Deploying and Monitoring the Model

Amber Israelsen
AUTHOR | DEVELOPER | TRAINER

www.amberisraelsen.com

Course Overview

- Course Introduction
- Identifying Opportunities for Machine Learning
- Defining Machine Learning Problems
- Fetching and Preparing Data
- Training and Evaluating the Model
- Deploying and Monitoring the Model
- The AWS Machine Learning Stack
- Next Steps

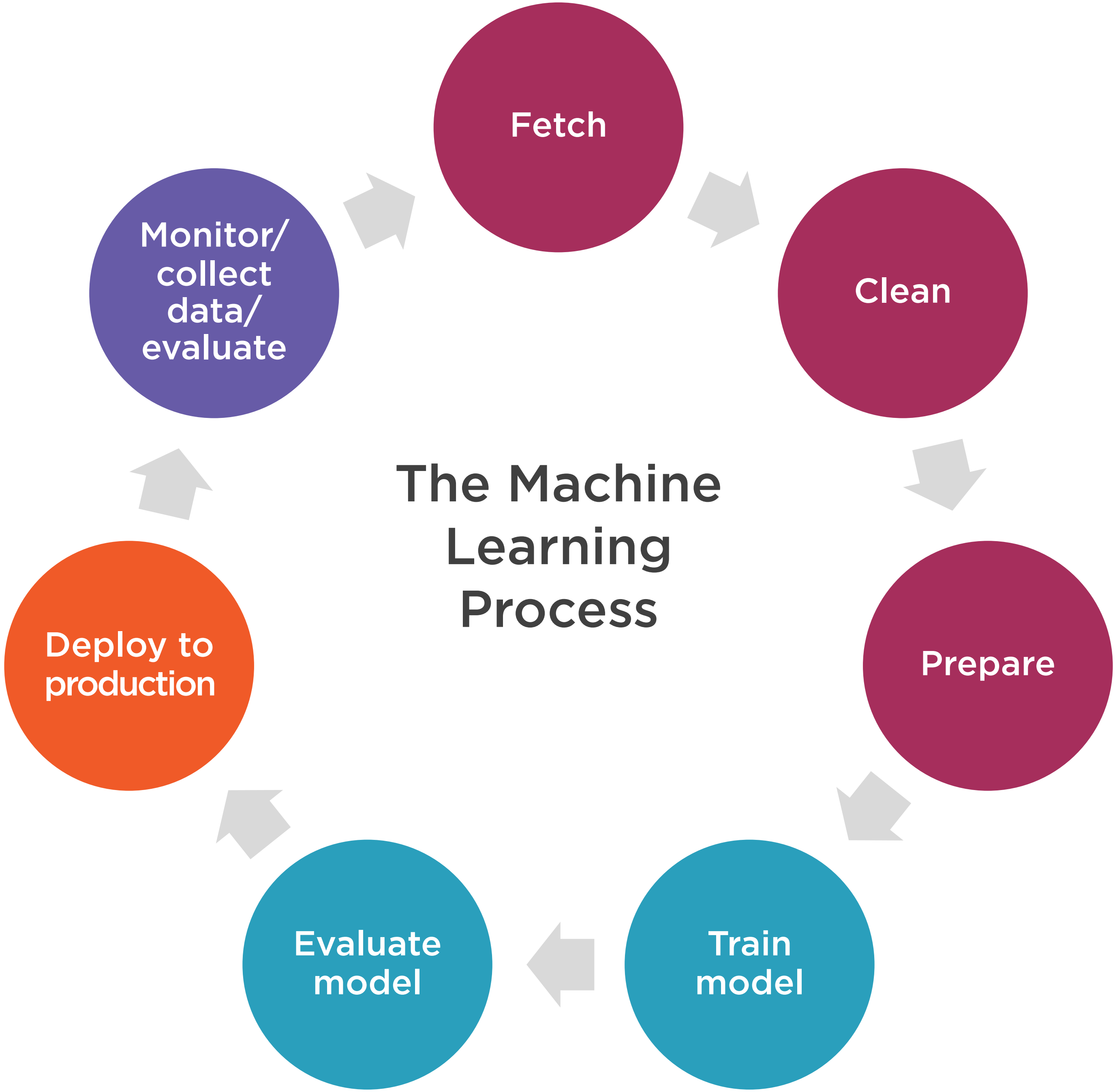
Module Overview

The machine learning process

- Deploying the model
  - Architecture and security best practices
- Monitoring the model
  - CloudWatch and CloudTrail

Demo in SageMaker Studio

Deploying the Model



The Machine Learning Process

Ensure Architectural Best Practices

High Availability

Fault Tolerance

Achieving Loose Coupling in AWS

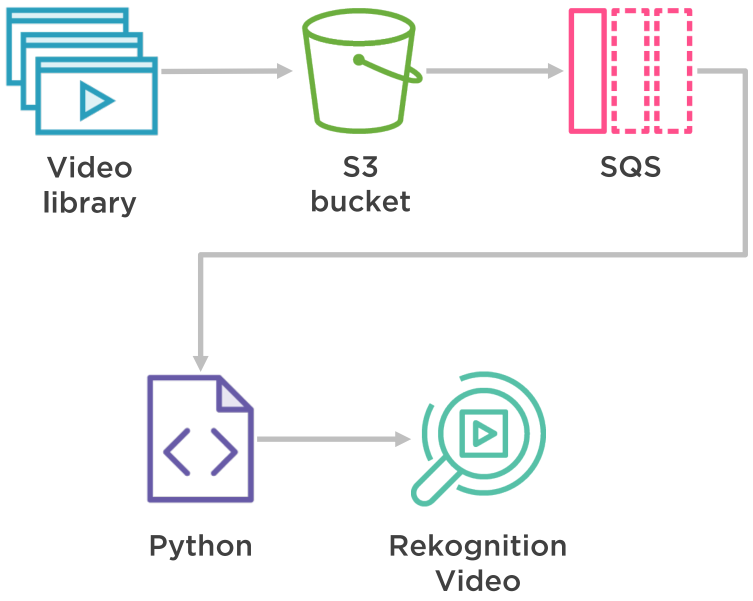
Amazon Simple Queue Service (SQS)

Message queuing service

AWS Step Functions

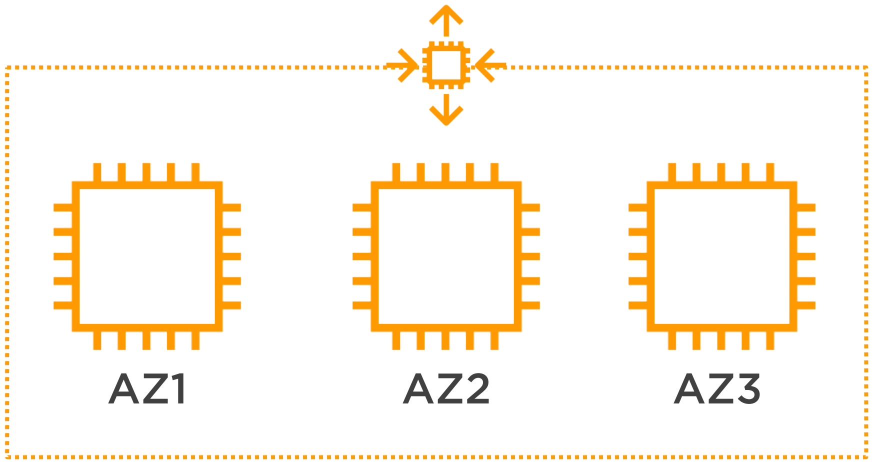
Serverless orchestrator

AN EXAMPLE

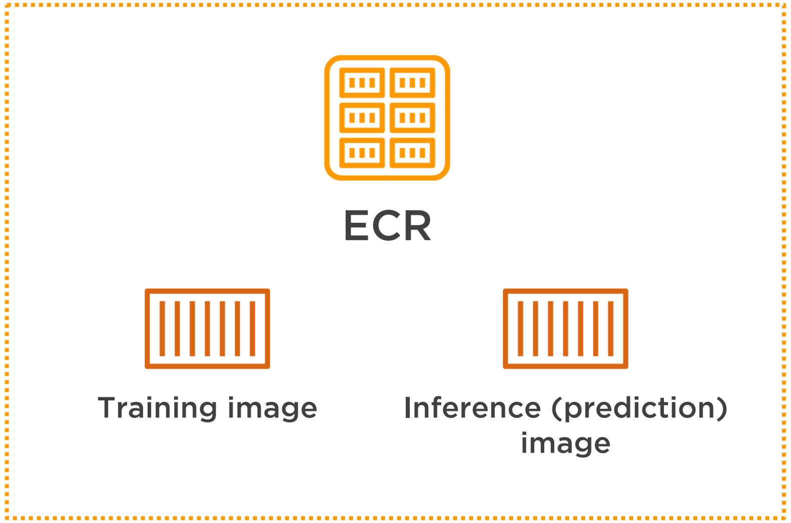


Best Practices for SageMaker

Deploy SageMaker endpoints to multiple Availability Zones (AZs), and use Auto Scaling



Use containers to achieve loose coupling



Security with SageMaker

Supports IAM role-based access

Data is encrypted

- **At rest** using Key Management Service (KMS) or a transient key
- **In transit** with TLS 1.2 encryption

Training data is stored and transferred in the customer's account

Instances can be launched in a customer-managed VPC

Deploying a Model Using SageMaker

#1

The model you previously created

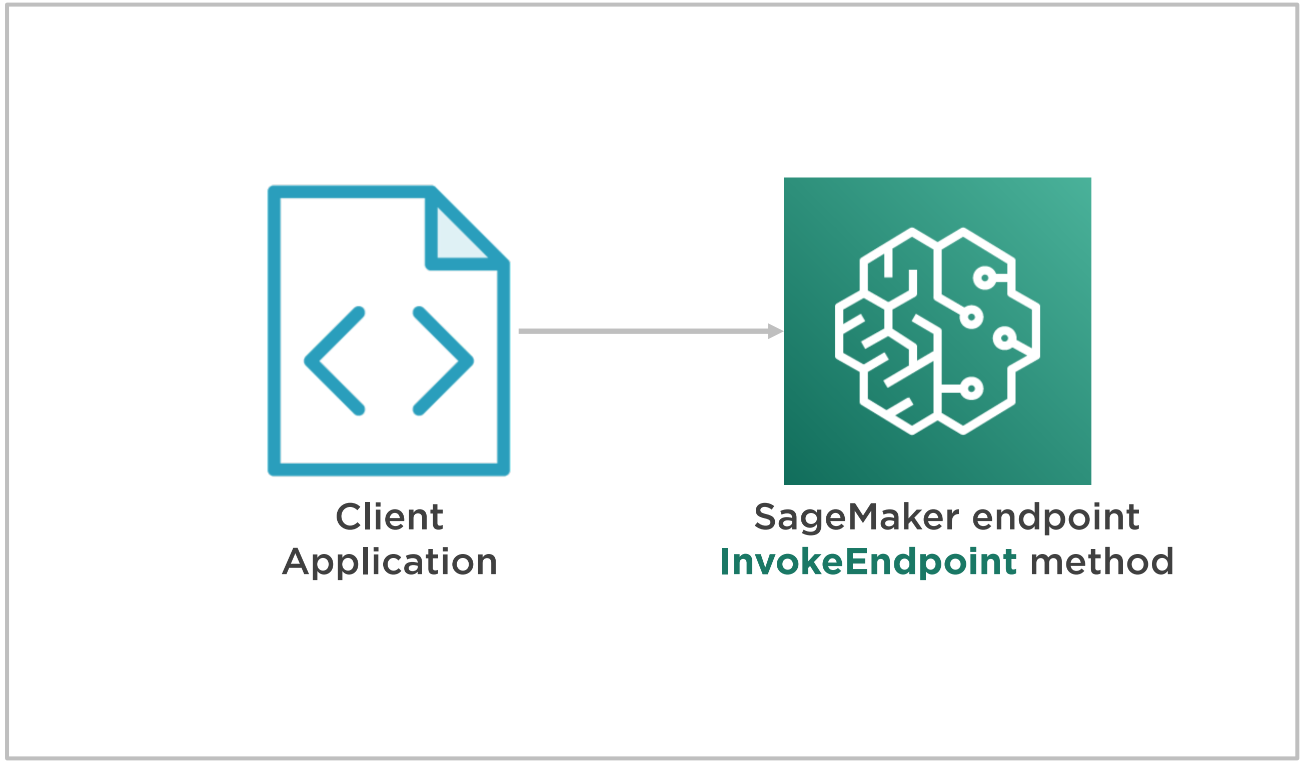
#2

Create an endpoint configuration for an HTTPS endpoint

#3

Create an HTTPS endpoint

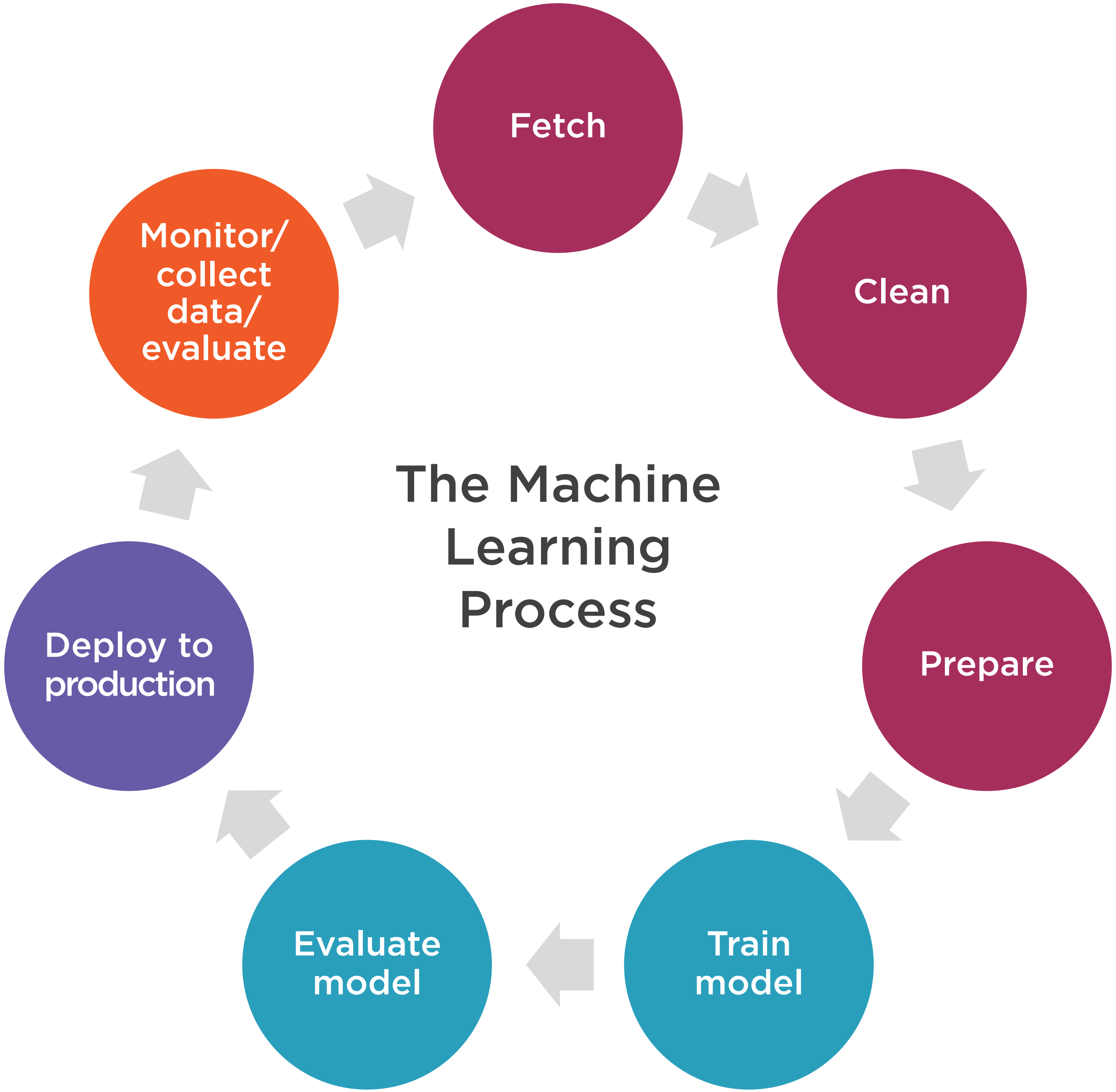
Using the Deployed Model



Client Application

SageMaker endpoint
InvokeEndpoint method

Monitoring and Collecting Data



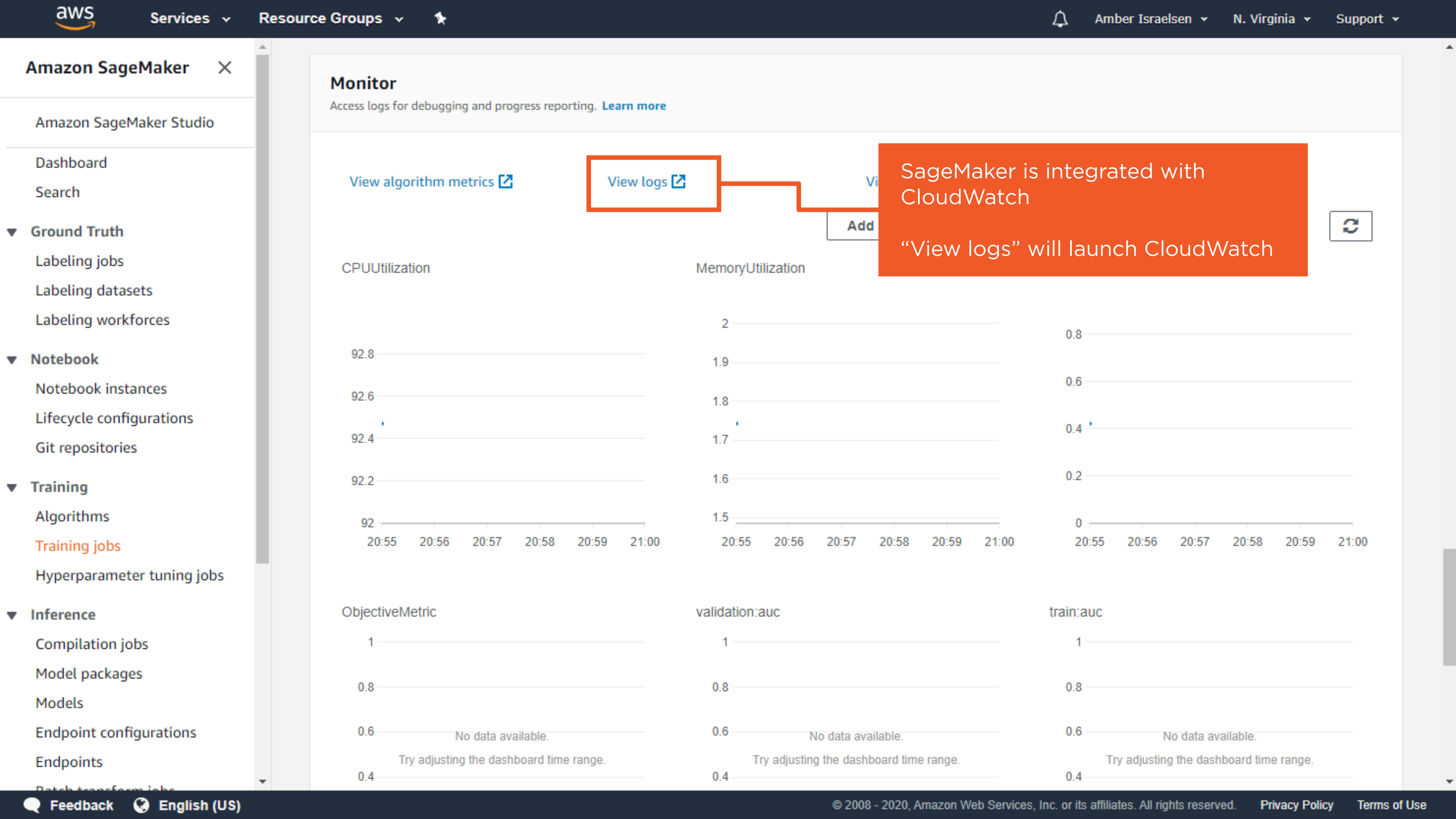
The Machine Learning Process

- Fetch
- Clean
- Prepare
- Train model
- Evaluate model
- Deploy to production
- Monitor/ collect data/ evaluate

Monitoring Services

CloudWatch

Near-real-time metrics such as CPU, memory, GPU utilization



Monitor

Access logs for debugging and progress reporting. Learn more

No data available.
Try adjusting the dashboard time range.

No data available.
Try adjusting the dashboard time range.

No data available.
Try adjusting the dashboard time range.

© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker is integrated with CloudWatch

“View logs” will launch CloudWatch

Monitoring Services

CloudWatch

Near-real-time metrics such as CPU, memory, GPU utilization

CloudTrail

Captures API calls and related events

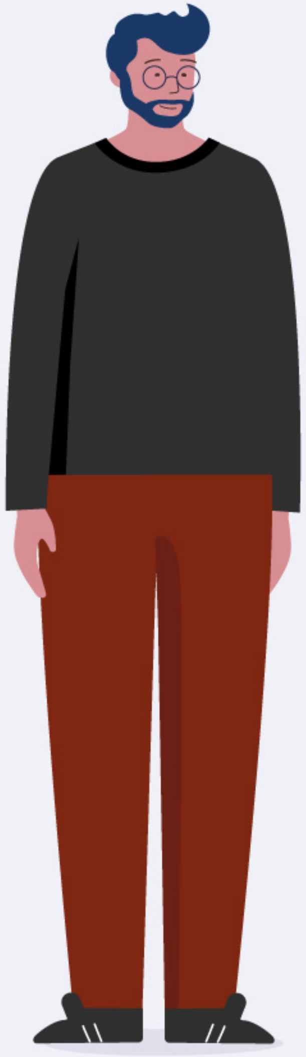
Monitoring Your Results

SageMaker Model Monitor

- Automatically collects data from your endpoints
- Detects changes in quality as compared to baseline
- Detects data drift
- Offers visualization tools to view results and statistics

Demo

Deploying and monitoring the model in SageMaker Studio



Ray

- VP of Customer Experience
- Concerned about customer churn

Key Points to Remember

Clean up your AWS resources!

S3 bucket
SageMaker endpoint
SageMaker application
Jupyter Notebook instance

Use architecture and security best practices

To deploy a model

- Create an endpoint configuration for HTTPS endpoint
- Create an HTTPS endpoint

Monitoring

- CloudWatch: performance metrics
- CloudTrail: auditing API activity
- Model Monitor: results from the model

Up Next:
The AWS Machine Learning Stack