & notebooks

Data labeling (GROUND TRUTH)

Algorithms & models (AWS MARKETPLACE FOR MACHINE LEARNING)

TRAIN

One-click model training & tuning

Optimization (NEO)

Reinforcement learning

DEPLOY

One-click deployment & hosting

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

TensorFlow

mxnet

PYTORCH

Interfaces

GLUON

Keras

Infrastructure


EC2 P3 & P3N


EC2 C5


FPGAs


GREENGRASS


ELASTIC INFERENCE

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

More machine learning happens on AWS than anywhere else

10,000+

customers have used machine learning on AWS

81% of deep learning in the cloud runs on AWS

85% of TensorFlow projects in the cloud run on AWS

AWS holds the top spots on Stanford's deep learning benchmark, DAWN, for fastest training time, lowest cost, lowest inference latency

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws

Thank you.

And please fill out the survey

http://bit.ly/2MJbGZF

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Trademark

aws