

SageMaker, AI and Machine Learning

with Python

Chandra Lingam

Cloud Wave LLC

Why

Focus on current best practices

Learn most useful algorithms

Easy to integrate

Elastic infrastructure

No upfront cost or commitment

How

Face comparison

Compare faces to see how closely they match based on a similarity percentage.

Reference face



Comparison faces



Done with the demo?

[Learn more](#)

▼ Results



=



Similarity

92.3 %

► Request

► Response

What

SageMaker

Artificial Intelligence

Integration

Data Lake

Certification

SageMaker

Jupyter Notebook on the Cloud

Learn popular algorithms

Examples: XGBoost, DeepAR, Factorization Machines, PCA and more

Train, Optimize and Deploy Models

Artificial Intelligence

Pre-trained service – Ready to use

Image and Video Analysis

Example: facial recognition, describe scene, text, track movement

Natural Language Processing

Example: sentiment, translation, chatbots

Integration

Integrate new capabilities in your application

Hosting, Scaling, Fault tolerance

Provide a Clean Interface with API Gateway, Lambda

Data Lake

Streamline data management

Ingest, Catalog, Transform, Secure and Visualize

Example: Kinesis Firehose, Glue, Athena, Redshift Spectrum, S3

AWS Certified Machine Learning - Specialty

Complete guide to AWS Machine Learning Specialty
Certification

Labs, Quiz, Timed Practice Exam

Source Code

<https://github.com/ChandraLingam/AmazonSageMakerCourse>

Access latest code

Ideal Student

Willing to Learn

Participate in Course Discussion Forums

Comfortable coding in Python

Chandra Lingam



50,000+ Students

Up-to-date Content



AWS Certified Machine Learning

Specialty – Exam Overview and Preparation

Chandra Lingam

Cloud Wave LLC

AWS Certified Machine Learning Specialty

Exam Guide and Sample Questions

<https://aws.amazon.com/certification/certified-machine-learning-specialty/>

Topics

Area	Number of Questions	Description
AWS	15%	<p>Questions check your understanding of AWS Cloud</p> <p>Example: IAM, CloudWatch, Regions, Availability Zones</p>
Machine Learning	50-60%	<p>Machine Learning Concepts – not specific to AWS</p> <p>Example: Missing data, Is ML the right choice for a problem, Evaluating Performance, Optimization</p>
AWS Machine Learning	25-35%	<p>Understanding of AWS ML Offerings and best practices</p> <p>Example: SageMaker, AI, Deploying model on AWS, Custom Algorithms</p>

AWS Machine Learning Specialty - Exam Details

Time: 170 minutes

Number of Questions: 65

Certification Exam Cost: USD 300

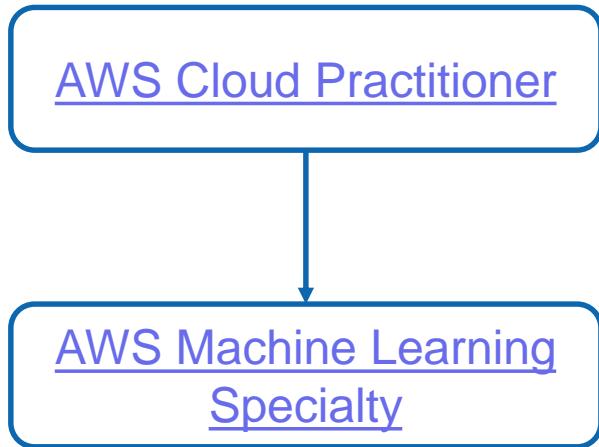
Practice Exam Cost: USD 40

Cost Saving Tip – After you complete a certification exam, AWS issues vouchers for future Practice and Certification Exams – This can give you access to free practice exams and 50% discount on certification exam fees

New to AWS - Preparation Strategy

- This course does not expect previous AWS experience
- Relevant concepts are introduced and explained
- Complete all the lectures and the labs in sequence

New to AWS - Certification



Recommended Track

- Certified Cloud Practitioner
- Covers the basics about AWS cloud
- Easy to prepare
- Take at test center or home
- USD 100
- Vouchers for Practice and Certification Exams

Whitepapers

- [Overview of Amazon Web Services](#)
- [Architecting for the Cloud: AWS Best Practices](#)
- [How AWS Pricing Works](#)
- [Cost Management in the AWS Cloud](#)
- [Compare AWS Support Plans](#)

Active Benefits

New to Machine Learning and/or AWS ML

- This course does not expect previous AWS experience
- Relevant concepts are introduced and explained
- Complete all the lectures and the labs in sequence

scikit-learn

Familiarity with scikit-learn - a popular machine learning library

Reference Book for scikit-learn:

Introduction to Machine Learning with Python by Andreas C. Müller and Sarah Guido

Digital version of book recommended (in color)

AWS Training

<https://aws.amazon.com/training/learning-paths/machine-learning/exam-preparation/>

- Free videos

Exams

1. Timed Practice Exam available in this course (included as part of this course)
2. AWS Practice Exam - USD 40

<https://www.aws.training/Certification>

Use free practice exam vouchers if you have taken another AWS certification exam

3. Register for Certification Exam – USD 300

Use 50% off discount vouchers if you have taken another AWS certification exam

Summary

Complete all the lectures and the labs in sequence

Supplemental resources

- Cloud Practitioner Whitepapers
- *scikit-learn - Introduction to Machine Learning with Python* by Andreas C. Müller and Sarah Guido

Chandra Lingam



50,000+ Students

Up-to-date Content



AWS Housekeeping

Account Setup, Support

Chandra Lingam

Cloud Wave LLC

Hands-on Experience

“Gain free, hands-on experience with the AWS platform, products, and services.”

<https://aws.amazon.com>

Three Types of Offers

- Always Free
- 12 months free and
- Trials

<https://aws.amazon.com/free>

<https://aws.amazon.com/free/free-tier-faqs/>

Billing

You are billed standard pay-as-you-go rates when -

- Usage exceeds free tier limits or
- Term expires

AWS requires a Credit or Debit card to sign-up for an account

Billing

Billing Alerts

Dashboard

- Free-Tier usage
- Monthly Charge Summary
- Itemized charges
- Past Bills and Usage

Free Support Center

- Account Issues
 - Billing Enquires
 - Service Limit Changes
-
- Technical Support – Part of paid plans

Service Quotas

Amazon SageMaker quotas for new accounts might be different from the default quotas listed here. If you receive an error that you've exceeded your quota, contact customer service to request a quota increase for the resources you want to use.

On-demand and Spot instance quotas are tracked and modified separately. For example, with the default quotas, you could run up to 20 training jobs with on-demand ml.m4.xlarge instances and up to 20 training jobs with Managed Spot ml.m4.xlarge instances simultaneously. Request quota increases for on-demand and spot instances separately.

Amazon SageMaker Notebooks	
Resource	Default
ml.t2.medium instances	20
ml.t2.large instances	20

https://docs.aws.amazon.com/general/latest/gr/aws_service_limits.html

<https://docs.aws.amazon.com/general/latest/gr/sagemaker.html>

Billing Alert - Best Practices

- Enable billing access to authorized users in your account
- Configure Free Tier Alerts
- Enable billing data collection for CloudWatch monitoring
- Configure Billing Alarms with CloudWatch
- Configure AWS Budget

User Accounts

Account/User	Purpose
Root Account (Highest Privilege)	Responsible for paying bills. Sign-in at https://aws.amazon.com/ Enable MFA
my_admin	IAM User with administrative access <u>Sign-in Link</u> <a href="https://<AccountId>.signin.aws.amazon.com/console">https://<AccountId>.signin.aws.amazon.com/console <a href="https://<Alias>.signin.aws.amazon.com/console">https://<Alias>.signin.aws.amazon.com/console
ml_user	Administrative user – Maintained for course backward compatibility
ml_user_predict	Limited privilege – Read only access to SageMaker, Invoke Prediction, S3 Read only access

MFA Setup

Recommended for root account

Login credentials + one-time passwords

- Google Authenticator App or similar

AWS Command Line Interface (CLI)

- Install [AWS CLI](#) in your laptop
- Used later for demonstrating invocation of prediction service from outside of AWS
- Configure CLI with ml_user_predict (region: us-east-1)

```
aws configure --profile ml_user_predict
```

- Verify Access – List S3 Buckets in your account

```
aws s3 ls --profile ml_user_predict
```

Summary

Account Setup

Types of free offers

Billing Dashboard and Alerts

Support

IAM Users

Chandra Lingam



50,000+ Students

Up-to-date Content



Cloud Computing Advantages

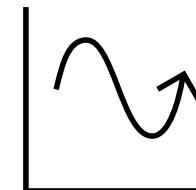
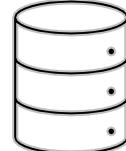
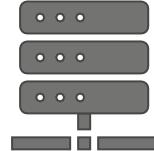
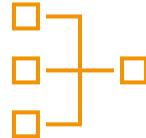
Chandra Lingam

Cloud Wave LLC

#1 - Trade Capital Expense for Variable Expense

No need to purchase expensive capital equipment

Pay only for what you consume and how much you consume



Consumption Based Pricing

Fear of billing surprises

AWS Tools

- Detailed break-down of charges
- Budget based alerts
- Restrict specific services and resources

Is consumption-based model cheaper than buying your own?

#2 - Benefit from Massive Economies of Scale

Massive scale

Shared infrastructure used by thousands of customers

Better utilized

Lower pay-as-go prices

#3 - Stop Guessing about Capacity

Long-term planning results in over capacity or under capacity

Eliminate guessing on your infrastructure capacity needs

Match infrastructure for actual need

Scale up or down with only a few minutes notice

#4 - Increase Speed and Agility

New resources are only a click away

Developers can get resources in minutes instead of waiting for weeks

Cost to experiment is significantly lower

Hourly pricing model – try new products at very low cost

#5 - Stop spending money running and maintaining data centers

Avoid undifferentiated heavy lifting

Focus on projects that differentiate your business, not infrastructure



#6 - Go Global in Minutes

Deploy application close to customer for lower latency and compliance requirements

Better end user experience at minimal cost



AWS Global Infrastructure

Region

Availability Zone

Edge Location

ABOUT AWS

About AWS >

Global Infrastructure >

What's New >

AWS in the News >

Events & Webinars >

RELATED LINKS

What is Cloud Computing?

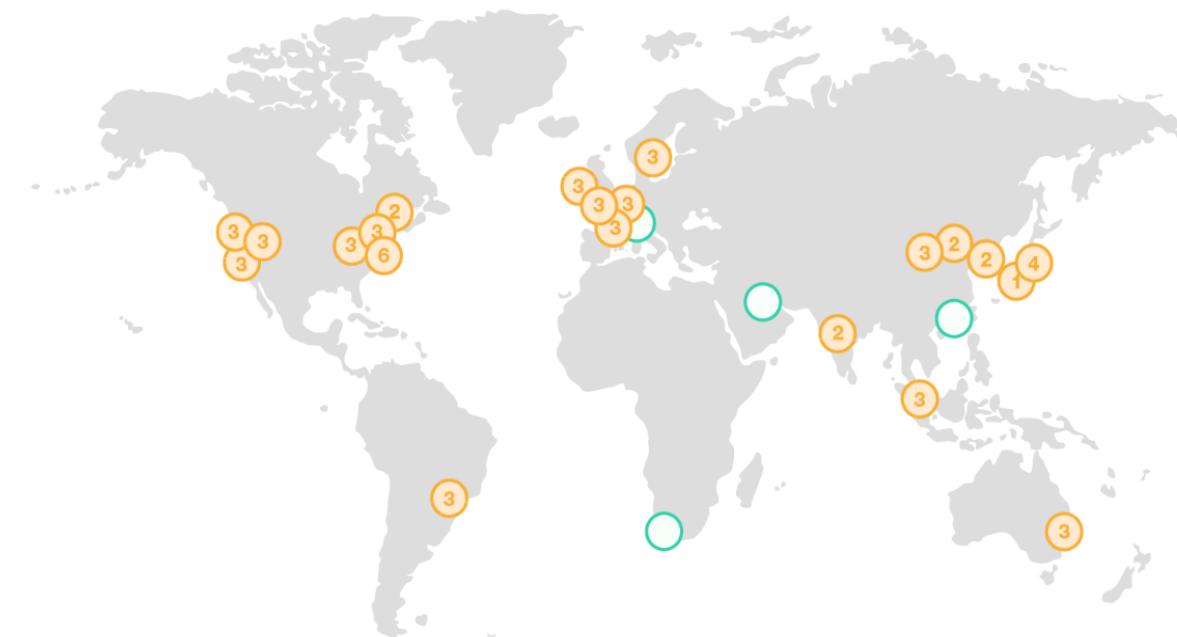
AWS Free Usage Tier

AWS Blog

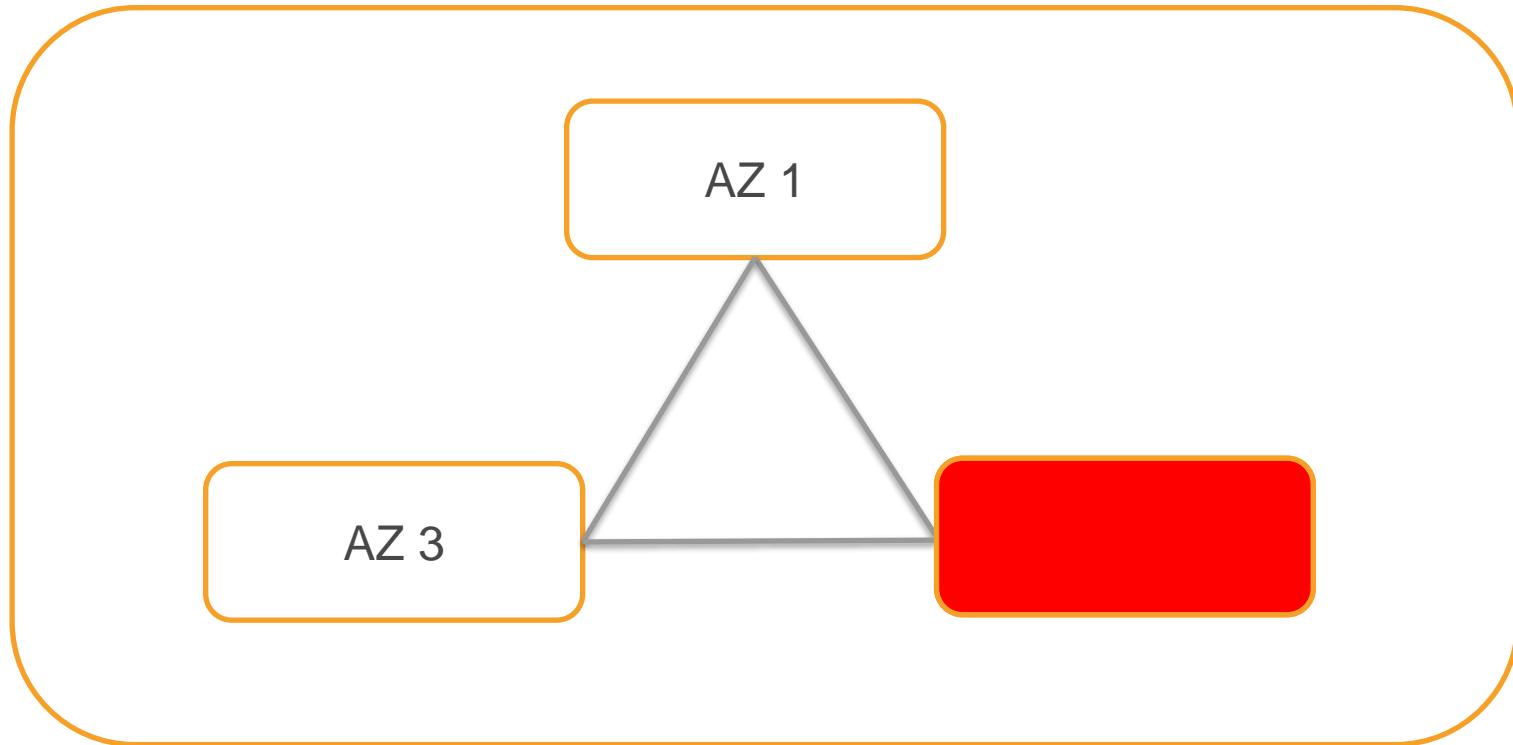
AWS Careers

AWS Training

Global Infrastructure



Region



Application should be spread across two or more availability zones

Azure Outage Proves the Hard Way that Availability Zones are a Good Idea

<https://www.datacenterknowledge.com/microsoft/azure-outage-proves-hard-way-availability-zones-are-good-idea>

“Lightning during a powerful storm caused a voltage swell in the utility feeds powering one of the Azure data centers in San Antonio, Texas, that overwhelmed the facility’s surge suppressors, knocking out its cooling systems”

“A significant number of storage servers were damaged, as well as a small number of network devices and power units.”

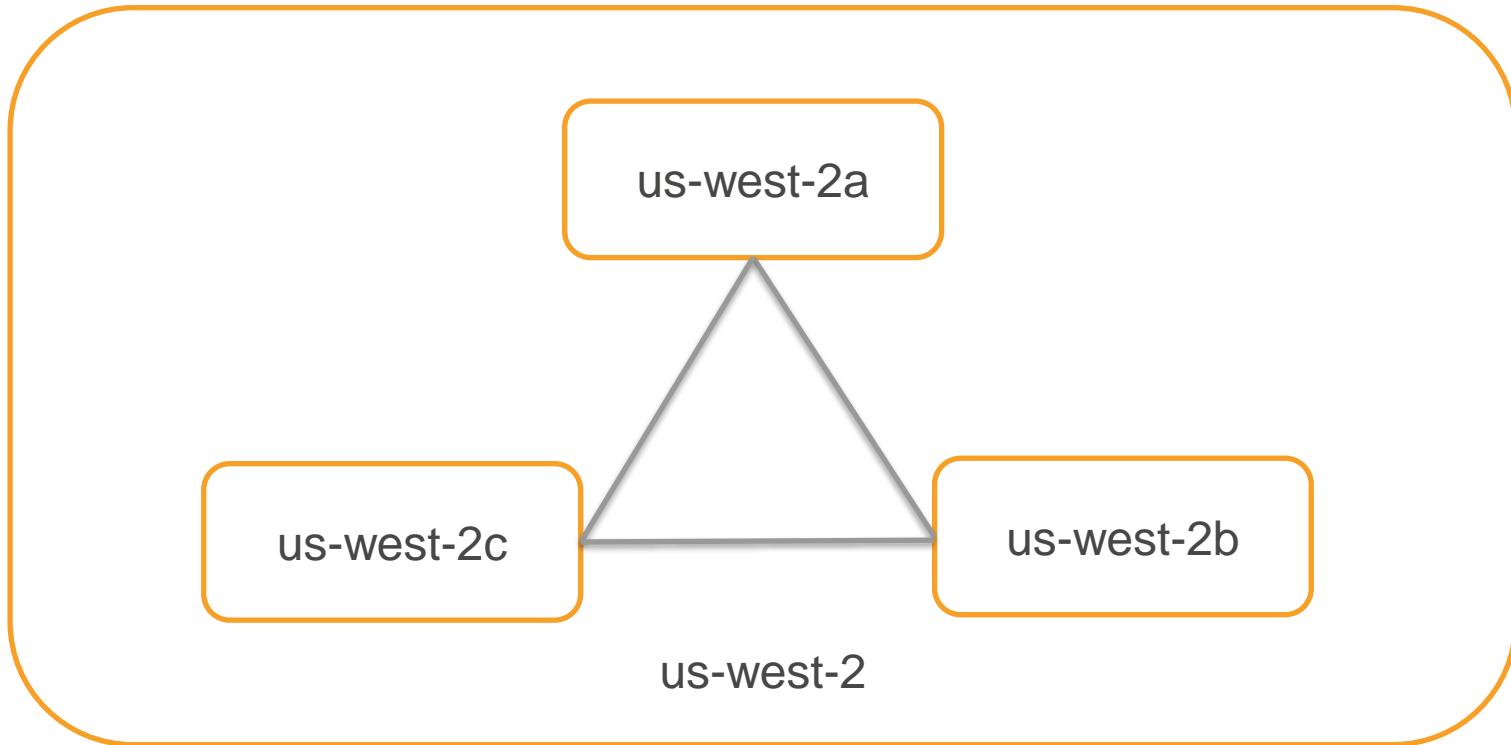
AWS Services and Multi-AZ

S3 maintains redundant copies across at least three AZs

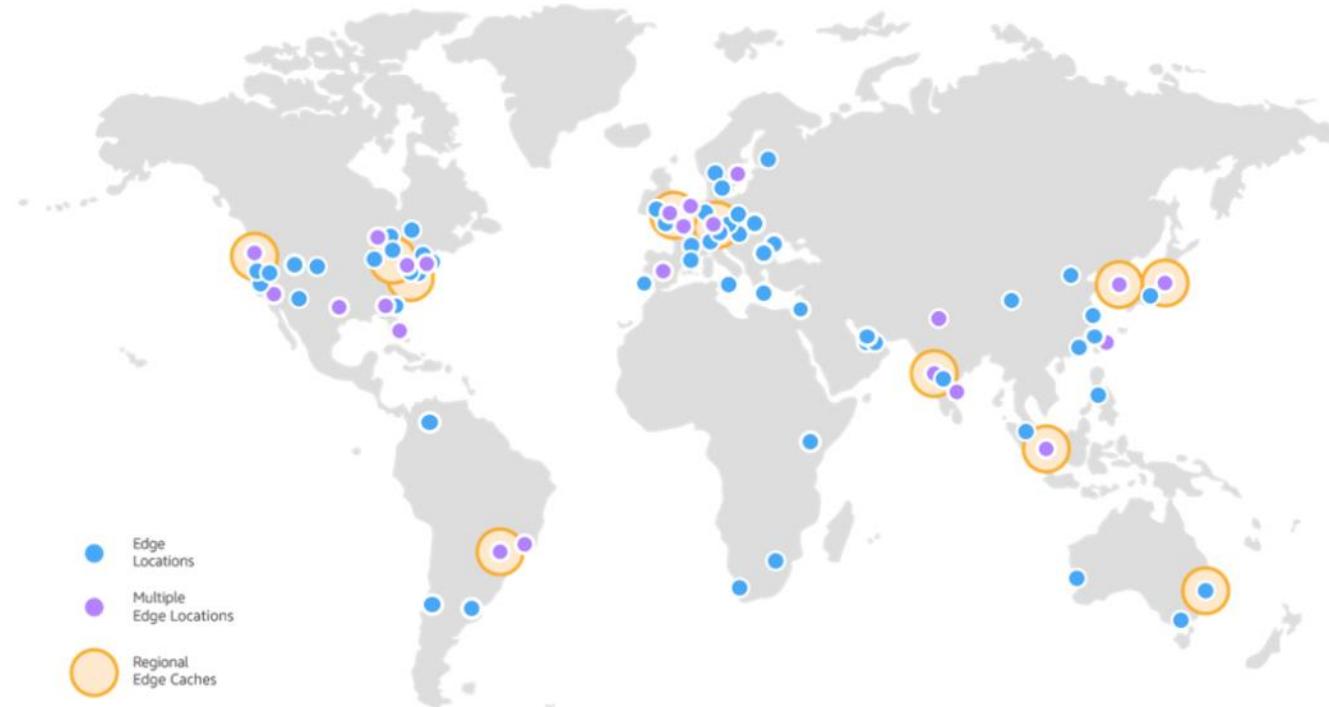
RDS in high availability mode maintains primary and standby in separate AZs

EC2 instance runs in a specific AZ. If AZ goes down, the instance is impacted

Oregon Region



Edge Locations



Source: <https://aws.amazon.com/cloudfront/features/>

Data Movement

AWS stores your data in a region that you choose

Data is not copied between regions – unless customer asks AWS to do so (including edge location)

Comply with data hosting requirement

Machine Learning

Concepts

Content Prepared By: Chandra Lingam, Cloud Wave LLC

Copyright © 2019 Cloud Wave LLC. All Rights Reserved.

All other registered trademarks and/or copyright material are of their respective owners



Machine Learning

Supervised

Unsupervised

Reinforcement

Supervised Learning

attributes

target

sepal_length	sepal_width	petal_length	petal_width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
5.5	2.3	4	1.3	Iris-versicolor
6.5	2.8	4.6	1.5	Iris-versicolor
5.8	2.8	5.1	2.4	Iris-virginica
6.4	3.2	5.3	2.3	Iris-virginica

Supervised Learning

What type of Iris plant is this?

attributes				
sepal_length	sepal_width	petal_length	petal_width	
6.5	3	5.5	1.8	

ML Predicted Answer: Iris-virginica

Model Comparison

Is this email a spam?

Buy a new home with a low down payment. Our 30 year fixed mortgage can give you the flexibility you need to ...

Model A

Yes, It is!

Model B

No, It is NOT!

Terminologies

Label or Target

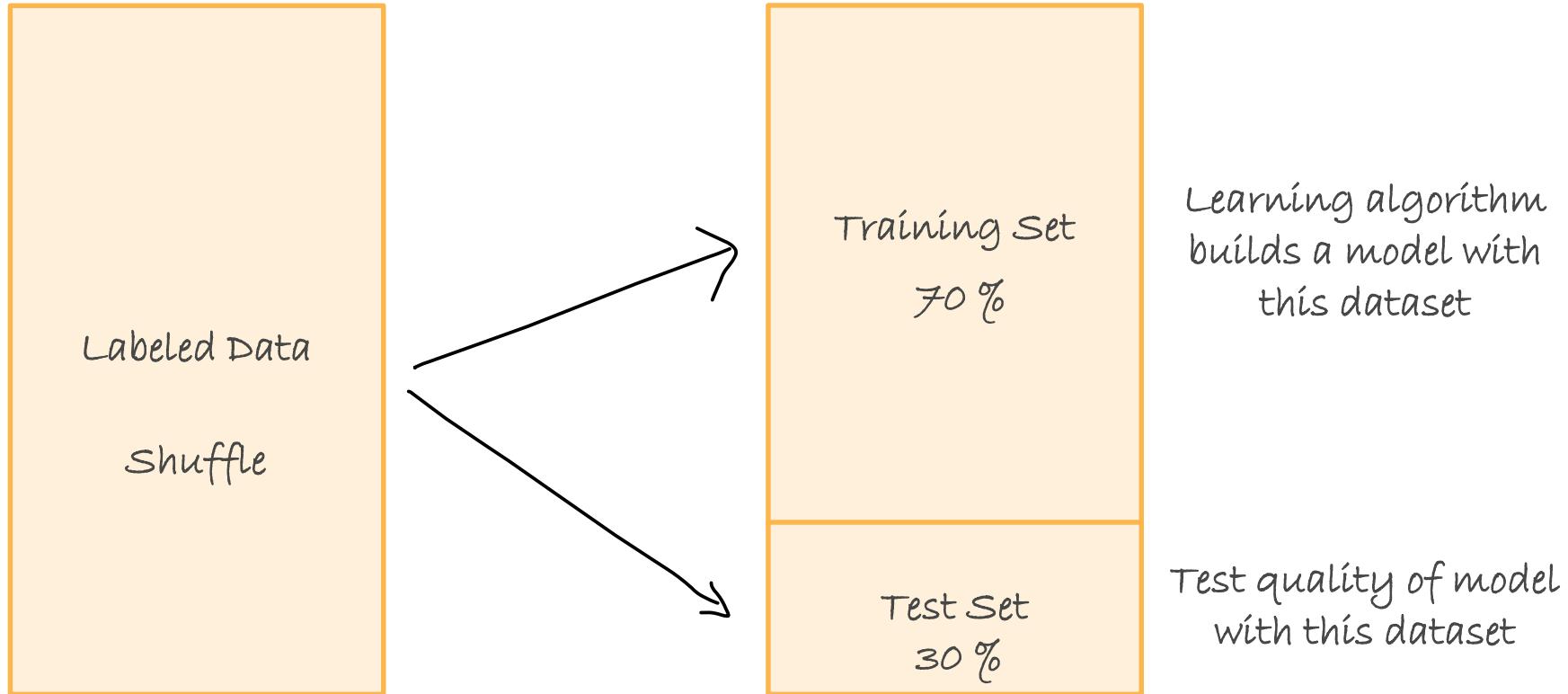
Features or Attributes

sepal_length	sepal_width	petal_length	petal_width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
5.5	2.3	4	1.3	Iris-versicolor
6.5	2.8	4.6	1.5	Iris-versicolor
5.8	2.8	5.1	2.4	Iris-virginica
6.4	3.2	5.3	2.3	Iris-virginica

Example, Sample, Instance, or Observation

Labeled Data

Terminologies



Supervised Algorithm Types

Regression

Binary Classification

Multi-Class Classification

Regression

Used for predicting numeric output

How much is my home worth?

How many passengers are going to travel by air this year?

Binary Classification

Used for predicting a binary output or two classes - We need a YES or NO answer

Is this email a spam?

Does this social media post require a follow-up?

Is this patient showing symptoms of a disease?

Multi-Class Classification

Used for predicting one out of several outcomes

How is the weather in NY tomorrow?

[Sunny, Windy, Cloudy, Rainy, Snow, ...]

What ad should be displayed for this search?

[Sport, Real Estate, Home Loan, Auto, ...]

Unsupervised Learning

Only data – There is no defined target

Group similar observations

Anomaly Detection

Words used in similar context

Unsupervised Algorithms

Clustering

Dimensionality Reduction

Group words that are used in similar context or have similar meaning

Reinforcement Learning

Decision Making under uncertainty

Autonomous Driving

Games

Reinforcement uses Reward Functions to reward correct decision and punish incorrect decision

Data Types

Data in Real Life

Numeric

Text

Categorical

Categorical

Day of Week	Sales
Sunday	100
Monday	50
Tues	20
Wed	30
Thursday	25
Fri	35
Sat	110

Categorical (Numeric Encoding)

Day of Week	Sales
1	100
2	50
3	20
4	30
5	25
6	35
7	110

Sunday=1, Monday=2, Tuesday=3, ...



Categorical – One Hot Encoding

Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sales
1	0	0	0	0	0	0	100
0	1	0	0	0	0	0	50
0	0	1	0	0	0	0	20
0	0	0	1	0	0	0	30
0	0	0	0	1	0	0	25
0	0	0	0	0	1	0	35
0	0	0	0	0	0	1	110

Categorical

Cartesian Transformation - Combine categorical features to form new features

Day of Week	Weather	Sales
Sunday	Sunny	100
Monday	Cloudy	50
Tues	Snow	20
Wed	Snow	30
Thursday	Snow	25
Fri	Rain	35
Sat	Sunny	110

Cartesian Transformation

Day of Week_Weather	Sales
Sunday_Sunny	100
Sunday_Cloudy	75
Sunday_Snow	15
Sunday_Rain	25

Text Data

Movie	Genre
Star Wars	Adventure
Notting Hill	Romance
Star Trek	Adventure

Star	Wars	Notting	Hill	Trek	Genre
1	1	0	0	0	Adventure
0	0	1	1	0	Romance
1	0	0	0	1	Adventure

Text Data

NGRAM Transformation

Orthogonal Sparse Bigram (OSB) Transformation

Lowercase Transformation

Remove Punctuation Transformation

Cartesian Transformation

Words Alter Meaning

“this is working. not disappointed”

“this is not working. disappointed”

After tokenization:

[‘disappointed’, ‘is’, ‘not’, ‘this’, ‘working’]

NGRAM

“this is working. not disappointed”

“this is not working. disappointed”

After NGRAM Transformation (window = 2):

[‘this is’, ‘is working’, ‘working not’, ‘not disappointed’]

*[‘this is’, ‘is not’, ‘**not working**’, ‘working disappointed’]*

Stemming

All these words are treated differently:

`['working', 'worked', 'works']`

After stemming - words have same root

`['work', 'work', 'work']`

Lower Case

How is the request rate LIMIT Determined?

HOW is the request rate limit determined?

After Lower Case Transformation:

how is the request rate limit determined?

Numeric Data

Numeric value as-is (for linear relationship)

Normalization Transformation (for linear relationship)

Binning Transformation - convert to categorical (for non-linear relationship)

Handling Missing Values

Impute - If there is a reasonable way to fill-in the missing-values, you should do it!

Drop missing features or observations

Knowledge about data and context is important!

References: Missing Values

Handling Missing Values in Time Series -

<https://www.kaggle.com/juejuewang/handle-missing-values-in-time-series-for-beginners>

Working with missing data - https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html

Substitute attribute – Boolean attribute that is set to 1 to indicate if another attribute has missing value: [Treatment of missing values](#)

Model Performance

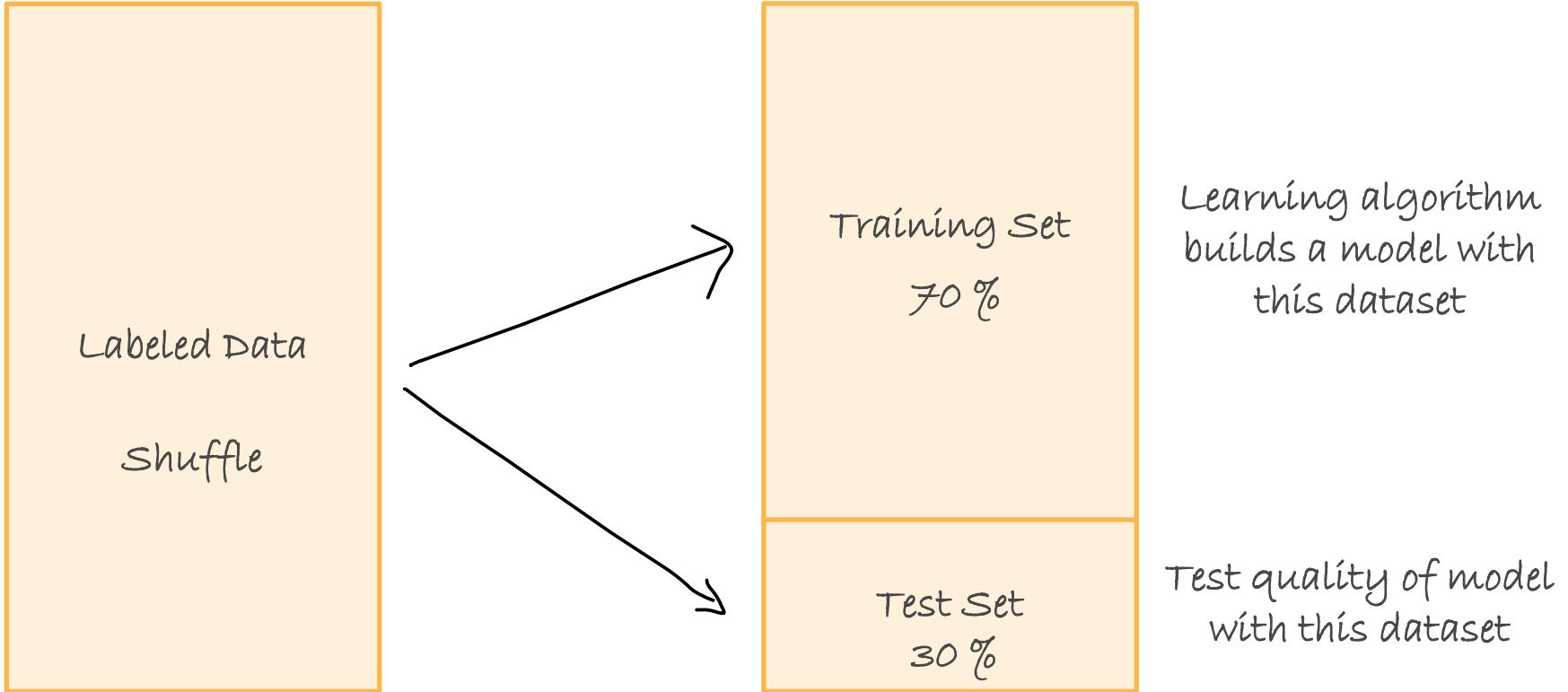
Quality Metrics

Content Prepared By: Chandra Lingam, Cloud Wave LLC

Copyright © 2020 Cloud Wave LLC. All Rights Reserved.

All other registered trademarks and/or copyright material are of their respective owners

Terminologies



Verify Model Fit

Fit	Issue
<u>Underfitting</u>	Poor performance on both training and test set
<u>Overfitting</u>	Good performance on training set Poor performance on test set
<u>Balanced</u>	Good performance on training set and test set

Supervised Algorithm Types

Algorithm	Output
Regression	Continuous Numeric
Binary	Binary
Multi-class Classification	Categorical – One of many possible known outcomes

Regression

Root Mean Square Error (RMSE)

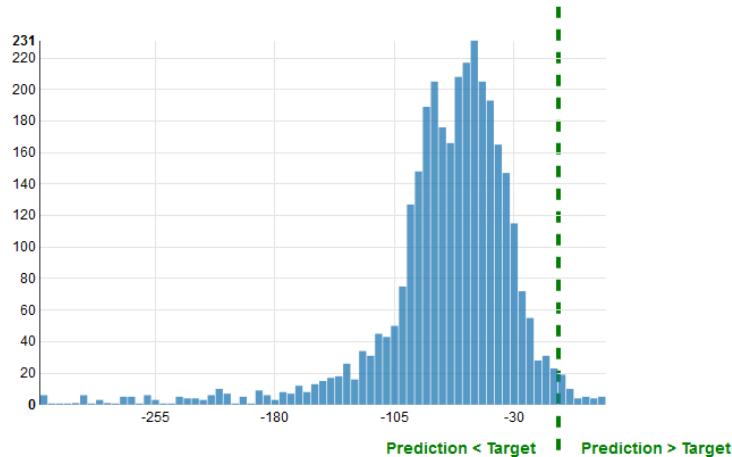
$$RMSE = \sqrt{1/N \sum_{i=1}^N (actual\ target - predicted\ target)^2}$$

<https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html>

Regression

Residual Histogram

Select Bin Width:



Difference = Predicted – Target

Plot the difference as a histogram

A good model has balanced over
and under predictions

<https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html>

Binary Classifier

Positive Class – Condition that we are interested in detecting

Admit Student?

Positive Class = *Admitted*

Negative Class = *Not-Admitted*

Binary Classifier

Positive Class – Condition that we are interested in detecting

Is this patient at risk of developing heart disease?

Positive Class = Heart Disease

Negative Class = Normal

Binary Classifier

Positive Class – Condition that we are interested in detecting

Is this email a SPAM?

Positive Class = Spam

Negative Class = Normal

Binary Classifier Output

Some Algorithms directly produce a binary output

Some Algorithms produce a raw-score that gives a probability of an example belonging to positive class

Raw Score to Binary Class

Raw Score	Class
0	Negative
0.1	Negative
0.2	Negative
0.3	Negative
0.4	Negative
0.5	Positive
0.6	Positive
0.7	Positive
0.8	Positive
0.9	Positive
1	Positive

← cut-off, Threshold

Email Spam Classifier

Raw Score	Class
0	Negative
0.1	Negative
0.2	Negative
0.3	Negative
0.4	Negative
0.5	Negative
0.6	Negative
0.7	Negative
0.8	Positive
0.9	Positive
1	Positive

Positive: Spam

Negative: Normal

Increasing Cut-off

1. Reduces possibility of a Normal email marked as spam (Reduces False Positive)
2. Increases possibility of a Spam marked as normal email (Increases False Negative)



Cut-off, Threshold

Identify Patients at risk for a disease

Raw Score	Class
0	Negative
0.1	Negative
0.2	Negative
0.3	Positive
0.4	Positive
0.5	Positive
0.6	Positive
0.7	Positive
0.8	Positive
0.9	Positive
1	Positive



Cut-off, Threshold

Positive: At-Risk

Negative: Normal

Lowering Cut-off

1. Reduces possibility of missing an at-risk patient (Increased Recall)
2. Increases possibility of normal person flagged as at-risk (Increase in False-Alarm)

Confusion Matrix - Concepts

Terminology	Description
Positive	Total Actual Positives = True Positive + False Negative
Negative	Total Actual Negatives = True Negative + False Positive
True Positive	How many samples were correctly classified as Positive
True Negative	How many samples were correctly classified as Negative
False Negative	How many positive samples were mis-classified as negative
False Positive	How many negative samples were mis-classified as positive

True Positive Rate

Fraction of Positives predicted correctly

$$TPR = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

TPR is also referred as Recall, Probability of detection

Model with Recall closer to 1 is good. Model with Recall closer to 0 is poor.

True Negative Rate

Fraction of Negatives predicted correctly

$$TNR = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

Model with TNR closer to 1 is good. Model with TNR closer to 0 is poor.

False Positive Rate

Fraction of negatives mis-classified as positives.

$$FPR = \frac{False\ Positive}{True\ Negative + False\ Positive}$$

FPR is also referred as Probability of false alarm

Model with FPR closer to 0 is good. Model with FPR closer to 1 is poor.

False Negative Rate

Fraction of positives mis-classified as negatives.

$$FNR = \frac{\text{False Negative}}{\text{True Positive} + \text{False Negative}}$$

FNR is also referred as misses.

Model with FNR closer to 0 is good. Model with FNR closer to 1 is poor.

Precision

Fraction of true positives among all predicted positives. Larger value indicates better predictive accuracy

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Good model has precision closer to 1. Poor model has precision closer to 0

Accuracy

Fraction of correct predictions. Larger value indicates better predictive accuracy

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Negative + False\ Positive + True\ Negative}$$

Good model has accuracy closer to 1

F1 Score

Harmonic Mean of Precision and Recall

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Good model has F1 score closer to 1

Area Under Curve (AUC)

AUC is the area of a curve formed by plotting True Positive Rate against False Positive Rate at different cut-off thresholds

Good model have AUC closer to 1

0.5 is considered random guess

Closer to 0 is unusual and it indicates model is flipping results

Summary

Don't depend on one metric

For classification, recall (true positive rate) and precision along with F1 Score are often used together

Corner Cases:

Recall = very high when a model classifies all samples as positive

Precision = very high when a model classifies one positive correctly and misclassifies all other samples as negative

F1 Score = balances these corner cases

Algorithms and Frameworks

AWS SageMaker

Chandra Lingam

Cloud Wave LLC

SageMaker – Training and Hosting

Options	Usage Scenario
Built-in Algorithms	Training algorithms provided by SageMaker Easy to use and scale Optimized for AWS Cloud
Pre-built Container Images	Supports popular frameworks like MxNet, TensorFlow, scikit-learn, PyTorch Flexibility to use wide selection of algorithms
Extend Pre-built Container Images	Extend pre-built container images to your needs
Custom Container Images	Custom algorithm, language and framework

Built-in Algorithms

- Training algorithms provided by SageMaker
- Easy to use and scale
- Optimized for AWS Cloud
- GPU Support

BlazingText

Type	Purpose	Use
Unsupervised	Convert Word to vector (Word2Vec)	<p>Word2Vec is a text preprocessing step for downstream NLP, Sentiment analysis, named entity recognition and translation.</p> <p>Words that are semantically similar have vectors that are closer to each other</p> <p>Example: All vegetable names are closer to each other in the vector space</p>
Supervised	Multi-class, Multi-label Classification	<p>Classification based on Text (single-label)</p> <p>Example: Spam Detection – Spam/Not-Spam</p> <p>A single instance can belong to many classes (multi-label)</p> <p>Example: movie can belong to multiple genre</p>

Reference: SageMaker BlazingText and <https://fasttext.cc/>

Object2Vec

Type	Purpose	Use
Supervised	Classification, Regression	Extends Word2Vec Captures structure of sentences Learns relationship between pair of objects Example: similarity search based on Customer-Product, Movie-Ratings and so forth

Factorization Machines

Type	Purpose	Use
Supervised	Regression, Classification	<p>Works very well with high dimensional sparse datasets</p> <p>Popular algorithm for building Recommender systems</p> <p>Collaborative Filtering</p> <p>Example: Movie Recommendation based on your viewing habits; Cross recommend based on similar users</p>

K-Nearest Neighbors

Type	Purpose	Use
Supervised	Regression, Classification	Classification – Queries K-Nearest Neighbors and assigns majority class for the instance Regression - Queries K-Nearest Neighbors and returns average value for the instance Does not scale well for large datasets

Linear Learner

Type	Purpose	Use
Supervised	Regression, Classification	Linear models for regression, binary classification and multi-class classification

XGBoost

Type	Purpose	Use
Supervised	Regression, Classification	Gradient Boosted Trees Algorithm Very Popular Algorithm - Won several competitions

DeepAR

Type	Purpose	Use
Supervised	Timeseries Forecasting	<p>Train multiple related time series using a single model</p> <p>Generate predictions for new, similar timeseries</p>

Object Detection

Type	Purpose	Use
Supervised	Classification	<p>Image Analysis Algorithm</p> <p>Detects and Classifies Objects in an Image</p> <p>Returns a bounding box of each object location</p>

Object Detection

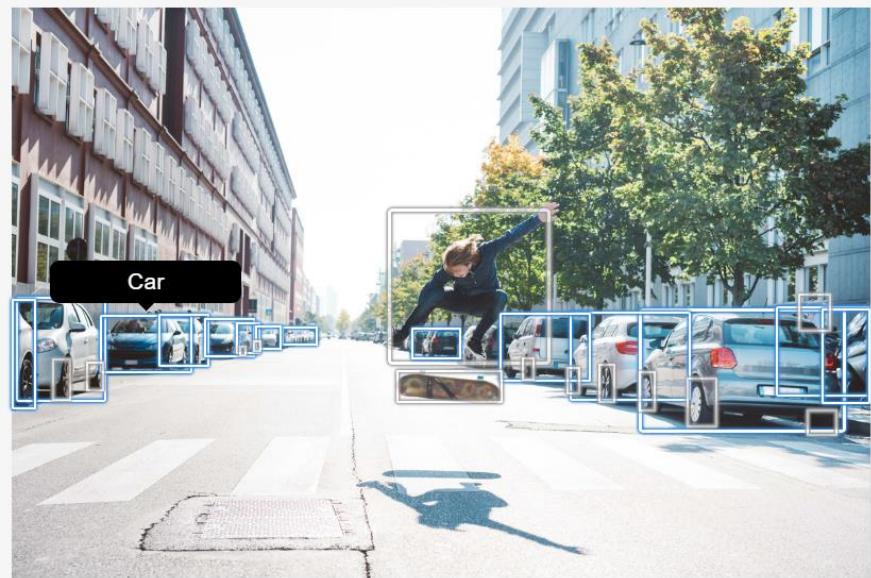
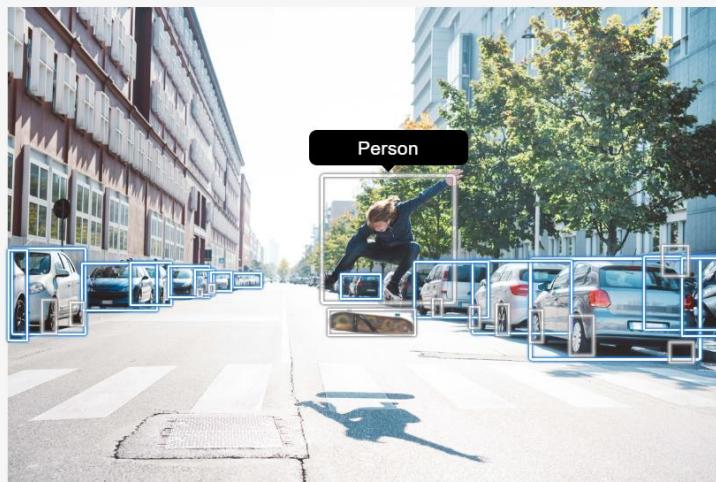


Image Courtesy: Amazon Rekognition

Image Classification

Type	Purpose	Use
Supervised	Classification	<p>Image Analysis Algorithm</p> <p>Classifies entire Image</p> <p>Supports multi-labels</p>

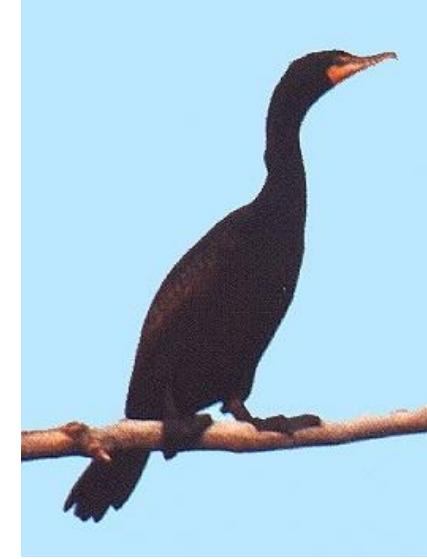
Image Classification



Bear



Butterfly



Bird

Image Courtesy: Caltech 256 Dataset

http://www.vision.caltech.edu/Image_Datasets/Caltech256/

Semantic Segmentation

Type	Purpose	Use
Supervised	Classification	<p>Image Analysis Algorithm for Computer Vision Applications</p> <p>Tags each pixel in an image with a class label</p> <p>Example: Identify shape of car</p>

Semantic Segmentation



Cars

Bus

People

Image Courtesy: COCO Dataset
<http://cocodataset.org/#explore>

Sequence to Sequence (seq2seq)

Type	Purpose	Use
Supervised	Convert sequence of tokens	<p>Input: Sequence of tokens Output: Another sequence of tokens</p> <p>Examples: Text Summarization, Language Translation, Speech to Text</p>

K-Means

Type	Purpose	Use
Unsupervised	Clustering	<p>Identify discrete groupings within data</p> <p>“Members of a group are as similar as possible to one another and as different as possible from members of other groups”</p>

Latent Dirichlet Allocation (LDA)

Type	Purpose	Use
Unsupervised	Topic Modeling	<p>Group documents by user specified “number” of topics</p> <p>For documents, it assigns a probability score for each topic</p>

Neural Topic Model (NTM)

Type	Purpose	Use
Unsupervised	Topic Modeling	Similar to LDA

Principal Component Analysis (PCA)

Type	Purpose	Use
Unsupervised	Dimensionality Reduction	<p>“Reduces dimensionality of a dataset while retaining as much information as possible”</p> <p>“Returns components – a new set of features that are composites of original features and that are uncorrelated to one another”</p> <p>Examples: Reduce the dimensions of a dataset, visualize high dimensional datasets, remove highly correlated features</p>

Random Cut Forest (RCF)

Type	Purpose	Use
Unsupervised	Anomaly Detection	<p>“Anomalous points are observations that diverge from otherwise well-structured or patterned data”</p> <p>For each data point, RCF assigns an anomaly score</p> <p>Low score indicates normal data and high score indicates an anomaly</p>

IP Insights

Type	Purpose	Use
Unsupervised	Detect unusual network activity	<p>Learns from (entity, IPv4 address) pairs</p> <p>Entity can be Account ID, User ID</p> <p>For a given entity, IPv4 address pair, it returns a score</p> <p>High score indicates unusual event – a website can trigger MFA</p>

SageMaker Ground Truth and Neo

SageMaker Ground Truth

Automatic Labeling

- Learns based on examples you provide
- Very cost-effective

Manual Labeling

- Human Labelers – Mechanical Turk
- Manages workflow

SageMaker Neo

- Run Machine Learning Algorithms anywhere in the Cloud and at Edge Locations
- Latency is critical
- Cross Compilation capability that can optimize your algorithms to run on:
 - Intel
 - NVIDIA
 - ARM
 - And other hardware

Support for ML Frameworks

- Use SageMaker to train and host models using popular frameworks
- SageMaker provides built-in container images for Apache MxNet, TensorFlow, scikit-learn, PyTorch, Chainer, SparkML and more

Containers

“Amazon SageMaker makes extensive use of *Docker containers* for build and runtime tasks”

“Amazon SageMaker provides pre-built Docker images for its built-in algorithms and the supported deep learning frameworks used for training and inference. By using containers, you can train machine learning algorithms and deploy models quickly and reliably at any scale”

Reference: <https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms.html>

Use Apache Spark with SageMaker



- SageMaker Apache Spark Library in Python and Scala
- Directly read DataFrames in Spark Clusters
- SageMakerEstimator automatically converts DataFrames to Protobuf format
- Train and Host using SageMaker

Popular Framework Support

- TensorFlow
- MxNet
- scikit-learn
- PyTorch
- Chainer
- SparkML

SageMaker provides SDKs and pre-built docker images to train and host models using above frameworks

Use Your own algorithms

- Host your custom algorithms on SageMaker
- Use a runtime and language of your choice
- Build Containers that conform to SageMaker Specification
- Train and Host on SageMaker

Deep Learning AMIs

- Launch EC2 instances preconfigured with all the tools and Deep Learning Framework
- Modify DL frameworks or extend them
- Contributors to DL frameworks
- Troubleshooting framework level issues

SageMaker – Training and Hosting

Options	Usage Scenario
Built-in Algorithms	Training algorithms provided by SageMaker Easy to use and scale Optimized for AWS Cloud
Pre-built Container Images	Supports popular frameworks like MxNet, TensorFlow, scikit-learn, PyTorch Flexibility to use wide selection of algorithms
Extend Pre-built Container Images	Extend pre-built container images to your needs
Custom Container Images	Custom algorithm, language and framework

Chandra Lingam



50,000+ Students

Up-to-date Content



SageMaker

Overview, Pricing, Data Format

Content Prepared By: Chandra Lingam, Cloud Wave LLC

Copyright © 2019 Cloud Wave LLC. All Rights Reserved.

All other registered trademarks and/or copyright material are of their respective owners



Introduction to SageMaker

Fully managed Cloud based machine learning service

Build – Jupyter Notebook development environment

Train – Managed Training infrastructure

Deploy – Scalable Hosting infrastructure

AWS SageMaker - Build

Managed Jupyter Notebook Environment

Extensive collection of popular Machine Learning Algorithms

Pre-configured to run TensorFlow and Apache MxNet

Bring-Your-Own Algorithm

AWS SageMaker - Train

Distribute training across one or many instances

Managed model training infrastructure

Scales to Petabyte datasets

Compute instances for training are automatically launched and released – Stores artifacts in S3

AWS SageMaker - Deploy

Realtime prediction

Batch Transform



Deploy for Realtime Predictions

Realtime Endpoint for interactive and low-latency use-cases

AutoScaling

- Maintain adequate capacity
- Replace unhealthy instances
- Dynamically scale-out and scale-in based on workload

Deploy for Batch Transforms

Batch Transform for non-interactive use-cases

Suitable for these scenarios:

- Inference for your entire dataset
- Don't need a persistent real-time endpoint
- Don't need sub-second latency performance

SageMaker manages resources for batch transform



SageMaker Instance Family

Instance Family	Strength/Uses
<u>Standard</u>	Balanced CPU performance
<u>Compute Optimized</u>	Highest CPU performance
<u>Accelerated Computing</u>	Graphics/GPU Compute
<u>Inference Acceleration</u>	Fractional GPUs (add-on)



Standard Instance Family

Balanced CPU, Memory and Network Performance

Example: T2, T3, M5

T type instances – Suitable for occasional burst. Perfect for Notebook and Development Systems

M type instances – Suitable for sustained load. Perfect for CPU intensive model training and hosting



Compute Optimized Family

Latest Generation CPUs. Higher Performance Systems

Example: C4, C5

Suitable for sustained load

Perfect for CPU intensive model training and hosting

Accelerated Computing Family

Powerful GPUs

Speed-up Algorithms optimized for GPUs

Example: P2, P3

Reduce time needed for training using GPUs. Perfect for GPU intensive model training and hosting

Inference Acceleration

Add-on Fractional GPUs

Some Algorithms are GPU intensive during Training but need only fractional GPU during Inference

Add GPU to lower cost Standard and Compute Optimized Instances

Perfect for speeding up inference using GPUs



Suggested Instance Types

Standard, Compute Optimized – Good for algorithms optimized for CPUs

Accelerated Computing – Good for algorithms optimized for GPUs

Choose a family first and then experiment various instance sizes – Simple AWS configuration change

Instance Type and Size



[ml.c5.2xlarge](#) = Compute Optimized, 5th generation, 2xlarge (8 vCPUs, 16 GB Memory)

SageMaker Pricing Components

Instance Type and Size

Fractional GPUs

Storage

Data Transfer

AWS Region



SageMaker Free Tier

Two months [free tier](#) – starts from the first month you create a SageMaker resource

Development – 250 Hours/Month t2.medium or t3.medium

Train – 50 Hours/Month m4.xlarge or m5.xlarge

Deploy – 125 Hours/Month m4.xlarge or m5.xlarge

Development – On Demand Pricing

Instance + Fractional GPU Hourly Cost (pro-rated to the nearest second with a 1 minute minimum)

Storage – USD 0.14 per GB/Month

Data Transfer IN, OUT – USD 0.016 per GB

Training – On Demand Pricing

Instance Hourly Cost (pro-rated to the nearest second with a 1 minute minimum)

Storage – USD 0.14 per GB/Month

Instances are automatically launched and terminated

You are charged only for the duration the training job ran

Realtime Inference – On Demand Pricing

Instance + Fractional GPU Hourly Cost (pro-rated to the nearest second with a 1 minute minimum)

Storage – USD 0.14 per GB/Month

Data Transfer IN, OUT – USD 0.016 per GB

Batch Transform – On Demand Pricing

Instance + Fractional GPU Hourly Cost (pro-rated to the nearest second with a 1 minute minimum)

Storage – USD 0.14 per GB/Month

Data Transfer IN, OUT – USD 0.016 per GB

You are charged only for the duration batch transform job ran

SageMaker Data Formats

Training Data Format

CSV

RecordIO

[Algorithm specific formats](#) (LibSVM, JSON, Parquet)

Data needs to be stored in S3

Inference Format

CSV

JSON

RecordIO

Data

Entire Dataset in a single file

Split across several files in a folder

Data Copy from S3 to Training Instance

File Mode:

- Training job copies entire dataset from S3 to training instance
- Space Needed: Entire data set + Final model artifacts

Pipe Mode:

- Training job streams data from S3 to training instance
- Faster start time and Better Throughput
- Space Needed: Final model artifacts



ML Terminology

Training Data – Used for training a model

Validation Data – Used for verifying training accuracy and for optimizing parameters

Test Data – Used for verifying accuracy of a built-up model
(last step)

Data needs to be stored in S3



S3 Data Source Configuration

Attribute	Values/Purpose
<u>S3DataDistributionType</u>	FullyReplicated – entire dataset is replicated on each training instance ShardedByS3Key – Subset of data is replicated on each training instance. If dataset is split across multiple S3 objects, then SageMaker will distribute equal number of S3 objects to each training node.
<u>S3DataType</u>	ManifestFile – S3Uri points to a file that in-turn contains a list of files to be used for training S3Prefix – S3Uri points to a prefix. SageMaker uses all the objects with the specified prefix
<u>S3Uri</u>	Identifies a Key name prefix or a manifest file

Cloud Practitioner Review – Infrastructure and Pricing

Prepared By: [Chandra Mohan Lingam](#), Cloud Wave LLC

Benefits of Cloud Computing

1. Trade capital expense for variable expense
 - a. No need to purchase expensive hardware in the cloud
 - b. Pay based on usage
2. Benefit from massive economies of scale
 - a. AWS buys in bulk as it needs to support 1000s of customers
 - b. Shared infrastructure – reduces idle resources
 - c. Better utilized
 - d. Lower pay as you go pricing
3. Stop Guessing about capacity
 - a. Scale up or down only with a few minutes notice
 - b. Match infrastructure with actual need
4. Increase speed and Agility
 - a. In the cloud, new resources are only a click away
 - b. Hourly pricing – try new products at a low cost
5. Stop spending money running and maintaining data centers
 - a. Avoid undifferentiated heavy lifting.
 - b. In a traditional on-premises data center, you need to spend upfront money on purchasing physical servers, storage, networking equipment, datacenters, and so forth. Whereas, in the cloud, you pay only for what you use
 - c. Focus on projects that differentiate your business
6. Go Global in Minutes
 - a. Deploy application close to the users
 - b. Use AWS regions, Edge Locations

AWS Global Infrastructure

AWS Global Infrastructure consists of multiple regions. There are 24 regions (and expanding) currently

For example, US West (Oregon), US West (Northern California), US East (Ohio), US East (N. Virginia), Europe (Frankfurt), Europe (Ireland), Asia Pacific (Tokyo), Asia Pacific (Mumbai) and so forth

All regions are global/public regions – that means you can spin up EC2 instances and other resources in any of the regions

However, there are two special regions

AWS GovCloud is only accessible to US entities that pass a screening process. Designed to host sensitive and regulated workloads to meet US Government security and compliance requirements

AWS China region complies with China's legal and regulatory requirements. This is a separate account that limits access only to AWS China regions

AWS Region

Each region consists of multiple, isolated, and physically separate Availability Zones within a geographic area.

For example, US West (Oregon) region has four availability zones within 60 miles of each other.

Oregon Region Name: us-west-2

Oregon Availability Zones: us-west-2a, us-west-2b, us-west-2c, us-west-2d

Availability Zone

An availability zone (AZ) consists of one or more data centers with redundant power, networking, and connectivity in an AWS region

All availability zones (AZ) in a region are interconnected with high speed, high bandwidth, and low latency networking

An architecture best practice is to deploy an application across multiple AZs. This will protect your application from issues such as power outages, lightning strikes, tornadoes, and more

Edge Locations

AWS also has 200+ edge locations worldwide. The Edge locations are used by CloudFront (Content Delivery Network) to cache content close to users

This allows you to deliver content to end-users with low latency

Data and Regions

All your data AWS is stored within a specific region that you choose.

To further increase redundancy and fault tolerance, you can replicate data across AWS regions.

To copy data between regions, you need to explicitly ask AWS service to do so.

For example, you can configure S3 to replicate the content of a bucket to another bucket in a different region

You can configure Relational Database Service (RDS) to maintain a read-replica in a different region

You can configure DynamoDB Global Tables to maintain a copy of data in multiple regions and automatically synchronize changes across regions

Note: CloudFront and Edge Cache also maintain copies of data in multiple edge locations. This is used primarily for performance reasons [and not for redundancy or fault tolerance]. When edge cache expires, it will go to origin to get the latest data

AWS Personal Health Dashboard

AWS personalized health dashboard provides alerts and remediation guidance when AWS is experiencing an outage.

The service health dashboard displays the status of AWS services, and the personalized health dashboard gives insight into your resources that are impacted due to AWS infrastructure issues

AWS Pricing

AWS pricing is based on

1. Compute (No. of hours of compute used per month)
2. How EC2 instance, Compute Capacity was purchased
3. Storage (GB of data stored per month)
4. Data Transfer (GB of data transferred out each month)

NOTE: AWS offers a free tier for new customers. This free-tier period is for 12-months and covers some specific services and usage. SageMaker comes with a 2-month trial period. Any use that exceeds the free-tier quota is billed at on-demand rates.

Compute Pricing

You can purchase compute capacity using

1. On-demand
 - a. No upfront payment or long-term commitment
 - b. Recommended for short-term, spiky, or unpredictable workloads that cannot be interrupted
 - c. Recommended when you are trying out AWS for the first time
2. Reserved Instances
 - a. Significant discount when compared to on-demand pricing
 - b. Standard Reserved Instances provides a discount of up to 72%, and it is region and instance family-specific
 - c. Convertible Reserved Instances provides a discount of up to 54%, and it is region-specific. However, you are free to change instance family
 - d. Requires 1-year or 3-year commitment
 - e. Recommended for steady-state or predictable usage
3. Savings Plan
 - a. Three types of savings plans: EC2 Instance Savings Plan, Compute Savings Plans, SageMaker Savings Plans
[My opinion – Compared to Reserved Instances, Savings plans are more flexible. This may be the future of AWS pricing]
 - b. Requires 1-year or 3-year commitment
 - c. EC2 Instance Savings plan applies to EC2 usage, and it is region and instance family-specific (up to 72% discount). This is similar to reserved instances
 - d. With SageMaker Savings Plans, you can reduce your SageMaker machine learning usage costs (up to 64% discount) irrespective of the region, instance family, and size.
 - e. With Compute Savings Plan, you can get a significant discount (up to 66%) on any compute service such as EC2, Lambda, Fargate Containers (applies to both server-based and serverless compute)
 - f. Compute savings plans automatically lowers cost irrespective of the region, instance family and size
4. Spot Instances

- a. Request for unused AWS capacity at a steep discount (up to 90% off on-demand pricing)
 - b. AWS can terminate an instance with 2-minute notice [i.e., your workload can be interrupted any time]
 - c. Recommended for applications that have flexible start and end times, urgent computing needs that require a large amount of compute capacity
 - d. Suitable for workloads that are fault-tolerant [i.e., workload that can safely continue in another instance when interrupted]
5. Spot-Block
- a. Spot-Block is for workloads that need to run continuously for 1 to 6 hours
 - b. Request for unused AWS capacity (discount up to 45% off on-demand pricing)
 - c. When spot instance capacity is available for the requested duration, the request is fulfilled
 - d. An instance is automatically terminated at the end of the time block
 - e. [My Opinion - It appears spot block is being phased out as new accounts are not eligible for spot blocks]
6. Dedicated Host
- a. Physical EC2 server dedicated for your use
 - b. Can be purchased on-demand or as a reserved instance
 - c. Useful when you need to bring your own software license to AWS cloud [such as SQL Server, SUSE Enterprise Linux Server, Windows Server]
 - d. Useful when your software license is tied to sockets or physical cores

Data Transfer Pricing

1. Same Availability Zone – free when using Private IP
2. Same Availability Zone – USD 0.01/GB when using Public or Elastic IP
3. Same Region – USD 0.01/GB
4. Another Region – USD 0.02/GB
5. From Internet – free
6. To Internet – USD 0.09/GB

Pricing Calculators

1. Simple Monthly Calculator – Helps estimate monthly AWS bills. Explore various AWS services, and create an estimate for your use case on AWS
2. AWS Pricing Calculator – Same capability as Simple Month Calculator [replacement for a simple monthly calculator with better User interface]
3. TCO Calculator – Total Cost of Ownership Calculator compares the cost of running your workload in an on-premises data center and AWS cloud. Useful when you need to plan for cloud migration

AWS Organizations

A lot of companies use several AWS accounts (some even have 100s of accounts). AWS Organizations helps you centrally manage all your accounts

1. You can organize various account as hierarchies
2. Centrally secure, control, and audit your accounts. For example, which region to use, what AWS services are allowed

3. Simplify billing by consolidating expenses across all accounts – single payment for all accounts
4. Aggregate usage across all accounts for volume discounts [AWS has a tiered pricing band. For example, if you store more data, you will get a better per GB storage pricing]
5. Share reserved instance discounts and Savings plan discounts across the entire organization

Cost Explorer and Usage Alerts

You can gain visibility of charges using these tools

1. AWS Budgets – Set your monthly budget and configure alerts when actual usage or forecasted usage exceeds your budget threshold
2. AWS Cost Explorer – Interactive tool to explore, understand, visualize, and manage costs over a time period. Useful to detect trends, pinpoint where you are spending money, and forecast future costs
3. AWS Bills – At the end of each billing cycle, you will receive a monthly invoice of usage charges. In addition, you can also view the past bills and current month bill from the Billing dashboard
4. CloudWatch Alarm – Estimated Charges are calculated several times daily and published to CloudWatch as a metric. You can also configure alarms based on this metric
5. Simple Notification Service (SNS) – SNS allows you to create an Email, SMS distribution list. You can configure budget alerts, CloudWatch Alarm to notify an SNS topic whenever the usage exceeds a configured threshold. SNS also allows you to invoke other services such as the Lambda function in response to an alert or event.

Support Plans

AWS offers the following support plans

1. Basic
 - a. Free and included for all customers
 - b. You can contact AWS support for questions related to billing, account, and for increasing service limits
 - c. Use the web-form to open a case
 - d. Access to 7-core Trusted Advisor Checks and Personal Health dashboard
2. Developer
 - a. Recommended if you are experimenting or testing in AWS
 - b. Pricing starts at \$29/month or 3% of monthly AWS usage (whichever is greater)
 - c. Email-based tech support during business hours
 - d. Response time is 24 hours. For system impaired issues, less than 12 hours
 - e. Access to 7-core Trusted Advisor Checks and Personal Health dashboard
3. Business
 - a. Recommended if you have production workloads in AWS
 - b. Pricing starts at \$100/month or 10% of monthly AWS usage (whichever is greater)
 - c. Contact Tech support - 24x7 Phone, email, and chat
 - d. Response time is less than 1 hour for Production system down
 - e. Access to the full set of Trusted Advisor Checks
 - f. Paid access to Infrastructure Event Management (additional planning and support for special high traffic events such as your new product launches, sports broadcast)
4. Enterprise

- a. Recommended if you have a business or mission-critical workload in AWS
- b. Pricing starts at \$15,000/month or 10% of monthly AWS usage (whichever is greater)
- c. Contact Tech support - 24x7 Phone, email, and chat
- d. Response time is less than 15 minutes for Business-critical system down
- e. Access to the full set of Trusted Advisor Checks
- f. Designated Technical Account Manager (TAM) – your single point of contact for AWS
- g. Infrastructure Event Management is included for free
- h. Concierge Support Team
- i. Well architected reviews, consultations, and guidance

Shared Responsibility Model

Security and compliance is a shared responsibility between AWS and the Customer. [Customer here refers to one who created an AWS account]

AWS is responsible for “security OF the cloud.”

Customer is responsible for “security IN the cloud” – another way to think about this is anything that you store and do in the cloud is your responsibility.

AWS Responsibility

AWS is responsible for protecting the infrastructure. For example, AWS is responsible for the physical security of the facilities. AWS is also responsible for protecting the cloud from infrastructure level attacks such as a distributed denial of service (DDOS).

The infrastructure is composed of hardware, software, networking, and facilities that run AWS cloud services

AWS is responsible for patching and fixing flaws in the infrastructure - Such as physical server security, patching the physical host operating system, and so forth

AWS also has increased responsibility in managed services such as S3 and DynamoDB. They are responsible for infrastructure, operating system, software

Customer Responsibility

Customer responsibility varies based on specific AWS Services that the customer selects

For example, if you spin up an EC2 instance, you are responsible for the guest operating system, patching the guest OS, any application or utilities installed in your instance, configuration of security group firewall, network ACL firewall, and so forth

Whereas AWS is responsible for an underlying physical server, virtualization software, and infrastructure

If you use a managed service such as S3 or DynamoDB, you are responsible for selecting the appropriate region, managing the data, encryption options, classifying assets, and using IAM to apply appropriate access controls [i.e., who can read, write, update, delete]

AWS trains AWS employees. Similarly, a customer must train their own employees

Cloud Practitioner Review – Infrastructure and Pricing

Prepared By: [Chandra Mohan Lingam](#), Cloud Wave LLC

Benefits of Cloud Computing

1. Trade capital expense for variable expense
 - a. No need to purchase expensive hardware in the cloud
 - b. Pay based on usage
2. Benefit from massive economies of scale
 - a. AWS buys in bulk as it needs to support 1000s of customers
 - b. Shared infrastructure – reduces idle resources
 - c. Better utilized
 - d. Lower pay as you go pricing
3. Stop Guessing about capacity
 - a. Scale up or down only with a few minutes notice
 - b. Match infrastructure with actual need
4. Increase speed and Agility
 - a. In the cloud, new resources are only a click away
 - b. Hourly pricing – try new products at a low cost
5. Stop spending money running and maintaining data centers
 - a. Avoid undifferentiated heavy lifting.
 - b. In a traditional on-premises data center, you need to spend upfront money on purchasing physical servers, storage, networking equipment, datacenters, and so forth. Whereas, in the cloud, you pay only for what you use
 - c. Focus on projects that differentiate your business
6. Go Global in Minutes
 - a. Deploy application close to the users
 - b. Use AWS regions, Edge Locations

AWS Global Infrastructure

AWS Global Infrastructure consists of multiple regions. There are 24 regions (and expanding) currently

For example, US West (Oregon), US West (Northern California), US East (Ohio), US East (N. Virginia), Europe (Frankfurt), Europe (Ireland), Asia Pacific (Tokyo), Asia Pacific (Mumbai) and so forth

All regions are global/public regions – that means you can spin up EC2 instances and other resources in any of the regions

However, there are two special regions

AWS GovCloud is only accessible to US entities that pass a screening process. Designed to host sensitive and regulated workloads to meet US Government security and compliance requirements

AWS China region complies with China's legal and regulatory requirements. This is a separate account that limits access only to AWS China regions

AWS Region

Each region consists of multiple, isolated, and physically separate Availability Zones within a geographic area.

For example, US West (Oregon) region has four availability zones within 60 miles of each other.

Oregon Region Name: us-west-2

Oregon Availability Zones: us-west-2a, us-west-2b, us-west-2c, us-west-2d

Availability Zone

An availability zone (AZ) consists of one or more data centers with redundant power, networking, and connectivity in an AWS region

All availability zones (AZ) in a region are interconnected with high speed, high bandwidth, and low latency networking

An architecture best practice is to deploy an application across multiple AZs. This will protect your application from issues such as power outages, lightning strikes, tornadoes, and more

Edge Locations

AWS also has 200+ edge locations worldwide. The Edge locations are used by CloudFront (Content Delivery Network) to cache content close to users

This allows you to deliver content to end-users with low latency

Data and Regions

All your data AWS is stored within a specific region that you choose.

To further increase redundancy and fault tolerance, you can replicate data across AWS regions.

To copy data between regions, you need to explicitly ask AWS service to do so.

For example, you can configure S3 to replicate the content of a bucket to another bucket in a different region

You can configure Relational Database Service (RDS) to maintain a read-replica in a different region

You can configure DynamoDB Global Tables to maintain a copy of data in multiple regions and automatically synchronize changes across regions

Note: CloudFront and Edge Cache also maintain copies of data in multiple edge locations. This is used primarily for performance reasons [and not for redundancy or fault tolerance]. When edge cache expires, it will go to origin to get the latest data

AWS Personal Health Dashboard

AWS personalized health dashboard provides alerts and remediation guidance when AWS is experiencing an outage.

The service health dashboard displays the status of AWS services, and the personalized health dashboard gives insight into your resources that are impacted due to AWS infrastructure issues

AWS Pricing

AWS pricing is based on

1. Compute (No. of hours of compute used per month)
2. How EC2 instance, Compute Capacity was purchased
3. Storage (GB of data stored per month)
4. Data Transfer (GB of data transferred out each month)

NOTE: AWS offers a free tier for new customers. This free-tier period is for 12-months and covers some specific services and usage. SageMaker comes with a 2-month trial period. Any use that exceeds the free-tier quota is billed at on-demand rates.

Compute Pricing

You can purchase compute capacity using

1. On-demand
 - a. No upfront payment or long-term commitment
 - b. Recommended for short-term, spiky, or unpredictable workloads that cannot be interrupted
 - c. Recommended when you are trying out AWS for the first time
2. Reserved Instances
 - a. Significant discount when compared to on-demand pricing
 - b. Standard Reserved Instances provides a discount of up to 72%, and it is region and instance family-specific
 - c. Convertible Reserved Instances provides a discount of up to 54%, and it is region-specific. However, you are free to change instance family
 - d. Requires 1-year or 3-year commitment
 - e. Recommended for steady-state or predictable usage
3. Savings Plan
 - a. Three types of savings plans: EC2 Instance Savings Plan, Compute Savings Plans, SageMaker Savings Plans
[My opinion – Compared to Reserved Instances, Savings plans are more flexible. This may be the future of AWS pricing]
 - b. Requires 1-year or 3-year commitment
 - c. EC2 Instance Savings plan applies to EC2 usage, and it is region and instance family-specific (up to 72% discount). This is similar to reserved instances
 - d. With SageMaker Savings Plans, you can reduce your SageMaker machine learning usage costs (up to 64% discount) irrespective of the region, instance family, and size.
 - e. With Compute Savings Plan, you can get a significant discount (up to 66%) on any compute service such as EC2, Lambda, Fargate Containers (applies to both server-based and serverless compute)
 - f. Compute savings plans automatically lowers cost irrespective of the region, instance family and size
4. Spot Instances

- a. Request for unused AWS capacity at a steep discount (up to 90% off on-demand pricing)
 - b. AWS can terminate an instance with 2-minute notice [i.e., your workload can be interrupted any time]
 - c. Recommended for applications that have flexible start and end times, urgent computing needs that require a large amount of compute capacity
 - d. Suitable for workloads that are fault-tolerant [i.e., workload that can safely continue in another instance when interrupted]
5. Spot-Block
- a. Spot-Block is for workloads that need to run continuously for 1 to 6 hours
 - b. Request for unused AWS capacity (discount up to 45% off on-demand pricing)
 - c. When spot instance capacity is available for the requested duration, the request is fulfilled
 - d. An instance is automatically terminated at the end of the time block
 - e. [My Opinion - It appears spot block is being phased out as new accounts are not eligible for spot blocks]
6. Dedicated Host
- a. Physical EC2 server dedicated for your use
 - b. Can be purchased on-demand or as a reserved instance
 - c. Useful when you need to bring your own software license to AWS cloud [such as SQL Server, SUSE Enterprise Linux Server, Windows Server]
 - d. Useful when your software license is tied to sockets or physical cores

Data Transfer Pricing

1. Same Availability Zone – free when using Private IP
2. Same Availability Zone – USD 0.01/GB when using Public or Elastic IP
3. Same Region – USD 0.01/GB
4. Another Region – USD 0.02/GB
5. From Internet – free
6. To Internet – USD 0.09/GB

Pricing Calculators

1. Simple Monthly Calculator – Helps estimate monthly AWS bills. Explore various AWS services, and create an estimate for your use case on AWS
2. AWS Pricing Calculator – Same capability as Simple Month Calculator [replacement for a simple monthly calculator with better User interface]
3. TCO Calculator – Total Cost of Ownership Calculator compares the cost of running your workload in an on-premises data center and AWS cloud. Useful when you need to plan for cloud migration

AWS Organizations

A lot of companies use several AWS accounts (some even have 100s of accounts). AWS Organizations helps you centrally manage all your accounts

1. You can organize various account as hierarchies
2. Centrally secure, control, and audit your accounts. For example, which region to use, what AWS services are allowed

3. Simplify billing by consolidating expenses across all accounts – single payment for all accounts
4. Aggregate usage across all accounts for volume discounts [AWS has a tiered pricing band. For example, if you store more data, you will get a better per GB storage pricing]
5. Share reserved instance discounts and Savings plan discounts across the entire organization

Cost Explorer and Usage Alerts

You can gain visibility of charges using these tools

1. AWS Budgets – Set your monthly budget and configure alerts when actual usage or forecasted usage exceeds your budget threshold
2. AWS Cost Explorer – Interactive tool to explore, understand, visualize, and manage costs over a time period. Useful to detect trends, pinpoint where you are spending money, and forecast future costs
3. AWS Bills – At the end of each billing cycle, you will receive a monthly invoice of usage charges. In addition, you can also view the past bills and current month bill from the Billing dashboard
4. CloudWatch Alarm – Estimated Charges are calculated several times daily and published to CloudWatch as a metric. You can also configure alarms based on this metric
5. Simple Notification Service (SNS) – SNS allows you to create an Email, SMS distribution list. You can configure budget alerts, CloudWatch Alarm to notify an SNS topic whenever the usage exceeds a configured threshold. SNS also allows you to invoke other services such as the Lambda function in response to an alert or event.

Support Plans

AWS offers the following support plans

1. Basic
 - a. Free and included for all customers
 - b. You can contact AWS support for questions related to billing, account, and for increasing service limits
 - c. Use the web-form to open a case
 - d. Access to 7-core Trusted Advisor Checks and Personal Health dashboard
2. Developer
 - a. Recommended if you are experimenting or testing in AWS
 - b. Pricing starts at \$29/month or 3% of monthly AWS usage (whichever is greater)
 - c. Email-based tech support during business hours
 - d. Response time is 24 hours. For system impaired issues, less than 12 hours
 - e. Access to 7-core Trusted Advisor Checks and Personal Health dashboard
3. Business
 - a. Recommended if you have production workloads in AWS
 - b. Pricing starts at \$100/month or 10% of monthly AWS usage (whichever is greater)
 - c. Contact Tech support - 24x7 Phone, email, and chat
 - d. Response time is less than 1 hour for Production system down
 - e. Access to the full set of Trusted Advisor Checks
 - f. Paid access to Infrastructure Event Management (additional planning and support for special high traffic events such as your new product launches, sports broadcast)
4. Enterprise

- a. Recommended if you have a business or mission-critical workload in AWS
- b. Pricing starts at \$15,000/month or 10% of monthly AWS usage (whichever is greater)
- c. Contact Tech support - 24x7 Phone, email, and chat
- d. Response time is less than 15 minutes for Business-critical system down
- e. Access to the full set of Trusted Advisor Checks
- f. Designated Technical Account Manager (TAM) – your single point of contact for AWS
- g. Infrastructure Event Management is included for free
- h. Concierge Support Team
- i. Well architected reviews, consultations, and guidance

Shared Responsibility Model

Security and compliance is a shared responsibility between AWS and the Customer. [Customer here refers to one who created an AWS account]

AWS is responsible for “security OF the cloud.”

Customer is responsible for “security IN the cloud” – another way to think about this is anything that you store and do in the cloud is your responsibility.

AWS Responsibility

AWS is responsible for protecting the infrastructure. For example, AWS is responsible for the physical security of the facilities. AWS is also responsible for protecting the cloud from infrastructure level attacks such as a distributed denial of service (DDOS).

The infrastructure is composed of hardware, software, networking, and facilities that run AWS cloud services

AWS is responsible for patching and fixing flaws in the infrastructure - Such as physical server security, patching the physical host operating system, and so forth

AWS also has increased responsibility in managed services such as S3 and DynamoDB. They are responsible for infrastructure, operating system, software

Customer Responsibility

Customer responsibility varies based on specific AWS Services that the customer selects

For example, if you spin up an EC2 instance, you are responsible for the guest operating system, patching the guest OS, any application or utilities installed in your instance, configuration of security group firewall, network ACL firewall, and so forth

Whereas AWS is responsible for an underlying physical server, virtualization software, and infrastructure

If you use a managed service such as S3 or DynamoDB, you are responsible for selecting the appropriate region, managing the data, encryption options, classifying assets, and using IAM to apply appropriate access controls [i.e., who can read, write, update, delete]

AWS trains AWS employees. Similarly, a customer must train their own employees

SageMaker Service, SDK Changes

<https://github.com/ChandraLingam/AmazonSageMakerCourse>

Model Training

AWS CLI

SageMaker Console x create-training-job — AWS CLI 2.x +

awscli.amazonaws.com/v2/documentation/api/latest/reference/sagemaker/create-training-job.html

AWS CLI Command Reference Home User Guide Forum GitHub Star 11,139

← create-project / create-transform-job →

[aws . sagemaker]

create-training-job

Description

Starts a model training job. After training completes, Amazon SageMaker saves the resulting model artifacts to an Amazon S3 location that you specify.

If you choose to host your model using Amazon SageMaker hosting services, you can use the resulting model artifacts as part of the model. You can also use the artifacts in a machine learning service other than Amazon SageMaker, provided that you know how to use them for inference.

In the request body, you provide the following:

- `AlgorithmSpecification` - Identifies the training algorithm to use.
- `HyperParameters` - Specify these algorithm-specific parameters to enable the estimation of model parameters during training. Hyperparameters can be tuned to optimize this learning process. For a list of hyperparameters for each training algorithm provided by Amazon SageMaker, see [Algorithms](#) .
- `InputDataConfig` - Describes the training dataset and the Amazon S3, EFS, or FSx location where it is stored.

Table of Contents

- create-training-job
 - Description
 - Synopsis
 - Options
 - Output

Quick search Search

Feedback

Did you find this page useful?
Do you have a suggestion?
[Give us feedback](#) or send us a [pull request](#) on GitHub.

User Guide

SageMaker Console

The screenshot shows the 'Create training job' page in the Amazon SageMaker console. The browser title bar reads 'Amazon SageMaker'. The address bar shows the URL 'us-east-2.console.aws.amazon.com/sagemaker/home?region=us-east-2#/jobs/create'. The AWS navigation bar includes 'Services ▾', a search bar with placeholder 'Search for services, features, marketplace products, and docs', and keyboard shortcut '[Alt+S]'. The main navigation path is 'Amazon SageMaker > Training jobs > Create training job'. The main heading is 'Create training job'. A descriptive text explains that when you create a training job, Amazon SageMaker sets up the distributed compute cluster, performs the training, and deletes the cluster when training has completed. It also mentions that resulting model artifacts are stored in the location specified when you created the training job, with a link to 'Learn more'. The 'Job settings' section contains fields for 'Job name' (a text input field), 'IAM role' (a dropdown menu currently set to 'sagemaker_lab_role'), and 'Algorithm options' (a collapsed section). At the bottom, there is a 'Algorithm source' section.

Amazon SageMaker

us-east-2.console.aws.amazon.com/sagemaker/home?region=us-east-2#/jobs/create

aws Services ▾ Search for services, features, marketplace products, and docs [Alt+S]

Amazon SageMaker > Training jobs > Create training job

Create training job

When you create a training job, Amazon SageMaker sets up the distributed compute cluster, performs the training, and deletes the cluster when training has completed. The resulting model artifacts are stored in the location you specified when you created the training job. [Learn more](#)

Job settings

Job name

Maximum of 63 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in an AWS Region.

IAM role

Amazon SageMaker requires permissions to call other services on your behalf. Choose a role or let us create a role that has the [AmazonSageMakerFullAccess](#) IAM policy attached.

sagemaker_lab_role

Algorithm options

Use an Amazon SageMaker built-in algorithm, your own algorithm, or a third-party algorithm from AWS Marketplace.

Algorithm source

Programmatic (SDK)

Build Model

```
In [ ]: # Configure the training job
# Specify type and number of instances to use
# S3 Location where final artifacts needs to be stored

# Reference: http://sagemaker.readthedocs.io/en/latest/estimators.html

# for managed spot training, specify the use_spot_instances flag, max_run, max_wait and checkpoint_s3_uri

# SDK 2.x version does not require train prefix for instance count and type
estimator = sagemaker.estimator.Estimator(
    container,
    role,
    instance_count=1,
    instance_type='ml.m5.xlarge',
    output_path=s3_model_output_location,
    sagemaker_session=sess,
    base_job_name = job_name,
    use_spot_instances=use_spot_instances,
    max_run=max_run,
    max_wait=max_wait,
    checkpoint_s3_uri=checkpoint_s3_uri)
```

```
In [ ]: # Specify hyper parameters that appropriate for the training algorithm
# XGBoost Training Parameter Reference
# https://github.com/dmlc/xgboost/blob/master/doc/parameter.rst#Learning-task-parameters
estimator.set_hyperparameters(max_depth=5,
                             objective="reg:squarederror",
                             eta=0.1)
```

Supported Channels and Input Mode

For built-in algorithms, this link has supported channel names, support for File and Pipe mode, and file format

<https://docs.aws.amazon.com/sagemaker/latest/dg/common-info-all-im-models.html>

Checkpoint



Save model state – protection from unexpected interruption to training job or instance



Resume training from existing checkpoint



Analyze model at intermediate stages



Use with Managed Spot Training

Managed Spot Training

Spot instance – AWS spare capacity at 90% discount over on-demand pricing

Use for Training and Hyperparameter tuning jobs

Spot instance may not be readily available

Spot instance can be terminated anytime by AWS with a two-minute notice. SageMaker handles this automatically

Cost Saving Tips

- With free-trial, you can complete most of the labs for free
 - Only on-demand instances are part of free-trial
- If you are no longer under free-trial, use spot-instances for training
- Inference Endpoints – Terminate after you are done with inference in the labs
- Use Batch Transform Jobs to generate predictions for large dataset
- Stop and Restart notebook instances – don't leave them running
- Setup billing and budget alerts



Chandra Lingam
70,000+ Students



AWS Certified Machine Learning Specialty

AWS SageMaker - XGBoost

Content Prepared By: Chandra Lingam, Cloud Wave LLC

Copyright © 2019 Cloud Wave LLC. All Rights Reserved.

All other registered trademarks and/or copyright material are of their respective owners



Algorithms Overview

Algorithm	Description
Linear Models	<ul style="list-style-type: none">+ Simple and easy to understand+ Performs surprisingly well for a variety of problems- Difficulty handling non-linear datasets- Features on similar scale, one-hot encoding, complex features
<u>Decision Tree</u>	<ul style="list-style-type: none">+ Can Handle Complex non-linear relationship+ Easily handles numeric categorical data, missing data- Prone to overfitting- Poor predictive accuracy
<u>Ensemble Methods</u>	<ul style="list-style-type: none">+ Combines multiple simple decision trees+ Addresses decision tree overfitting problem+ Much better predictive performance- More complex to understand

Must Watch Videos

Gradient Boosting Machine Learning by Trevor Hastie

Learning Decision Tree by Alexander Ihler

Ensembles (Bagging) by Alexander Ihler

Lab: Compare XGBoost and Linear Regression

Compare XGBoost and Linear Regression Algorithm

Using a Simple Dataset

Understand how these algorithms learn from data

Train locally on Notebook instance

Lab: Compare XGBoost and Linear Regression

Using a non-linear dataset

Understand how these algorithms learn from data

Train locally on Notebook instance

Lab: Forecast Bike Rental Count

Old Kaggle Competition Problem

Complex dataset

Need to forecast hourly rentals

Lab: Forecast Bike Rental Count - Optimization

Transform Count to *Log (Count)*

A technique used when model needs to predict positive integers

Use inverse transform *Exp (Count)* on predicted value

Smoothen effect of seasonality and trend, brings count to a similar scale

Lab: Train using SageMaker's XGBoost

Upload Train and Validation files to S3

Specify Algorithm and Hyperparameters

Configure type of server and number of servers to use for Training

Create a real-time Endpoint for interactive use case

Lab: Prediction using SageMaker's XGBoost

Invoke Endpoint for interactive use cases

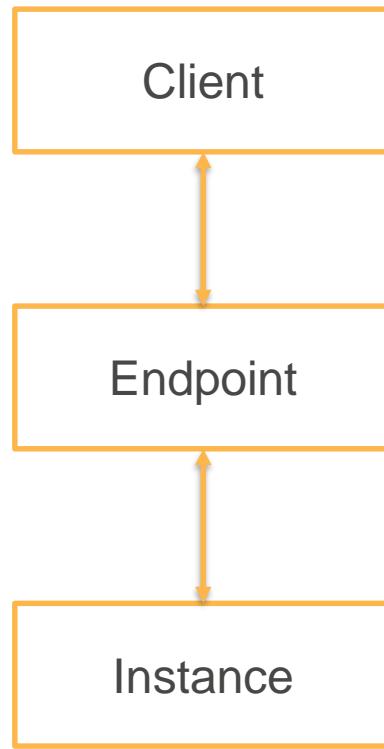
Connect to an existing Endpoint

Endpoint Security

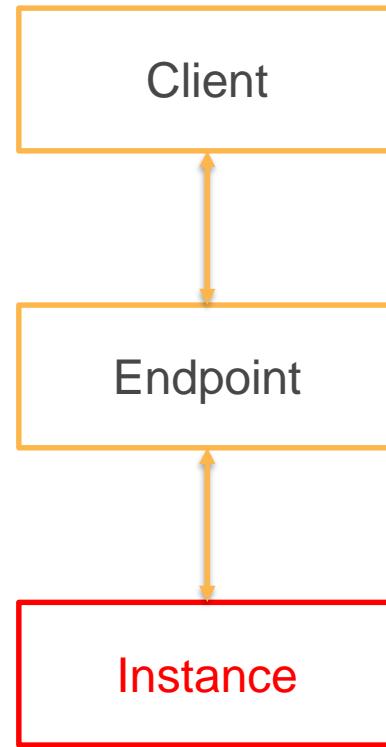
Multiple observations in a single round trip

Model Hosting

Model Hosting



Model Hosting



Single Instance Hosting = Single Point of Failure

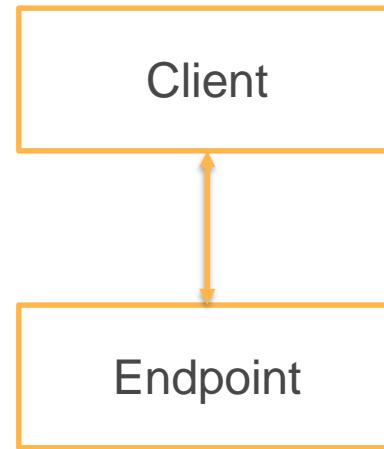


Monitoring and Scaling

CloudWatch – Monitoring Service

AutoScaling – Take automated scaling actions to maintain capacity

Model Hosting



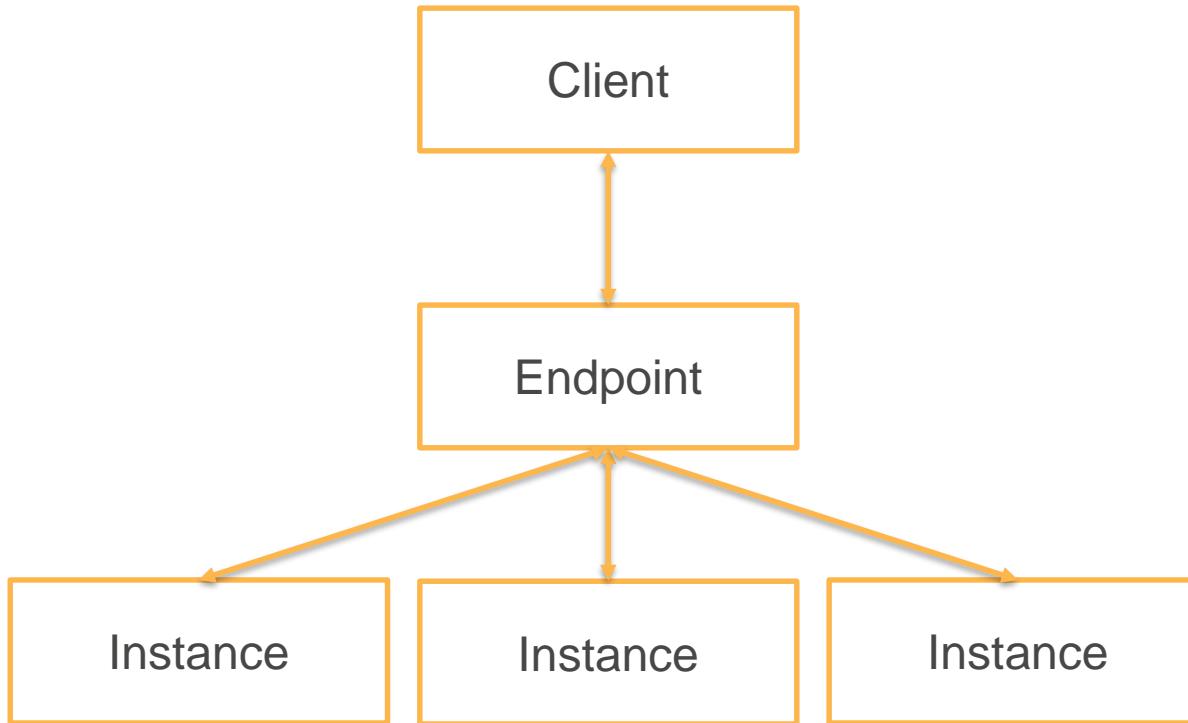
Instance

Instance

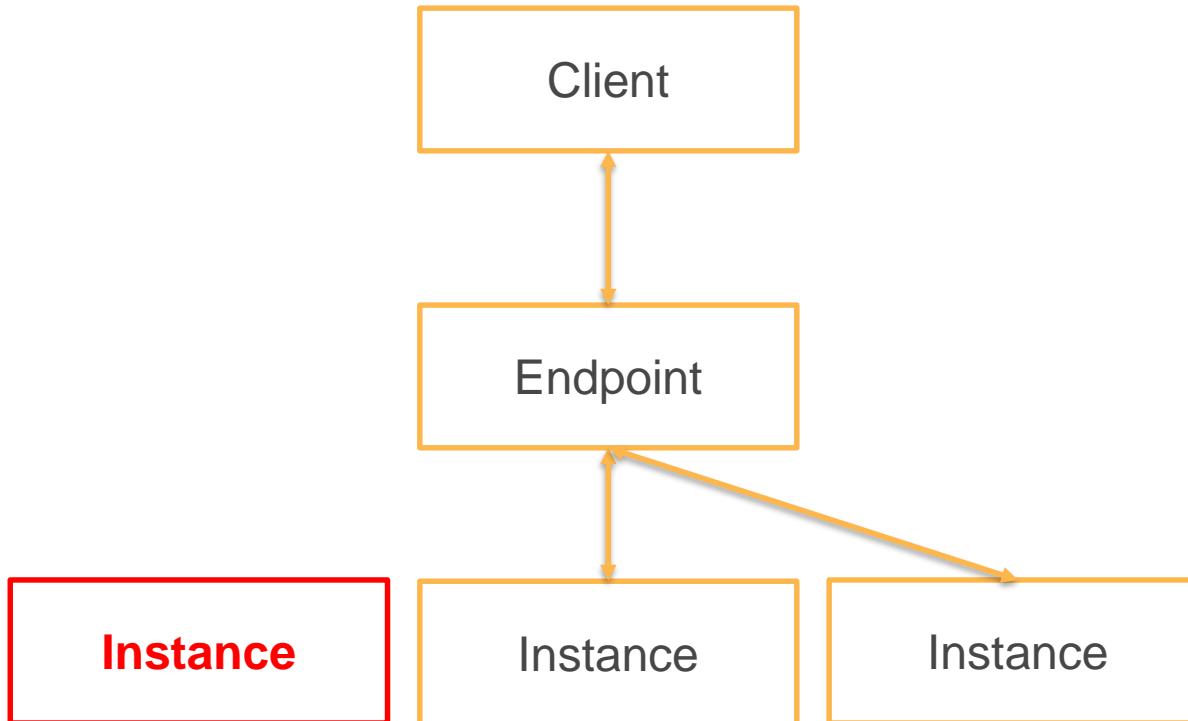
New instance launched to replace a failed instance



Model Hosting – Multiple Instance

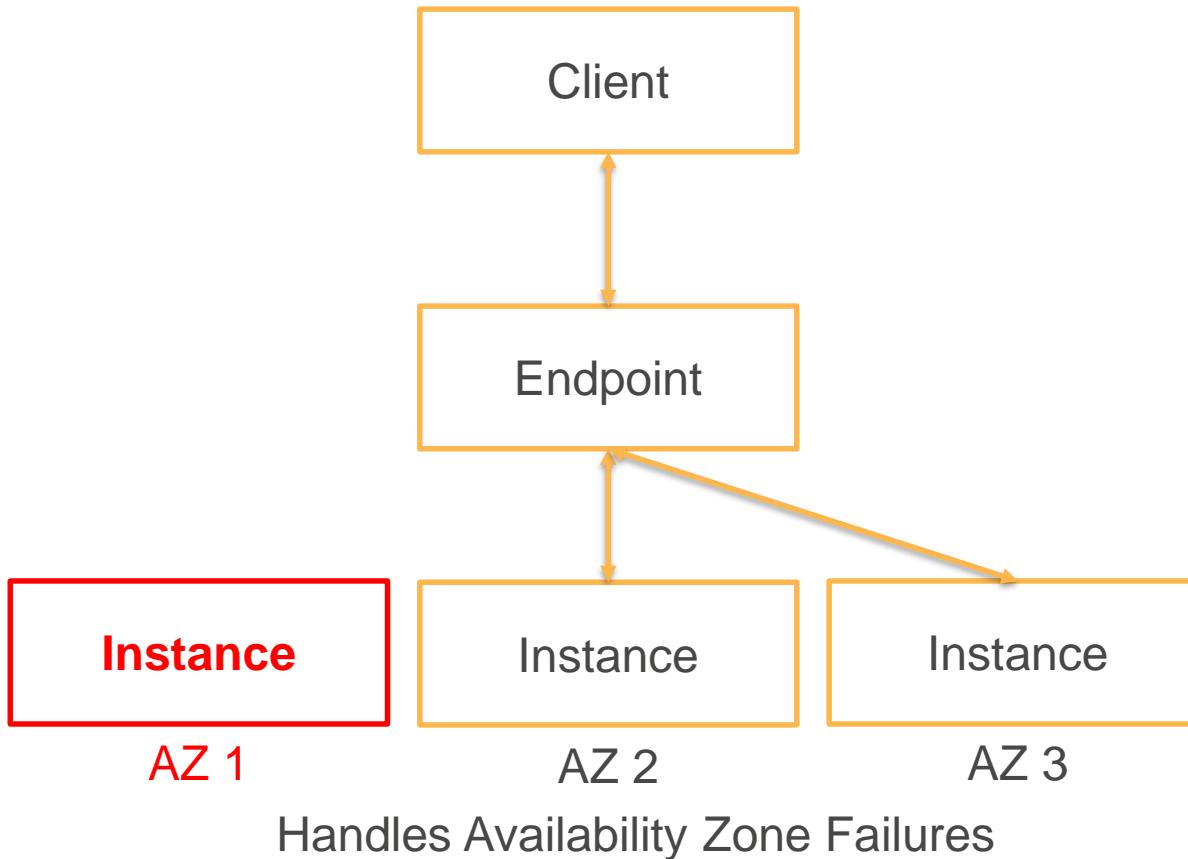


Model Hosting – Multiple Instance

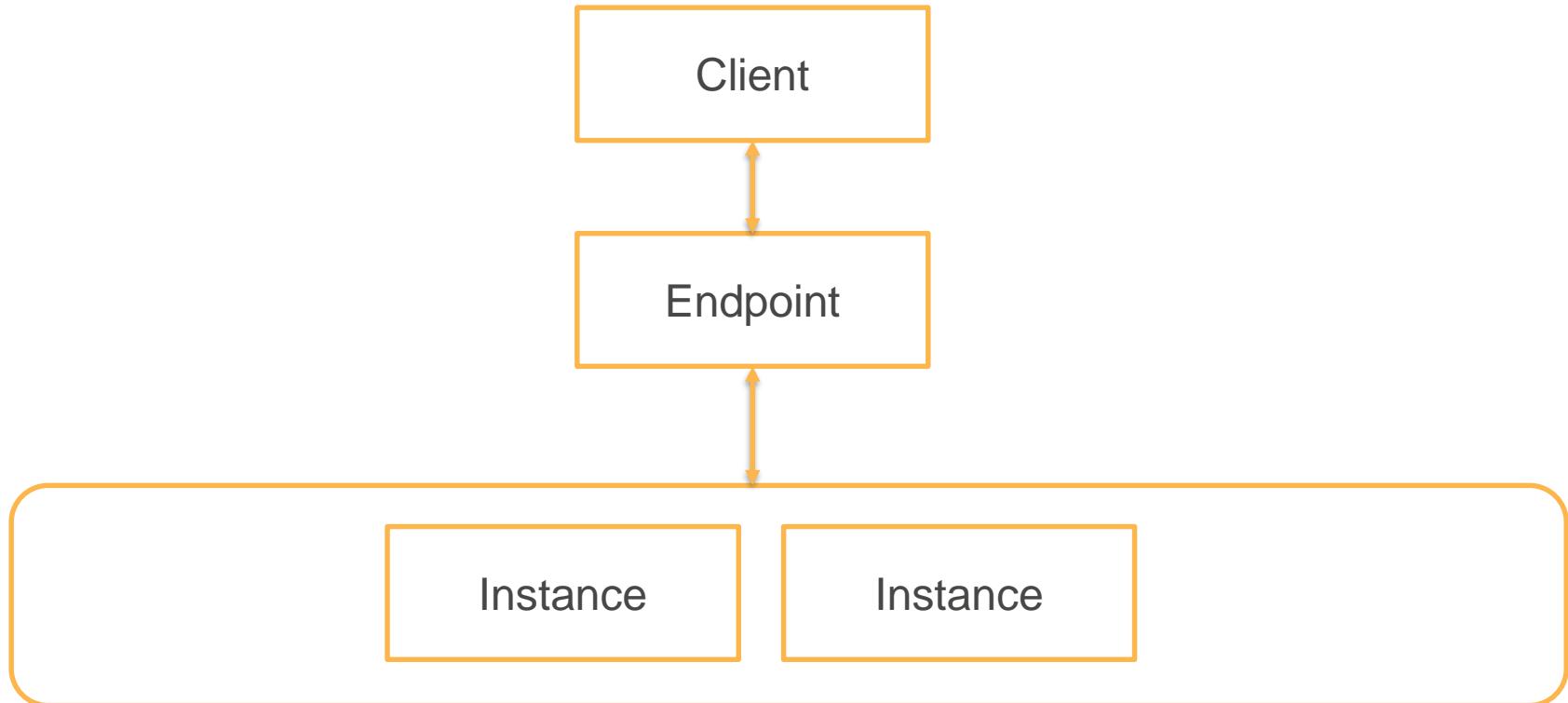


Requests load balanced across healthy instances

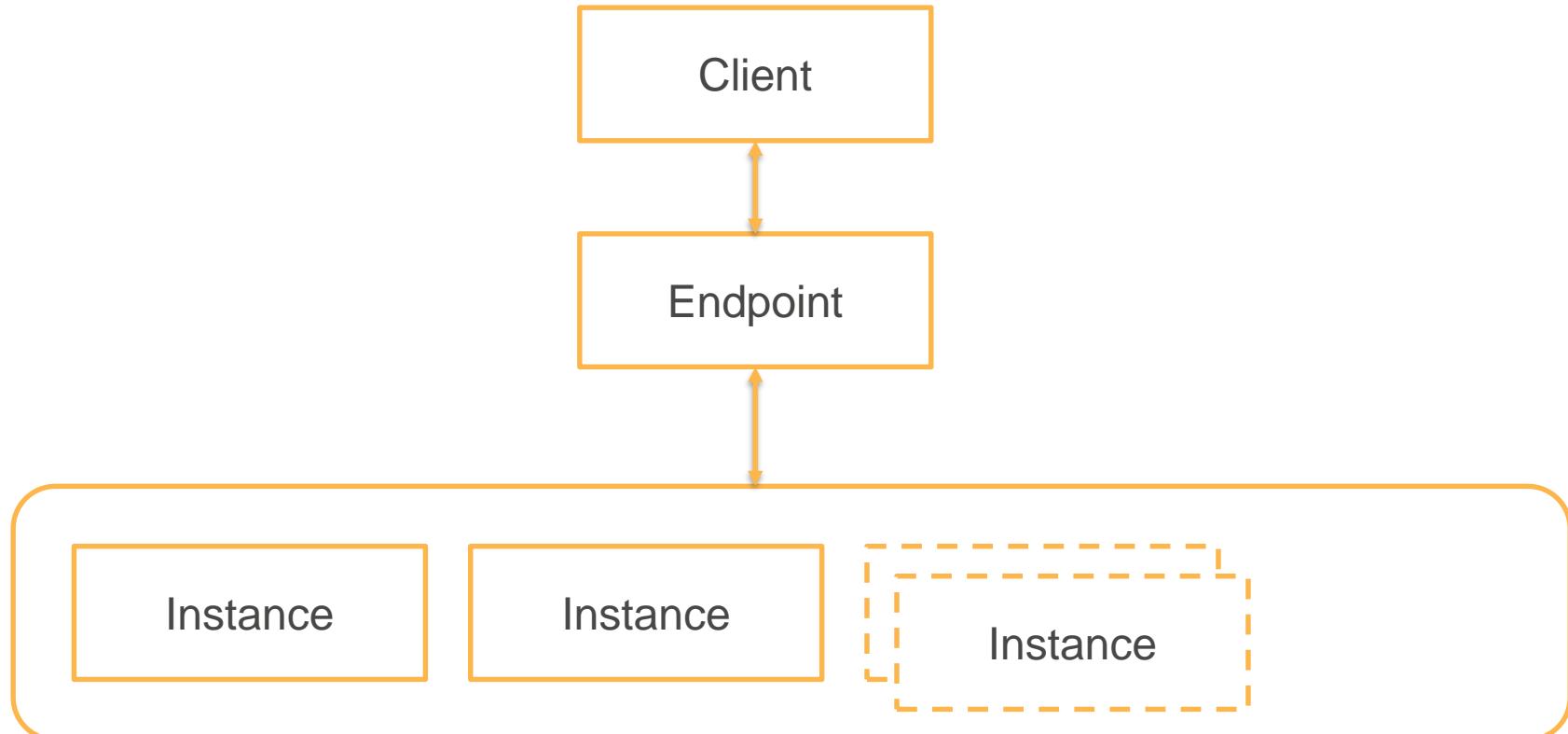
Model Hosting – Multiple Instance



Model Hosting – Scale on Demand



Model Hosting – Scale on Demand



Scaling based on Invocations

SageMakerVariantInvocationsPerInstance = Metric that records average number of requests per minute per instance

Use this metric for AutoScaling - For example,

Instance max load (MAX RPS) = 100 requests per second

Safety Factor = 0.5

Target SageMakerVariantInvocationsPerInstance =

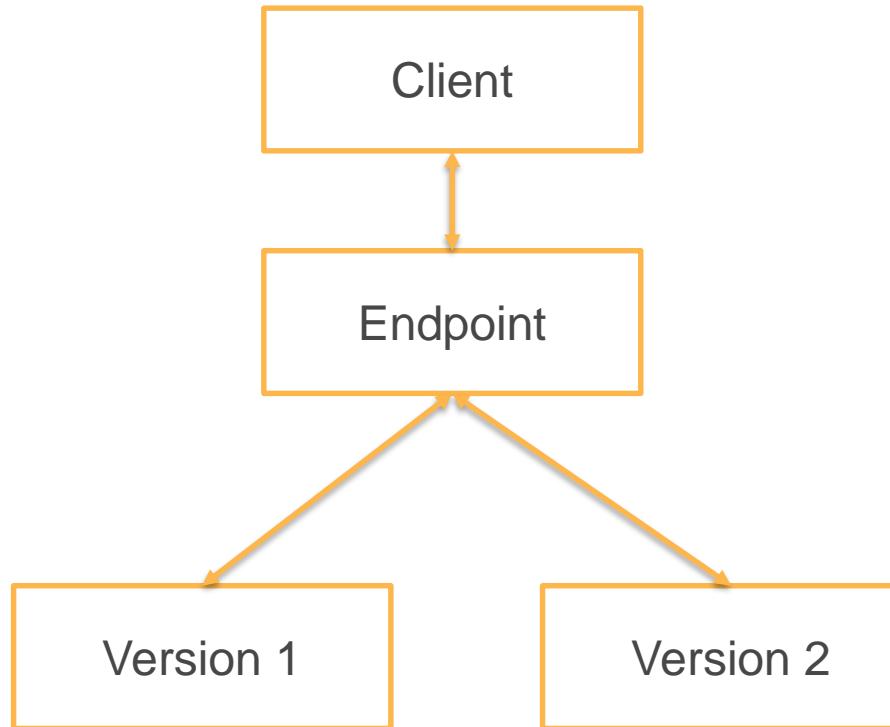
$$(\text{MAX_RPS} * \text{SAFETY_FACTOR}) * 60$$

$$\text{SageMakerVariantInvocationsPerInstance} = 100 * 0.5 * 60 = 3,000$$

Add additional instance when the metric crosses 3,000



Model Hosting – Variants of Algorithm



SageMaker Hosting

Automatically Replace Unhealthy Instances

Scale number of instances based on workload

Test Multiple Variants of Model

Hyperparameters

Training Objective

objective

Regression - “reg:linear”

Binary Classification - “binary:logistic”

Multiclass Classification - “multi:softmax”

Parameter References:

[SageMaker Documentation](#)

[XGBoost Documentation](#)

Bias and Variance



Photo Credit : [Ugrashak](#)

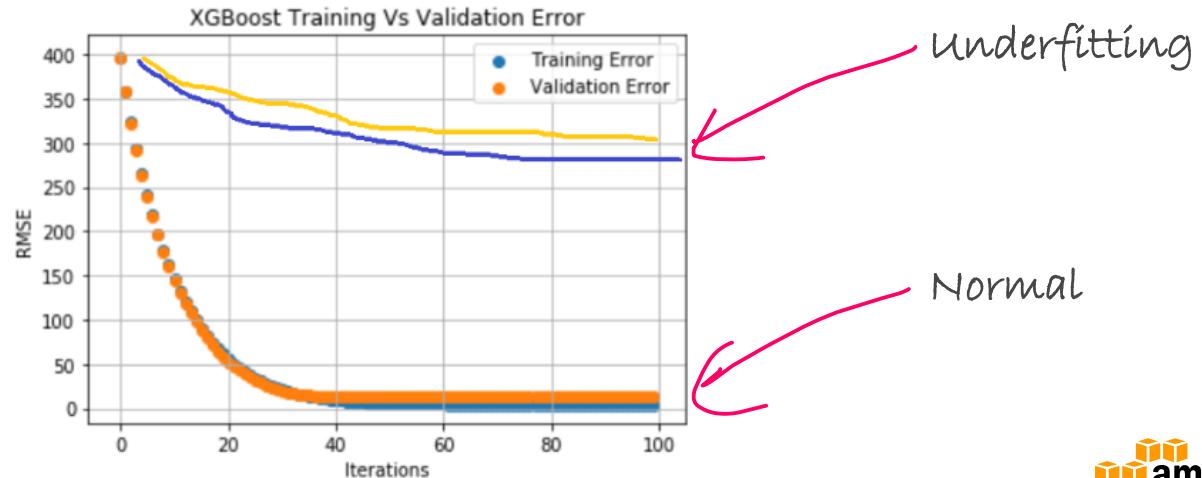
Biased => Does not match reality

High Bias

Model is not learning from data

Translates to large training and validation errors

Underfitting



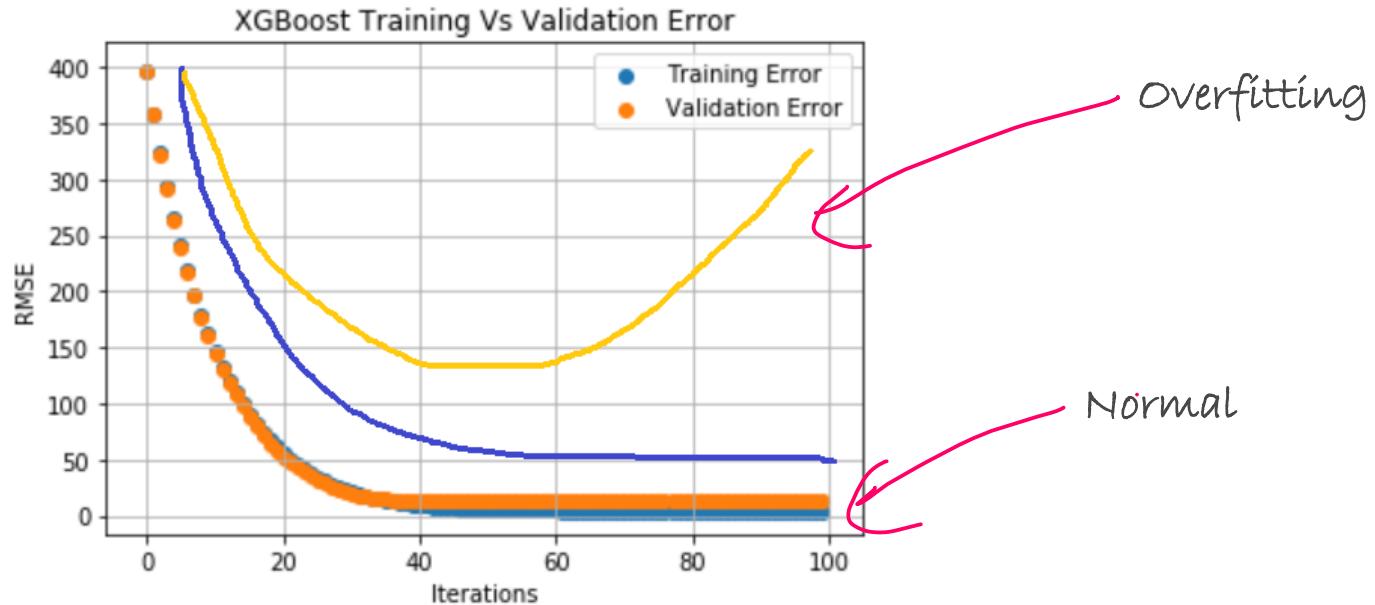
Variance

Measures how well the algorithm generalizes for unseen data

Difference between Validation Error and Training Error

High Variance – validation error is high; but training error is small. Overfitting

High Variance



Strategies to handle High Bias (Underfitting)

Add relevant features

Combine features (Example: area = length * width)

Create higher order features

Train longer (more iterations)

Decrease Regularization

Strategies to handle High Variance – Overfitting

Use fewer features

Use straightforward features (instead of higher order features)

Reduce Training iterations

Increase Regularization

Regularization

Many features are equally good at predicting outcome

Which combination of features is the model going to use?

Feature selection depends on algorithm and regularization parameters

Regularization

Regularization – Tone down overdependence of specific features

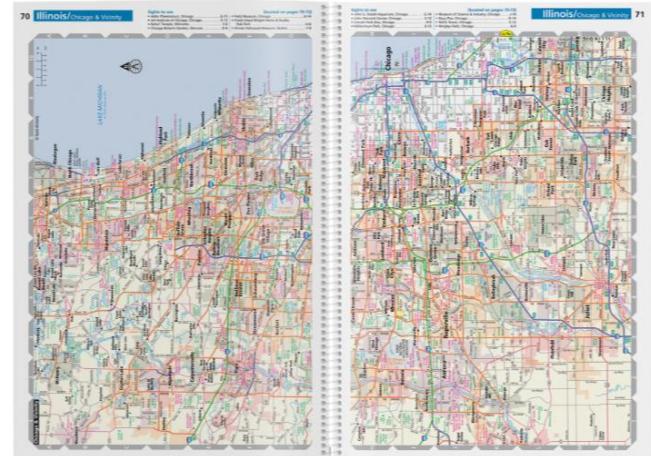
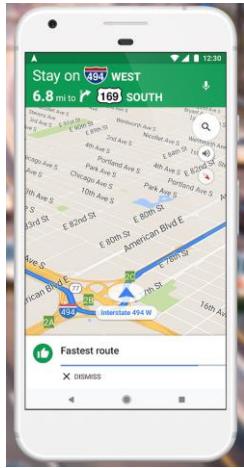


Photo Credit : Google, Garmin, Rand McNally

L1 Regularization

Algorithm aggressively eliminates features that are not important

Example:

Phone GPS = Substantial weight

Standalone GPS = Zero weight

Paper Map = Zero Weight

Useful in large dimension dataset – reduce the number of features

L2 Regularization

Algorithm simply reduces weight of features

Allows other features to influence outcome

L2 Regularization is a good starting point

Example:

Phone GPS = Larger weight

Standalone GPS = Medium weight

Paper Map = Smaller weight

XGBoost Regularization

alpha – L1 Regularization. Default 0

lambda – L2 Regularization. Default 1

Hyper Parameter Tuning

XGBoost Parameter Tuning

SageMaker XGBoost Hyper Parameter Documentation

SKLearn - Automatic Tuning

GridSearch – Exhaustive search using specified lower and upper bound of parameter values

RandomSearch – Random Search of parameters from specified lower and upper bound

SageMaker - Automatic Tuning

Bayesian Search – Smart Search. Treats hyperparameter tuning as a machine learning problem. Often converges faster

Random Search – Random Search of parameters from specified lower and upper bound

Hyper Parameter Tuning

n_estimators (in XGBRegressor) is same as num_round (in XGBoost and SageMaker documentation)

This parameter controls number of rounds of boosting i.e. total number of trees.

Make sure you use correct parameter depending on the library. SKLearn XGBRegressor *silently ignores parameters it does not understand* 😞

Model Endpoint Integration

Content Prepared By: Chandra Lingam, Cloud Wave LLC

Copyright © 2019 Cloud Wave LLC. All Rights Reserved.

All other registered trademarks and/or copyright material are of their respective owners



Client to Model Endpoint Integration

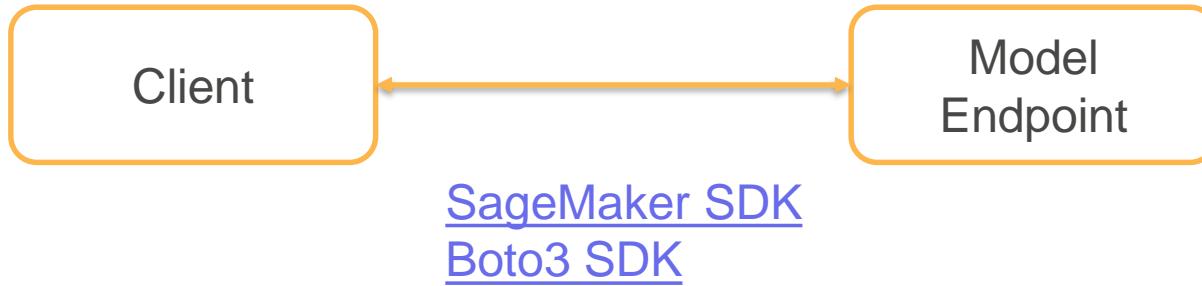
SageMaker SDK

Boto3 SDK

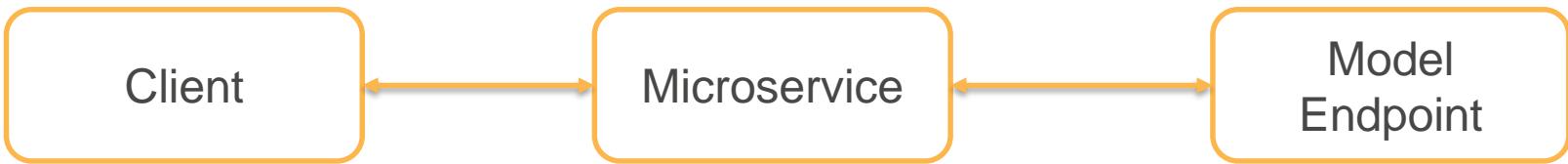
Microservice - Lambda

Microservice - API Gateway and Lambda

Client to Model Endpoint



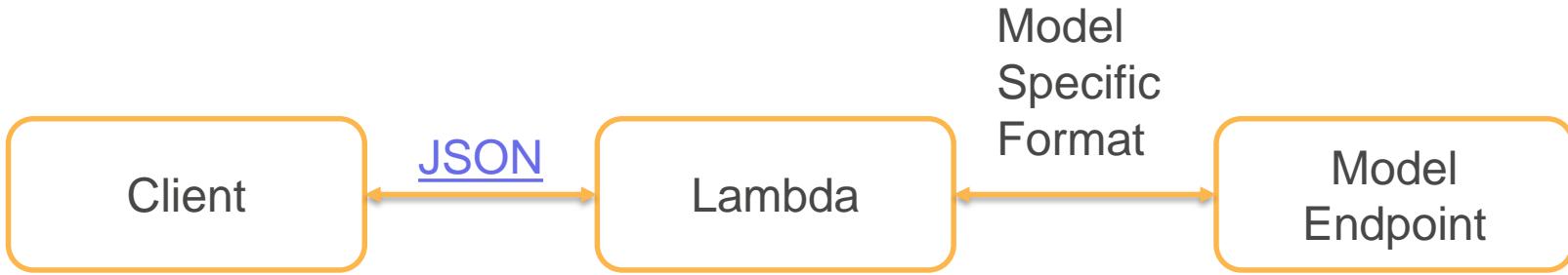
Microservice



Microservice – Lambda Function



Lambda Integration



Inference Formats

text/csv

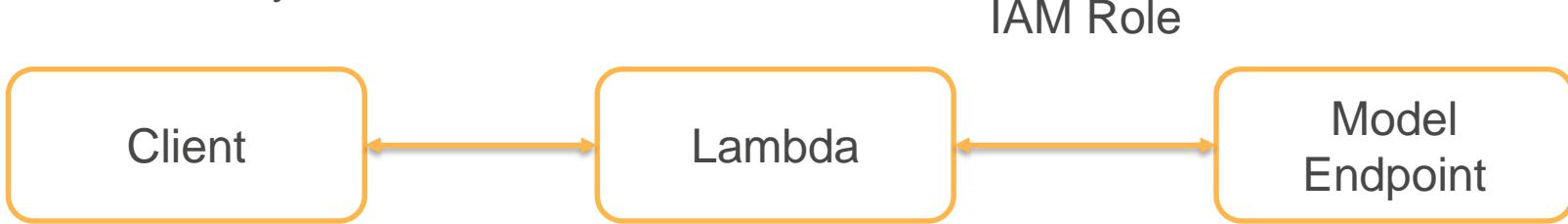
application/json

application/x-recordio-protobuf

text/x-libsvm

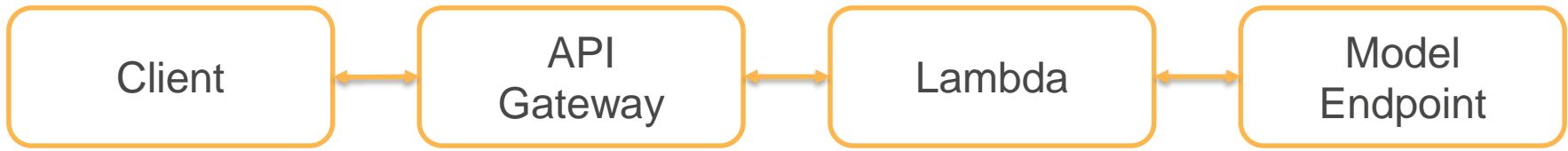
Lambda Permissions

Who can call your function?



What service actions can your lambda function perform?

Microservice - API Gateway and Lambda



Summary

Architecture	Summary
Client – Model Endpoint	<p>Client has too much knowledge about the model and feature transformation</p> <p>Not flexible from long term maintenance</p>
Lambda Microservice	<p>Lambda hides model related details</p> <p>Handles basic feature transformation and processing</p> <p>Numpy Support</p> <p>More complex feature transformation with SKLearn or Pandas is challenging in Lambda</p>
API Gateway and Lambda	<p>Recommended Approach – RESTful API is easy to integrate with variety of clients</p> <p>Same limitation of Lambda applies</p> <p>API Gateway as a gate keeper</p>

Housekeeping – Local Machine Setup

Chandra Lingam, Cloud Wave LLC

Revised: JAN, 2021

Read entire document and complete all the steps (5 steps in all)

1. Anaconda Python

Download and Install Anaconda Python in your local machine

Recommended: Python 3.6 or later

<https://www.anaconda.com/distribution/>

Install Instructions: [Windows](#), [macOS](#), [Linux](#)

To Test the installation, from **Anaconda Prompt**, run the command:

```
python --version
```

2. SageMaker SDK Install

Install SageMaker SDK in your local machine.

From **Anaconda Prompt**, run the command:

```
pip install sagemaker
```

<https://github.com/aws/sagemaker-python-sdk>

3. Git Client

If you don't have GIT installed, please install Git Client from

<https://git-scm.com/downloads>

4. Clone Repository

You can now clone the source code for this course to your local machine

From your local directory, run this command

```
git clone https://github.com/ChandraLingam/AmazonSageMakerCourse.git
```

AmazonSageMakerCourse folder now contains all the source code for this course.

If you need to get updated files from github, simply run the command:

`git pull` inside your **AmazonSageMakerCourse** folder

5. Launch Jupyter

From **Anaconda Prompt**, change to **AmazonSageMakerCourse** folder

Run the command

```
jupyter notebook
```

SageMaker Endpoints

<https://github.com/ChandraLingam/AmazonSageMakerCourse>

SageMaker SDK

`estimator.fit(...)`

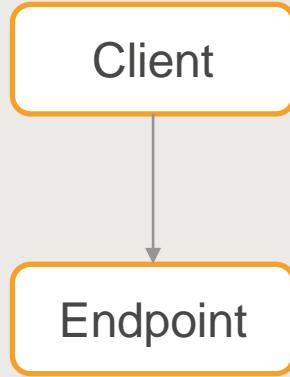
- Stores model artifacts in S3

`estimator.deploy(...)`

- Create Model
- Create Endpoint Configuration
- Create Endpoint

Deployment Flexibility!

Motivation

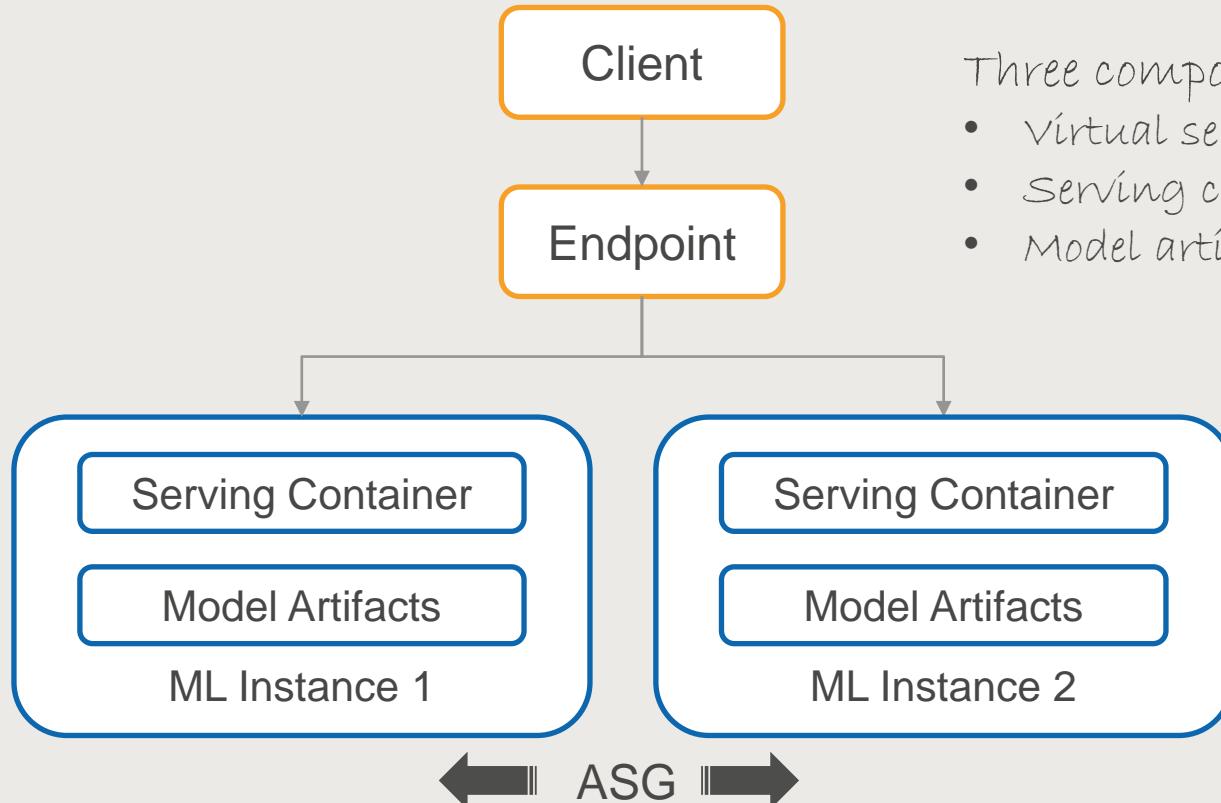


Continuous Improvement

- New model versions
- New algorithm versions
- Different algorithms

How to incorporate these changes in a live system with zero-downtime?

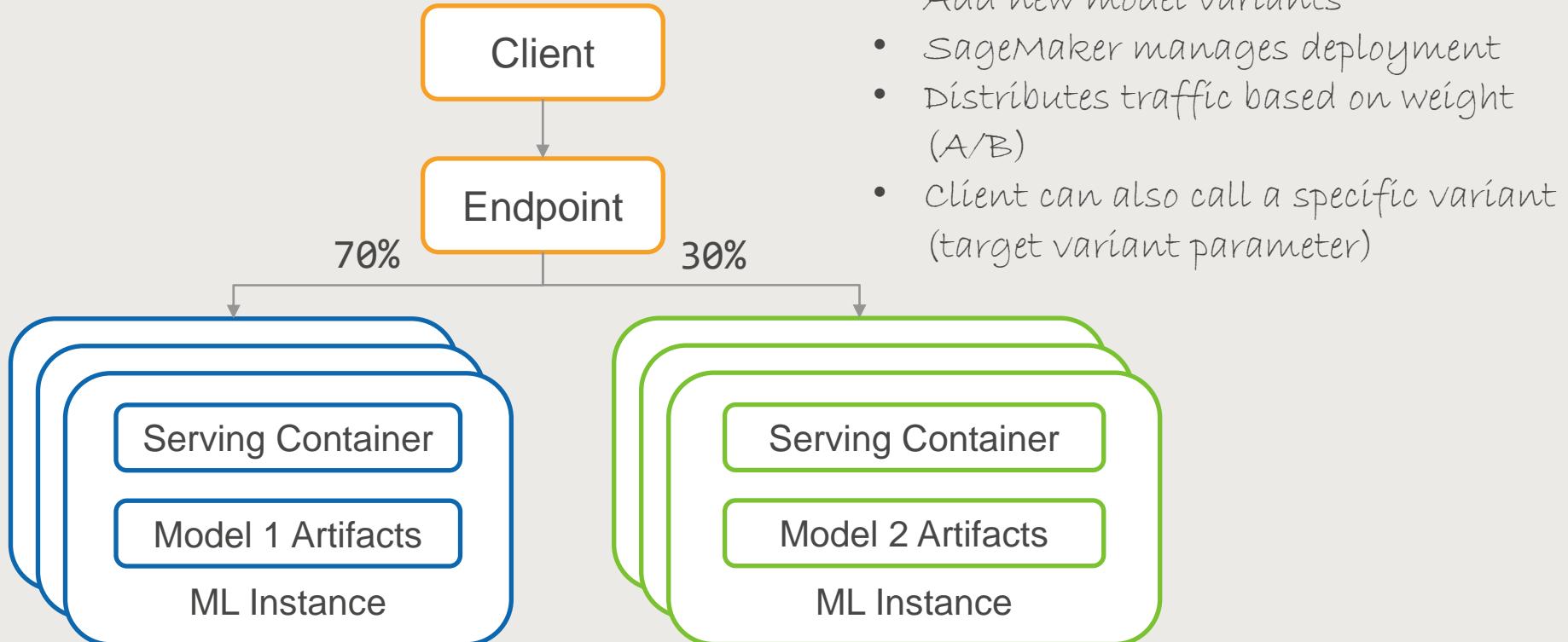
Single-model Endpoint



Three components

- Virtual server instance
- Serving container
- Model artifacts

Multiple Production Variants



Multiple Production Variants (with single-model endpoint)

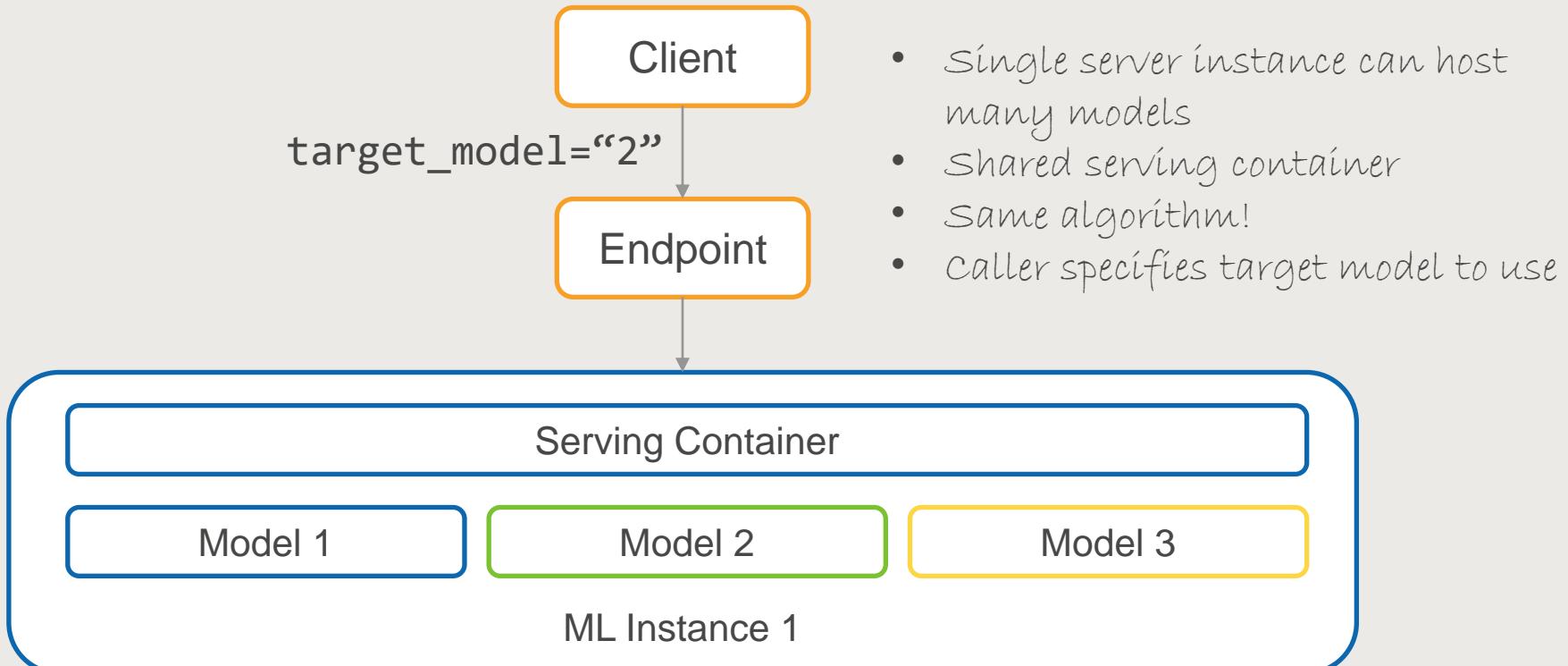
Advantages

- Make model changes with zero-downtime
- Distribute traffic based on weight
- Client can also use a specific variant
- AutoScaling rules by variant
- Mix and Match Algorithms

Disadvantages

- Each model is hosted in a separate server instance
- One serving container and model artifact per instance
- Too many servers!

Multi-model Endpoint



Multi-model Endpoint

Advantages

- Single serving container to host many models
- Reduce infrastructure costs
- Automatically host new models when you add new model artifacts in S3

Disadvantages

- Cold start delay when a model is invoked for the first time
- Container needs to download artifacts from S3 and host
- Models need to be algorithm compatible
- Caller must keep track of models and their S3 location
- Only some of the algorithms support multi-model hosting

Multi-model algorithm support (2021)

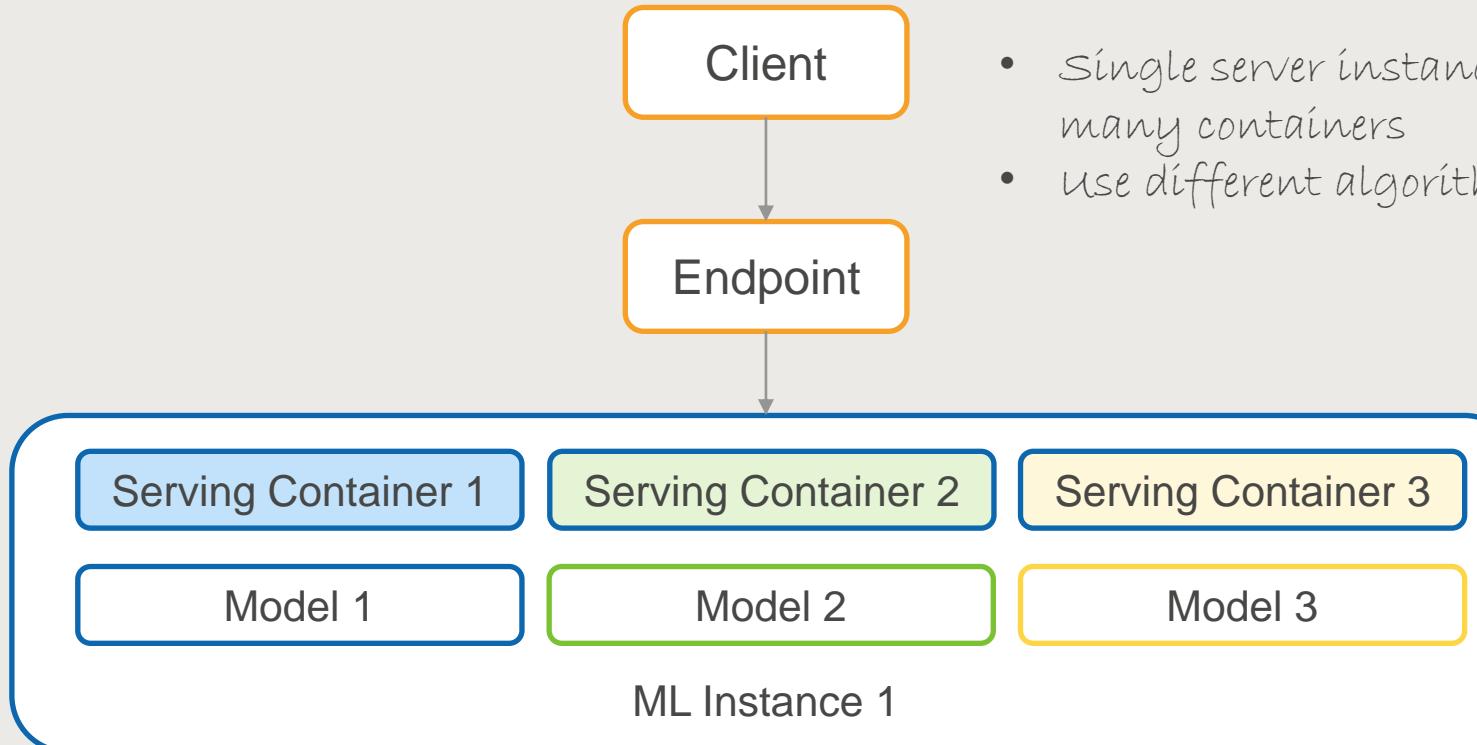
*** Check AWS documentation for updated list ***

The screenshot shows a web browser displaying the AWS Documentation for Amazon SageMaker. The URL is docs.aws.amazon.com/sagemaker/latest/dg/multi-model-endpoints.html. The page title is "Host Multiple Models with Multi-Model Endpoints". The left sidebar contains a navigation tree under "Multi-Model Endpoints", including "Create a Multi-Model Endpoint", "Invoke a Multi-Model Endpoint", "Add or Remove Models", "Bring Your Own Container" (with sub-links for "Security", "CloudWatch Metrics for Multi-Model Endpoint Deployments"), "Deploy multi-container endpoints", "Automatically Scale Models", "Host Instance Storage Volumes", and "Test models in production". The main content area is titled "Supported Algorithms and Frameworks" and lists the following supported items:

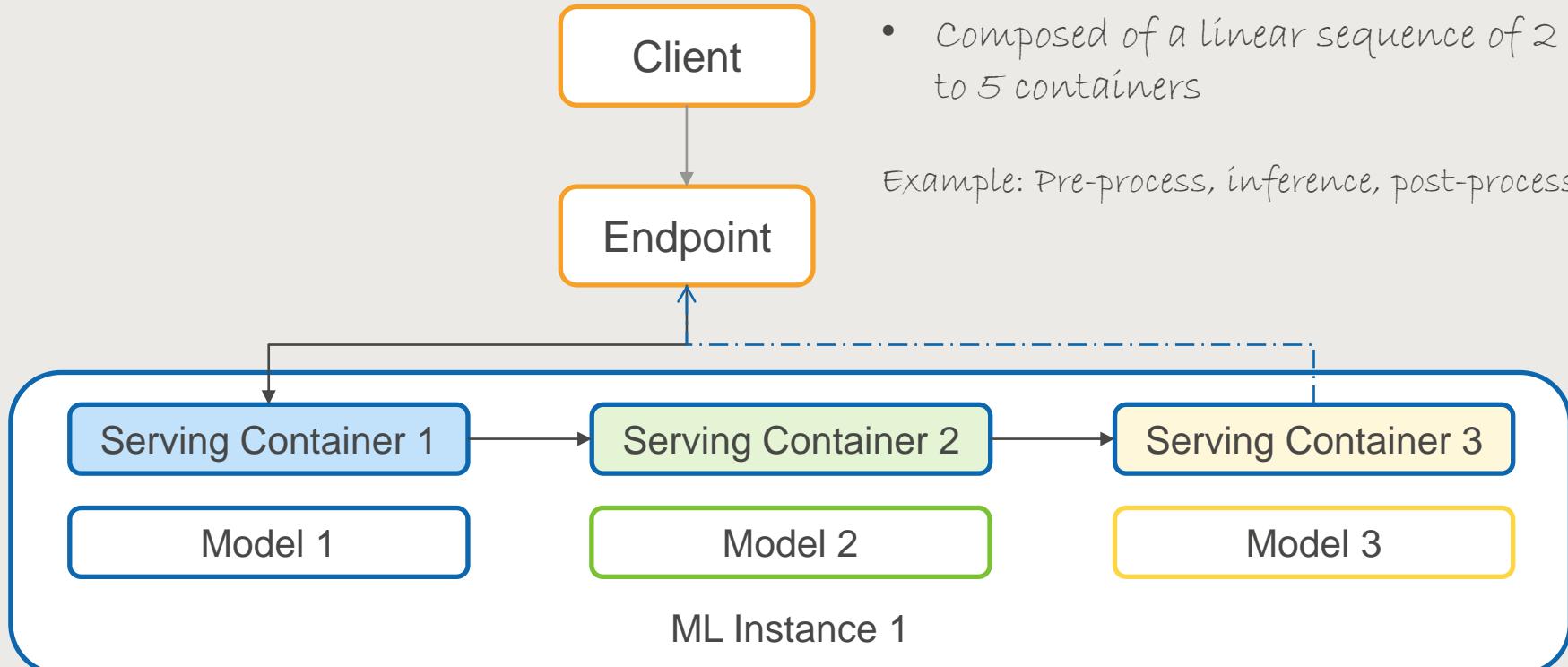
- [XGBoost Algorithm](#)
- [K-Nearest Neighbors \(k-NN\) Algorithm](#)
- [Linear Learner Algorithm](#)
- [Random Cut Forest \(RCF\) Algorithm](#)
- [Use TensorFlow with Amazon SageMaker](#)
- [Use Scikit-learn with Amazon SageMaker](#)
- [Use Apache MXNet with Amazon SageMaker](#)
- [Use PyTorch with Amazon SageMaker](#)

Below this list, a note states: "To use any other framework or algorithm, use the SageMaker inference toolkit to build a container that supports multi-model endpoints. For information, see [Build Your Own Container with Multi Model Server](#)".

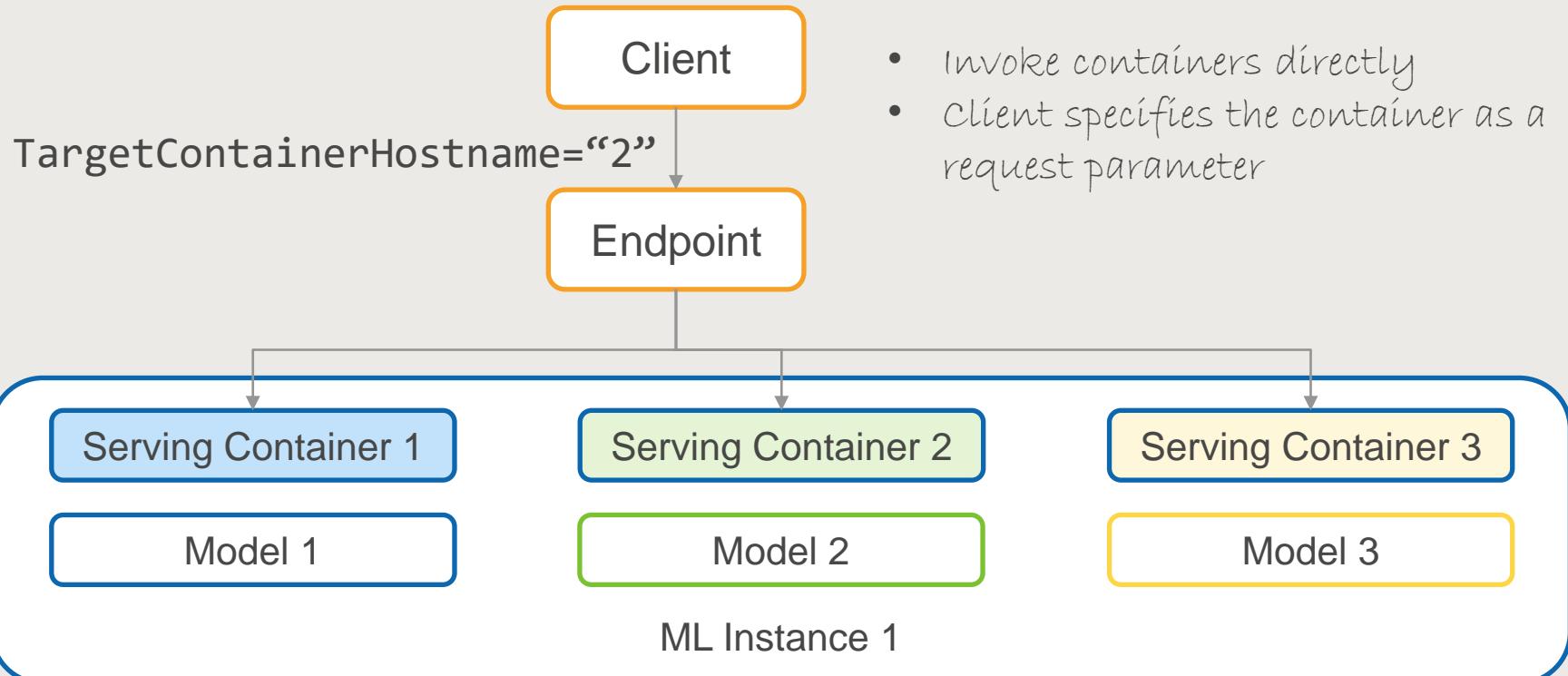
Multi-container Endpoint



Inference Pipeline Mode



Direct Mode



Multi-container Endpoint

Advantages

- Host multiple models and algorithms in an instance
- Chain containers as an inference pipeline
- Invoke containers directly
- Reduce infrastructure costs

Disadvantages

- Memory and storage are shared among containers
- Cross-interference, one bad serving container can affect stability
- May need more powerful instances
- Limit of 5 containers that can be co-hosted

Summary

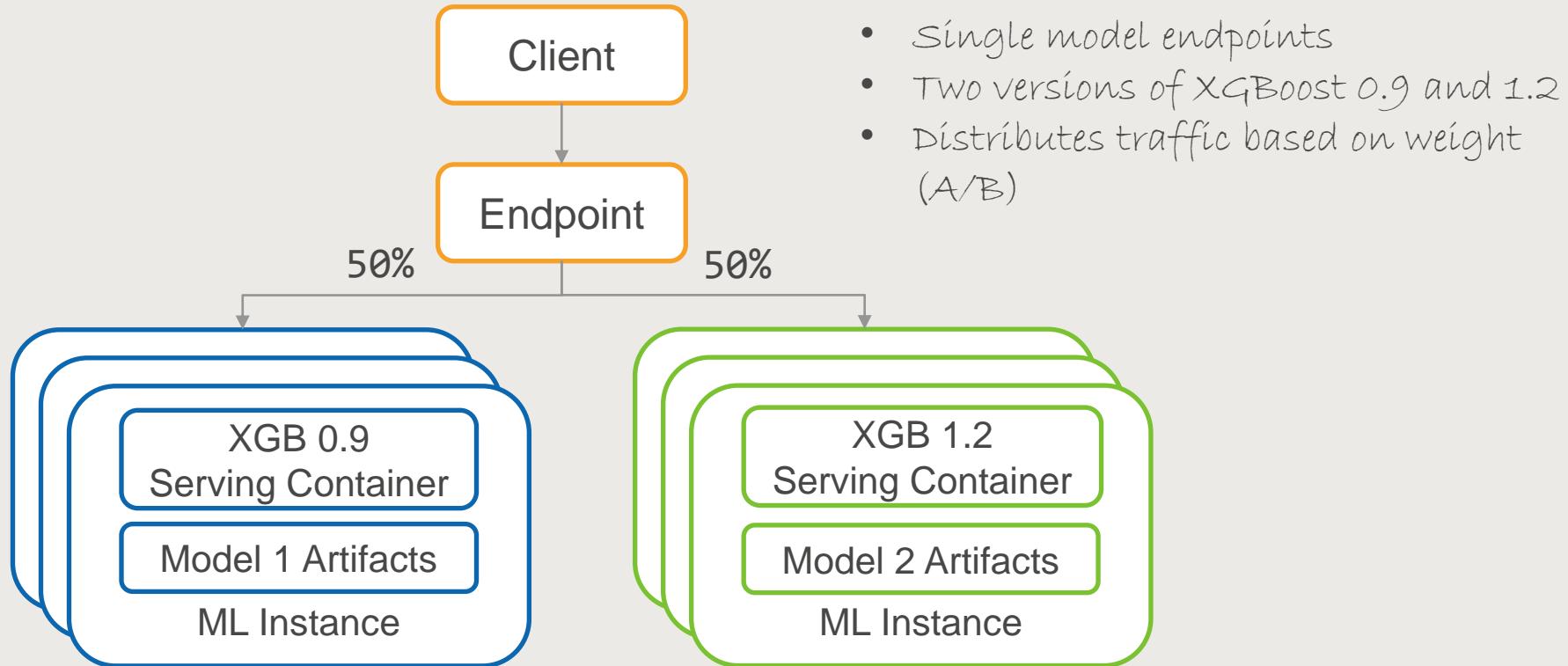
1 Single-model endpoint

2 Multiple production variants

3 Multi-model endpoint

4 Multi-container endpoint

Lab – A/B Testing Multiple Production Variants



XGBoost Versions

The screenshot shows a web browser displaying the AWS Documentation for the Amazon SageMaker Developer Guide. The URL in the address bar is docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html. The page content is about the XGBoost Algorithm, specifically its supported versions.

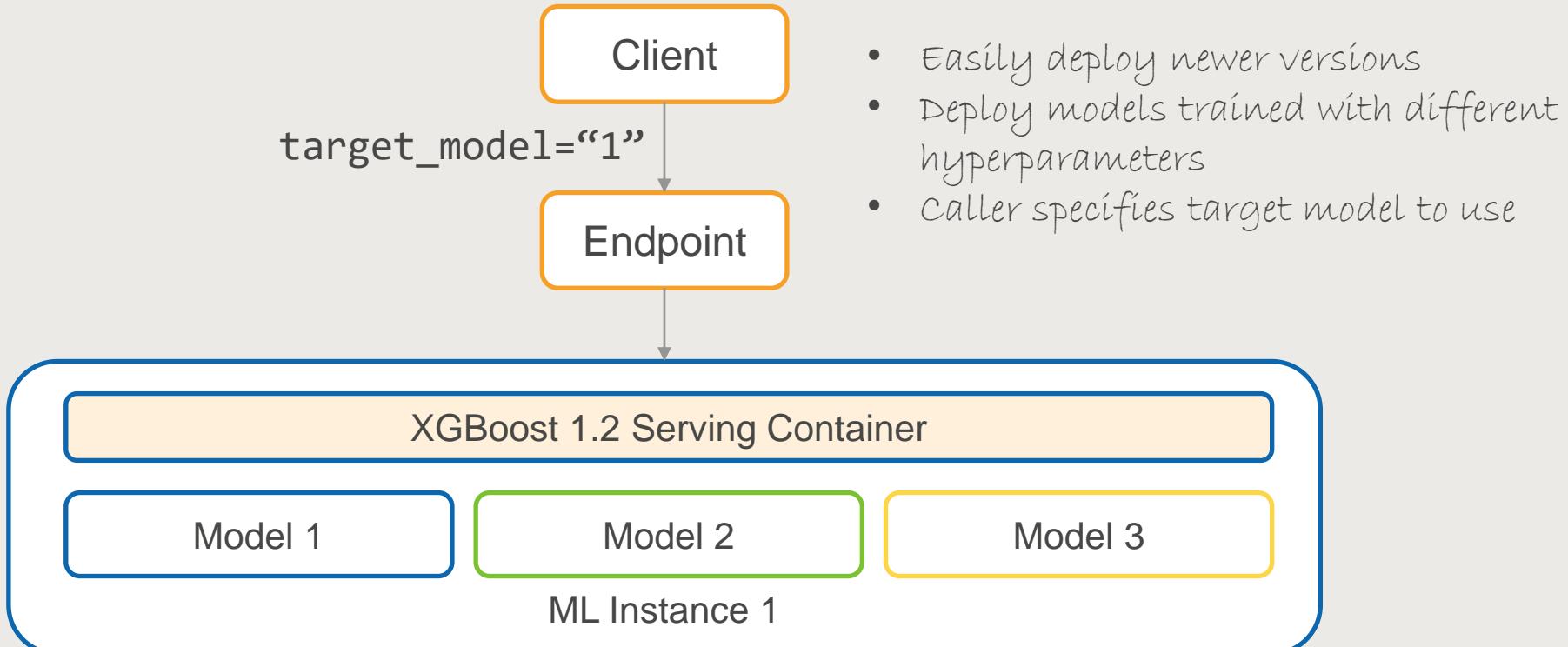
Supported versions

- Framework (open source) mode: 0.90-1, 0.90-2, 1.0-1, 1.2-1, 1.2-2, 1.3-1
- Algorithm mode: 0.90-1, 0.90-2, 1.0-1, 1.2-1, 1.2-2, 1.3-1

Note

XGBoost 1.1 is not supported on SageMaker because XGBoost 1.1 has a broken capability to run prediction when the test input has fewer features than the training data in LIBSVM inputs. This capability has been restored in XGBoost 1.2. Consider using SageMaker XGBoost 1.2-2 or later.

Lab – Multi-model Endpoint





Chandra Lingam

70,000+ Students



AWS Certified Machine Learning Specialty

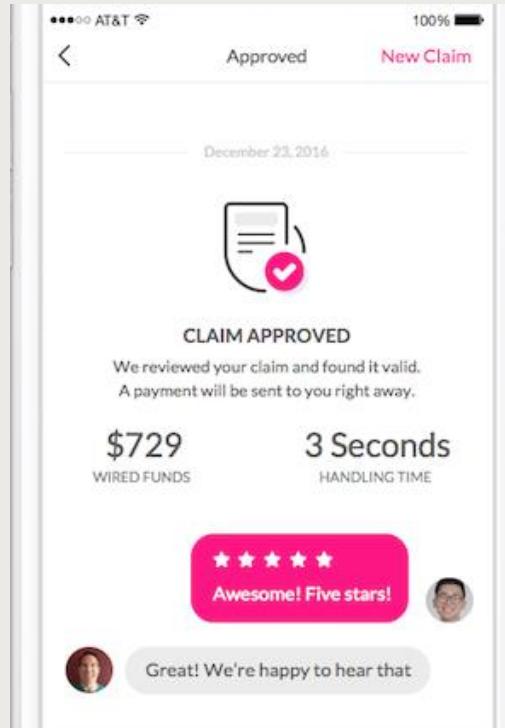
Fairness and Model Explainability

Fairness

“A computer system might be considered biased if it discriminates against certain individuals or groups of individuals.”

<https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-detect-data-bias.html>

AI in Action - Insurance



Fastest insurance claim paid out to a customer: world record set by Lemonade's claims bot AI Jim

"reviewed a customer's claim, cross referenced it with the policy, ran 18 anti-fraud algorithms on it, approved the claim, sent wiring instructions to the bank, and informed the client the claim was closed - all within three seconds and zero paperwork"

Accusations of Bias and Discrimination

“Lemonade tweeted about what it means to be an AI-first insurance company. It left a sour taste in many customers’ mouths”

<https://www.vox.com/recode/22455140/lemonade-insurance-ai-twitter>

Lemonade  @Lemonade_Inc · 1T A typical homeowners policy form has 20-40 fields (name, address, bday...), so traditional insurers collect 20-40 data points per user.

AI Maya asks just 13 Q's but collects over 1,600 data points, producing nuanced profiles of our users and remarkably predictive insights. (2/7)

15 51 36 

Lemonade  @Lemonade_Inc · 1T This data helps us understand the level of risk each customer brings, which improves our underwriting, customer acquisition, and fraud detection. (3/7)

5 20 36 

Lemonade  @Lemonade_Inc · 1T For example, when a user files a claim, they record a video on their phone and explain what happened.

Our AI carefully analyzes these videos for signs of fraud. It can pick up non-verbal cues that traditional insurers can't, since they don't use a digital claims process. (4/7)

914 2.044 174 

Lemonade  @Lemonade_Inc So, we deleted this awful thread which caused more confusion than anything else.

TL;DR: We do not use, and we're not trying to build AI that uses physical or personal features to deny claims (phrenology/physiognomy) (1/4)

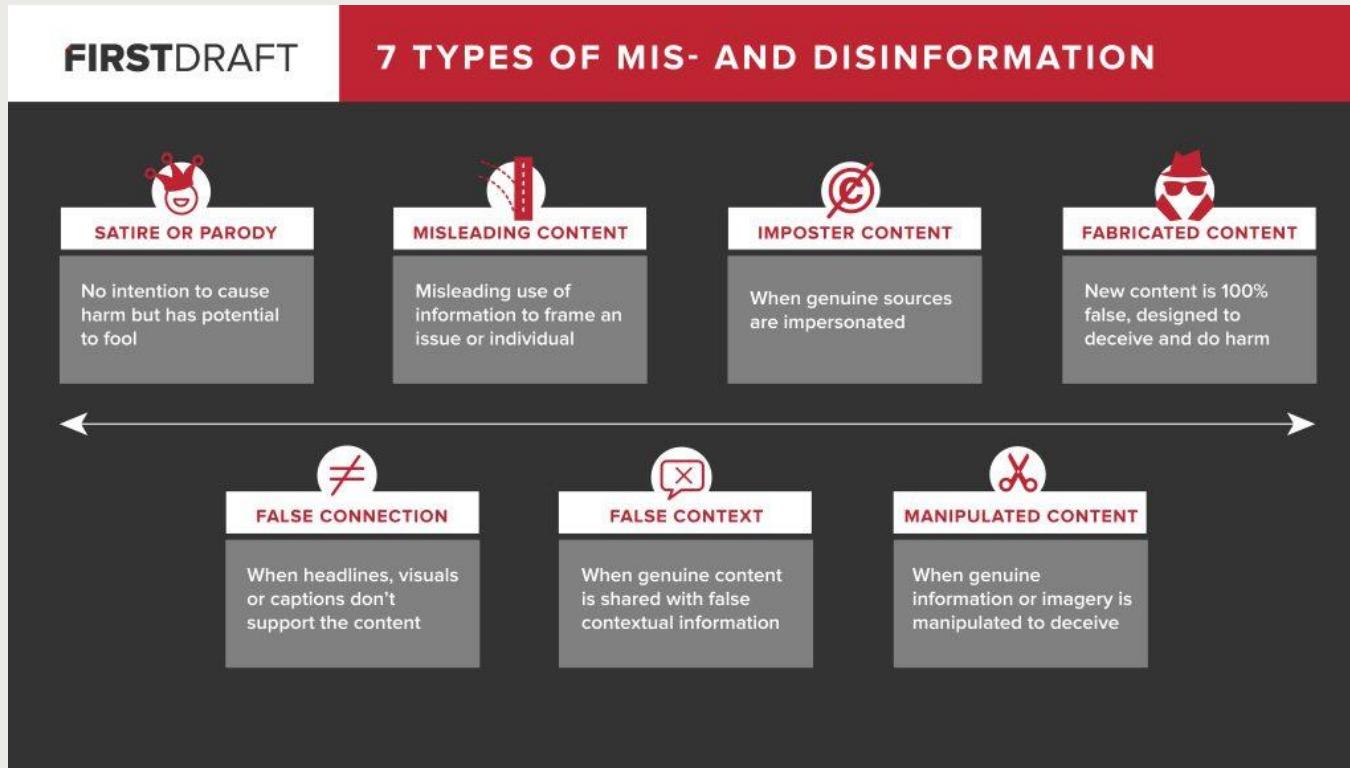
10:45 AM · May 26, 2021 

313 367  Share this Tweet

Fake News on Social Media

Fake news. It's complicated. By Claire Wardle

<https://firstdraftnews.org/articles/fake-news-complicated/>



AI in Action – Social Media

**Twitter promises to fine-tune its
5G coronavirus labeling after
unrelated tweets were flagged**

*Tweets with the words “oxygen”
and “frequency” were being
tagged with a fact-check label*

<https://www.theverge.com/2020/6/27/21305503/twitter-labels-5g-conspiracy-coronavirus>



Twitter Support  @TwitterSupport

In the last few weeks, you may have seen Tweets with labels linking to additional info about COVID-19. Not all of those Tweets had potentially misleading content associating COVID-19 and 5G. We apologize for any confusion and we're working to improve our labeling process. (1/4)

7:39 PM · Jun 26, 2020

 1.3K  See the latest COVID-19 information on Twitter

AI in Action – Image Cropping

Twitter says its image-cropping algorithm was biased, so it's ditching it

“when tested on randomly linked images of people of various races and genders, favored White people over Black people and women over men”

<https://www.cnn.com/2021/05/19/tech/twitter-image-cropping-algorithm-bias/index.html>

AI in Action – Credit Card

The Apple Card Didn't 'See' Gender—and That's the Problem

- “*users noticed that it seemed to offer smaller lines of credit to women than to men*”
- “*No one from the company seemed able to describe how the algorithm even worked, let alone justify its output.*”

<https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/>

Algorithm is gender-blind

“Goldman landed on what sounded like an ironclad defense: The algorithm, it said, has been vetted for potential bias by a third party; moreover, it doesn’t even use gender as an input. How could the bank discriminate if no one ever tells it which customers are women and which are men?”

<https://www.wired.com/story/the-apple-card-didnt-see-gender-and-thats-the-problem/>

Proxies

The idea that removing an input eliminates bias is "*a very common and dangerous misconception*," says [Rachel Thomas](#), a professor at the University of San Francisco and the cofounder of [Fast.ai](#), a project that teaches people about AI

<https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/>



I'm an AI researcher, and here's what scares me about AI



Rachel Thomas

fast.ai co-founder & professor
USF Data Institute | twitter:
@math_rachel

[Follow](#)

1.5K



8



AI is being increasingly used to make important decisions. Many AI experts (including [Jeff Dean](#), head of AI at Google, and [Andrew Ng](#), founder of Coursera and deeplearning.ai) say that warnings about sentient robots are overblown, but other harms are not getting enough attention. I agree. I am an AI researcher, and [I'm worried](#) about some of the societal impacts that we're already seeing. In particular, these 5 things scare me about AI:

1. Algorithms are often implemented without ways to address mistakes.
2. AI makes it easier to not feel responsible.
3. AI encodes & magnifies bias.
4. Optimizing metrics above all else leads to negative outcomes.
5. There is no accountability for big tech companies.

At the end, I'll briefly share some positive ways that we can [try to address these](#).

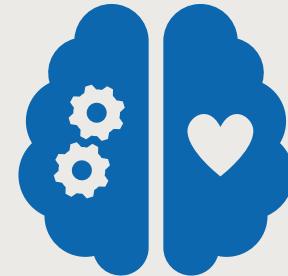
<https://twitter.com/janellecshane/status/1405598023619649537>



Challenges - What is the definition of fairness?



How to build models that
are fair?



How would you prove that
your model is not biased?

Types of Bias

1. Data Bias
2. Model Bias
3. Inference Bias

<https://aws.amazon.com/sagemaker/clarify/>

Data bias

If your data prefers a particular ethnic group or race, or age or accent, then the model trained on that data will also reflect or even amplify that bias

Similarly, a dataset that contains too many negative samples for one group may train a model to discriminate against that group

<https://aws.amazon.com/sagemaker/clarify/>

Model bias

A model can introduce bias if the prediction behavior is not consistent across different groups such as age, or gender, or income brackets

This behavior could be due to data or from bias introduced by the algorithm

"For instance, if an ML model is trained primarily on data from middle-aged individuals, it may be less accurate when making predictions involving younger and older people."

<https://aws.amazon.com/sagemaker/clarify/>

Inference bias

The deployed model is showing signs of bias. The training data and the model were okay.

This can happen if the training data distribution and production data distribution are different

"For example, the outputs of a model for predicting home prices can become biased if the mortgage rates used to train the model differ from current, real-world mortgage rates"

<https://aws.amazon.com/sagemaker/clarify/>

SageMaker Tools

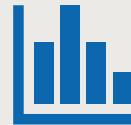
Detect bias and explain the model behavior



Clarify



Experiments



Model
Monitor



Augmented
AI

SageMaker Clarify

“Detect bias in ML models and understand model predictions”

- Unified capability (consolidates data from other tools)
- Detect bias
 - During data preparation
 - After model training
 - Deployed Models
- Tools to help explain model predictions

Explainability

Clarify uses model-agnostic feature-attribution approach to explain predictions

Uses game-theory to assign each feature an importance value [Shapley values]

- Why did the model reject a particular loan application?
- How does the model make predictions?
- Why did this model make an incorrect prediction?
- Which feature has the most significant influence on the behavior of the model?

Fairness metrics

How would you define fairness? How would you measure it?

Clarify Metrics

- At least eight different metrics for data bias
- Eleven other metrics for model bias
- Metrics to measure drift in live data

Metrics to quantify data bias

The screenshot shows a browser window displaying the AWS SageMaker Developer Guide. The URL is docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html. The page title is "Pretraining Bias Metrics". The left sidebar has a tree view under "Bias" with the "Measure Pretraining Bias" node expanded, highlighted with a red box. The main content area shows a table with four columns: "Bias metric", "Description", "Example question", and "Interpreting metric values". The first row in the table corresponds to the "Class Imbalance (CI)" metric from the sidebar.

Bias metric	Description	Example question	Interpreting metric values
Class Imbalance (CI)	Measures the imbalance in the number of members between different facet values.	Could there be age-based biases due to not having enough data for the demographic outside a middle-aged facet?	Normalized range: [-1,+1] Interpretation: <ul style="list-style-type: none">Positive values indicate the facet a has more training samples in the dataset.Values near zero indicate the facets are balanced in the number of training samples in the dataset.

Metrics to quantify post training model bias

The screenshot shows a web browser displaying the AWS SageMaker Developer Guide. The URL is docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-post-training-bias.html. The page title is "Posttraining Bias Metrics". On the left, there is a sidebar with a red box highlighting the "Measure Posttraining Data and Model Bias" section, which lists various metrics: DPPL, DI, DCAcc, DCR, RD, DAR, DRR, and Accuracy.

Posttraining Bias Metrics			
posttraining bias metric	Description	Example question	Interpreting metric values
Difference in Positive Proportions in Predicted Labels (DPPL)	Measures the difference in the proportion of positive predictions between the favored facet a and the disfavored facet d .	Has there been an imbalance across demographic groups in the predicted positive outcomes that might indicate bias?	Range for normalized binary & multiclass facet labels: $[-1, +1]$ Range for continuous labels: $(-\infty, +\infty)$ Interpretation: <ul style="list-style-type: none">Positive values indicate that the favored facet a has a higher proportion of predicted positive outcomes.Values near zero indicate a more equal proportion of predicted positive outcomes between facets.Negative values indicate the disfavored facet d has a higher proportion of predicted positive outcomes.

Metrics to quantify drift in production model

The screenshot shows a browser window displaying the AWS Documentation for Amazon SageMaker. The URL is docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-monitor-feature-attribution-drift.html. The page title is "Monitor Feature Attribution Drift for Models in Production". On the left, there is a navigation sidebar with a red box highlighting the "Monitor Feature Attribution Drift" section under "Monitor Model Quality". The main content area describes how drift in live data distribution can lead to feature attribution drift and provides a hypothetical scenario for college admissions.

A drift in the distribution of live data for models in production can result in a corresponding drift in the feature attribution values, just as it could cause a drift in bias when monitoring bias metrics. Amazon SageMaker Clarify feature attribution monitoring helps data scientists and ML engineers monitor predictions for feature attribution drift on a regular basis. As the model is monitored, customers can view exportable reports and graphs detailing feature attributions in SageMaker Studio and configure alerts in Amazon CloudWatch to receive notifications if it is detected that the attribution values drift beyond a certain threshold.

To illustrate this with a specific situation, consider a hypothetical scenario for college admissions. Assume that we observe the following (aggregated) feature attribution values in the training data and in the live data:

College Admission Hypothetical Scenario		
Feature	Attribution in training data	Attribution in live data
SAT score	0.70	0.10
GPA	0.50	0.20
Class rank	0.05	0.70

Complex collection of metrics!

"single, universal definition of fairness or a metric to measure it will probably never be possible. Instead, different metrics and standards will likely be required, depending on the use case and circumstances."

Tackling bias in artificial intelligence (and in humans)

<https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans>

SageMaker Experiments

Need to optimize for predictive quality, and fairness of predictions!

May have to train 1000s of models

Hard to track best-performing models, and their input configurations

SageMaker Experiments

SageMaker Experiments automatically tracks the input, parameters, configurations, and results as trials

clarify consolidates data to provide a feature importance graph to explain model's overall decision-making process after the model has been trained.

SageMaker Model Monitor

Continuously monitor quality of models in production

Configure alerts when deviations in model quality, bias drift

Model monitor is integrated with Clarify

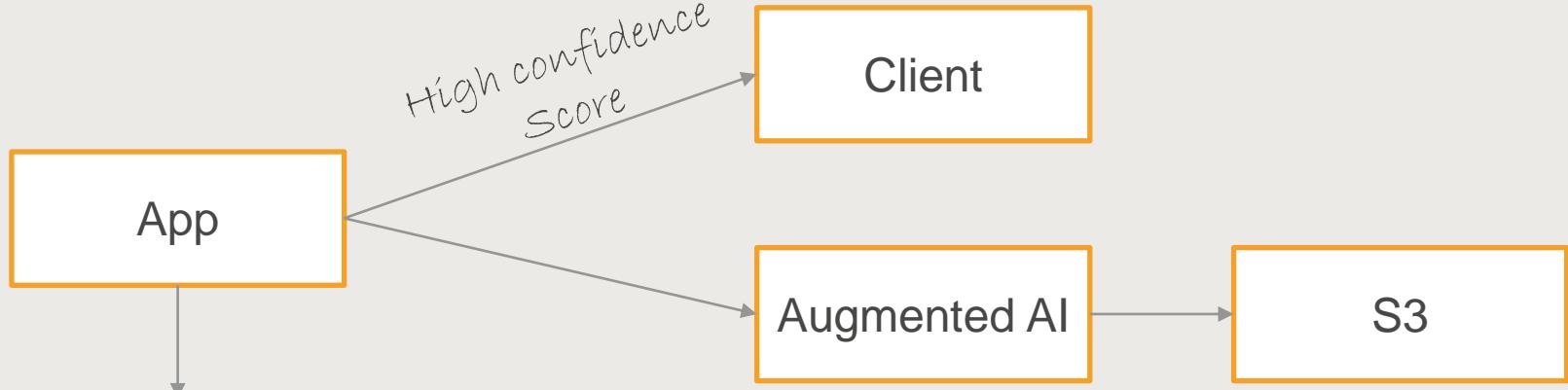
Amazon Augmented AI

Bring human-in-the-loop

Human oversight for ML predictions

Combine the benefits of ML and human-review

Amazon Augmented AI



Human Review

- Low confidence score
- Random sampling
- Single reviewer or multiple-reviewers

Augmented AI – Workforce options

Amazon
Mechanical Turk

A crowd sourced marketplace of reviewers

Suitable for public-data, non-confidential data

Private
workforce

Reviewers are your own employees

Ideal for customer confidential data

Labeling Service
Providers
(AWS
Marketplace)

Suitable for customer confidential data

Service agreement and clauses to protect customer data

SageMaker Tools

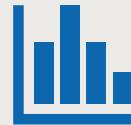
Detect bias and explain the model behavior



Clarify



Experiments



Model
Monitor



Augmented
AI

Summary



Regulatory Compliance

Policymakers, Regulators, Advocates
Ethics and policy challenges posed by AI
Companies may have to explain how AI makes decision



Internal Reporting and Compliance

Adoption of AI requires Trust
Explain behavior of trained models, How they make predictions



Customer Service

Financial advisors, Loan officers may review predictions made by AI system
Communicate to customers



Chandra Lingam

70,000+ Students



AWS Certified Machine Learning Specialty

SageMaker Tools

SageMaker is expanding rapidly with the addition of several tools!

But do remember that, like any other AWS service, not all these tools will gain traction and usage. Usually, new services will come in, and existing services will also silently disappear. For now, getting basic familiarity is sufficient for the exam.

You may see a couple of questions on the below topics.

SageMaker Studio

SageMaker Studio is an AWS developed fully integrated development environment (IDE) for Machine Learning

The Studio extends the standard Jupyter notebook instances (referred to as instance-based notebooks) in a few different ways

1. You can quickly launch Studio notebooks without requiring to provision an instance manually
2. Studio notebook startup-time is 5-10 times faster than instance-based notebooks
3. Flexibility to choose from an extensive collection of instance types
4. Each user gets an isolated home directory in EFS (Elastic File System)
5. The user home directory is automatically mounted into all studio notebooks that you use. So, you can access your files from any instance
6. You can easily share your notebook with your peers and colleagues
7. Studio is integrated with AWS Single Sign-On (SSO). You can use your corporate credentials to access your Studio notebook (no need for AWS user credentials)
8. Studio has a **JumpStart** capability that includes 150+ pre-trained open-source models from PyTorch Hub and TensorFlow Hub.
9. JumpStart also includes example solutions for image processing, natural language processing, demand forecasting, fraud detection, and so forth
10. Studio has visualization capabilities for SageMaker Clarify, SageMaker Experiments, Model Monitor, Debugger, Data Wrangler, Pipelines [Looks like AWS is moving towards using Studio as the primary IDE for interacting and integrating with other SageMaker tools]

SageMaker Debugger

Debugger help you track your training job and identify issues such as CPU, GPU, Disk IO, Network, Memory bottlenecks, Model issues such as Overfitting

Metrics are visualized using SageMaker Studio along with remediation advice

Data Wrangler

Data Wrangler is a data analysis and preparation tool for machine learning applications. You can think of this as an extract-transform-load (ETL) type tool.

"You can create a data flow to combine datasets from different data sources, identify the number and types of transformations you want to apply to datasets, and define a data prep workflow that can be easily integrated into an ML pipeline."

1. You can import data from one or more sources such as S3, Athena, Redshift
2. Transform the data for your needs. For example, string, number and date formatting, feature engineering, categorical encoding, and so forth
3. Analyze features in your dataset with built-in visualization such as histogram, scatter plots, correlation
4. Export data

Auto Pilot

Before SageMaker was born, AWS had another Machine Learning capability simply named AWS Machine Learning Service that can automatically create machine learning models and tune them. One fine morning, AWS announced they were phasing out that product (I got a panic attack and had to rewrite the entire course – six months gone)

Auto Pilot is the latest incarnation of that tool!

You simply provide a tabular dataset and specify the target column to predict (regression, binary classification, and multi-class classification)

Auto Pilot will automatically explore the dataset, do feature engineering, and explore different solutions to find the best model

It will automatically try multiple algorithms such as linear models, XGBoost, and Deep Learning

You can directly deploy the model to production

<https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-model-support-validation.html>

<https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-automate-model-development.html>

Ground Truth

Ground Truth is an automatic labeling service.

Labeled data with correct answers are required for many supervised learning problems (classification, object detection, sentiments, and so forth)

However, labeling the dataset is very time consuming and labor-intensive

Ground Truth automates this process by using a combination of machine learning and human-in-the-loop workflow.

Ground Truth may sound similar to the Augmented AI capability that we saw in the previous lecture. Augmented AI is used for building a custom workflow for your machine learning solution. Whereas Ground Truth is explicitly intended for labeling

You need to provide sample data with correct labels

Ground Truth will use machine learning to learn from the sampled data and attempt to label the rest of the data

For low-confidence scores, Ground Truth will send the data to human labelers for review

The data labeled by humans are used to improve the model

This process repeats until all the raw data are labeled.

For human labelers, you can use:

- Mechanical Turk for crowdsourced labelers
- Private Workforce (your own employees) or
- AWS Marketplace labeling service providers

<https://aws.amazon.com/sagemaker/groundtruth/faqs/>

Distributed Training

When working with large datasets, you may run into a situation where the time to train a model is too long

One solution is to increase the number of CPUs and GPUs in the training instance

Another option is to scale by using multiple instances

AWS recommends that you try a larger instance before trying to increase the number of instances

If the problem requires multiple instances, you would need to ensure all the instances are in the same region and availability zone. This will ensure network latency is kept to a minimum as instances will frequently share the data and learned parameters.

SageMaker SDK ensures all training instances are deployed in the same availability zone

Many of the SageMaker built-in algorithms automatically support distributed training

If you write a custom algorithm, AWS recommends that you use SageMaker Distributed Data Parallel library and SageMaker Distributed Model Parallel library

You can approach distributed training in two ways: Data Parallel and Model Parallel

“Data parallel is the most common approach to distributed training: You have a lot of data, batch it up, and send blocks of data to multiple CPUs or GPUs (nodes) to be processed by the neural network or ML algorithm, then combine the results”

“A model parallel approach is used with large models that won’t fit in a node’s memory in one piece; it breaks up the model and places different parts on different nodes. In this situation, you need to send your batches of data out to each node so that the data is processed on all parts of the model.”

<https://docs.aws.amazon.com/sagemaker/latest/dg/distributed-training.html#distributed-training-scenarios>

FAQ

Please review SageMaker FAQs before your exam: <https://aws.amazon.com/sagemaker/faqs/>

Identity and Access Management

IAM

Chandra Lingam

Cloud Wave LLC

AWS Cloud Security

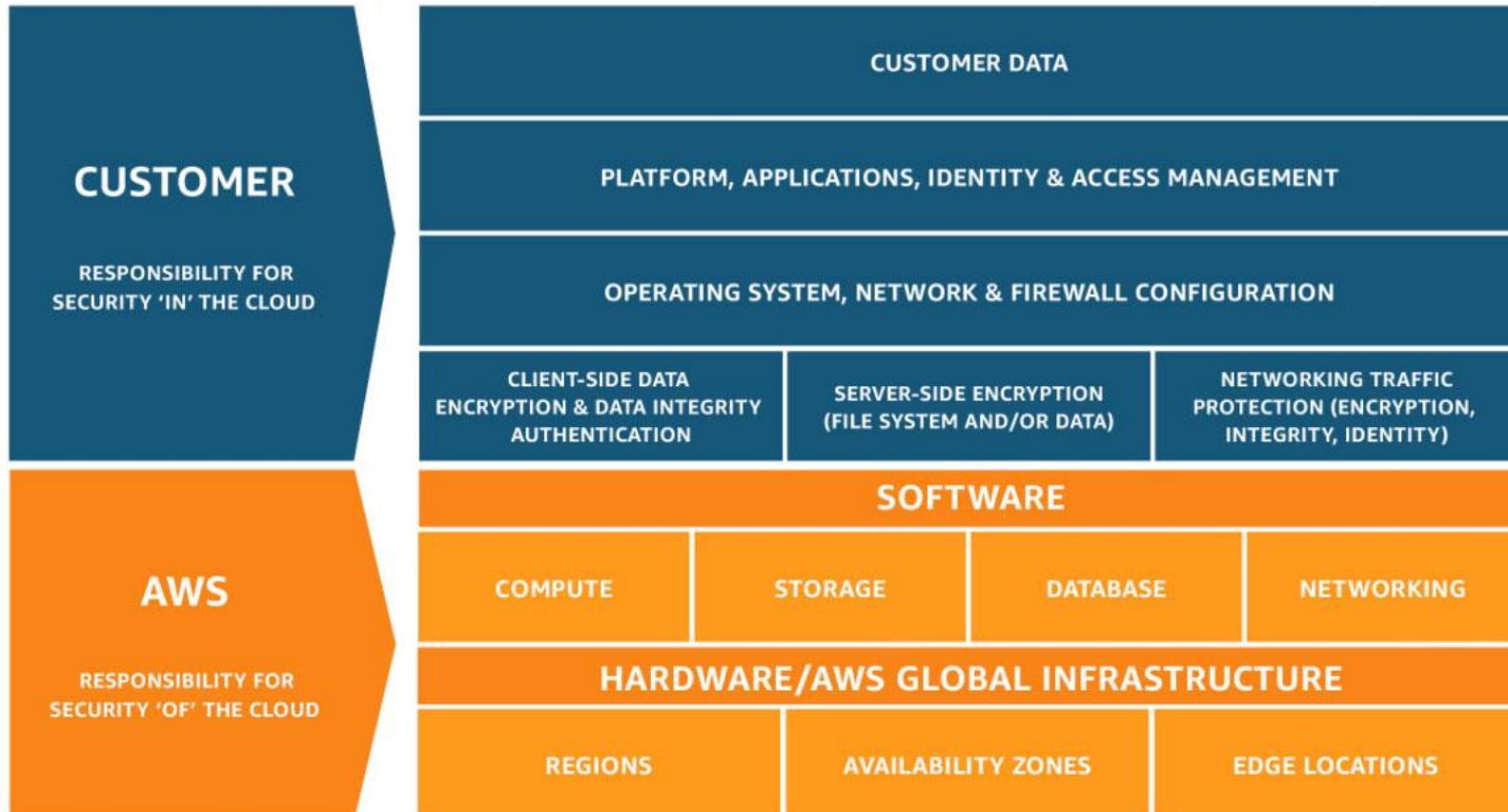


Image Source: [Shared Responsibility Model](#)

EC2

EC2

Customer Responsibilities

- Guest OS, Patching
- Firewalls (Security Group, Network ACL)
- Availability, Scalability, Monitoring

Physical Host

AWS Responsibilities

- Physical Host
- Virtualization

S3

Bucket

Customer Responsibilities

- Storage Class
- Access Controls
- Data Encryption

S3

AWS Responsibilities

- Hardware, Software
- Scalability

AWS Compliance Programs

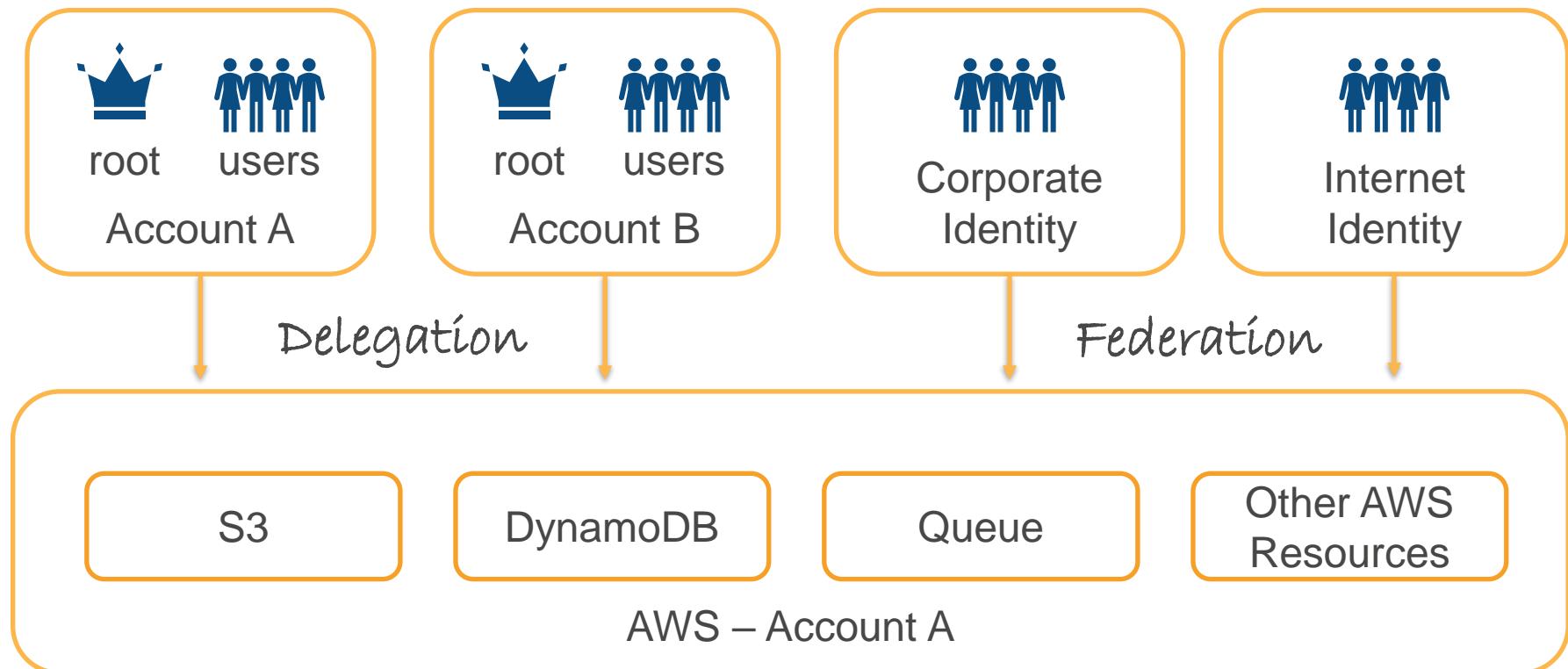
Useful Resources

[Security is Job Zero | AWS Public Sector Summit 2016](#)

Steve Schmidt

CISO, AWS

Types of Identities



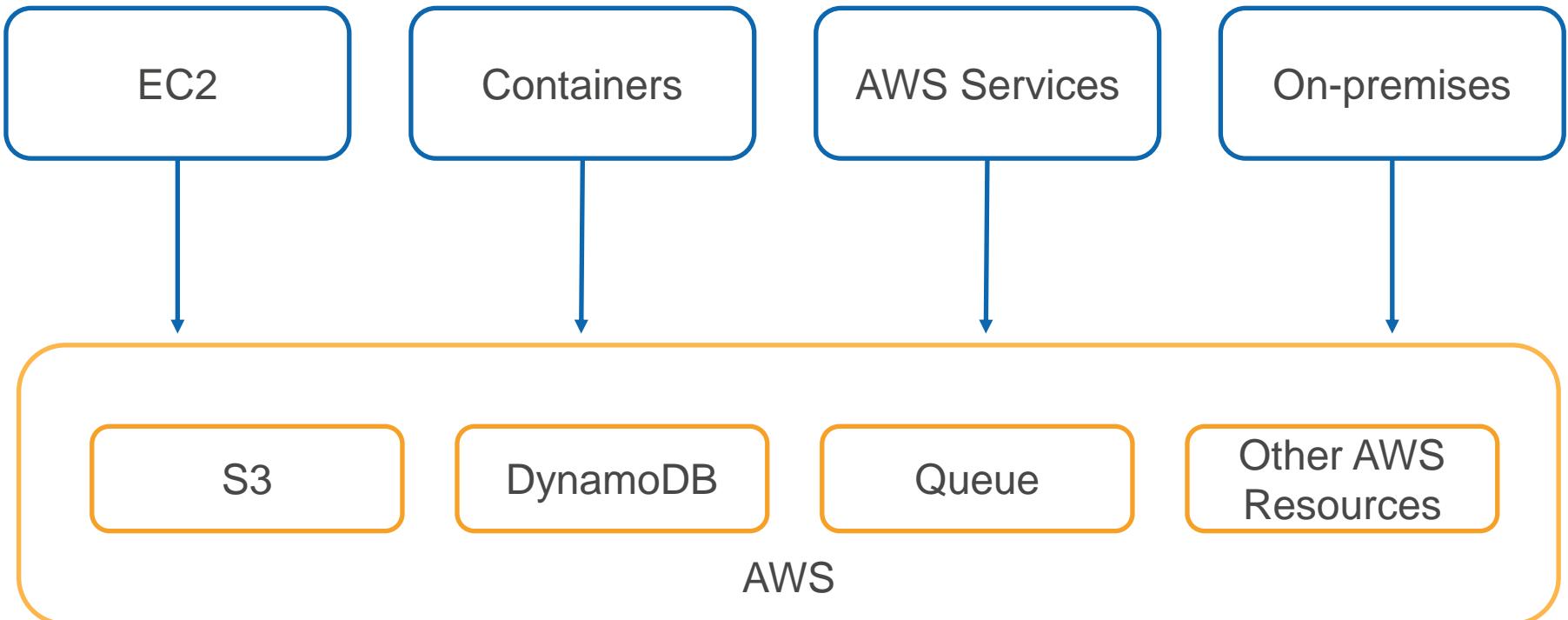
Zero Trust

AWS operates on principle of zero trust

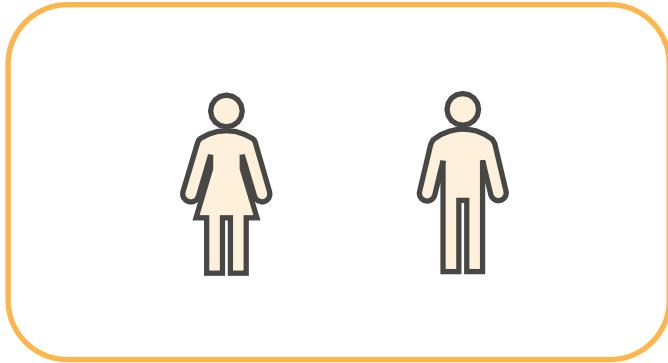
- Authentication – caller needs to prove identity
- Authorization – caller needs permission
- Some services allow anonymous access
 - S3 bucket with Public Access

Reference: [AWS re:Invent 2019: Getting started with AWS identity \(SEC209-R1\) by Becky Weiss](#)

Types of Applications

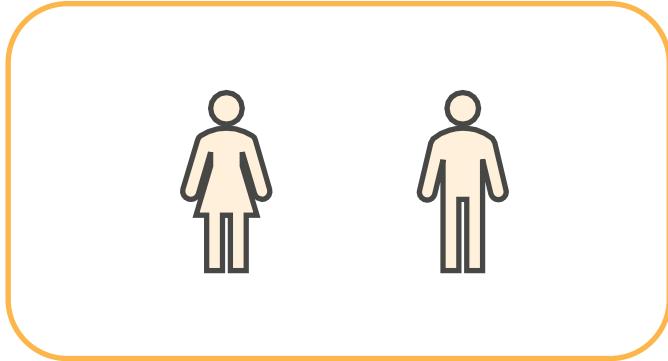


IAM User Sign-in Credentials



- No sharing of credentials
- User ID and Password for Management Console Access
- Access Key and Secret Access Key for CLI, Programmatic Access
- Optional Multi-Factor Authentication (MFA)

IAM User Access Management



- No resource access by default
- Identity-based policy - Attach Policy to the User
- Resource-based policy - Attach Policy to the resource (only for supported AWS resources like S3, SQS, Lambda and so forth)
- IAM Roles - Grant temporary access to resources (and user gains privileges assigned to that Role)

Sample Identity-based Policy

```
{  
    "version": "2012-10-17",  
    "Statement": [  
        {  
            "Effect": "Allow",  
            "Action": [  
                "s3:Get*",  
                "s3>List*"  
            ],  
            "Resource": "*"  
        }  
    ]  
}
```

Reference: <arn:aws:iam::aws:policy/AmazonS3ReadOnlyAccess>

Sample Resource-based Policy

```
{"version": "2012-10-17",
"statement": [
{
    "Effect": "Allow",
    "Action": ["s3:Get*", "s3>List*"],
    "Resource": [
        "arn:aws:s3:::bucket_name",
        "arn:aws:s3:::bucket_name/*"],
    "Principal": {
        "AWS": [
            "arn:aws:iam::AWS-account-ID:user/alice",
            "arn:aws:iam::AWS-account-ID:user/bob"]}
    }
]
}
```

} NOTE: [Groups](#) are not supported as principals in any policy

Access Management Concepts

Permissions

Policy

Attach

User

Group

Resource

Role

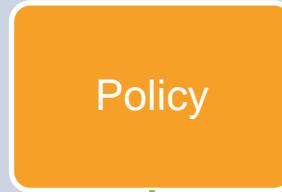
Inherit

User

User

Policy Types

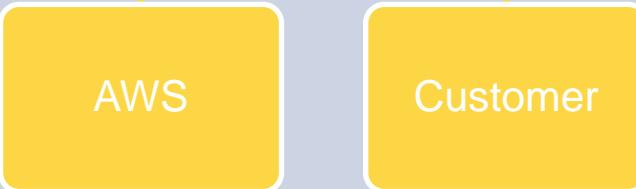
Permissions



Types



Owner



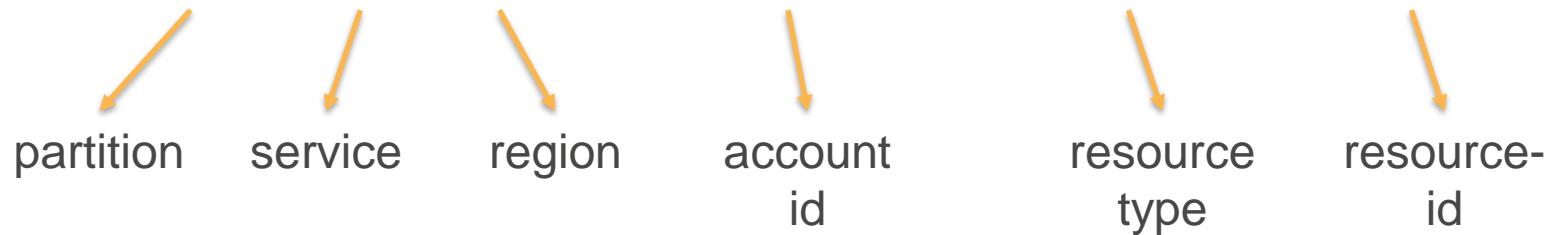
Policy Development and Testing

- Policy Visual Editor
- Policy Examples
https://docs.aws.amazon.com/IAM/latest/UserGuide/access_policies_examples.html
- Policy Simulator to Test Policy
https://docs.aws.amazon.com/IAM/latest/UserGuide/access_policies_testing-policies.html

Amazon Resource Name (ARN)

Uniquely identify resource and principal in AWS

arn:aws:iam::123456789012:user/alice



Amazon Resource Name (ARN)

Uniquely identify resource and principal in AWS

arn:aws:iam::123456789012:user/alice

arn:aws:iam::123456789012:policy/db_admin

arn:aws:s3:::my_bucket

arn:aws:sqs:us-east-2:123456789012:order

ARN – Structure and Examples

Structure

arn:partition:service:region:account-id:resource-id

arn:partition:service:region:account-id:resource-type/resource-id

arn:partition:service:region:account-id:resource-type:resource-id

Examples

Policy Document Structure (1/3)

Element	Description
Version	Current version of the policy language. You should always set the version element. Current version is “2012-10-17”.
Statement	Permissions are allowed or denied using Statements. A policy document can have one or more statements
Effect	Specifies if a statement allows or denies permission Valid Values: Allow, Deny
Principal	Identity for which the statement applies. Implied when attached to the identity-based policy (for example, a user) Needs to be specified in resource-based policy and IAM Role trust policy Groups are not supported as principals in any policy

Example of Principal

Principal	Example
AWS account, root user	"Principal": {"AWS": "arn:aws:iam::123456789012:root"} "Principal": {"AWS": "123456789012"}
IAM Users	"Principal": {"AWS": "arn:aws:iam::123456789012:user/alice"}
IAM Roles	"Principal": {"AWS": "arn:aws:iam::123456789012:role/cross-acct"}
AWS Services	"Principal": {"Service": "elasticmapreduce.amazonaws.com"}
Anonymous users	"Principal": "*"
Federated Users	"Principal": {"Federated": "www.amazon.com"}
<u>Assumed-role sessions</u>	Use the role session name to uniquely identify a session when the same role is assumed by different principals or for different reasons "Principal": {"AWS": "arn:aws:sts::123456789012:assumed-role/role-name/role-session-name"}

Policy Document Structure (2/3)

Element	Description
Action	<p>Service API actions for which statement applies</p> <p>Example:</p> <p>“Action” : “s3:Get*” - All S3 API Calls that have a Get Prefix</p> <p>“Action” : “s3:*” - All S3 API Calls</p>
Resource	<p>Resource for which the statement applies</p> <p>Example:</p> <p>“Resource” : “arn:aws:s3:::bucket_name” – Actions apply to specified bucket</p> <p>“Resource” : “arn:aws:s3:::bucket_name/*” – Actions apply to objects stored in the specified bucket</p>
Conditions	<p>Additional conditions for fine grained access</p> <p>Example: IP Address, Authentication Mechanism</p>

Sample Identity-based Policy

```
{  
    "version": "2012-10-17",  
    "Statement": [  
        {  
            "Effect": "Allow",  
            "Action": [  
                "s3:Get*",  
                "s3>List*"  
            ],  
            "Resource": "*"  
        }  
    ]  
}
```

Reference: <arn:aws:iam::aws:policy/AmazonS3ReadOnlyAccess>

Sample Resource-based Policy

```
{"version": "2012-10-17",
"statement": [
{
    "Effect": "Allow",
    "Action": ["s3:Get*", "s3>List*"],
    "Resource": [
        "arn:aws:s3:::bucket_name",
        "arn:aws:s3:::bucket_name/*"],
    "Principal": {
        "AWS": [
            "arn:aws:iam::AWS-account-ID:user/alice",
            "arn:aws:iam::AWS-account-ID:user/bob"]}
    }
]
}
```

} NOTE: [Groups](#) are not supported as principals in any policy

Policy Document Structure (3/3)

Element	Description
NotAction	<p>Matches everything except specified API action</p> <p>Example:</p> <p>“NotAction”：“s3:DeleteBucket” – All S3 actions except for deleting a bucket</p>
NotResource	<p>Match everything except the specified resource</p> <p>Example:</p> <p>“NotResource”:[“arn:aws:s3:::bucket_name”, “arn:aws:s3:::bucket_name/*”]</p>
NotPrincipal	<p>Match all principals except for specified principal</p> <p>Example:</p> <p>“Effect”：“Deny”, “NotPrincipal”:[“AWS”：“123456789012”]</p>

Sample – Limit access to S3 by Source IP

```
"Statement": [
{
  "Effect": "Deny",
  "Action": "s3:*",
  "Resource": [
    "arn:aws:s3:::bucket_name",
    "arn:aws:s3:::bucket_name/*"],
  "Principal": "*",
  "Condition": {
    "NotIpAddress": {
      "aws:SourceIp": "151.29.0.0/16"
    }
  }
}
```

Reference: https://docs.aws.amazon.com/IAM/latest/UserGuide/reference_policies_examples_aws_deny-ip.html

Sample Conditional Variables – VPC Endpoint Only Access

```
"Statement": [
{
    "Effect": "Deny",
    "Action": "s3:*",
    "Resource": [
        "arn:aws:s3:::bucket_name",
        "arn:aws:s3:::bucket_name/*"],
    "Principal": "*",
    "Condition": {
        "StringNotEquals": {
            "aws:sourceVpce": "vpce-0f143973exyzabcdef"
        }
    }
}
```

Reference: <https://docs.aws.amazon.com/AmazonS3/latest/dev/example-bucket-policies-vpc-endpoint.html>

Improved VPC Endpoint Only Access

```
"Statement": [
{
    "Effect": "Deny",
    "Action": "s3:*",
    "Resource": [
        "arn:aws:s3:::bucket_name",
        "arn:aws:s3:::bucket_name/*"] ,
    "NotPrincipal": {
        "AWS": [
            "arn:aws:iam::AWS-account-ID:root",
            "arn:aws:iam::AWS-account-ID:user/myadmin"] } ,
    "Condition": {
        "StringNotEquals": {
            "aws:sourceVpce": "vpce-0f143973exyzabcdef"
        }
    }
}
```

Reference: <https://docs.aws.amazon.com/AmazonS3/latest/dev/example-bucket-policies-vpc-endpoint.html>

Global Environment Data (request context)

Key	Description
aws:CurrentTime	Data and time of the request – use it for timebased restrictions
aws:RequestedRegion	AWS Region used in the request – use it to limit access only to specific region(s)
aws:PrincipalTag	Compare the tag attached to the principal making the request
aws:SecureTransport	Check if request was sent using SSL. Values: True or false
aws:SourceIP, aws:SourceVpc, aws:SourceVpce	Check if request comes from whitelisted IP, VPC or VPC Endpoint

Attribute Based Access Control (ABAC)

Allow actions only when cost center match

```
"Statement": [
{
    "Effect": "Allow",
    "Action": ["ec2:startInstances", "ec2:stopInstances"],
    "Resource": "*",
    "Condition": {
        "StringEquals": {
            "ec2:ResourceTag/CostCenter":
                "${aws:PrincipalTag/CostCenter}"
        }
    }
}
```

Reference: https://docs.aws.amazon.com/IAM/latest/UserGuide/reference_policies_examples_ec2-start-stop-match-tags.html

Role Based Access Control (RBAC)

- “Traditional authorization model used in IAM is called role-based access control. This is based on a person's job-role.” [in AWS context, role refers to IAM-role]
- “In RBAC model, you implement different policies for different job functions.”
- “As best policy, you grant the minimum permissions necessary for the job function. this is known as "granting least privilege".”
- “The disadvantage of RBAC model is that when employees add new resources, you must update policies to allow access to those resources.”

Attribute Based Access Control (ABAC)

- “ABAC is an authorization strategy that defines permissions based on attributes (also known as Tags)”
- “Tags can be attached to IAM Principals (users or roles) and to AWS resources”
- “With ABAC, policies can be designed to allow operations when the principal's tag matches the resource tag.”
- “ABAC is useful in environments that are growing rapidly and helps with situations where policy management becomes cumbersome”
- “With ABAC, permissions scale - it is no longer necessary for administrator to update existing policies.”
- “ABAC requires fewer policies”

Policy Evaluation

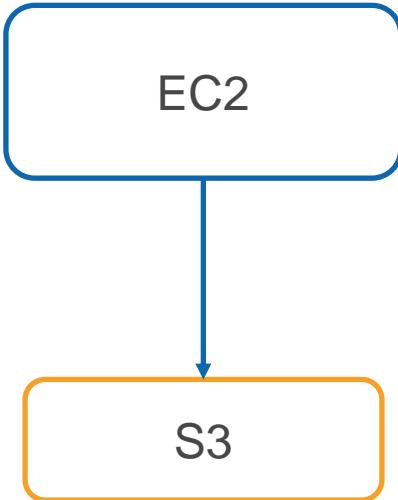
- Implicit Deny – By default, all requests are implicitly denied
- Explicit Allow – Overrides this default
- Explicit Deny – Overrides explicit allow
- Permissions Boundary, Organization Service Control Policy, Session Policy – it might override the allow with an implicit deny

Maximum Permissions

- IAM Permission Boundaries – Sets maximum permissions that an identity-based policy can grant to an entity (user or role)
 - When Set, an entity can perform only the actions that are allowed by identity based policies and permission boundaries
- Session Policy – When you assume role for temporary credentials, you can include a policy as a parameter – this is helpful when you want to limit what the temporary credential can do

IAM Roles

Application Access to AWS Resources



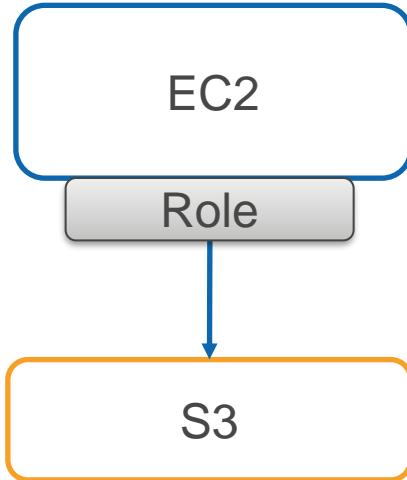
Access Key Credentials (Treat the server as a user):

- Generate Access Key, Secret Access Key
- Store the key in the EC2 instance
- EC2 uses the credentials to talk to services

Issues:

- Access Key Credentials are long-term (several days to months)
- Security risks due to long term credentials (accidental leakage, malicious users)
- Credential rotation is a problem at scale

Application Access to AWS Resources



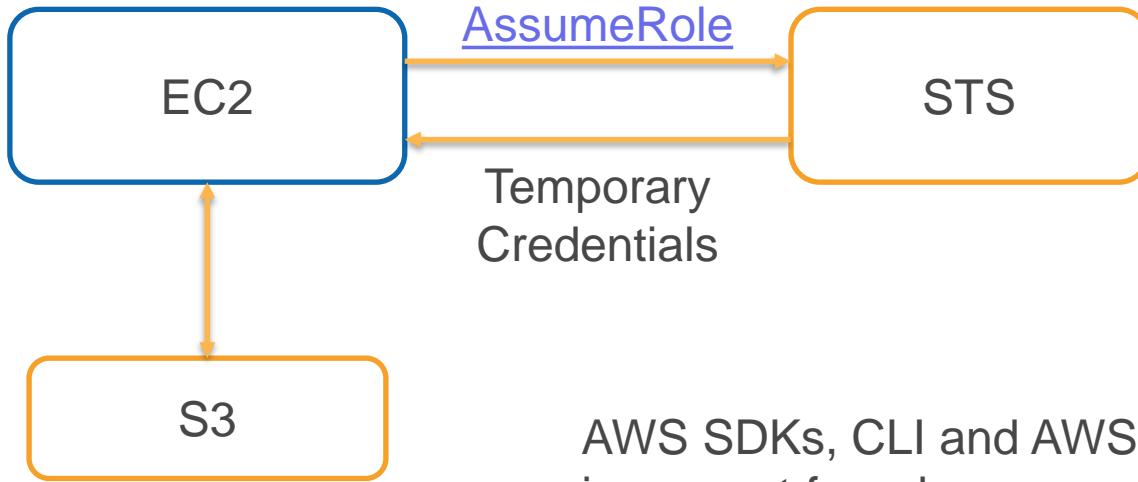
IAM Roles:

- Attach an IAM Role to the instance
- EC2 instance talks to metadata service to get temporary credentials for the role
- EC2 uses the temporary credentials to talk to services

Benefits:

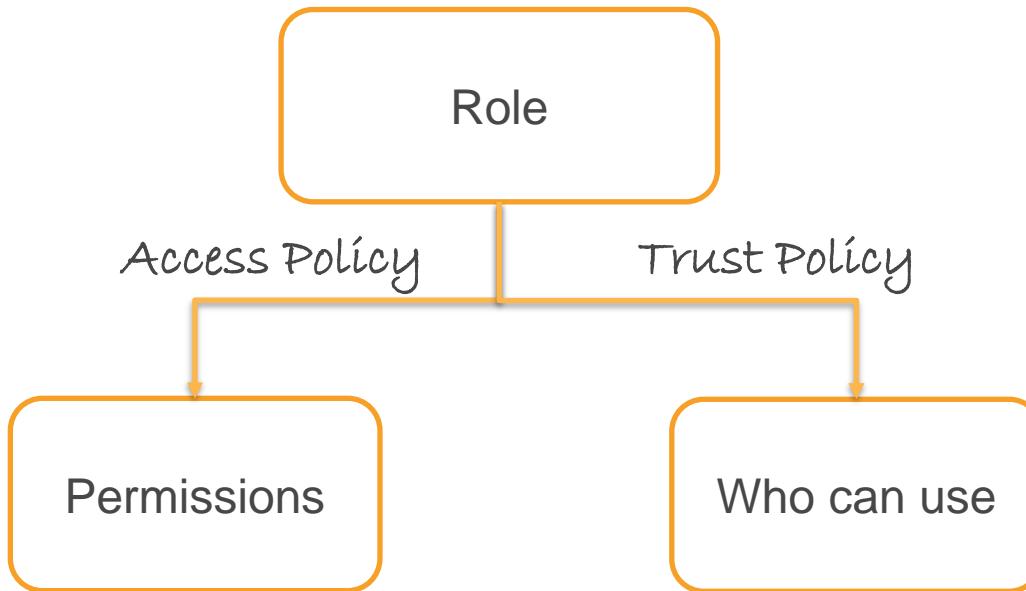
- No need to maintain credentials in the server
- Automatic Credential rotation – Credentials are valid only for a few hours (configurable)
- Reduced impact due to accidental leakage (credential is replaced every few hours)

Security Token Service (STS)

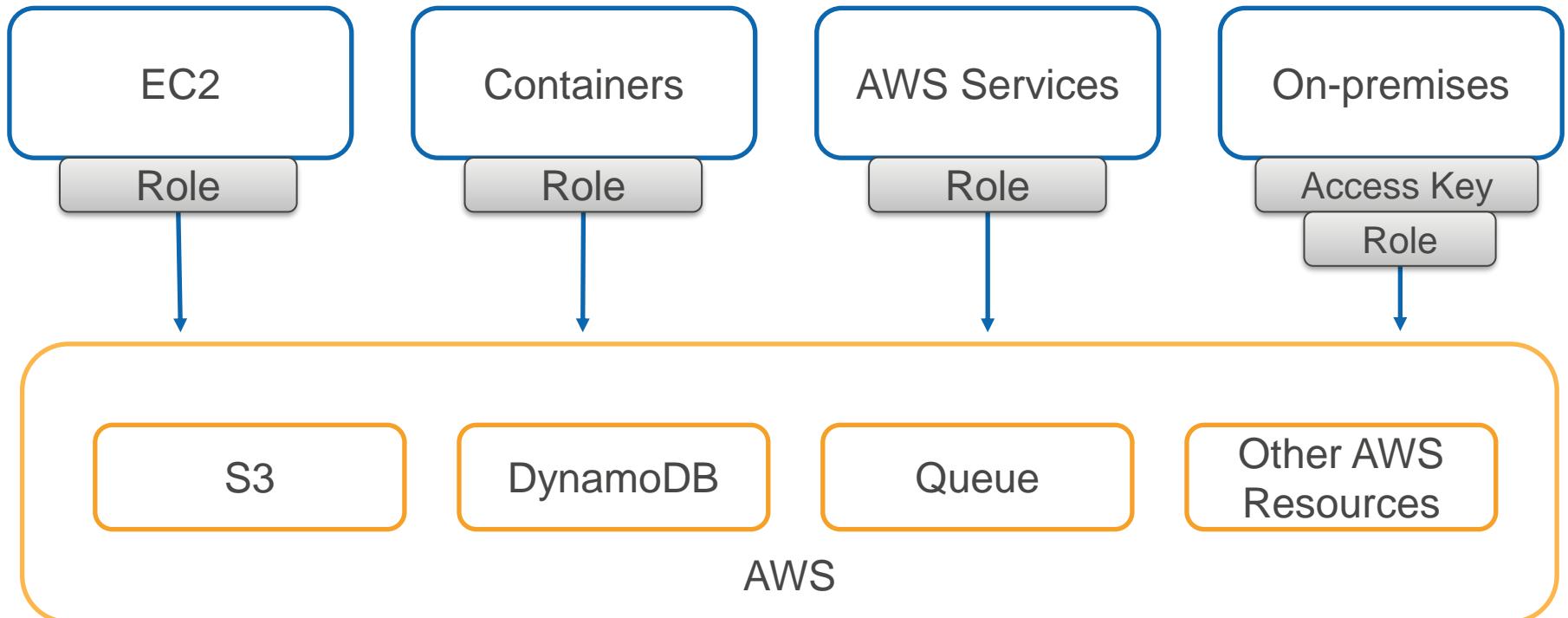


AWS SDKs, CLI and AWS Services have built-in support for roles

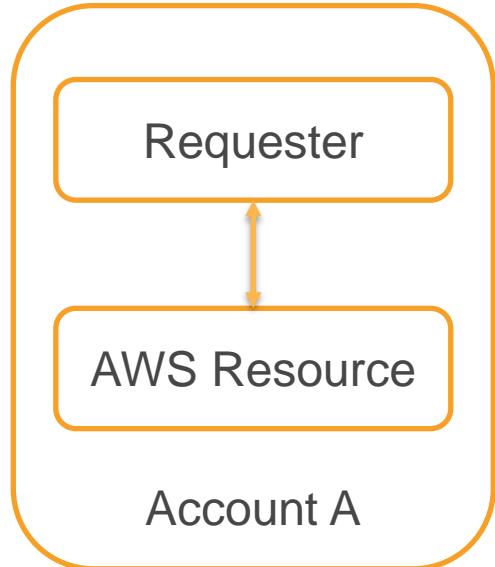
Role Concepts



Application Access To Resources

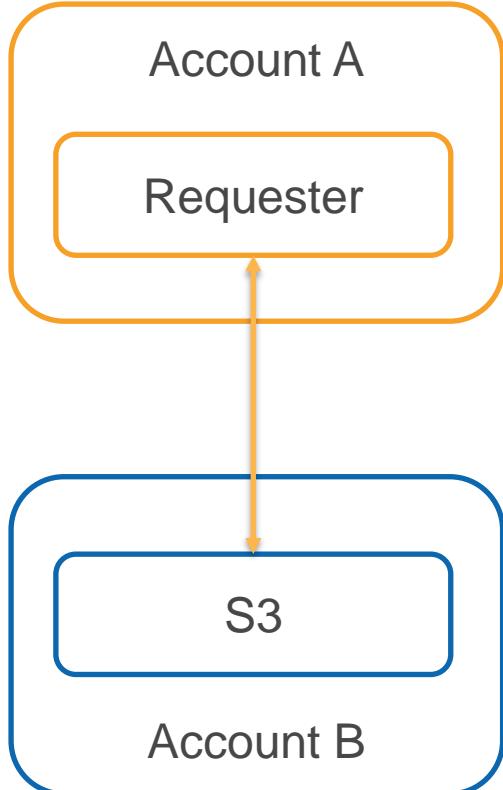


Same Account Access



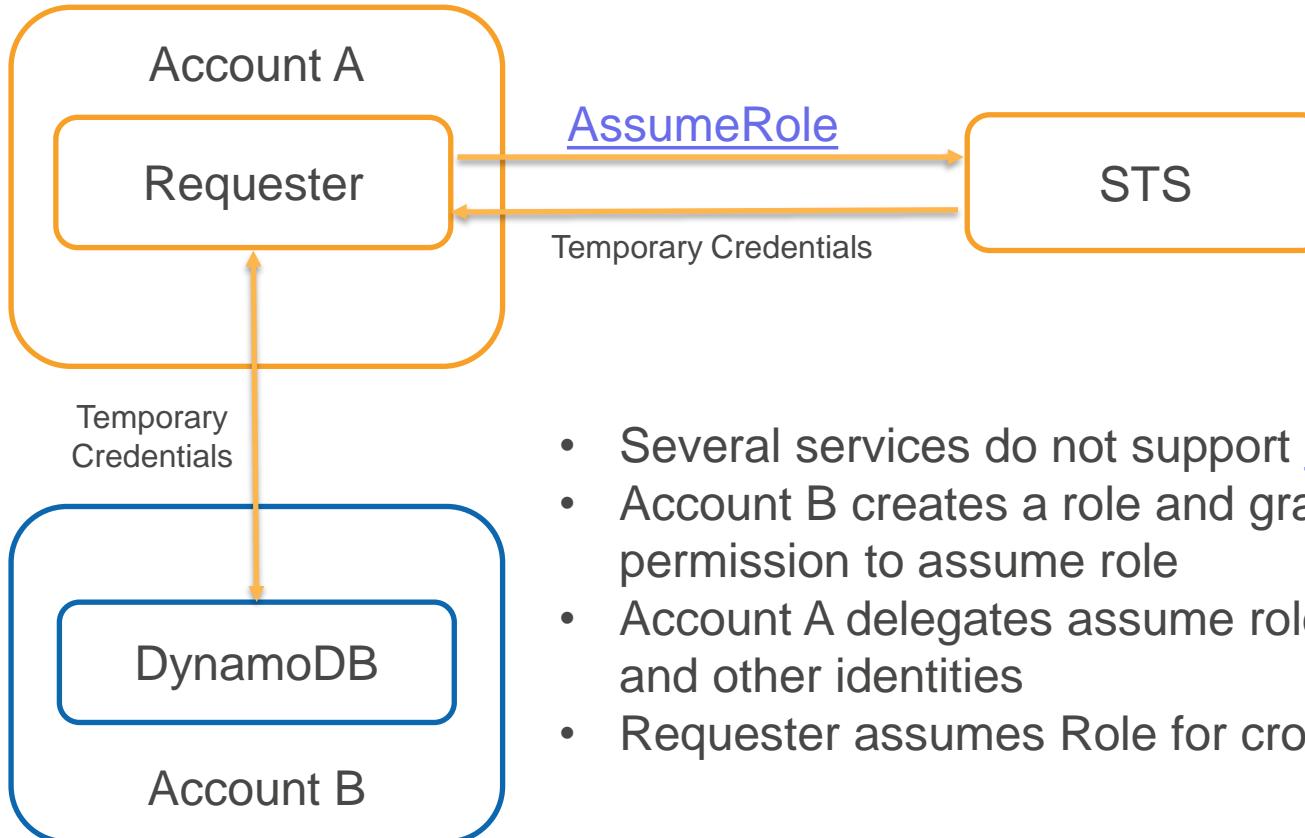
- Identity-based policies – for access to any AWS resource
- Resource-based policies – for access to supported resources
- Roles - for EC2 and Service integration

Cross Account Access using Resource Based Policy



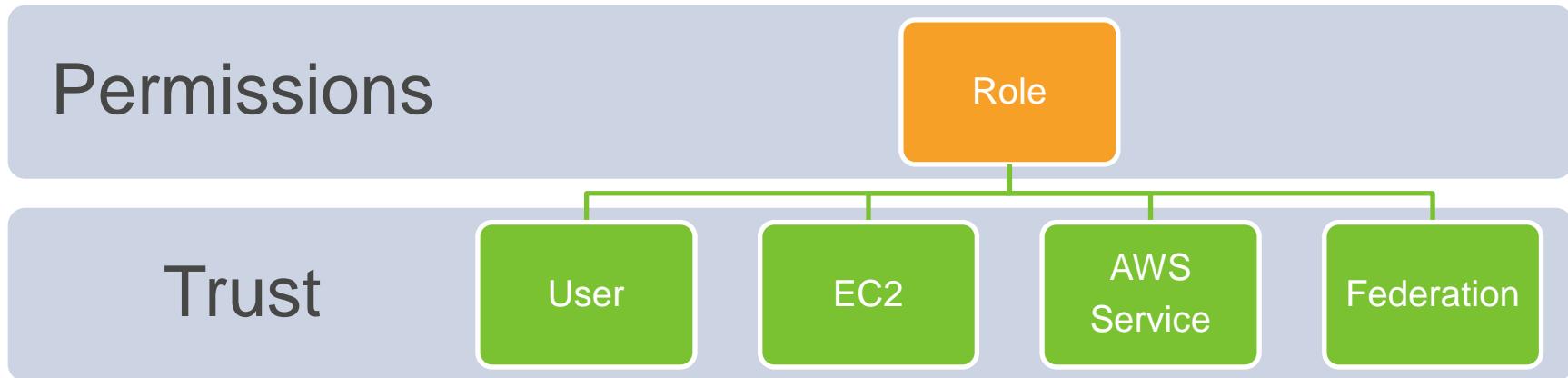
- [Resource Based Policies](#) (for supported resources like S3, SQS, SNS, Lambda)
- Identity in Account A needs access to resource in Account B
 - Resource owner (Account B) grants permission to Account A
 - Account A delegate's permission to the other identities in the account (user, role)

Cross Account Access Using Roles



- Several services do not support [Resource Based Policy](#)
- Account B creates a role and grants Account A permission to assume role
- Account A delegates assume role permission to users, and other identities
- Requester assumes Role for cross account access

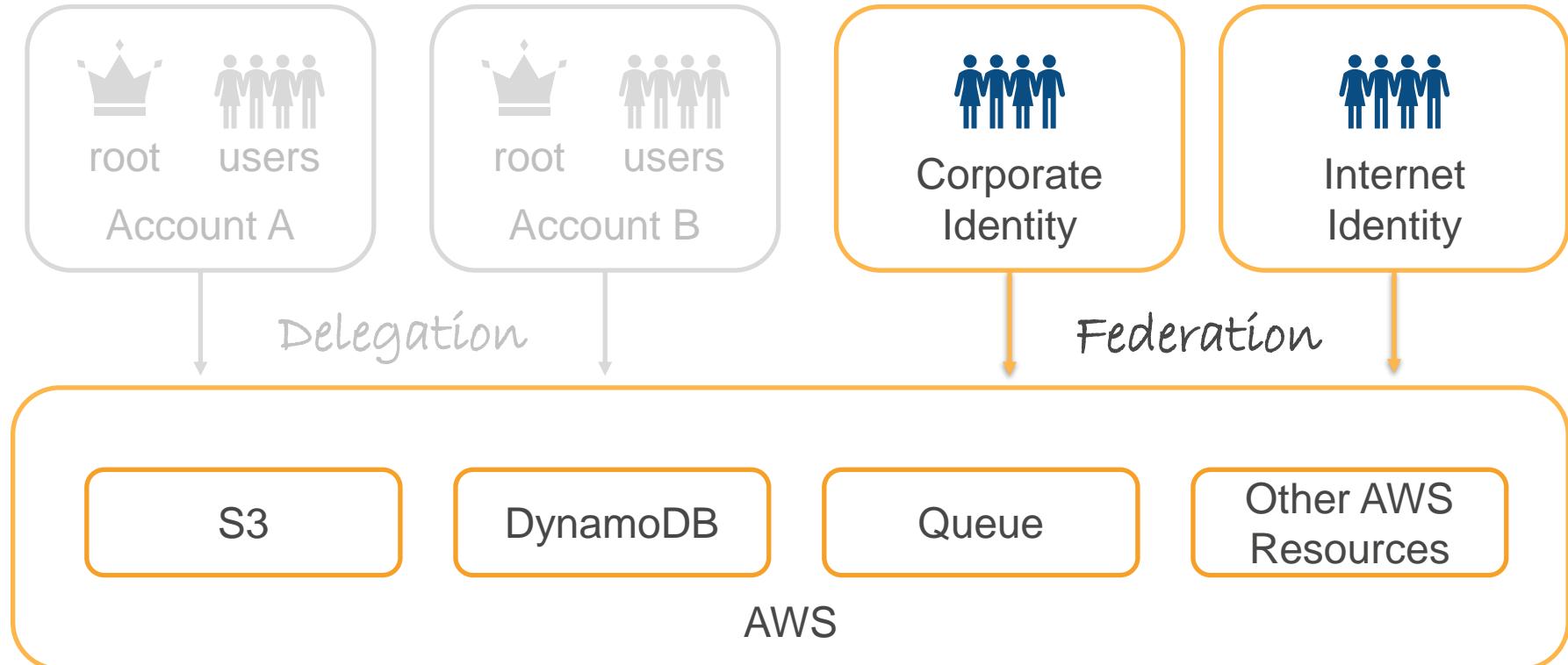
Role



Role has two parts:

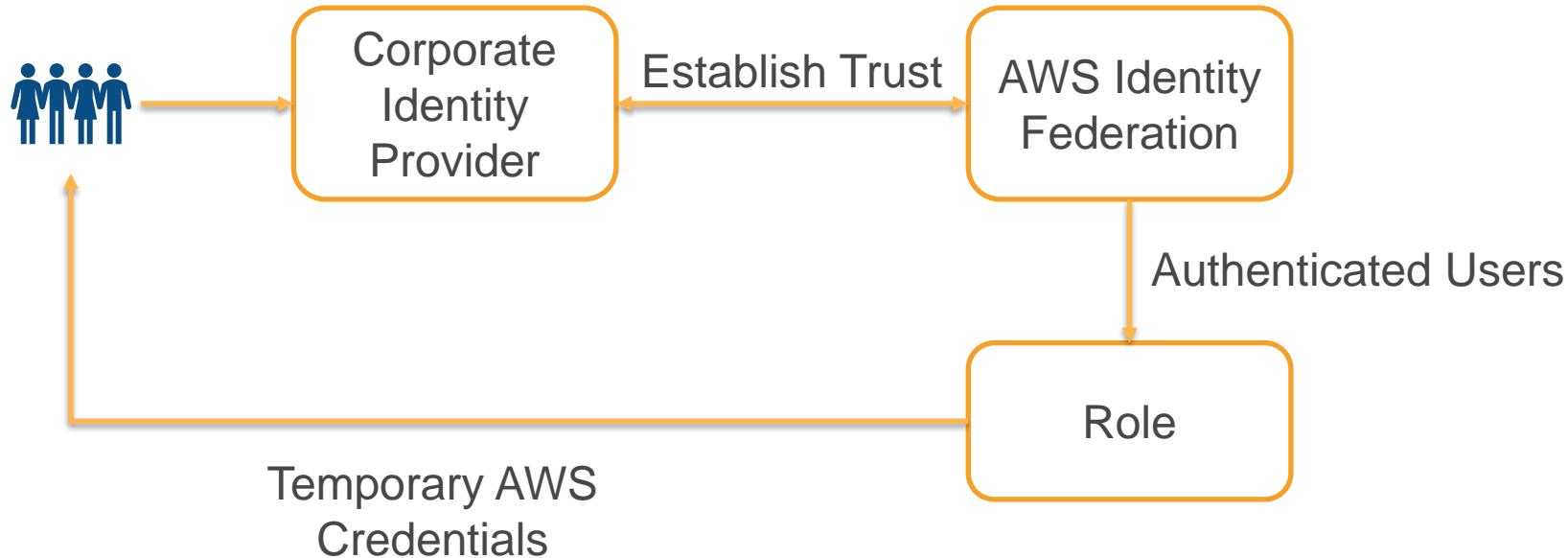
1. Who can assume the role (trust relationship) and
2. What access is allowed (permissions)

Types of Identities



Corporate Identity Federation

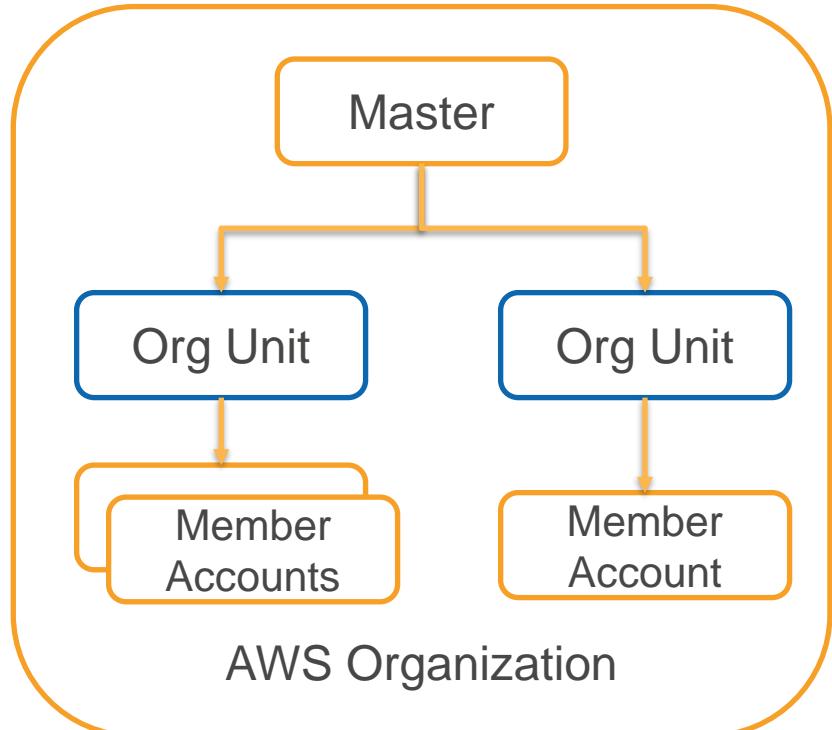
SAML 2.0, Microsoft Active Directory



Corporate Identity Federation

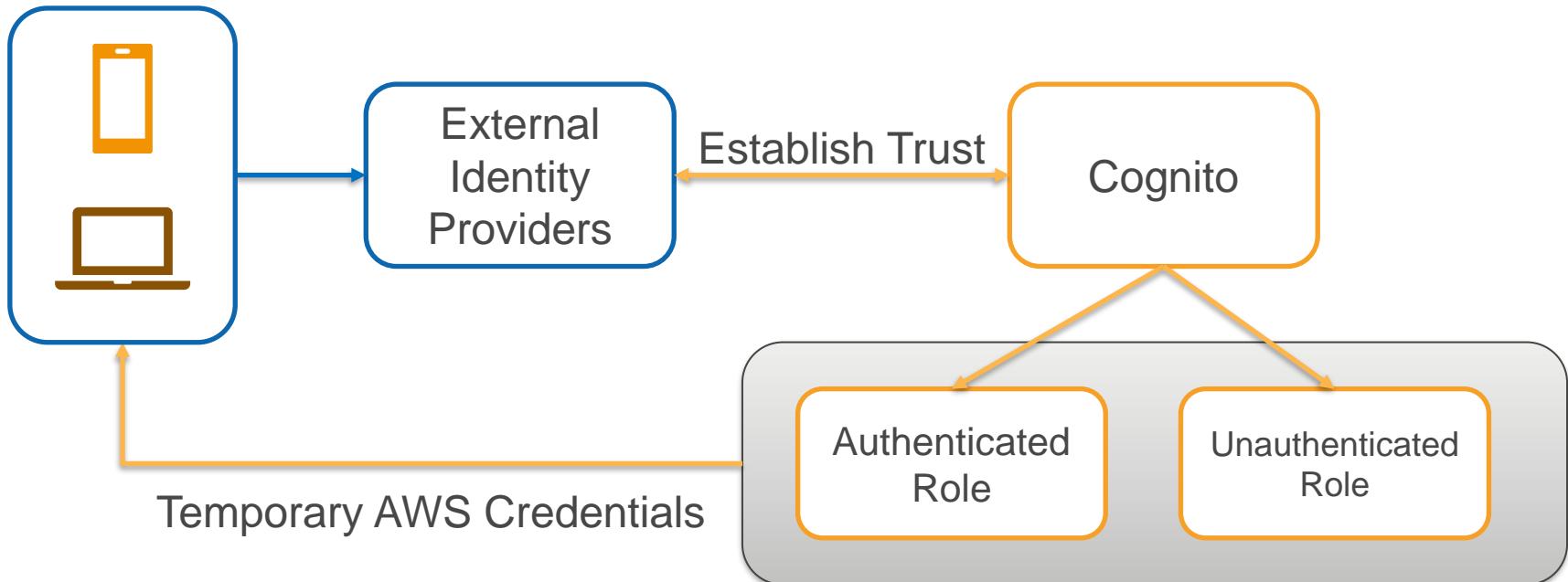
- SAML 2.0 (Security Assertion Mark Up Language) to exchange identity and security information between identity provider and application
- AWS IAM Federation – enables users sign-in to their AWS account with existing corporate credentials
- Non-SAML options – AWS Directory Service for Microsoft Active Directory
- AWS Organizations – Use AWS Single Sign On (SSO) to scale to multiple AWS Accounts (centrally manage access)

AWS Organizations



- Centrally manage costs and billing
- Service Control Policy – control services, resources, regions used by member account
- Share resources across accounts
- AWS Single Sign-on – manage access to employees and accounts
- Centralize identity management and federation

Internet Identity Federation



SAML 2.0, OAuth 2.0, OpenID Connect

Cognito Identity Federation

- Users can sign-in to mobile and web apps using social identity providers like Facebook, google, amazon
- Support for corporate identity federation using SAML 2.0
- Open Standards Support - Oauth 2.0, SAML 2.0, OpenID Connect
- Map users to roles and limit access to resources

Useful Resources

[AWS re:Invent 2019: Getting started with AWS identity \(SEC209-R1\) by Becky Weiss](#)

[AWS re:Invent 2015: How to Become an IAM Policy Ninja in 60 Minutes or Less \(SEC305\) by Jeff Wierer](#)

Lab – Identity-based Policy

- IAM Users, Groups
- Programmatic Access and Management Console Access
- Implicit Deny
- Explicit Allow
- Managed Policy

Lab – Resource-based Policy

Explore resource-based policy

Configure principals, resources and actions

Lab – IP Based Restriction

Policy with conditional variable

Limit access to S3 Bucket based on requester IP

For this lab, we use the `IPRestrictionPolicy` file available under resources

Reference

https://docs.aws.amazon.com/IAM/latest/UserGuide/reference_policies_examples_aws_deny-ip.html

Lab – VPC Endpoints

Allow access to the bucket only from VPC

Route traffic internally on AWS network using VPC Endpoint

For this lab, we use the policy file available under resources: `VPCEndPointRestrictionPolicy`

Reference: <https://docs.aws.amazon.com/vpc/latest/userguide/vpc-endpoints-s3.html#vpc-endpoints-policies-s3>

Lab – Cross Account Access using Resource Policy

- Bucket and the user are in separate accounts
- You need two accounts for this lab
- Manage permission using resource-based policy

For this lab, we use the policy file:

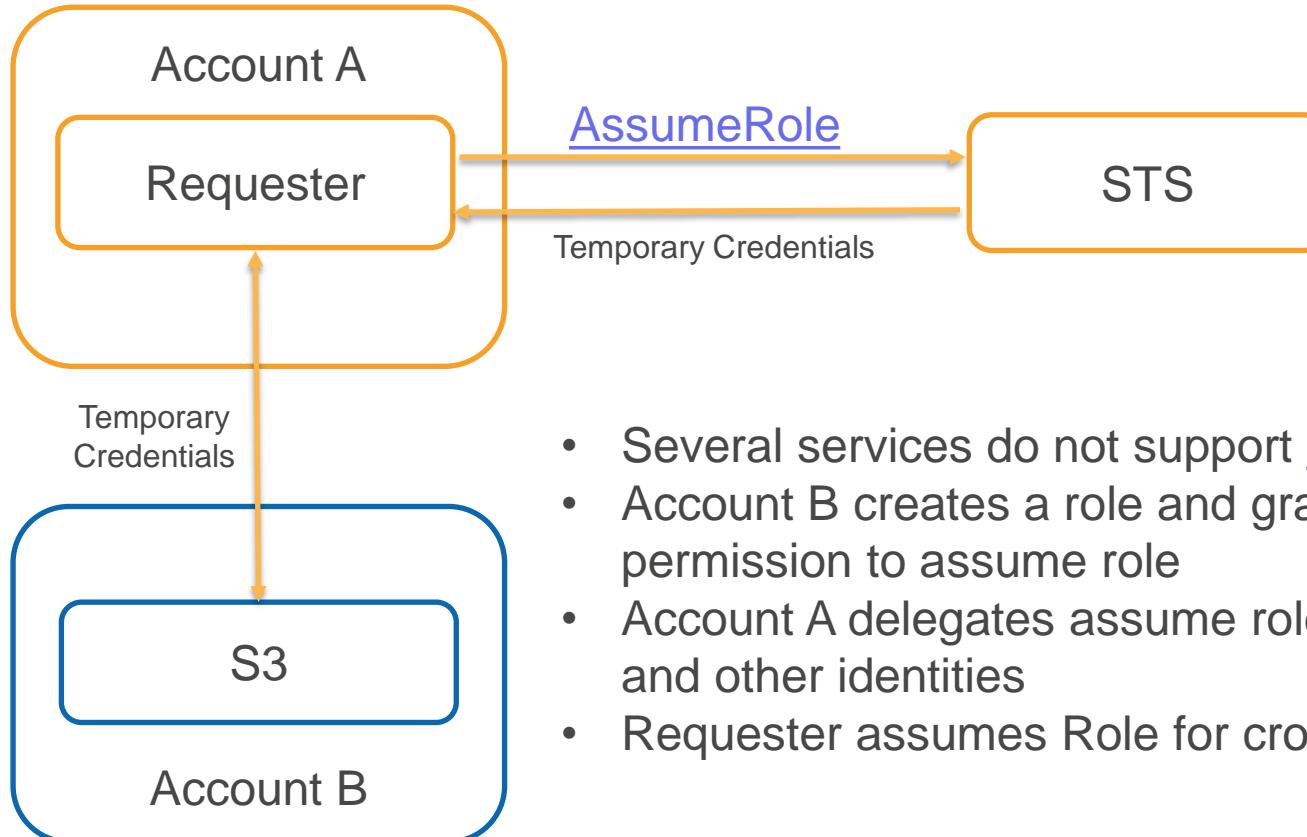
`ResourceBasedCrossAccountAccessPolicy`

Lab – Cross Account Access using IAM Roles

- Explore cross-account resource sharing using IAM Roles
- Temporary Credentials using Assume Role
- Demonstrate access from browser, laptop and from EC2 instance

For this lab, we use the policy file:
`RoleBasedCrossAccountAccessPolicy`

Lab - Cross Account Access Using Roles



- Several services do not support Resource Based Policy
- Account B creates a role and grants Account A permission to assume role
- Account A delegates assume role permission to users, and other identities
- Requester assumes Role for cross account access



Chandra Lingam

57,000+ Students



Copyright © 2019 ChandraMohan Lingam. All Rights Reserved.

For AWS self-paced video courses, visit:

<https://www.cloudwavetraining.com/>

