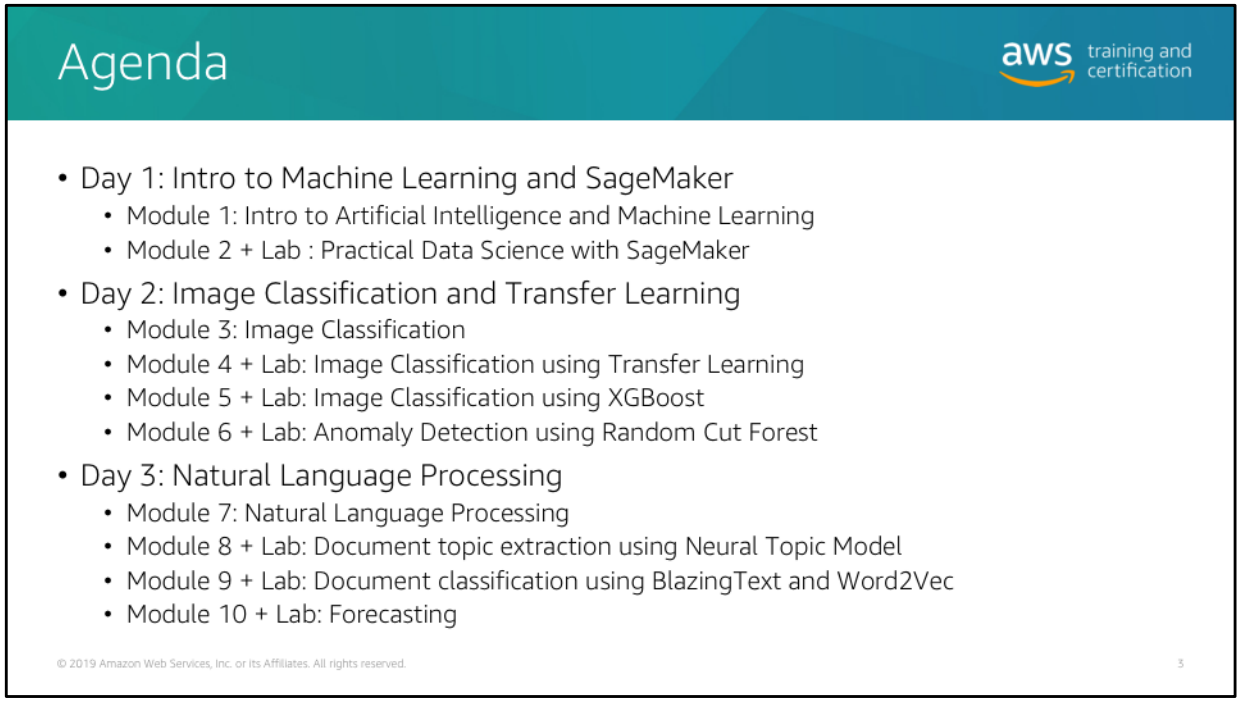
AWS Artificial Intelligence Machine Learning Workshop

aws training and certification

Module 0: Workshop Overview

aws training and certification



Welcome to AWS AI-ML Workshop. Here is an overview of the class. As you can see this class has a substantial hands-on component. As we cover the material in the class you will have an opportunity to try out the machine learning concepts in the labs and see how well they work.

Day 1:

- We are going to start out with an overview of Machine Learning and of SageMaker. Amazon SageMaker is a managed service that AWS provides for making it easy to build machine learning models on large datasets and then to deploy these models at scale.
- After this we will work through a machine learning problem using XGBoost for figuring out how we can prevent customers of a cell phone provider from leaving.

Day 2:

- The focus of the second day will be image recognition and classification. We will look at the Rekognition service of AWS. Then we will look at several techniques for building image recognition systems on custom images and categories that are specific to our organization.
- After image recognition we will have a module on anomaly detection.

Day 3:

- The focus on the third day will be on natural language processing. We will look at several machine learning models for classifying documents and for extracting key topics from them.
- After this we will have a module on forecasting.

This will end the class.

Intended audience

aws training and certification

- Data Scientists
- Machine Learning Engineers
- Developers and Engineers interested in Machine Learning
- Systems Architects

Prerequisites

aws training and certification

We recommend that attendees of this course have the following prerequisites:

- Familiarity with Python programming language
- Familiarity with NumPy and Pandas Python libraries is a plus

Hands-on activity

aws training and certification

This course allows you to test new skills and apply knowledge to your working environment through a variety of practical exercises.

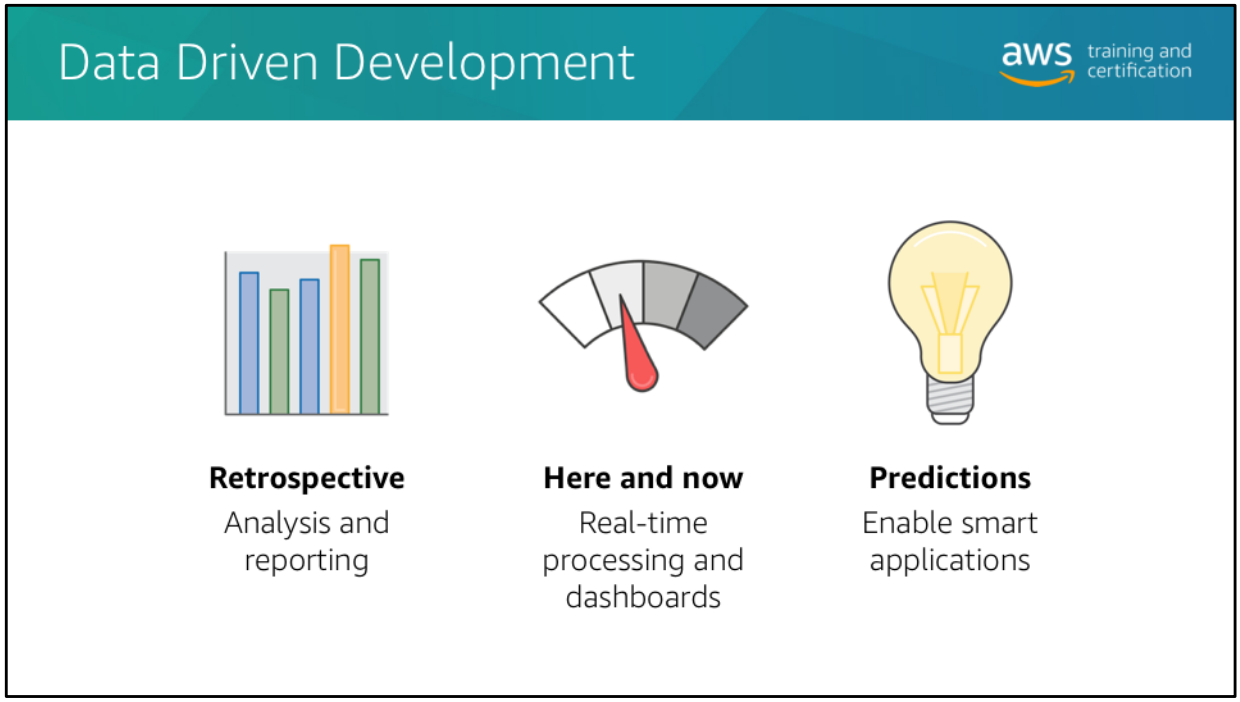
- Analyzing and visualizing a dataset
- Preparing the data and feature engineering
- Model building, training, tuning and deployment

Module 1: Machine Learning Overview

aws training and certification

A Flywheel for Data

aws training and certification



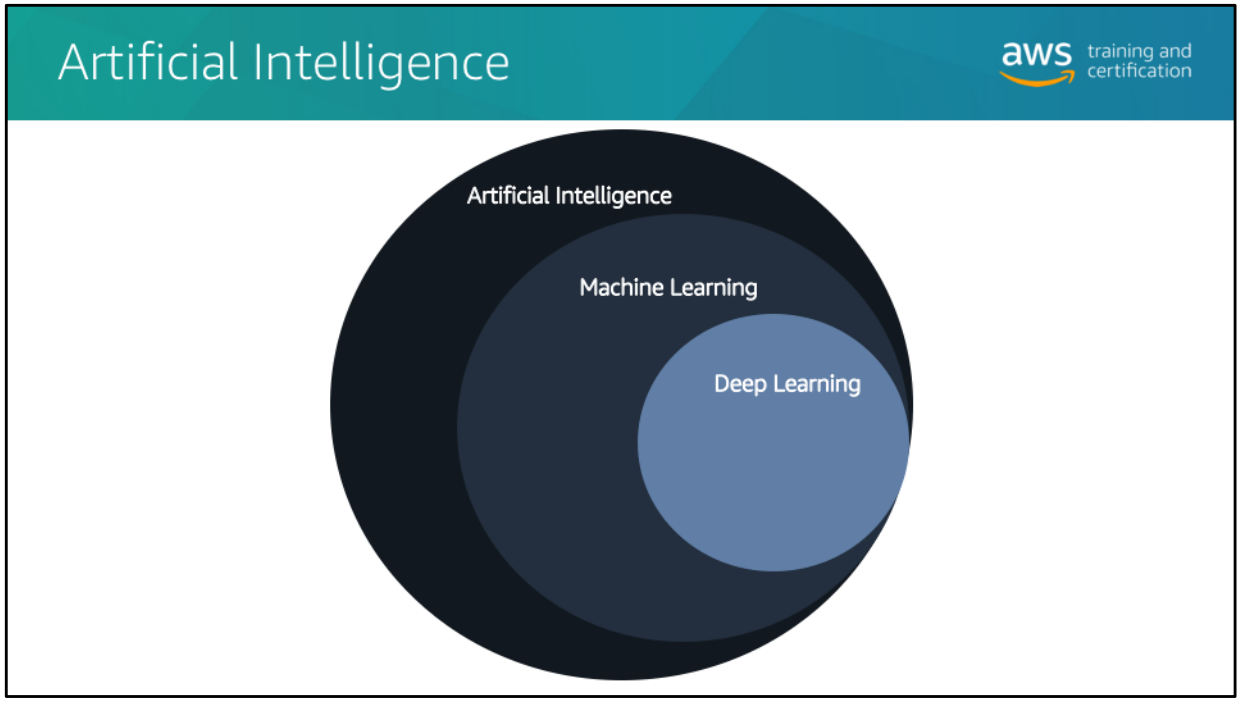
The data collected as part of the businesses falls into three major categories.

1. **Retrospective Data Analysis**: Data is collected, stored, and analyzed to answer specific questions about what happened in the past. Many companies use Amazon S3 to store large volumes of structured data, and then process it with EMR, using tools like Hive and Pig, while other companies store data in Amazon Redshift or in traditional database systems that are managed by Amazon RDS. No matter the mechanism, this approach is valuable for understanding how customers use your applications, or how data flows through them, and using this information to improve your businesses. For example, gathering information about how long customers spend in each part of your application and how often they are able to successfully complete their task can lead to identifying and improving problematic flows in these applications.

2. **Real-Time Processing and Analysis**: Capture the information as it comes in, and use it to immediately take action. Using services like Amazon Kinesis, Amazon EC2, and AWS Lambda, customers are able to immediately extract value from incoming data streams – for example, using dashboards for the real-time state of their applications, or by building applications that immediately respond to events

as they happen. It is very powerful to be able to not only know what has happened to your business in the past but also what is happening here and now.

3. **Predictions**: There is one more step beyond knowing what is happening here and now. You can use the data you already have to make accurate, actionable predictions about what will happen in the future. You can build a new breed of smart applications using these predictions.



To begin, let's establish a brief history of the topics we're going to discuss. Artificial intelligence (AI) incorporates any technique that enables computers to mimic human behavior. Machine learning (ML) is a subset of AI that enables machines to improve at tasks with repeated experience. Deep learning (DL) is a subset of ML in which software can train itself to perform tasks based on vast amounts of data it is exposed to.

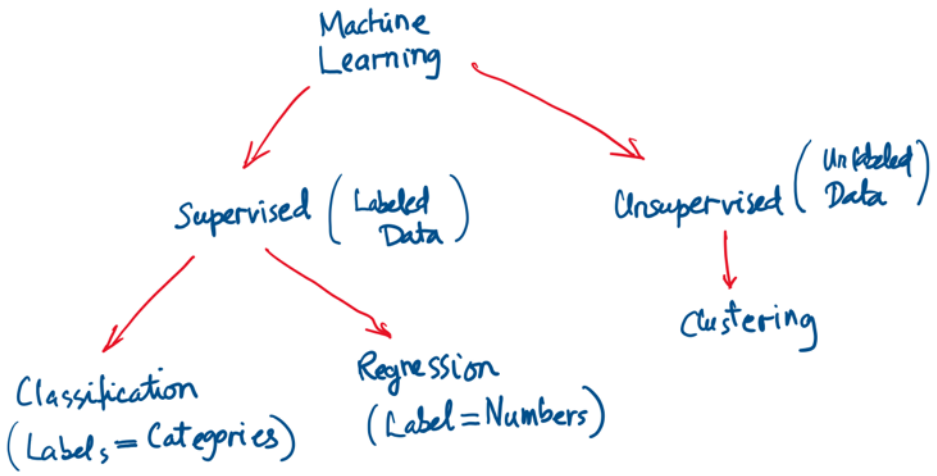
Artificial Intelligence

aws training and certification

	Features	Algorithms
AI	Human	Human
ML	Human	Machine
DL	Machine	Machine

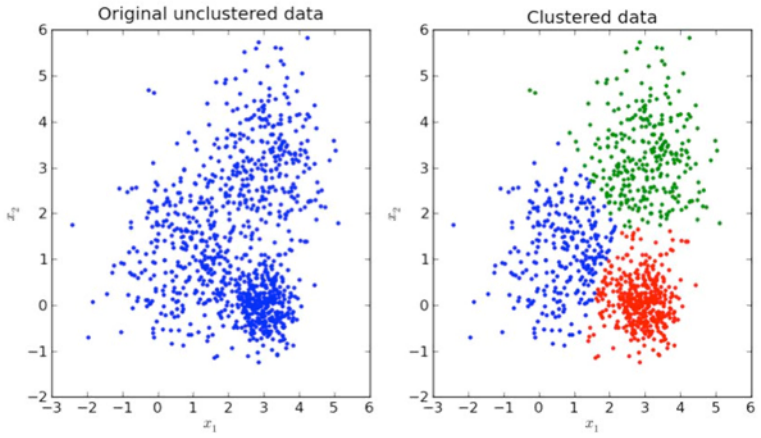
Machine Learning Example

Machine Learning Algorithms



Unsupervised Learning: Clustering

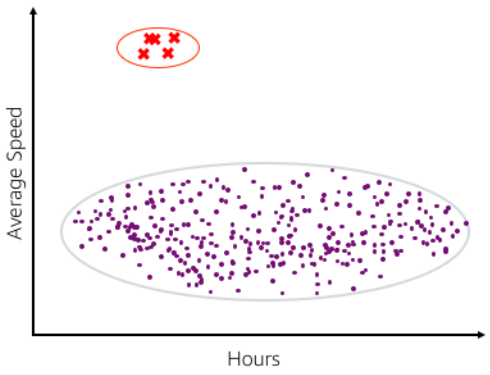
aws training and certification



Unsupervised learning uses no external teacher and is based on only local information. It is also referred to as self-organization. In unsupervised learning, the model uses only the data presented to the network, without any labels, and detects the emerging properties of the whole dataset. The model then constructs patterns from the available information. This is particularly useful in domains where a data point is checked to match previous scenarios, e.g., detecting credit card fraud. In the example, the model finds patterns and classifies the data into three clusters without any pre-trained data.

Unsupervised Learning: Anomaly Detection

aws training and certification



- Data from freeway traffic sensors
- Sensor data includes
  - Sensor ID
  - Year and day of the year Day of the week
  - Time
  - Occupancy (count)
  - Average speed (mph)
- Using clustering to predict which sensor might go bad

The example shows a case study on a network of traffic sensors. These sensors can be of different types, such as radar reading, temperature, and in-road. For this type of data, an anomaly can be a defective sensor.

For a sensor network such as ours, an anomaly may be a single variable with an unreasonable reading (speed = 250 m.p.h.; for a thermometer, air temperature = 200° F). One complication, however, is that each sensor reading has several variables. An anomaly may be a highly unlikely combination of the individual variable readings, despite each reading itself having a reasonable value. In the case of traffic sensors, a speed of more than 100 m.p.h. is possible during times of low congestion (that is, low occupancy and low volume) but extremely unlikely during high congestion.

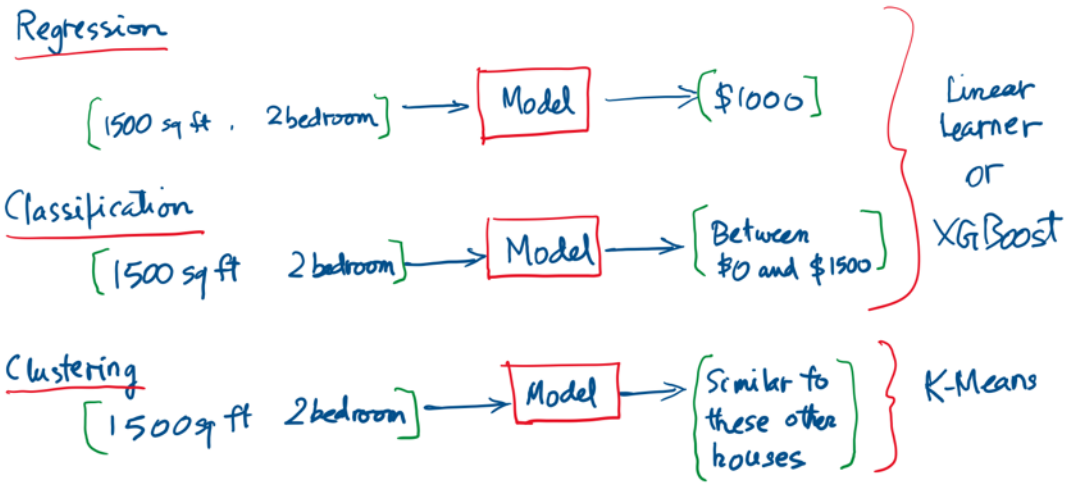
Pop Quiz

- Which approach should we use to predict rent?
  - Classification
  - Regression
  - Clustering

Pop Quiz

- Which approach should we use to predict rent?
  - Classification
  - Regression
  - Clustering
- We can use all three.

Machine Learning Algorithms



Pop Quiz

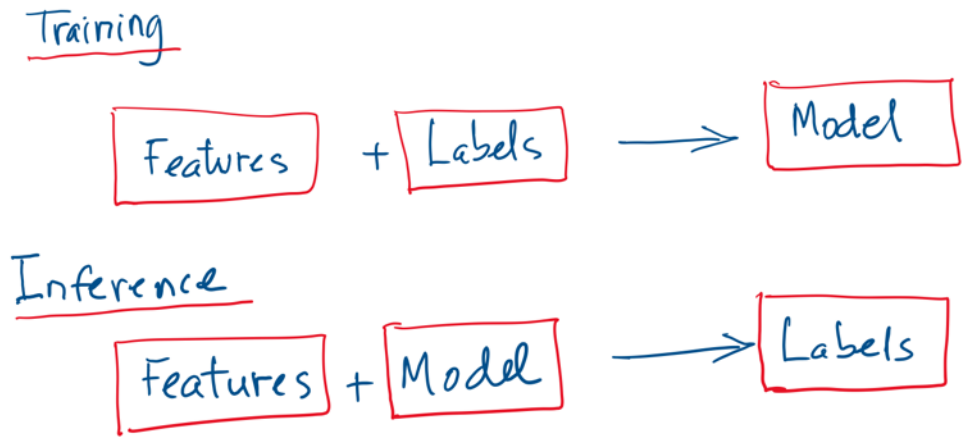
aws training and certification

- Is this next machine learning example using regression, classification, or clustering?

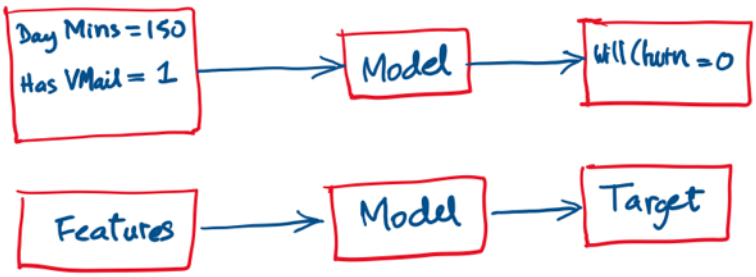
Machine Learning Example

1 = True / Yes
0 = False / No

Training vs Inference



Inference: Features to Target



Training vs Inference

aws training and certification

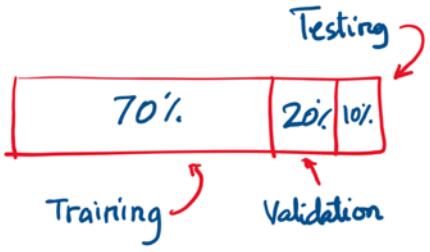
Training

Inference

Evaluating Models

aws training and certification

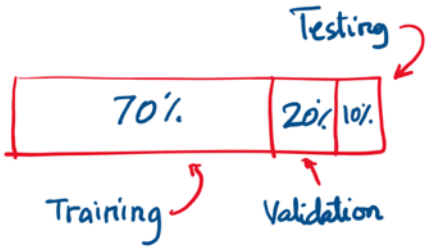
- How can we evaluate our model?
- Set aside some data from the training data set.
- Test the model on that data.
- Commonly, we use 70% for training and 30% for testing and validating.



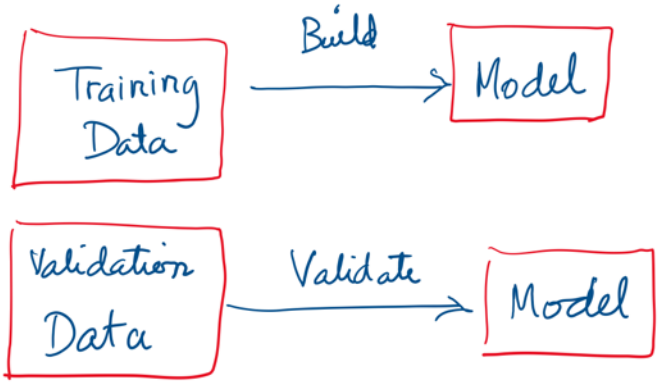
Evaluating Models

aws training and certification

- What is the benefit of evaluating models?
- We have a way to score our models.
- We can find out how accurate our model will be on new data.
- New data is data that it has not been trained on.

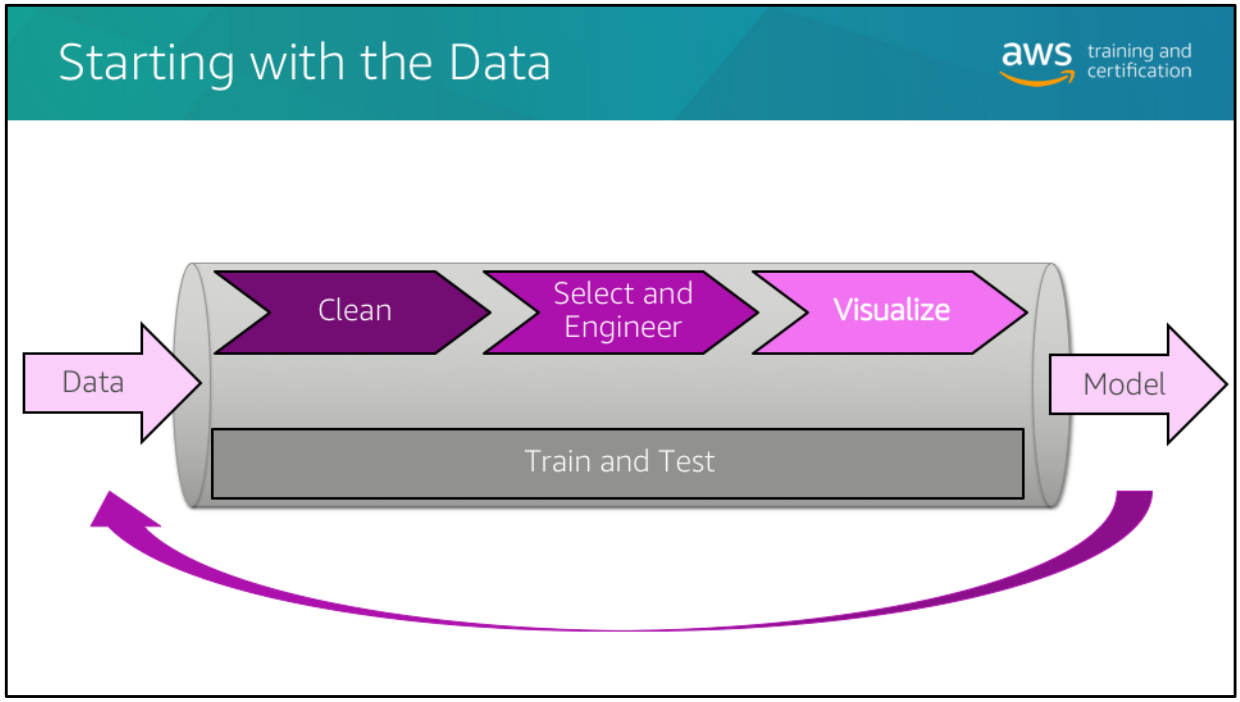


Training vs Validation



aws training and certification

The Business Importance of Machine Learning



In machine learning, one of the most important (if not *the* most important) things you'll need is data. You need to collect as much data as possible. This raw input data will be transformed into features that are formatted for compatibility with algorithms or models.

To begin, you should form a hypothesis and say to yourself, "I think these features would be relevant to help me assess my business problem." Then you'll gather that data, clean it, select and engineer your features, get your results, and adjust your features accordingly, deciding what data helps your model and what is a distractor. And you do this repeatedly. It's an extremely iterative process because reducing the amount of fragmented data will ultimately produce the best model results.

A Business Example: Building Smart Applications

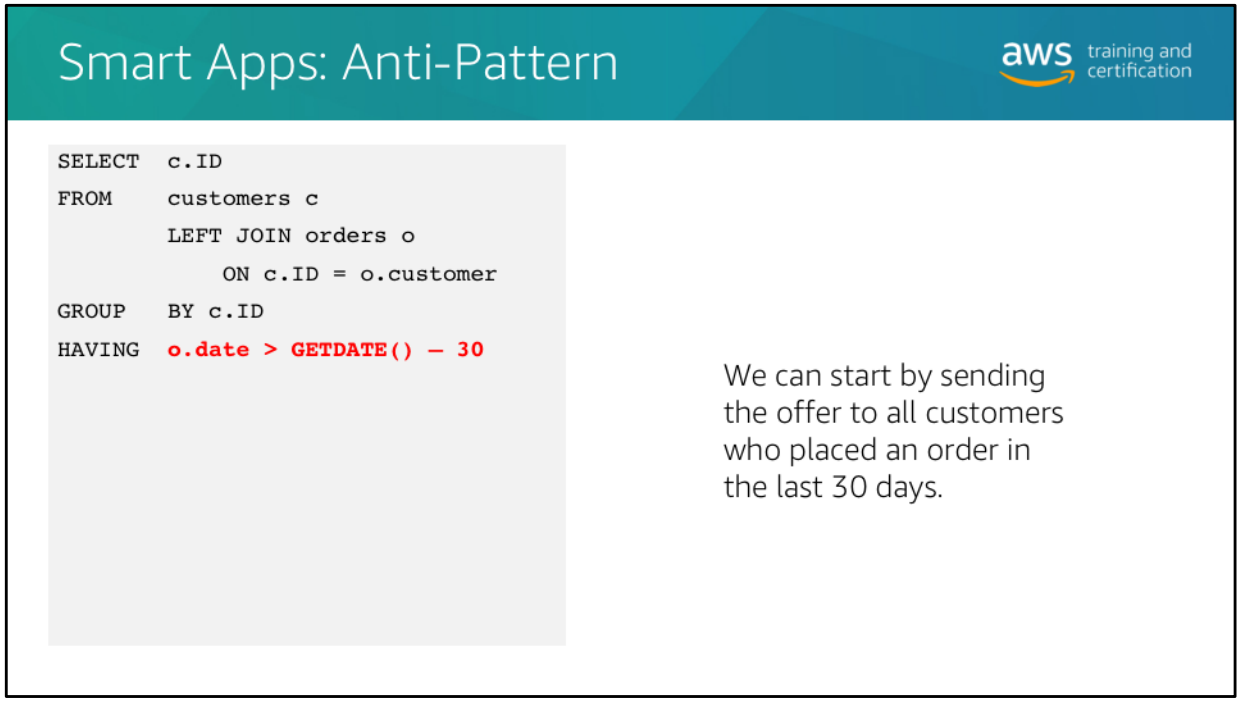
aws training and certification

The Anti-Pattern

Dear John,

This awesome quadcopter is on sale for just $49.99!

Let’s take a look at the pipeline in action. The slide shows an example of a smart application anti-pattern. You can get an email from various companies that highlights their products (in this example, a quadcopter). These emails are targeted at an audience who either use products that are similar or who are very interested in these products. This example shows how these smart applications target users.



The targeting starts with a simple SQL query that sends this email to all customers who placed an order in the last 30 days.

Smart Apps: Anti-Pattern (Contd.)

aws training and certification

```
SELECT   c.ID
FROM     customers c
         LEFT JOIN orders o
             ON c.ID = o.customer
GROUP    BY c.ID
HAVING   O.CATEGORY = 'TOYS'
   AND   o.date > GETDATE() – 30
```

Let's narrow it down to only customers who bought toys in the last 30 days.

Then, to further narrow the query, you include only customers who bought toys in the last 30 days.

Smart Apps: Anti-Pattern (Contd.)

aws training and certification

```
SELECT  c.ID
FROM    customers c
        LEFT JOIN orders o
            ON c.ID = o.customer
        LEFT JOIN products p
            ON p.ID = o.product
GROUP   BY c.ID
HAVING  o.category = 'toys'
   AND ((p.description LIKE
'%COPTER%'
         AND o.date > GETDATE() - 60)
     OR (COUNT(*) > 2
         AND  SUM(o.price) > 200
         AND  o.date > GETDATE() -
30)
       )
```

But what about quadcopters?

"Helicopter" doesn't cover everyone you want to reach. Luckily, you can do a fuzzy match on just "copter." That's better, right?

Smart Apps: Anti-Pattern (Contd.)

aws training and certification

```
SELECT  c.ID
FROM    customers c
        LEFT JOIN orders o
            ON c.ID = o.customer
        LEFT JOIN products p
            ON p.ID = o.product
GROUP   BY c.ID
HAVING  o.category = 'toys'
   AND ((p.description LIKE
'%copter%'
         AND o.date > GETDATE() -
120)
     OR (COUNT(*) > 2
         AND  SUM(o.price) > 200
         AND  o.date > GETDATE() –
30)
       )
```

Maybe we should go back further in time…

Maybe you think of looking further back in time; it's not every day that one buys an expensive flying toy.

Smart Apps: Anti-Pattern (Contd.)

aws training and certification

```
SELECT  c.ID
FROM    customers c
        LEFT JOIN orders o
            ON c.ID = o.customer
        LEFT JOIN products p
            ON p.ID = o.product
GROUP   BY c.ID
HAVING  o.category = 'toys'
   AND ((p.description LIKE
'%copter%'
         AND o.date > GETDATE() -
120)
     OR (COUNT(*) > 2
         AND  SUM(o.price) > 200
         AND  o.date > GETDATE() -
40)
       )
```

… and tweak the query more…

Smart Apps: Anti-Pattern (Contd.)

aws training and certification

```
SELECT  c.ID
FROM    customers c
        LEFT JOIN orders o
            ON c.ID = o.customer
        LEFT JOIN products p
            ON p.ID = o.product
GROUP   BY c.ID
HAVING  o.category = 'toys'
   AND ((p.description LIKE
'%copter%'
         AND o.date > GETDATE() -
120)
     OR (COUNT(*) > 2
         AND  SUM(o.price) > 150
         AND  o.date > GETDATE() –
40)
       )
```

… and tweak it again…

Two hundred dollars is a lot to spend on a single purchase, so you lower that to 150 dollars.

Smart Apps: Anti-Pattern (Contd.)

aws training and certification

```
SELECT  c.ID
FROM    customers c
        LEFT JOIN orders o
            ON c.ID = o.customer
        LEFT JOIN products p
            ON p.ID = o.product
GROUP   BY c.ID
HAVING  o.category = 'toys'
   AND ((p.description LIKE
'%copter%'
         AND o.date > GETDATE() - 90)
     OR (COUNT(*) > 2
         AND  SUM(o.price) > 150
         AND  o.date > GETDATE() -
40)
       )
```

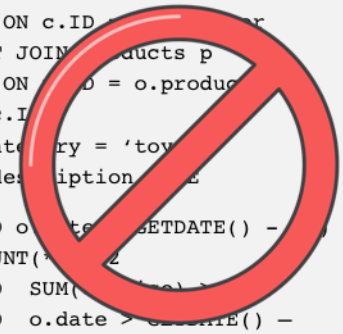
… and again…

That query may give you too many results, so you try to reduce the number of days again.

Smart Apps: Anti-Pattern (Contd.)

aws training and certification

```
SELECT  c.ID
FROM    customers c
        LEFT JOIN orders o
            ON c.ID
        LEFT JOIN
            ON
GROUP   BY c.I
HAVING  o.cat
   AND ((p.des
'%copter%'
         AND o
     OR (COUNT(
         AND  SUM(
         AND  o.date >
40)
       )
```



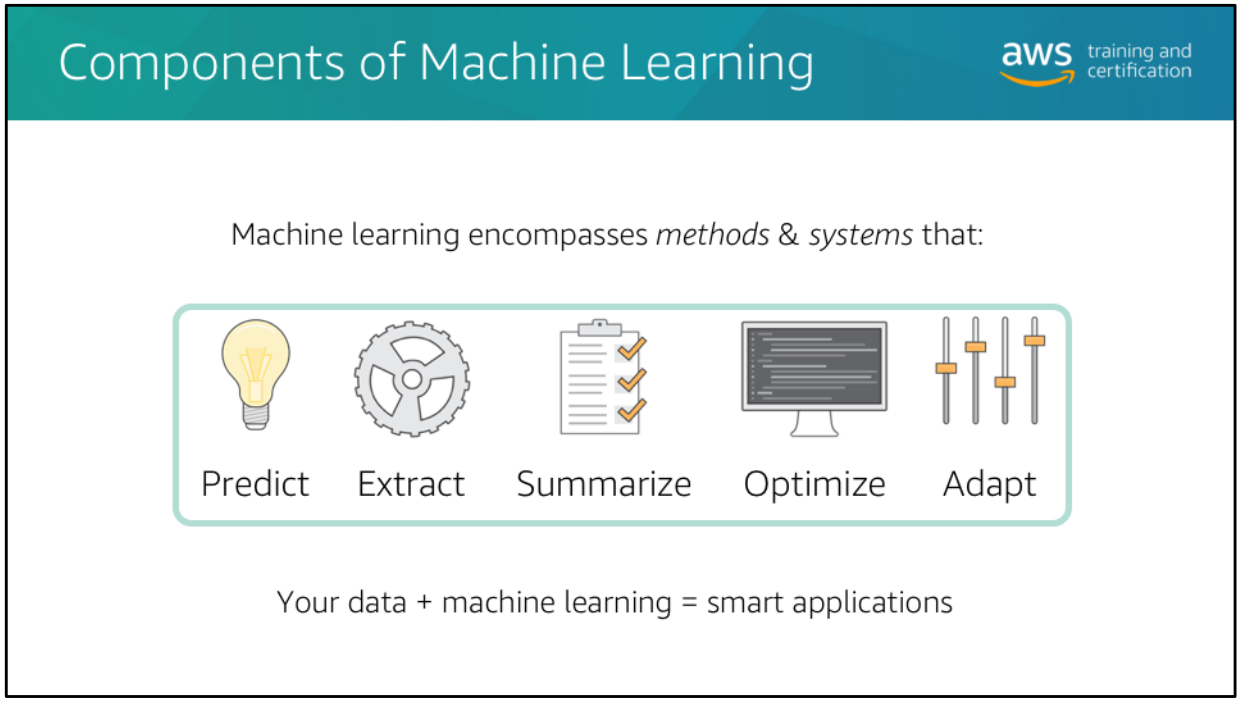
Or, instead, use machine learning technology to **learn** your business rules from data!

You won't find the right business rule by tweaking SQL queries. And even if we do find a good rule by sheer luck, it will become obsolete immediately. By using machine learning, you can build smart applications by finding patterns in existing data and making them actionable as predictions.

Machine Learning Approach

aws training and certification

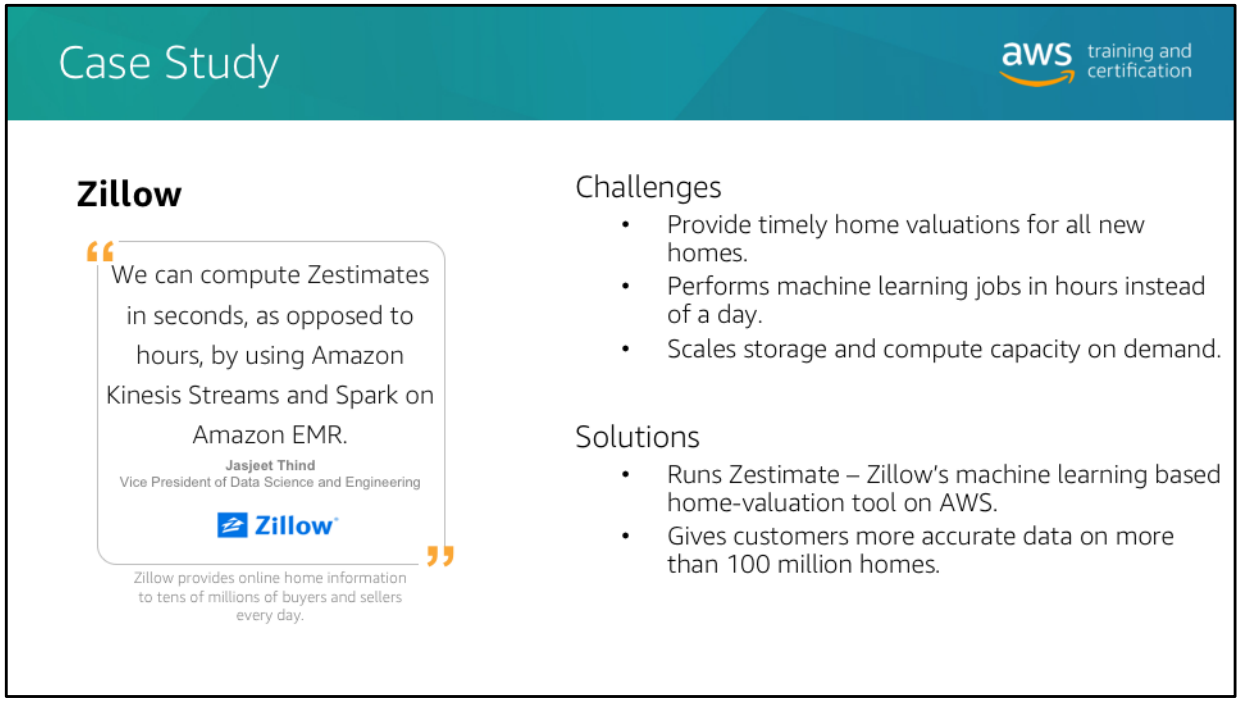
	Revenue	Age	Toy Purchases	Will Buy?
Customer 1	$200	10	90%	Y
Customer 2	$20	30	0%	N
Customer 3	$300	15	20%	N



As we just saw, machine learning (ML) systems discover hidden patterns in data and leverage these patterns to predict patterns in the future. If the product name contains words like “jeans” or “jacket”, then this product category likely belongs to “apparel.”

ML systems learn from examples similar to how children learn language or patterns. The patterns discovered in data by ML algorithms can be thought of as data summary. So ML algorithms can also be used to obtain concise descriptions of data. ML can also be thought of as a combination of methods and systems that:

1. Predict new data based on observed data.
2. Extract hidden structure from the data.
3. Summarize data into concise descriptions.
4. Optimize an action, given a cost function and observed data.
5. Adapt based on observed data.



This case study highlights the owner of a portfolio of the largest online real-estate and home-related brands, including the Zillow website.

The Challenge:
The customer wanted to
- Provide timely home valuations for all new homes.
- Perform machine-learning jobs in hours instead of a day.
- Scale storage and compute capacity on demand.

The Solution:
Zillow uses a wide variety of public-record data—including tax assessments, sales transactions, images of homes, MLS listing data, and other information provided by homeowners—as inputs to its Zestimate algorithm.
- Runs Zestimate, its home-valuation tool on AWS
- Performs machine-learning tasks in hours instead of one day.
- Provides more accurate home valuation data.

For more information see, https://d0.awsstatic.com/case-studies/PDF%20Case%20Studies/AWS-Casestudy_Zillow.pdf

Application Areas

aws training and certification

Application Domain	Use Cases
Personalization	• Recommending content • Loading Predictive content • Improving user experience
Fraud detection	• Detecting fraudulent transactions • Filtering spam emails • Flagging suspicious reviews
Targeted marketing	• Matching customers and offers • Choosing marketing campaigns • Cross-selling and up-selling

With those components in mind, let's take a look at some use cases and where it makes sense to apply machine learning.

Application Areas (Contd.)

aws training and certification

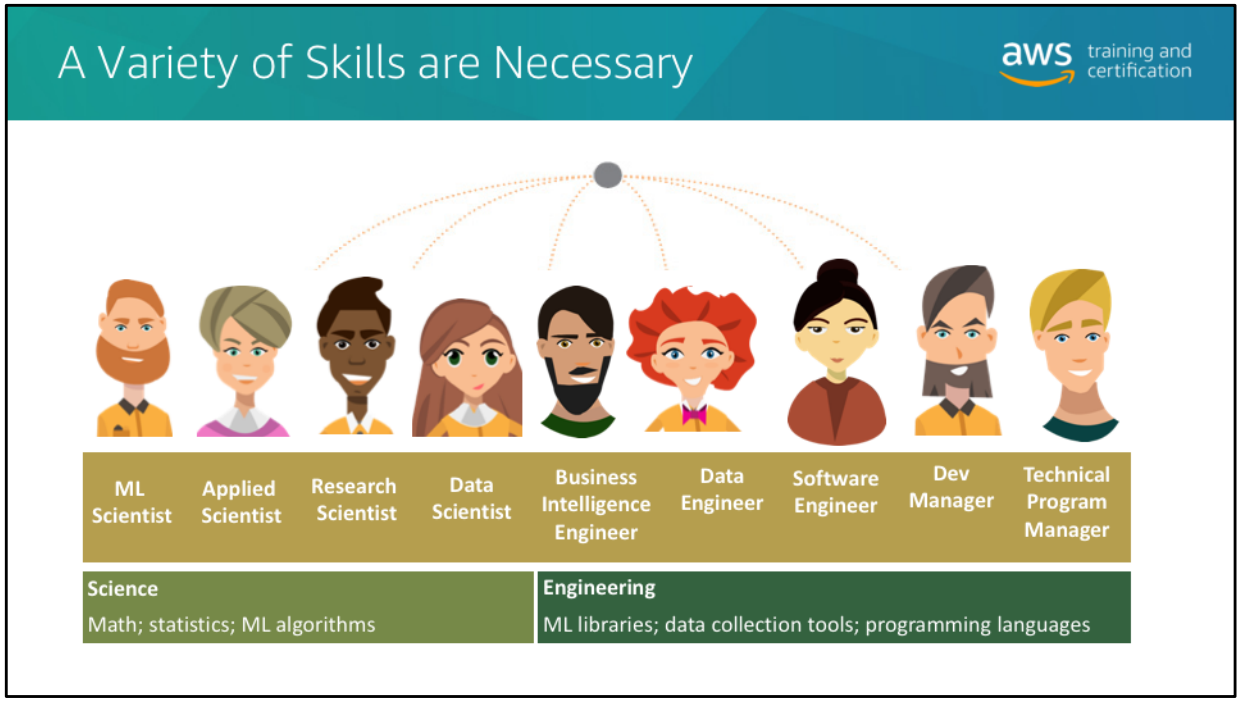
Application Domain	Use Cases
Customer support	• Predictive routing of customer emails • Social media listening
Content classification	• Categorizing documents • Matching hiring managers and resumes
Churn prediction	• Finding customers who are likely to stop using the service • Upgrading targeting

aws training and certification

Discuss: What are some ML/AI use cases in your industry?

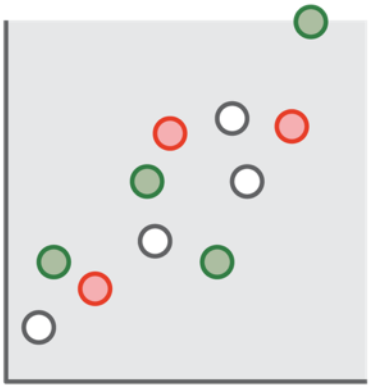
aws training and certification

Common Challenges in Machine Learning



All of this data, time, and modeling – it costs money. But where does the money go? **Primarily to the people needed to perform the necessary ML functions.** You're going to need people with a variety of **skills.**

Challenge 1: Expertise



- Limited supply of domain experts (Data scientists and Machine Learning engineers) who can extract meaningful information from the data.
- Expensive to hire or outsource.

Securing adequate expertise is one of the three primary barriers to implementing and executing ML problems.

You need experts to use ML technology, and there are not enough experts to keep up with the demand. Companies of all sizes and in all industries are feeling the talent shortage. Outsourcing ML model and smart application development is an expensive option.

Particular roles might include:

- ML Scientist
- Applied Scientist
- Research Scientist
- Data Scientist
- Business Intelligence Engineer
- Data Engineer
- Software Engineer
- Dev Manager
- Technical Program Manager

And you can break the skill sets needed into two larger categories:

Science skills

- Math
- Statistics
- ML Algorithms

Engineering skills

- ML Libraries
- Data Collection Tools
- Programming Languages
- Reporting Tools
- Analytics Tools
- Visualization Tools

Challenge 2: Scale

aws training and certification

Building and scaling machine learning technology

- Has many choices for tools, but few are good.
- Is difficult to use and scale.
- Has many moving pieces, which leads to custom solutions ***every time.***

To build machine learning software that scales to your business requires significant time and engineering expertise. Although, there are many available tools, these tools are often oriented at machine learning experts, and are not built for the demands of large datasets with real-world data types. And there is a tremendous fragmentation of tooling.

For example, to build an end-to-end ML application requires not just the learning technology but also data connectors and transformers to get data to the learning algorithms, deployment technology to make models available, and metric computation and evaluation scripts. Frequently, no matter which technology you start with, you can end up doing a lot of the engineering work yourself, just to connect all the different pieces together.

Challenge 3: Operationalization

aws training and certification

- Working with complex and error-prone data workflows and custom platforms and API's.
- Spending significant time on managing the model lifecycle.

The ability to tie machine learning to business applications, and to go from having a model to having an application that is using that model is important. This is the lesson many early adopters of ML technology learned the hard way. Much of the current ML software focuses on making ML models, not on using these models to make predictions.

Often, generating predictions is overlooked when designing the ML workflow. Generating prediction runtimes can only run on a single machine, are not robust. These predictions are very limited in volume and the latency at which these predictions can arrive. In addition to predictions, work is required for automating data pipelines, and managing, updating, and retiring models as they are deployed to production. These processes can turn into a lot of work.

Challenge 4: Complexity of Data

- Large volumes of high quality data
- Manual process of cleaning, labeling, and engineering

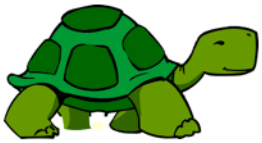
Challenge 5: Cost

- Purchasing and maintaining hardware is complex and typically comes at a fixed cost
- Varying CPU, GPU, memory, and networking capacity

The Result

aws training and certification

Building a machine learning application can be...



Slow

Expensive

Risky

...but what if there were a better way?

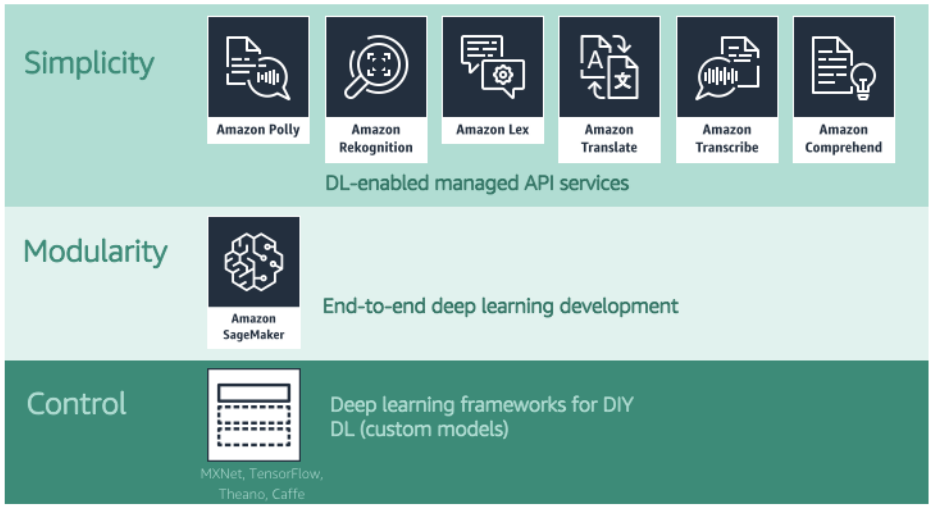
To develop ML today requires teams of engineers and data scientists working together, and much of what they build can end up being custom code that will not be reused on your next ML application. It is slow, expensive, and risky.

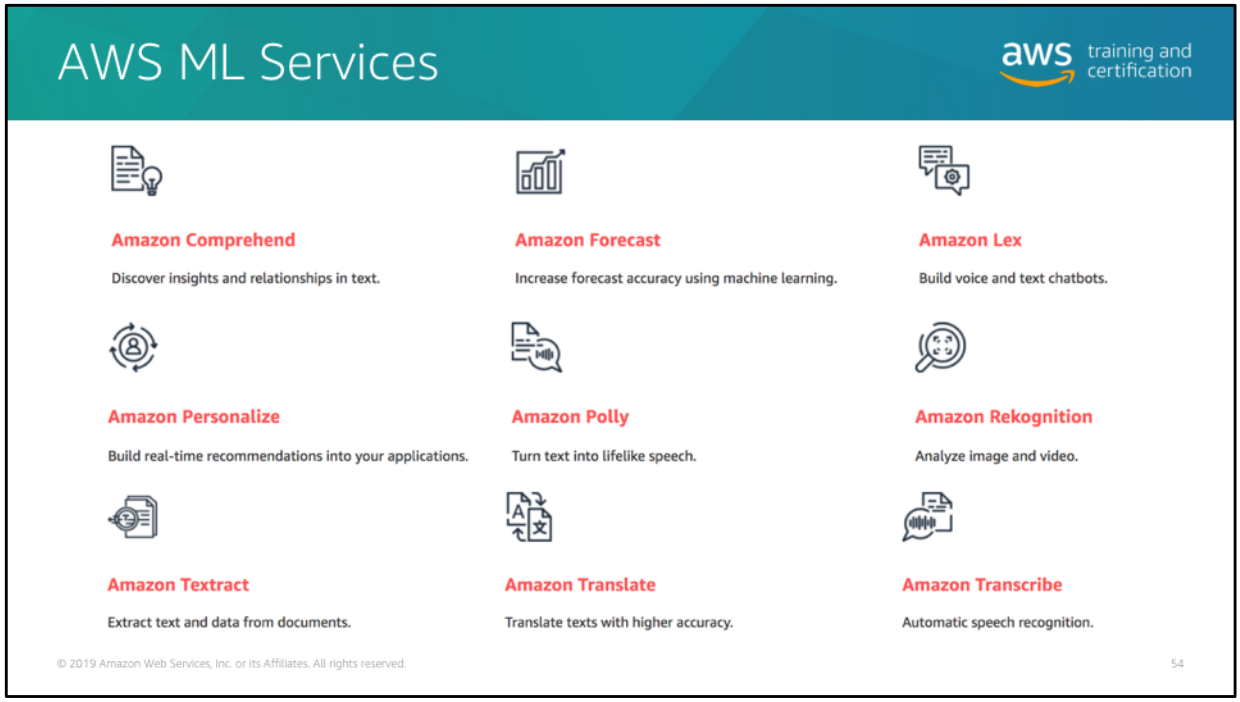
aws training and certification

AWS ML Solutions

Machine Learning on AWS

aws training and certification





These are managed services AWS provides for machine learning.

- Comprehend: Extracts insights from text.
- Forecast: Predicts future values of time series data.
- Lex: Enables building voice and text chatbots.
- Personalize: Enables recommendations.
- Polly: Converts text to speech.
- Rekognition: Classifies images and videos.
- Textract: Extracts text and data from documents.
- Translate: Translate text to different languages.
- Transcribe: Converts speech to text.

Module 2: Practical Data Science with Amazon SageMaker

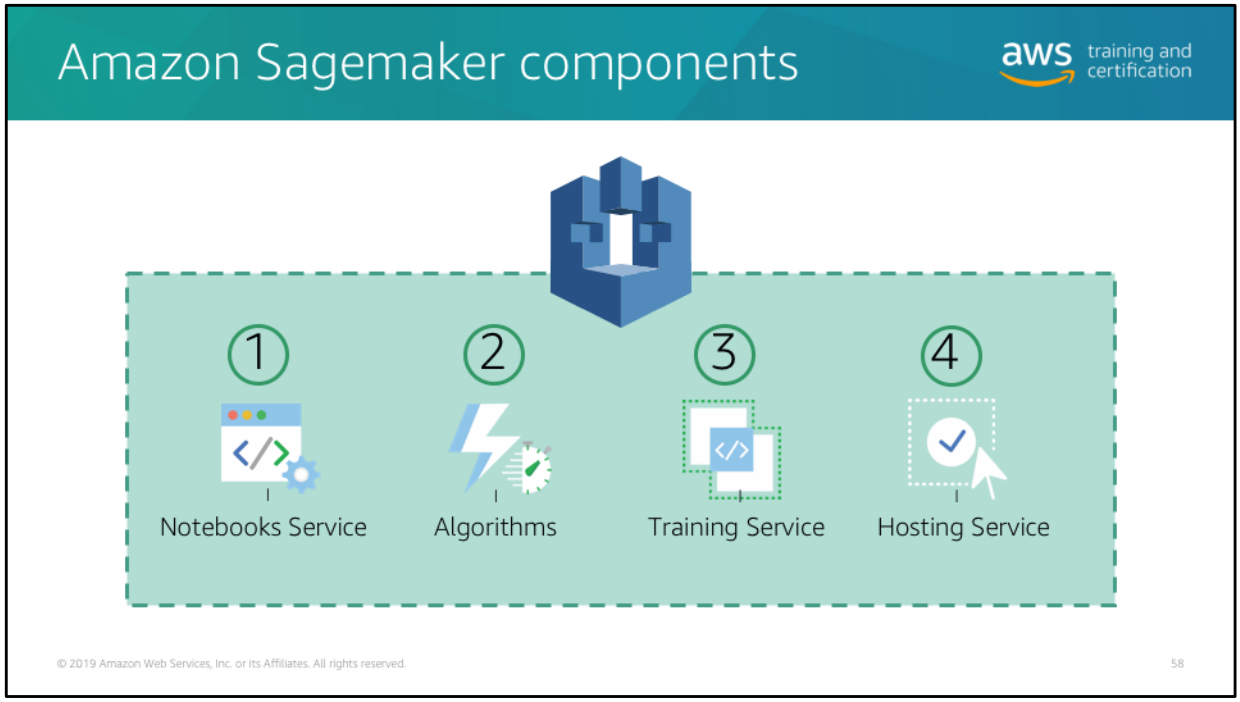
aws training and certification

Note: Play this video in incognito mode in the browser; otherwise it does not start and end correctly.

https://www.youtube.com/embed/lM4zhNO5Rbg?start=24&end=381&version=3

aws training and certification

Amazon SageMaker



The next handful of slides will cover Amazon SageMaker's 4 main components.

1 – Amazon SageMaker Notebooks Service
Amazon SageMaker provides a managed environment for Jupyter notebooks. It's an EC2 machine of the size you pick, accessible after a few clicks.

2. Amazon SageMaker Algorithms
SageMaker contains ready to train algorithms in the form of a Docker container. You just have to bring your data. These are the common ML algorithms that were optimized for **10x speed** and scale to **large datasets**.

3 – Amazon SageMaker Training Service
A fully managed algorithms training service. You can use Amazon provided ML algorithms, or bring your own algorithm (BYOA) in the form of a Docker container. Your ML algorithm could be written in any language or framework, SageMaker training is flexible enough to train the model at scale. Out of the box, SageMaker supports distributed training for MXNet- and TensorFlow-based models, as well as Amazon-provided algorithms. Amazon SageMaker also support Spark ML compatible Estimators (SageMaker Estimators) which enable Spark ML practitioners to launch a

training job into Amazon SageMaker from Apache Spark.

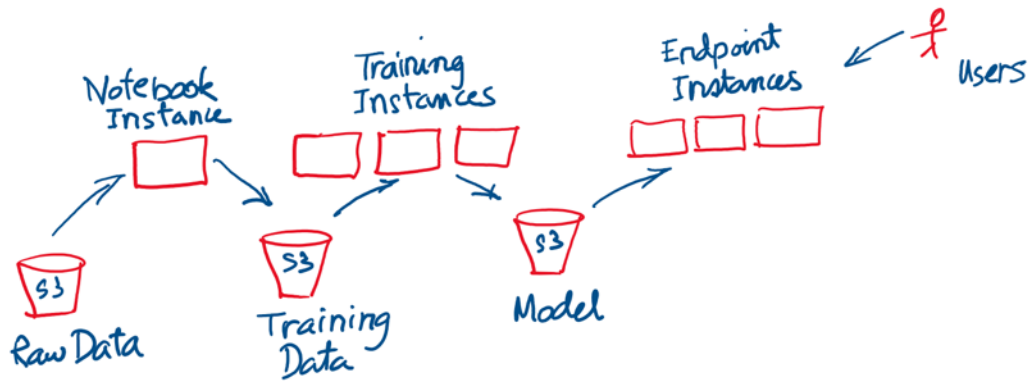
4 – Amazon SageMaker Hosting Service
A fully managed model hosting service that supports:

- Amazon SageMaker trained models or BYOM (Bring Your Own Model)
- For SageMaker provided algorithms, MXNet- and TensorFlow-based training, the hosting image is provided.
- For BYOM deployments or BYOA trained models, you have to provide a Docker container with your inference code and the location of your trained model artifacts on Amazon S3 (see docs for more details >>).
- Hosting API Versioning
- Weighted A/B deployments
- Auto scaling
- Low latency and high throughput

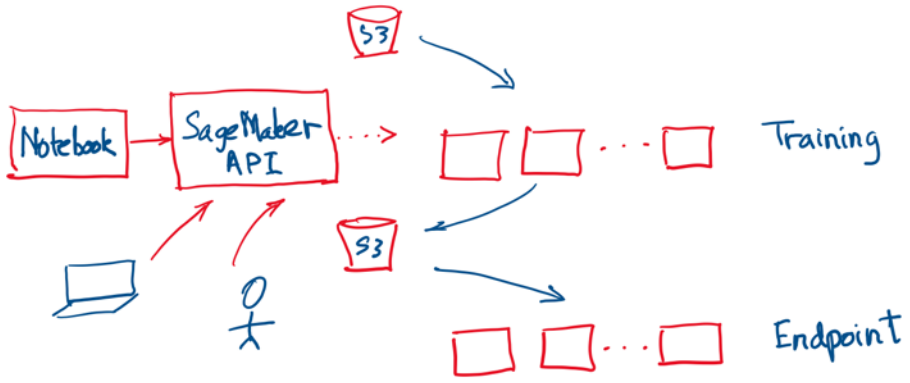
The Benefits of all of this:

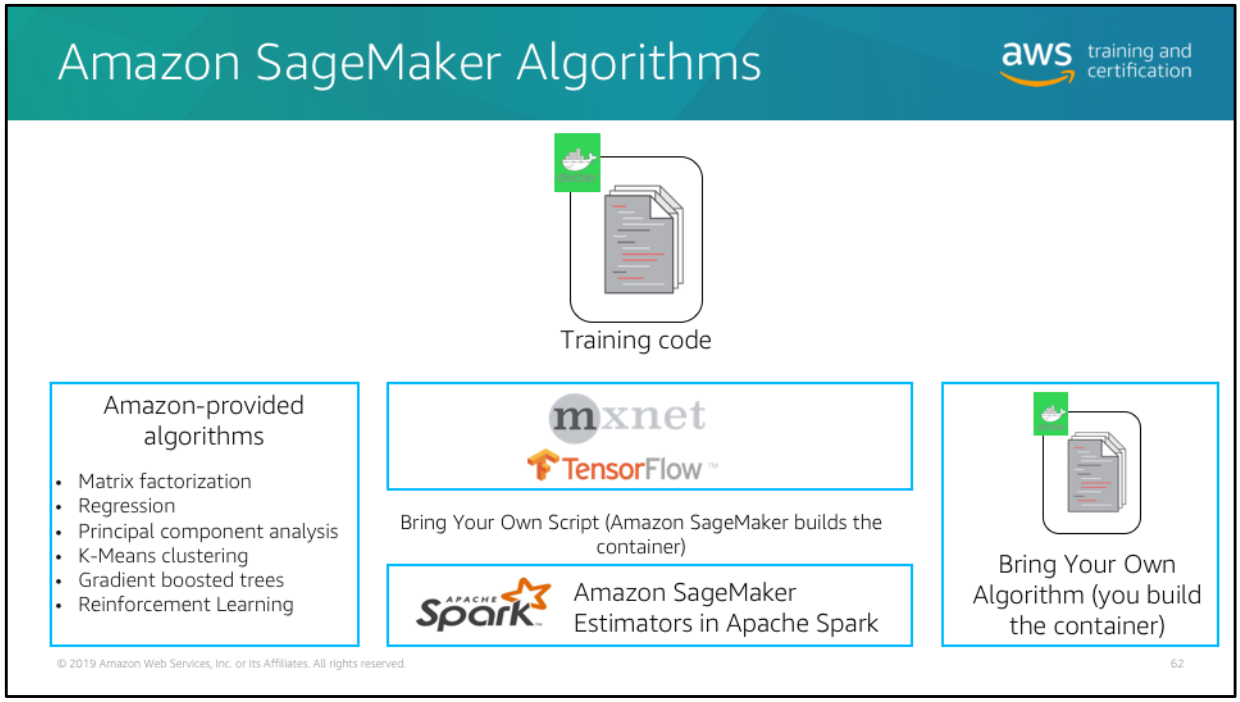
- Agile, Reliable, GPU powered, and Productivity Ready Workspaces for Data Scientist and Developers.
- High Performance Web-Scale Algorithms Out Of The Box.
- Managed Distributed model training service.
- Production Ready Model Hosting Requiring No Engineering.

SageMaker Components



SageMaker Components





The atomic unit of machine learning algorithm training on Amazon IM is a Docker container. After you explore the Jupyter notebooks, you will find the training cove in a Docker container image that is available on Amazon ECR. It will be shipped to the Amazon SageMaker Algorithms Training Service.

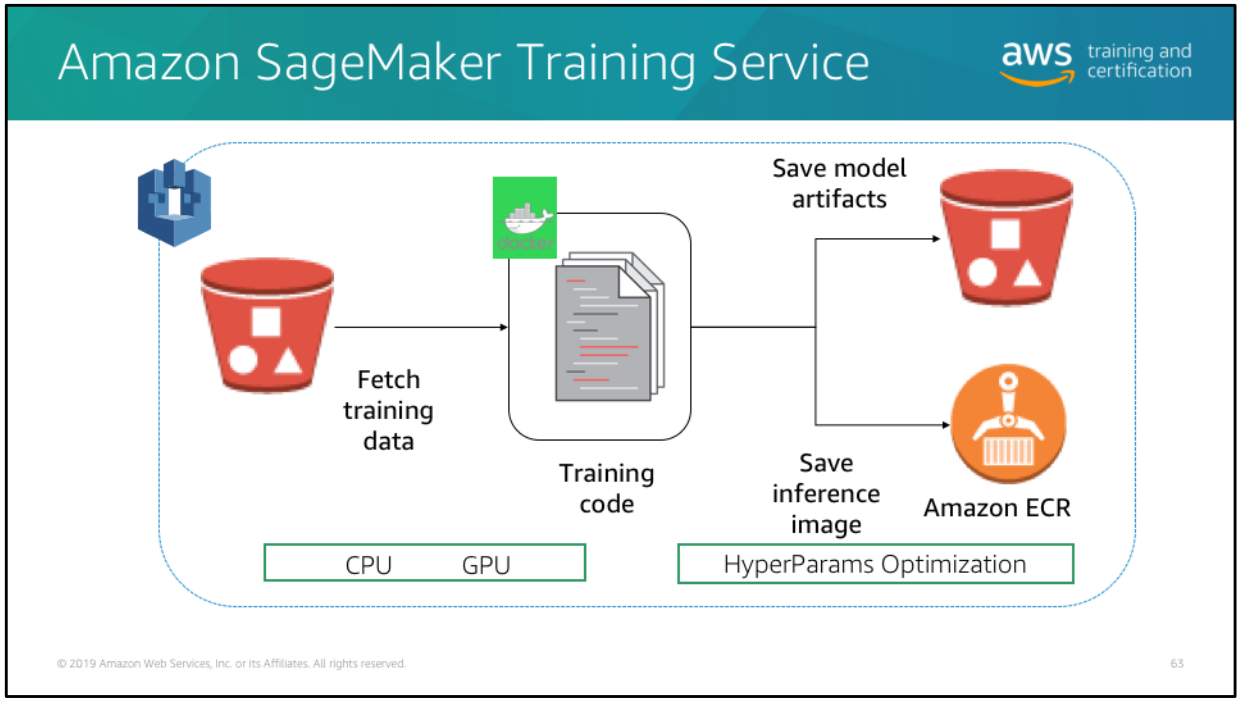
Amazon provided Algorithms support:

- Matrix Factorization
- Regression
- Principal Component Analysis
- K-Means Clustering
- Gradient Boosted Trees
- Reinforcement Learning

For Deep Learning, Amazon SageMaker provides the container image for MXNet and TensorFlow. This model is called **BYOS** (Bring Your Own Script). Just submit your training, and Amazon SageMaker will build an Algorithm training image required to launch and distribute training workloads for MXNet and TensorFlow. For Algorithms built in other machine learning or deep learning frameworks, regardless of the programming language or framework, Amazon SageMaker supports a **BYOA** (Bring

Your Own Algorithm) model through which you can package your own algorithm in a Docker container image, and submit it to Amazon IM for training.

The container must expose a REST API to which Amazon SageMaker will submit training data observations for the number of training iterations. Finally, Amazon SageMaker support SparkML-like estimators. This makes it easy to start a pre-processing job in a Spark cluster, and submit the machine learning training part of the code to Amazon SageMaker. The Spark pipeline has a data transformation stage, and a model estimation (training) stage where models are fit over the transformed dataset. This feature enables data scientists and developers familiar with the Spark way of building ML pipelines, to leverage Amazon SageMaker without changing the familiar syntax or the environment they are use do.



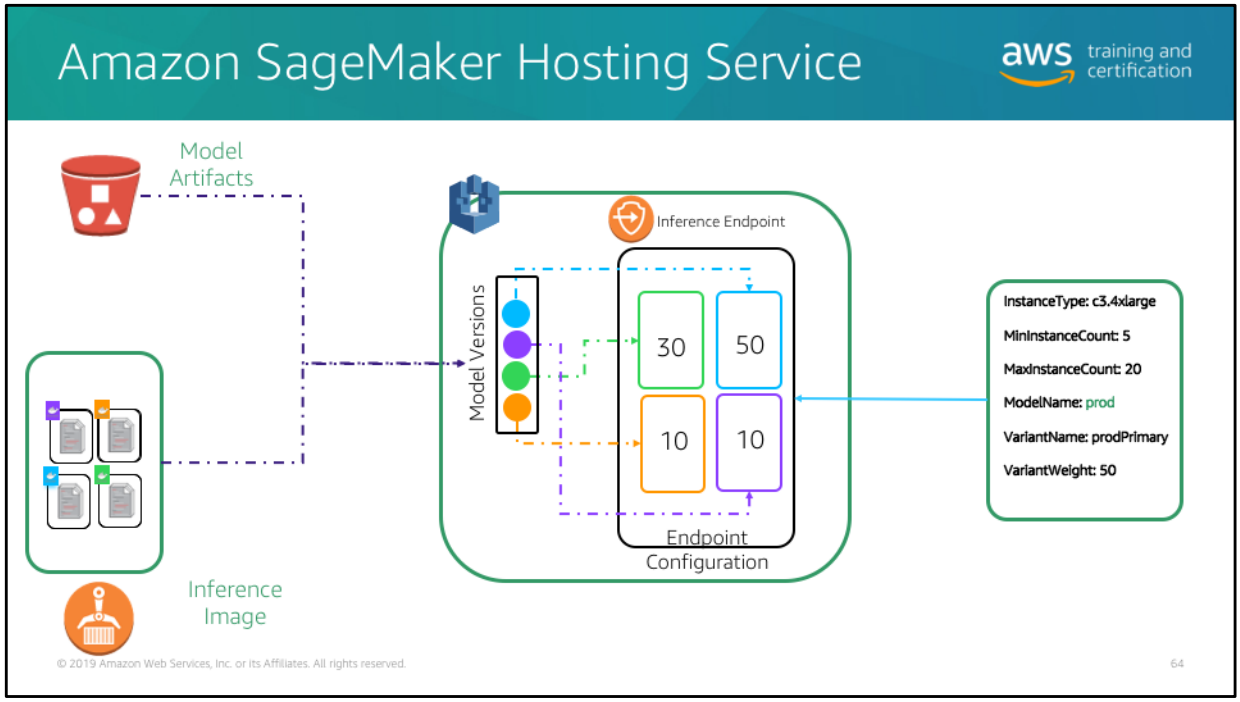
Managed Distributed Training With Flexibility

Training machine learning algorithms that are wrapped in a Docker image requires:

- Launching a number of containers from the images (there could be helper code available in a secondary container)
- Connecting the Docker containers to source data on Amazon S3. Download or stream the data to the container
- Train the algorithms for iterations set in the hyper parameters, or for a certain time duration
- Save model artifacts back to Amazon S3
- Save an inference image to Amazon ECR

Training these algorithms can be done on CPU or GPU infrastructure.

Amazon IM fully manages the training process in a secured environment. The training cluster is destroyed once training is over, which makes the training experience cost-effective.



To host a model on Amazon SageMaker, you need to two things:

1- **Model artifacts saved on Amazon S3**: These are the data-dependent definitions of a model. In the case of a deep neural network, this will be the neural network configuration (number of layers, size of input, etc.) associated with the weights values of the connections between neurons.
This is the final state of a trained algorithm. For A/B testing, you will likely have different model artifacts for different versions of the same model. Also, to try different models on the same problem, you will have different model artifacts on Amazon S3.

2- **Model inference images on Amazon ECR**: These are the container images Amazon SageMaker will launch in order to serve predictions for a specific model.

An Amazon SageMaker model is a entity that **links** the inference image to the model artifacts**.** You can host multiple versions of the same model on Amazon SageMaker by creating multiple Amazon SageMaker models, linking inference images to model artifacts.

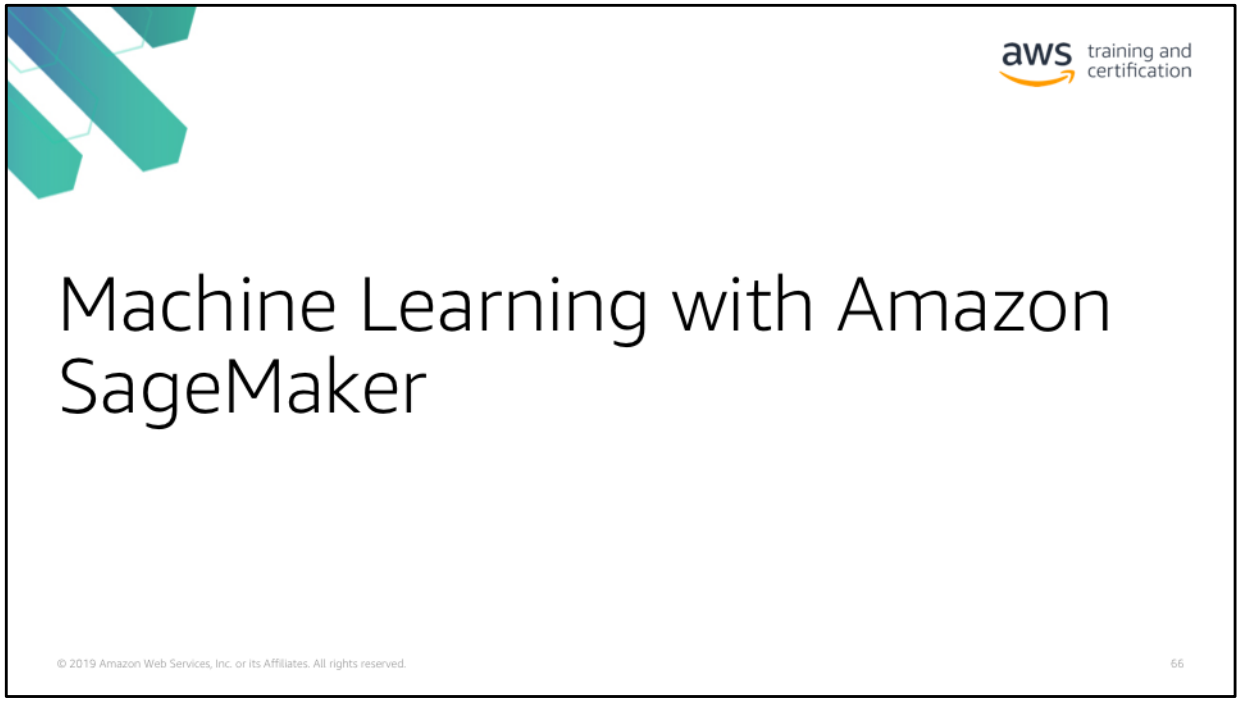
With one or many versions of a model, you create an endpoint configuration that defines the weight associated with each model. This is the essence of workload distribution to many versions of the model. From the Endpoint Configuration, Amazon SageMaker knows how to distribute prediction/inference traffic to the different models in the configuration.

For A/B testing for example, you will have the primary model handling 80 percent of the traffic, and the model you're testing handling 20 percent. That equates to a weight of 80 for the primary model, and 20 for the model you are testing. This is done using a ProductionVariant command. A ProductionVariant command identifies the model that you want to host and resources that you want to deploy for hosting the model. In the slide, four models are hosted in an endpoint configuration, with the blue model supporting 50% of the overall traffic.

For Amazon-provided algorithms, MXNet and TensorFlow algorithms, it's as easy as a one-click deployment. Amazon SageMaker provides the inference containers.

How SageMaker Works

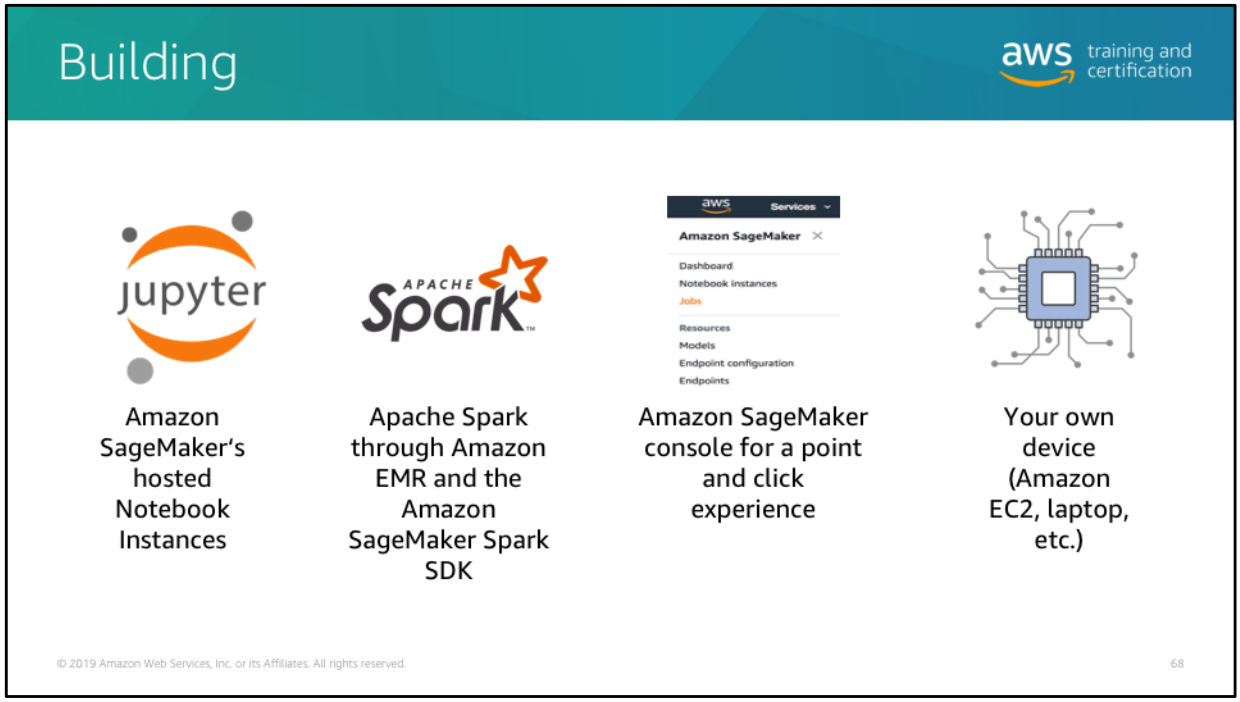
aws training and certification



This section describes a typical machine learning workflow and summarizes how you accomplish those tasks with Amazon SageMaker.

In machine learning, you "teach" a computer to make predictions, or inferences. First, you use an algorithm and example data to train a model. Then you integrate your model into your application to generate inferences in real time and at scale. In a production environment, a model typically learns from millions of example data items and produces inferences in hundreds to less than 20 milliseconds.

The following diagram illustrates the typical workflow for creating a machine learning model:



When it comes to machine learning, your model is only as good as your data. Data can come from multiple places. It can be internal, external, public, purchased, or something else. No matter its origin, however, the data needs to be cleaned. Once the data is ready, choose an algorithm and then train the model to make predictions.

Amazon SageMaker helps you choose algorithms and frameworks, and has integrated the field's most common and popular algorithms.

Amazon SageMaker also enables you to use your own algorithms and frameworks. Once you've selected your algorithm, training is easy.

Training and Testing

aws training and certification

Zero setup

Streaming datasets and distributed compute

Docker/ Amazon ECS

Deploy trained models locally or to Amazon SageMaker, Greengrass, DeepLens

We need text to describe this slide.

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

-Ashok Srivastava

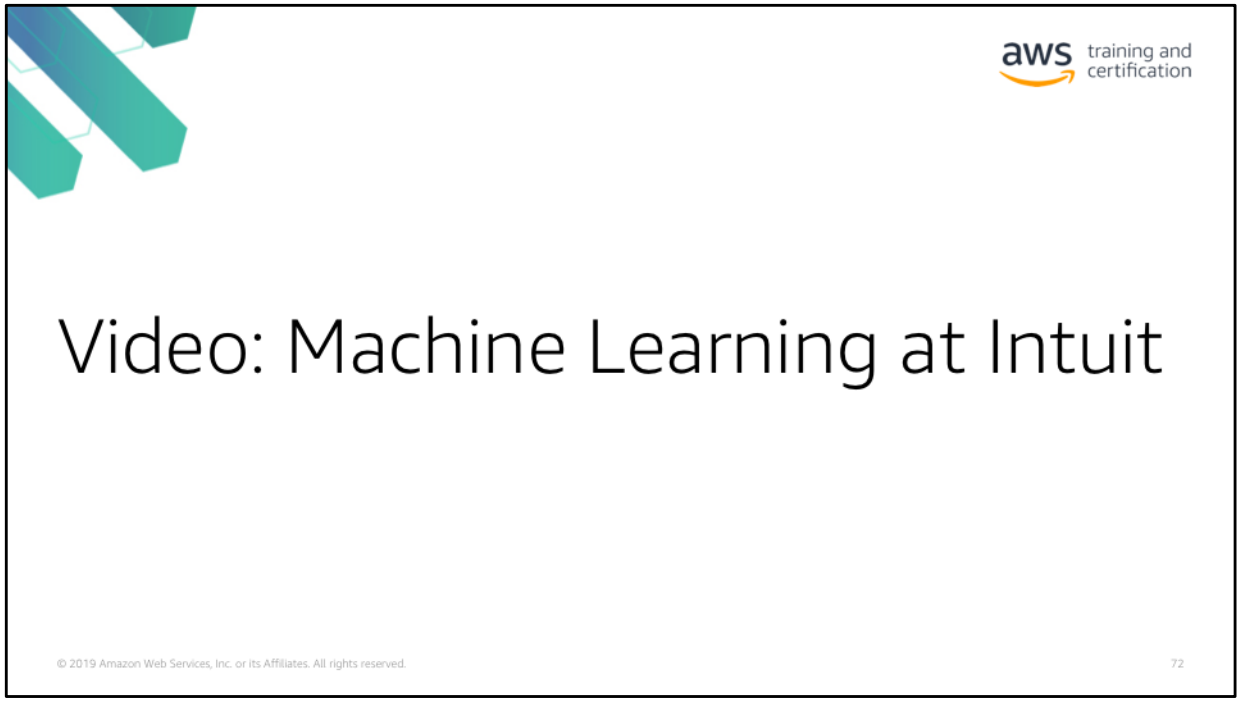
Chief Data Officer at Intuit

aws training and certification

Before SageMaker

aws training and certification

- What were things like before SageMaker?
  - Deployments took longer.
  - Required infrastructure expertise on teams.
- What are primary infrastructure components?
  - Training infrastructure: Scales to handle large amounts of data.
  - Inference infrastructure: Scales to handle large number of inference requests.
- Let's compare the before and after by looking at Intuit.



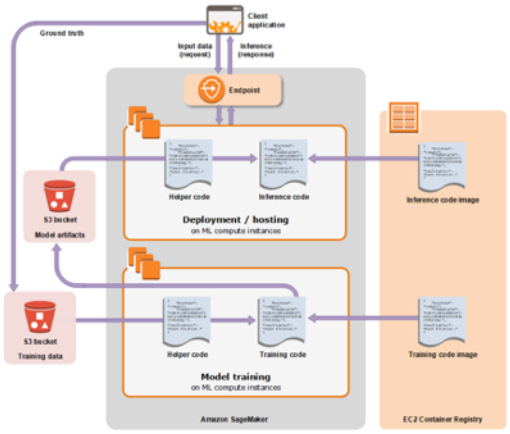
https://www.youtube.com/embed/0NJJAqRtGiA?start=219&end=513&version=3

Reminder to do it through the console.

Amazon SageMaker workflow

aws training and certification

- Prepare training data
- Train model
- Tune hyperparameters
- Deploy as real-time endpoint or batch transform



- Data pre-processing: Extract and pre-process data from Amazon S3 to prepare the training data.
- Prepare training data: To build the recommender system, we'll use the Amazon SageMaker built-in algorithm, Factorization machines. The algorithm expects training data only in recordIO-protobuf format with Float32 tensors. In this task, pre-processed data will be transformed to RecordIO Protobuf format.
- Training the model:Train the Amazon SageMaker built-in factorization machine model with the training data and generate model artifacts. The training job will be launched by the Airflow Amazon SageMaker operator.
- Tune the model hyperparameters:A conditional/optional task to tune the hyperparameters of the factorization machine to find the best model. The hyperparameter tuning job will be launched by the Amazon SageMaker Airflow operator.
- Batch inference:Using the trained model, get inferences on the test dataset stored in Amazon S3 using the Airflow Amazon SageMaker operator.

Note: Play this video in incognito mode in the browser; otherwise it does not start and end correctly.

https://www.youtube.com/embed/jWWyKE5ApqI?start=321&end=660&version=3

aws training and certification

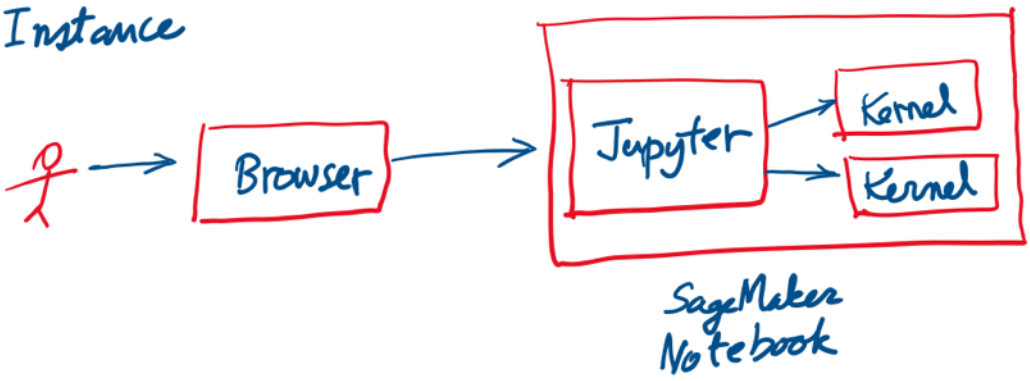
Getting Started with QwikLabs

Setting up QwikLabs

aws training and certification

1. Open a browser and navigate to **aws.qwiklabs.com**
2. Create a new account or log in with your existing account.
3. Open the **Classrooms** tab on the left to find the courses you enrolled in.
4. Click the course name to view each of the labs for your class.
5. Click the lab and follow the steps to launch the selected labs.
6. On the Labs page, click the **Start Lab** button to launch your lab.
7. Under Connection Details, copy the **username** and **password.**
8. Click **Open Console** to open the AWS Management Console.
9. Enter the username and password you copied from Connection Details.
10. Click **Sign In**.

Jupyter Notebooks



aws training and certification

Lab: Launch Jupyter Notebook

aws training and certification

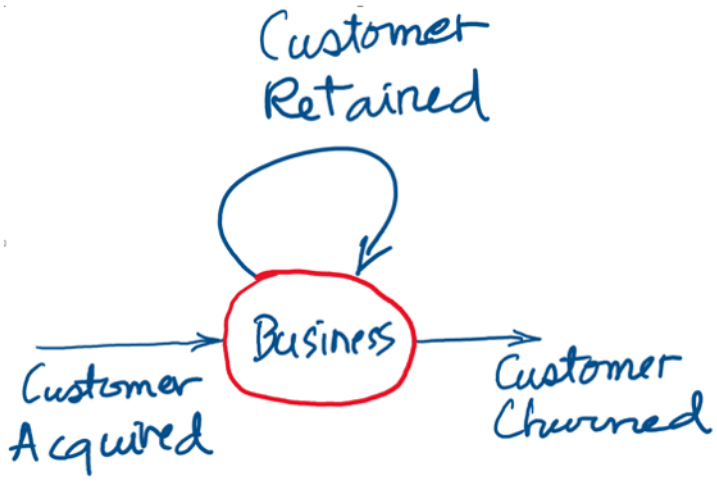
Business Challenge and an Introduction to DataSets

The Business Challenge: Leaving a Mobile Phone Operator

aws training and certification

- Background information: If a provider knows that a particular customer is thinking of leaving, the provider can offer timely incentives.
- For example, one incentive might be a phone upgrade, which may encourage the customer to stay with the provider.
- Offering an incentives can be more cost-effective than losing and then reacquiring a customer.
- ***Important:*** ML models ***rarely*** give perfect predictions. We will also show how to incorporate the relative costs of prediction mistakes.

Mobile operator customer churn



Dataset introductions

aws training and certification

- Mobile operators have historical records on customers who churn, and who end up continuing the service.
- Use this historical information to construct an ML model. After training the model, pass the profile information of an arbitrary customer through it and create a prediction. Will this customer churn or not?
- ***Important:*** Expect the model to make mistakes. Use this back-propagation to optimize your model.
- The dataset we are going use is publicly available and is mentioned in the book *Discovering Knowledge in Data* by Daniel T. Larose. The author attributes the dataset to the University of California Irvine Repository of Machine Learning Datasets.

The dataset we use is publicly available and was mentioned in the *book Discovering Knowledge in Data* by Daniel T. Larose. It is attributed by the author to the University of California Irvine Repository of Machine Learning Datasets.

Download and read that dataset now.

Training Data Set

1 = True / Yes
0 = False / No

aws training and certification

Demo: Visualizing Data

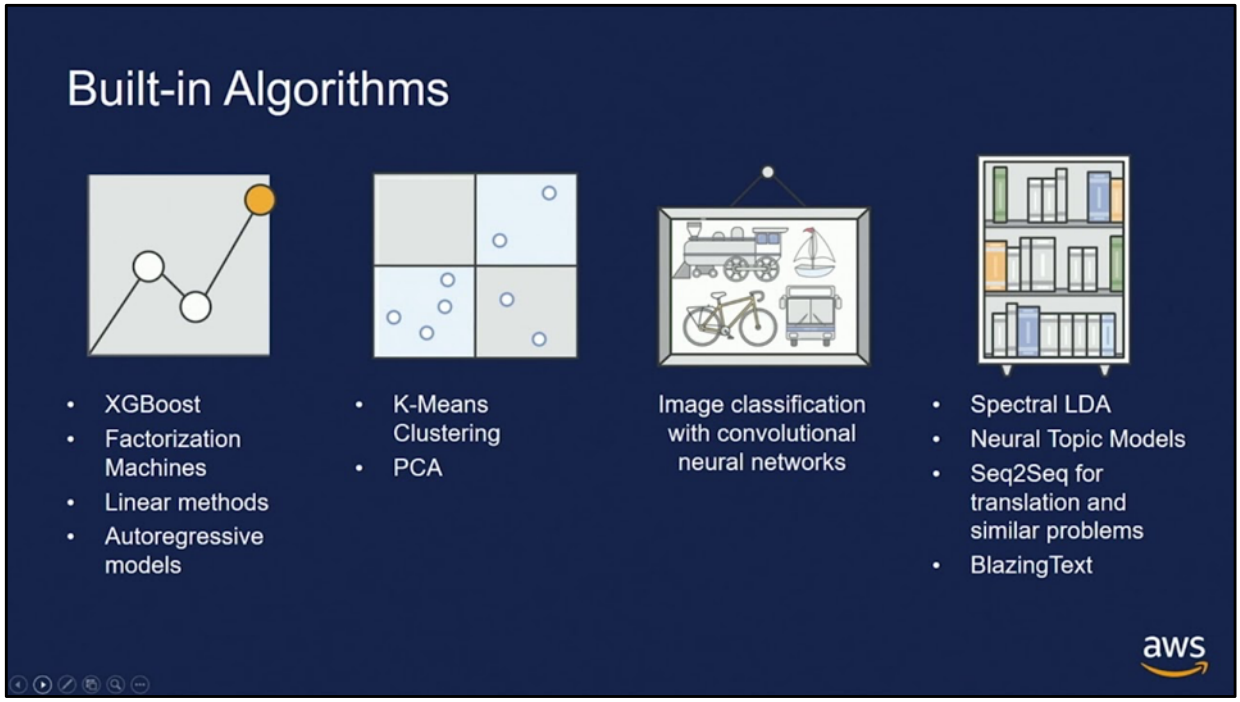
Cleaning the data

aws training and certification

- In preparation for model training, first remove the columns observed as useless for our purposes.
- Then remove one feature from each of the highly-correlated pairs:
  - Day Charge from the pair with Day Mins, Night
- ***Important:*** We see features that essentially have 100% correlation with one another.
- Once the dataset is cleaned up, we can determine which algorithm to use.

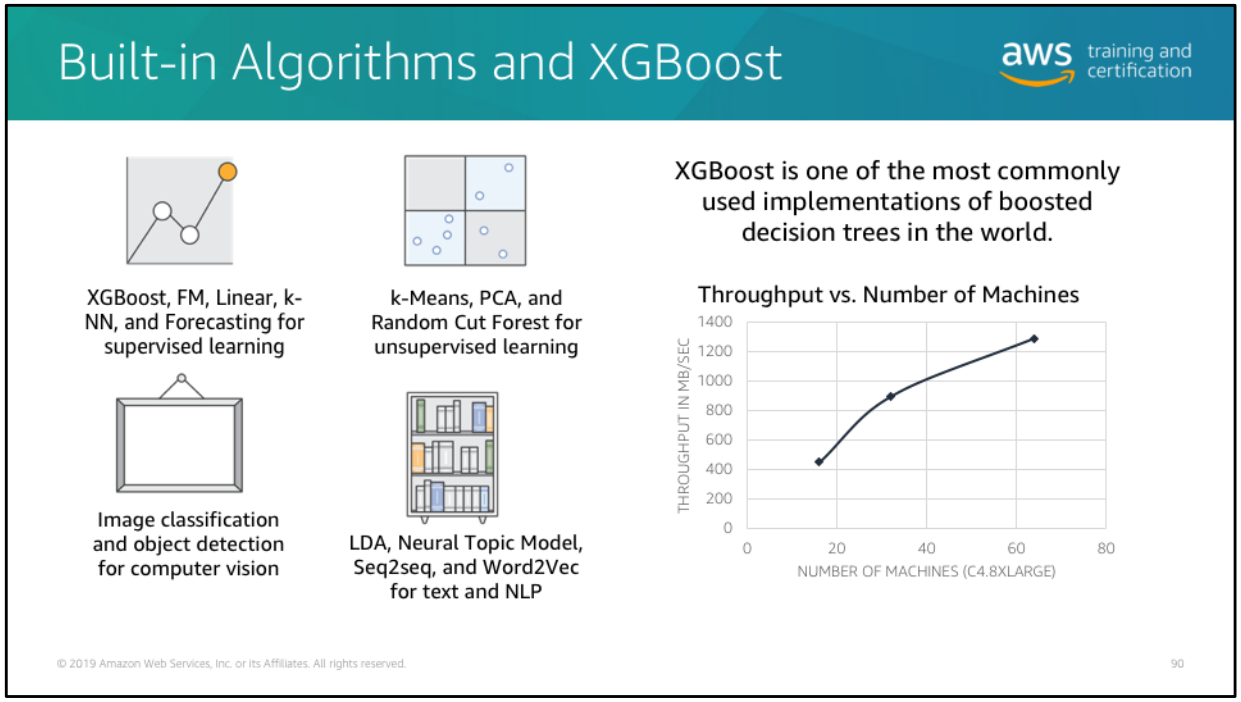
aws training and certification

Preparing Data: Amazon SageMaker's Built-in Algorithms and XGBoost



SageMaker algorithms

- SageMaker algorithms are broadly provided in 4 categories.
    - Supervised
    - Unsupervised
    - Image
    - Natural language processing

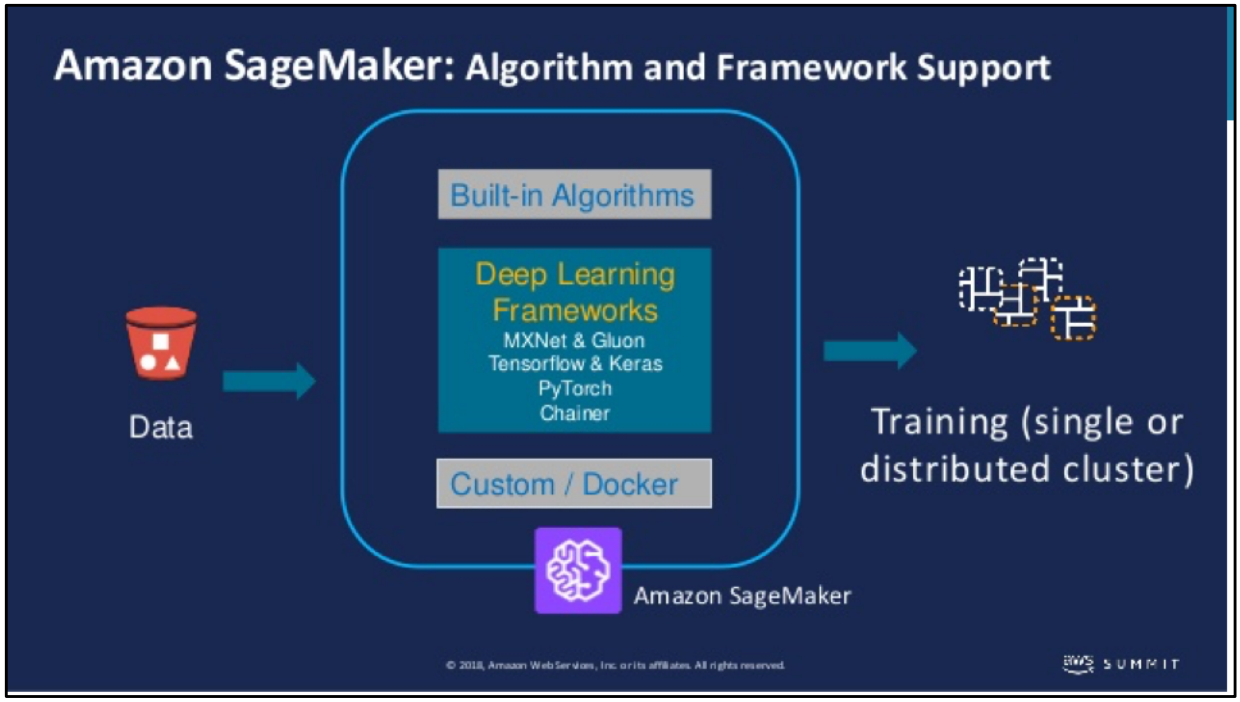


Built-in algorithms can be divided into four categories:
- Supervised Learning: Classification, Regression
- Unsupervised Learning
- Image Classification
- Natural Language Processing

XGBoost (eXtreme Gradient Boosting) is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models. XGBoost has done remarkably well in machine learning competitions because it robustly handles a variety of data types, relationships, and distributions, and the large number of hyperparameters that can be tweaked and tuned for improved fits. This flexibility makes XGBoost a solid choice for problems in regression, classification (binary and multiclass), and ranking.

When using gradient boosting for regression, the weak learners are regression trees, and each regression tree maps an input data point to one of its leafs that contains a continuous score. XGBoost minimizes a regularized (L1 and L2) objective function that

combines a convex loss function (based on the difference between the predicted and target outputs) and a penalty term for model complexity (in other words, the regression tree functions). The training proceeds iteratively, adding new trees that predict the residuals or errors of prior trees that are then combined with previous trees to make the final prediction. It's called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.


Amazon SageMaker: Algorithm and Framework Support
Built-in Algorithms
Deep Learning Frameworks
MXNet & Gluon
Tensorflow & Keras
PyTorch
Chainer
Custom / Docker
Amazon SageMaker
Data
Training (single or distributed cluster)
© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
SUMMIT

Use an algorithm that you subscribe to from AWS Marketplace

With AWS Marketplace, you can browse and search for hundreds of machine learning algorithms and models in a broad range of categories, such as computer vision, natural language processing, speech recognition, text, data, voice, image, video analysis, fraud detection, predictive analysis, and more.

Algorithms for Discrete Classifications

aws training and certification

- "Based on past customer responses, should I mail this particular customer?"
  - Answers to this question fall into two categories, "yes" or "no."
  - In this case, you use the answer to narrow the recipients of the mail campaign.
- "Based on past customer segmentation, which segment does this customer fall into?"
  - Answers might fall into categories such as "empty nester," "suburban family," or "urban professional."
  - You could use these segments to decide who should receive the mailing.
- For these type of discrete classification problem, Amazon SageMaker provides two algorithms: Linear Learner Algorithm and the XGBoost Algorithm.

Algorithms for Quantitative Problems

aws training and certification

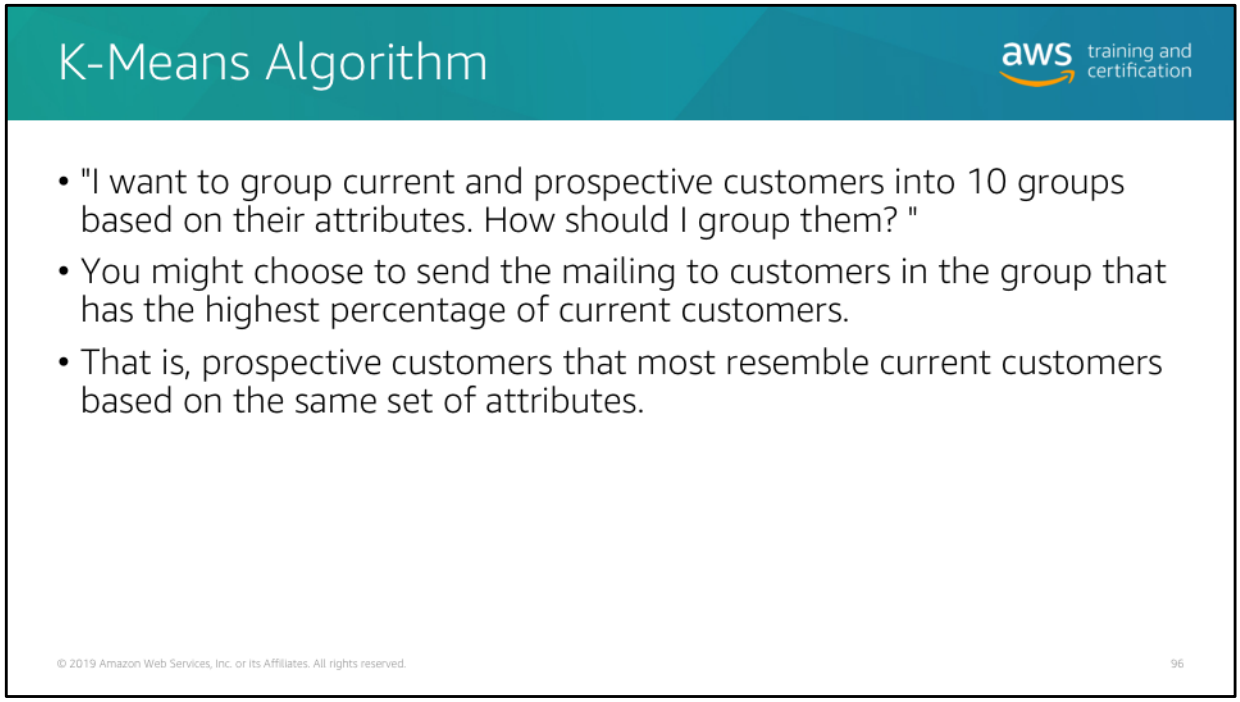
- "Based on the return on investment (ROI) from past mailings, what is the ROI for mailing this customer?"
- In this case, you use the ROI to target customers for the mail campaign.
- For these quantitative analysis problems, you can also use the Linear Learner Algorithm or the XGBoost Algorithm algorithms.

Algorithms for Recommendations

aws training and certification

- "Based on past responses to mailings, what is the recommended content for each customer?"
  - In this case, you are looking for a recommendation on what to mail, not whether to mail, the customer.

For this problem, Amazon SageMaker provides the Factorization Machines Algorithm algorithm.



For this type of question, Amazon SageMaker provides the K-Means Algorithm.

Principal Component Analysis (PCA) Algorithm

aws training and certification

- "What are the attributes that differentiate these customers, and what are the values for each customer along those dimensions."
- You use these answers to simplify the view of current and prospective customers, and, maybe, to better understand these customer attributes.

For this type of question, Amazon SageMaker provides the Principal Component Analysis (PCA) Algorithm algorithm.

Use Case Specific Algorithms

aws training and certification

- Image Classification Algorithm—Use this algorithm to classify images. It uses example data with answers (referred to as supervised algorithm).
- Sequence-to-Sequence Algorithm—This supervised algorithm is commonly used for neural machine translation.
- Latent Dirichlet Allocation (LDA) Algorithm—This algorithm is suitable for determining topics in a set of documents. It is an unsupervised algorithm, which means that it doesn't use example data with answers during training.
- Neural Topic Model (NTM) Algorithm—Another unsupervised technique for determining topics in a set of documents, using a neural network approach.

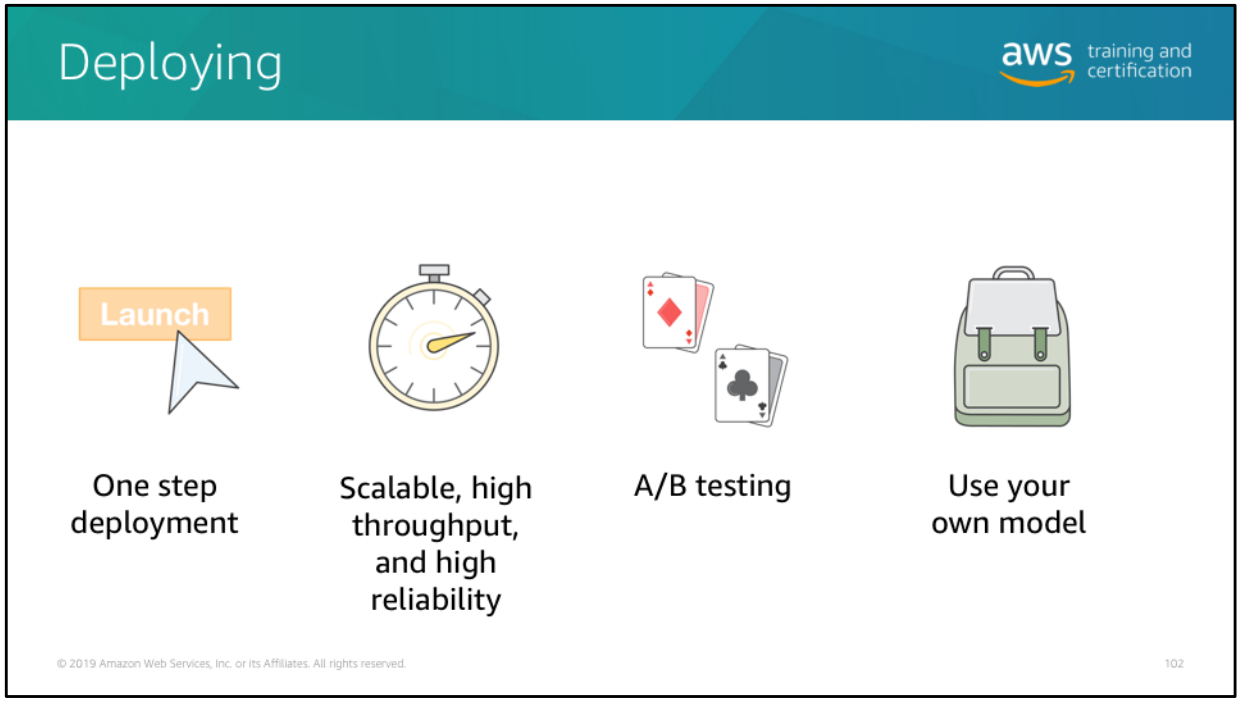
In addition to these general-purpose algorithms, Amazon SageMaker provides algorithms that are tailored to specific use cases. These include:

https://www.youtube.com/embed/ami2RFJBrsI?start=823&end=1053&version=3

https://www.youtube.com/embed/ami2RFJBrsI?start=1095&end=1374&version=3

aws training and certification

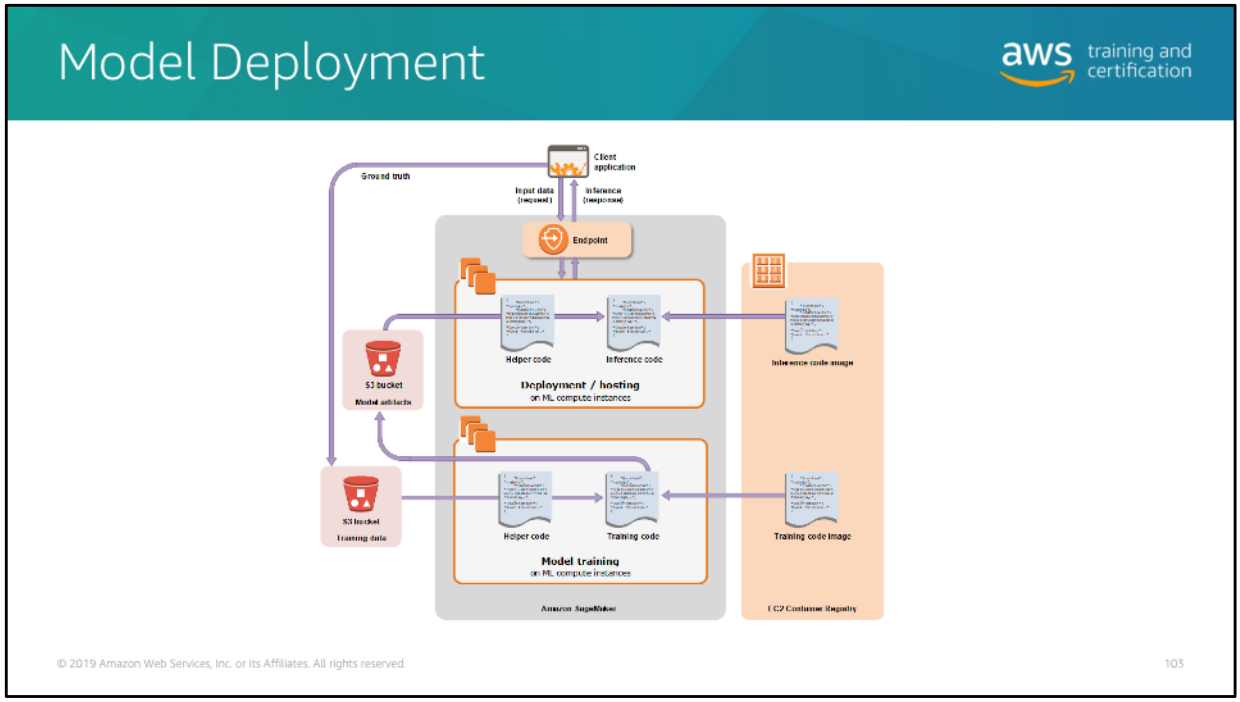
Deployment



Deploying a machine learning model to production requires lots of iterations and in-depth knowledge of devops. This is especially true to run a scalable API endpoint serving millions of requests and conducting AB testing. Amazon SageMaker allows you to do this with minimal friction.

Once your model is trained and tuned, Amazon SageMaker makes it easy to deploy in production so you can start generating predictions (a process called *inference*) for real-time or batch data. Amazon SageMaker deploys your model on automatically scaling clusters of Amazon SageMaker ML instances that are spread across multiple Availability Zones to deliver both high performance and high availability. Amazon SageMaker also includes built-in A/B testing capabilities to help you test your model and experiment with different versions to achieve the best results.

Amazon SageMaker eliminates the heavy lifting of machine learning, so that you can build, train, and deploy machine learning models quickly and easily.



Create a model in Amazon SageMaker—By creating a model, you tell Amazon SageMaker where it can find the model components. This includes the S3 path where the model artifacts are stored and the Docker registry path for the image that contains the inference code. In subsequent deployment steps, you specify the model by name.
Create an endpoint configuration for an HTTPS endpoint—You specify the name of one or more models in production variants and the ML compute instances that you want Amazon SageMaker to launch to host each production variant.

When hosting models in production, you can configure the endpoint to elastically scale the deployed ML compute instances. For each production variant, you specify the number of ML compute instances that you want to deploy. When you specify two or more instances, Amazon SageMaker launches them in multiple Availability Zones. This ensures continuous availability. Amazon SageMaker manages deploying the instances. For more information, see the CreateEndpointConfig (p. 596) API.

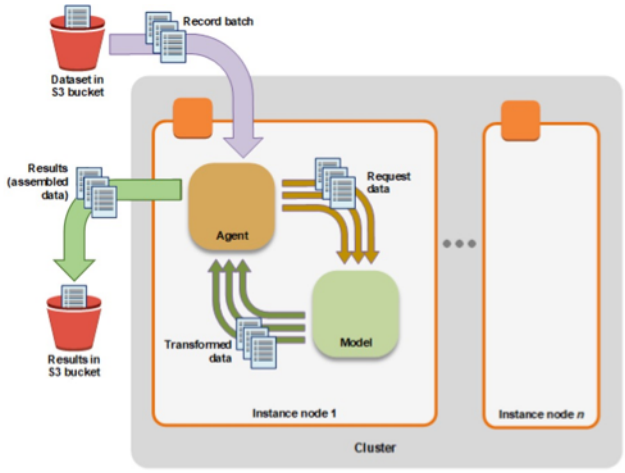
Create an HTTPS endpoint—Provide the endpoint configuration to Amazon SageMaker. Theservice launches the ML compute instances and deploys the model or models as specified in the configuration. For more information, see the

CreateEndpoint (p. 593) API. To get inferences from the model, client applications send requests to the Amazon SageMaker Runtime HTTPS endpoint. For more information about the API, see the InvokeEndpoint (p. 812) API.

When you create an endpoint, Amazon SageMaker attaches an Amazon EBS storage volume to each ML compute instance that hosts the endpoint. The size of the storage volume depends on the instance type. For a list of instance types that Amazon SageMaker hosting service supports, see AWS Service Limits. For a list of the sizes of the storage volumes that Amazon SageMaker attaches to each instance, see Hosting Instance Storage Volumes (p. 365).

To increase a model's accuracy, you might choose to save the user's input data and ground truth, if available, as part of the training data. You can then retrain the model periodically with a larger, improved training dataset.

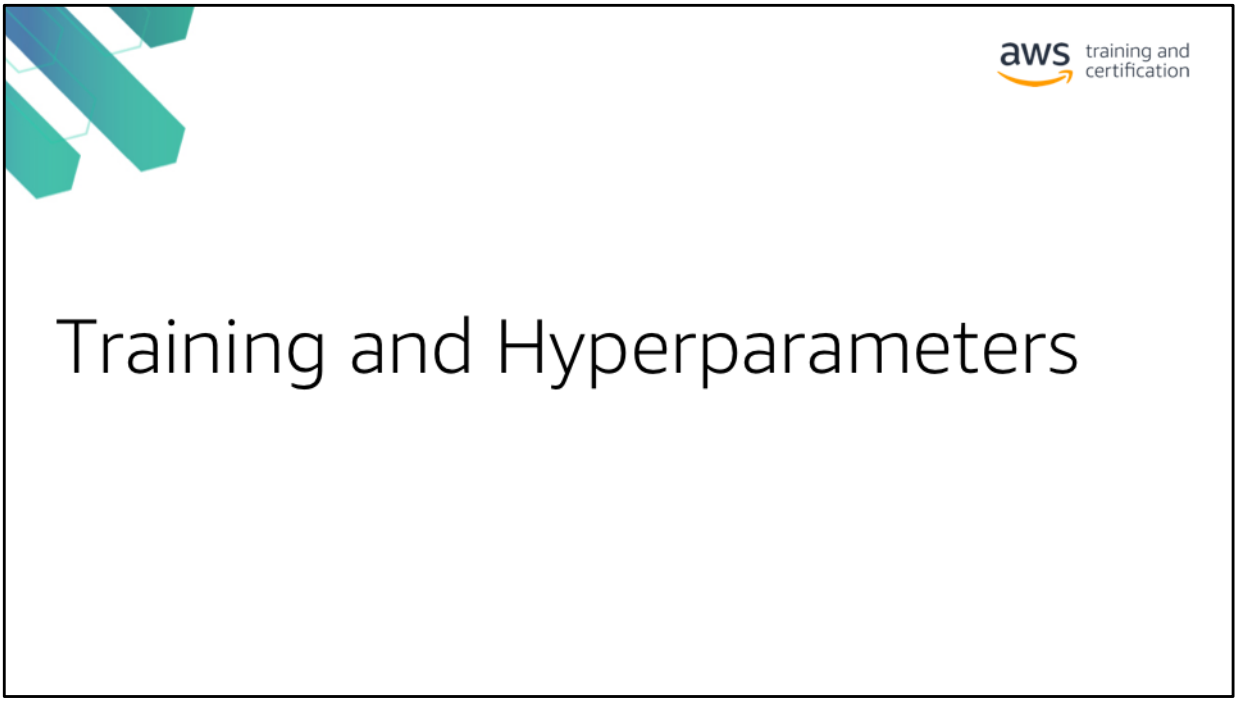
Batch Transform



With batch transform, you create a batch transform job using a trained model and the dataset, which must be stored in Amazon S3. Amazon SageMaker saves the inferences in an S3 bucket that you specify when you create the batch transform job. Batch transform manages all of the compute resources required to get inferences. This includes launching instances and deleting them after the batch transform job has completed. Batch transform manages interactions between the data and the model with an object within the instance node called an agent.

Use batch transform when you:

- Want to get inferences for an entire dataset and index them to serve inferences in real time
- Don't need a persistent endpoint that applications (for example, web or mobile apps) can call to get inferences
- Don't need the subsecond latency that Amazon SageMaker hosted endpoints provide
- You can also use batch transform to preprocess your data before using it to train a new model or generate inferences.


aws training and certification
Training and Hyperparameters

Training Data

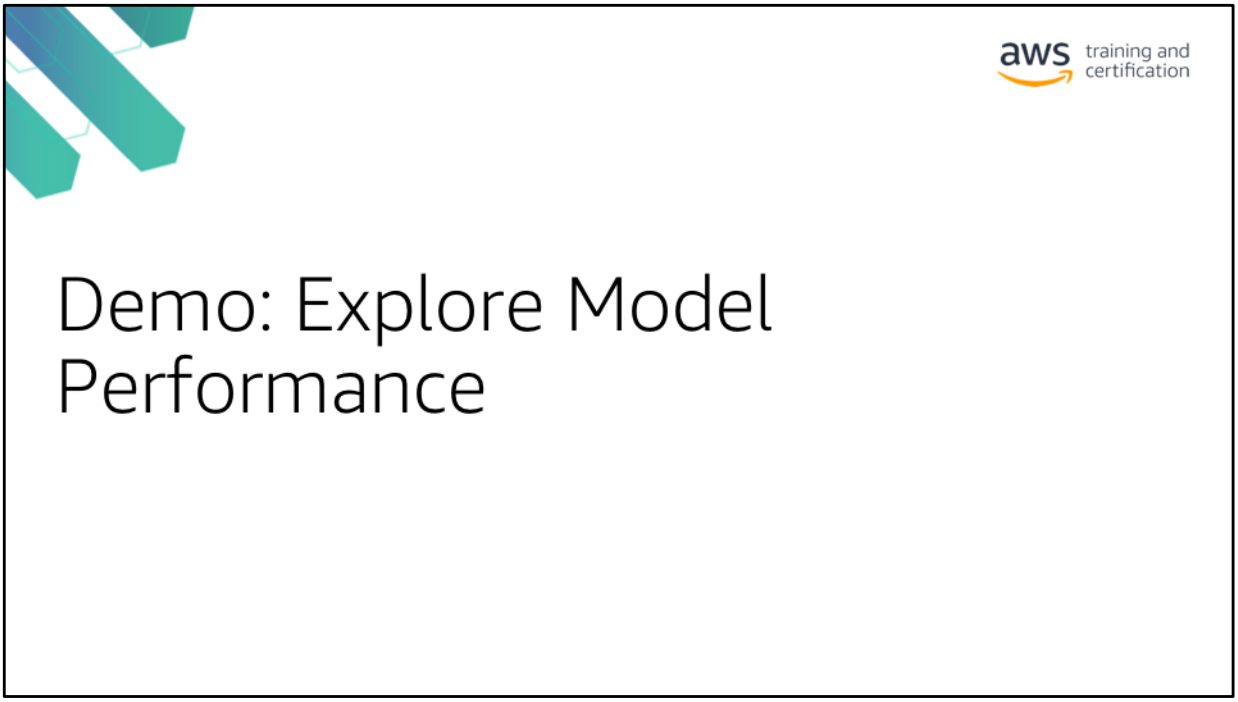
aws training and certification

- Amazon SageMaker XGBoost can train on data in either a CSV or LibSVM format.
- For this example, use CSV.
  - It should have the predictor variable in the first column.
  - It should ***not*** have a header row.
- First we'll convert our categorical features into numeric features, then we'll split the data into training, validation, and test sets.

Hyperparameters

aws training and certification

- An ML algorithm is configured and tuned based on its hyperparameters.
- Hyperparameters change the way the algorithm works.
- For this example, the required hyperparameters for XGBoost are:
  - objective - Specifies the learning task and the corresponding learning objective.
  - Use **binary:logistic** for the binary classification task.
  - num_round - Controls the number of boosting rounds. This is essentially the subsequent models that are trained using the residuals of previous iterations.


aws training and certification
Demo: Explore Model Performance

Creating and Evaluating a Model

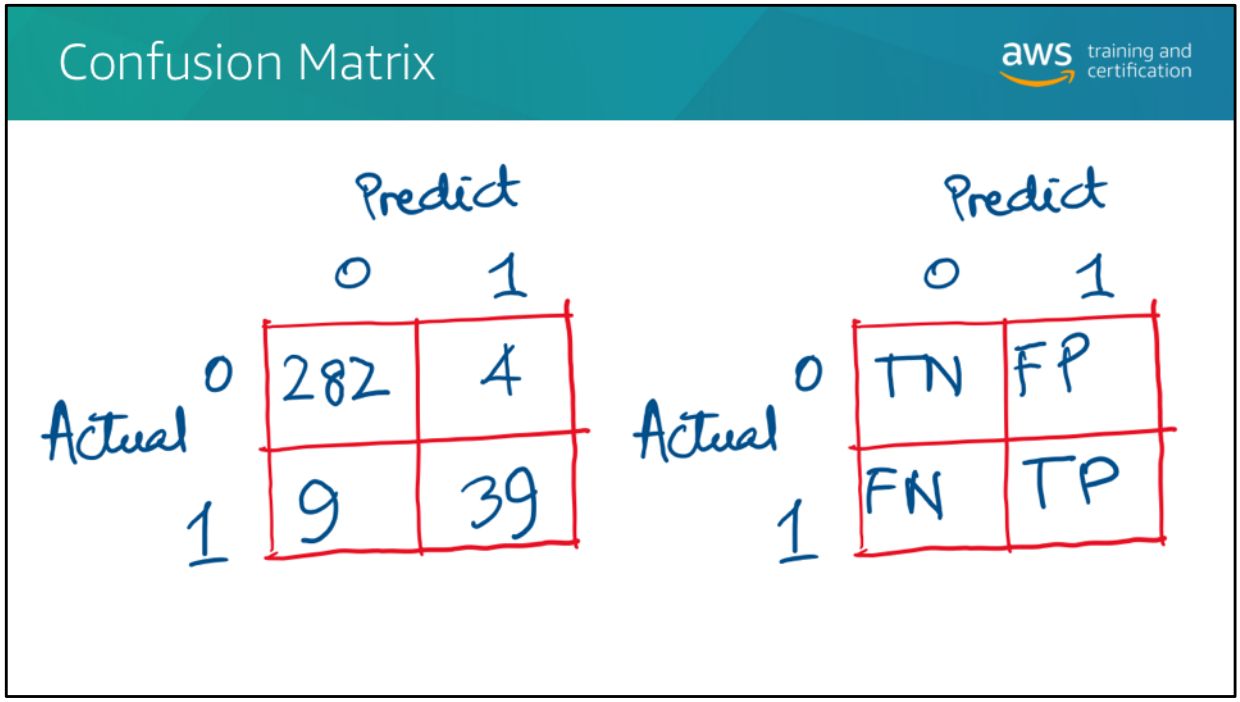
aws training and certification

- Now that we've trained the algorithm, let's create a model and deploy it to a hosted endpoint.
- We'll do this using the deploy API of Amazon SageMaker estimator.
- Once we have a hosted endpoint running, we can make real-time predictions from our model very easily, simply by making an http POST request.
  - But first, we'll need to set up serializers and deserializers for passing our test_data NumPy arrays to the model behind the endpoint.

Model Performance

aws training and certification

- There are many ways to compare the performance of a machine learning model, but let's start by simply by comparing to predicted values.
- In this case, we're simply predicting whether the customer churned (1) or not (0), which produces a simple confusion matrix



Amazon ML provides a *confusion matrix* as a way to visualize the accuracy of multiclass classification predictive models. The confusion matrix illustrates in a table the number or percentage of correct and incorrect predictions for each class by comparing an observation's predicted class and its true class.

For example, if you are trying to classify a movie into a genre, the predictive model might predict that its genre (class) is Romance. However, its true genre actually might be Thriller. When you evaluate the accuracy of a multiclass classification ML model, Amazon ML identifies these misclassifications and displays the results in the confusion matrix, as shown in the following illustration.

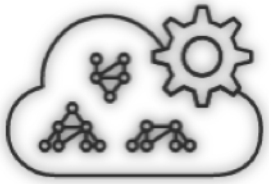
aws training and certification

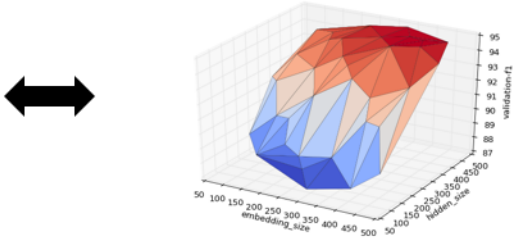
Hyperparameter Tuning

Hyperparameter Tuning

aws training and certification

- Automatic model tuning, also known as *hyperparameter tuning*, finds the best version of a model
- by running many jobs that test a range of hyperparameters on your dataset. You choose:
  - Tunable hyperparameters (a range of values)
  - The objective metric (from the metrics that the algorithm computes)





Hyperparameter Optimization

aws training and certification

Technique	Summary	Pros/Cons
Random Search	Randomly try different combinations	Economical, Not Optimal
Grid Search	Try all possible combinations	Expensive, Optimal
Bayesian Search	Use random sampling then leverage past results	Economical, Nearly Optimal

Bayesian Search Concepts

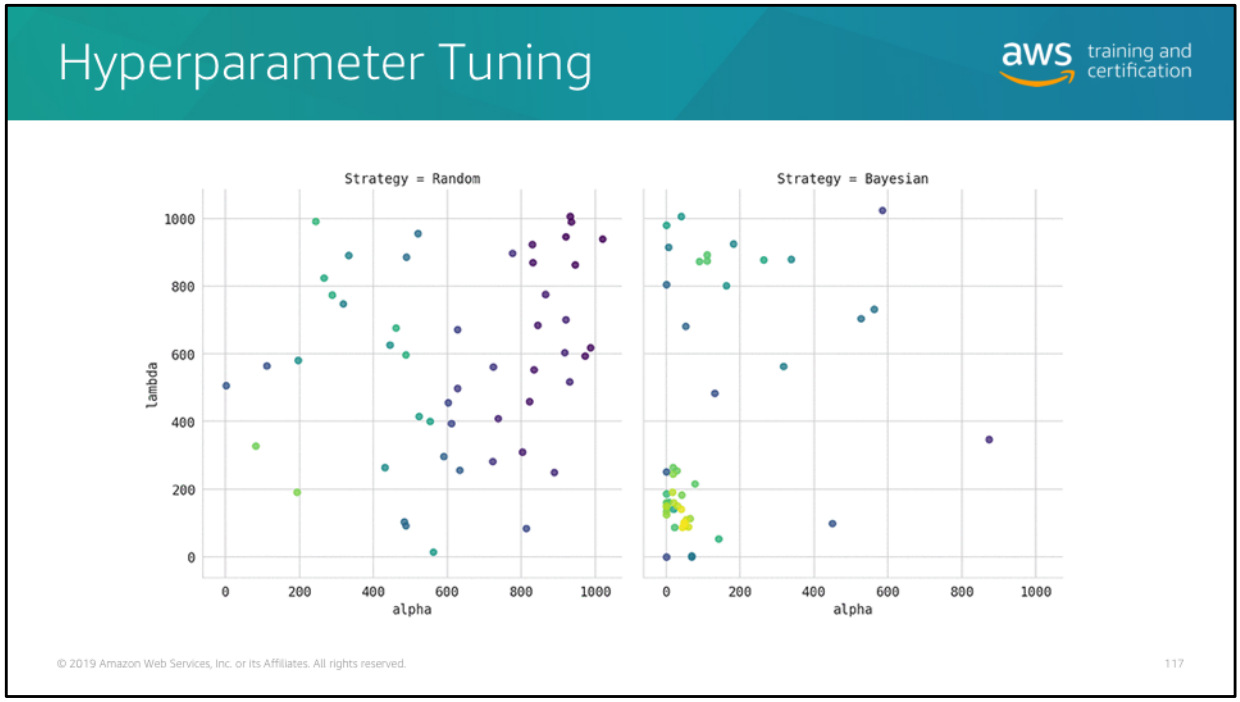
aws training and certification

Concepts	Meaning
Exploit	Choose hyperparameter values close to previous optima to improve incrementally
Explore	Choose hyperparameter values far from previous optima to find new unexplored areas
Explore/Exploit Trade-off	How to distribute resources between explore and exploit
Regression	Extrapolate error in nearby neighborhoods using linear regression

Bayesian Search Issues

aws training and certification

Question	Answer
Should I run all jobs in parallel?	Parallel jobs leverage explore, sequences of parallel jobs can leverage results from previous parallel jobs.
Does Bayesian search always find the optimal hyperparameters?	Bayesian search is a stochastic process and it is possible that it will fail to converge on the best answer.



Hyperparameter Optimization: Finding the best hyperparameters is like doing a search through a grid of all possible combinations of hyperparameters and finding the best. It is like searching for a needle in a haystack.

Grid Search: The brute force approach is to test every combination of hyperparameters. This is called Grid Search. Testing each combination of hyperparameters requires creating a model with that set of hyperparameters, training that model, then testing it to see how well it does.

Random Search: Another option is to randomly test some of the combinations of hyperparameters. This is not as expensive. But it is likely to miss the best combination of hyperparameters.

Can we do better than grid search and random search?

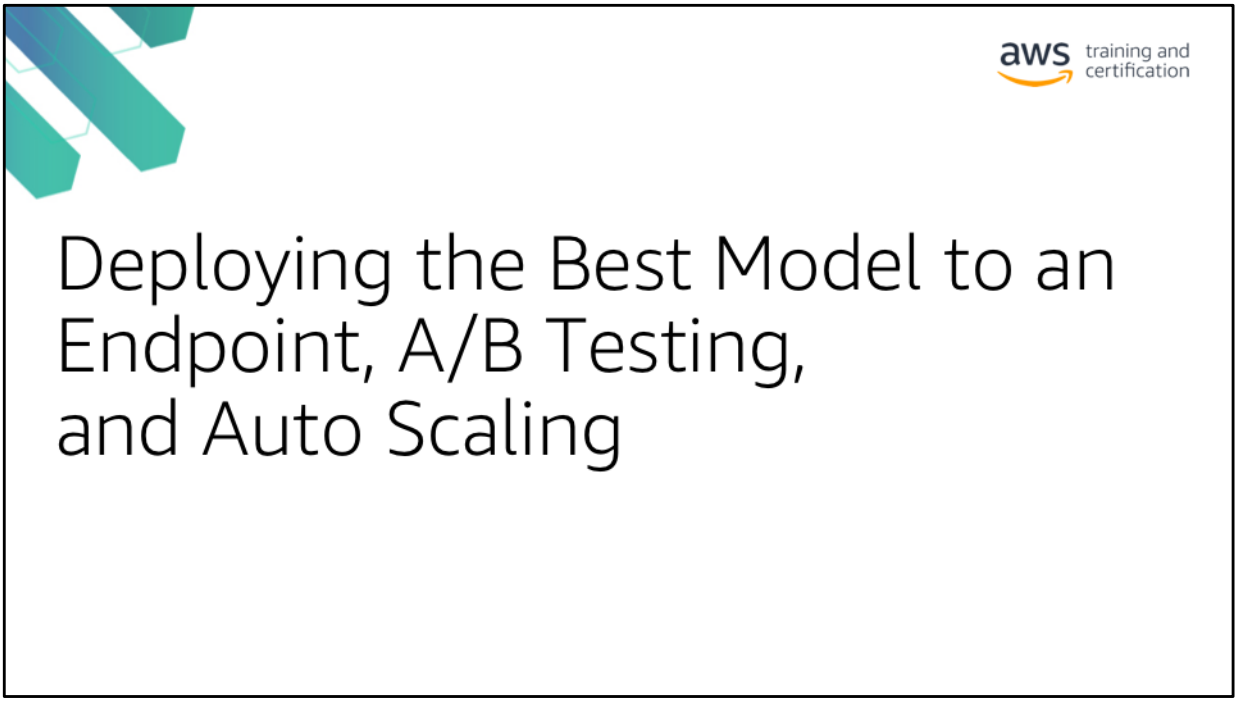
This is an active area of research. There is considerable room for improvement in the state of the art.

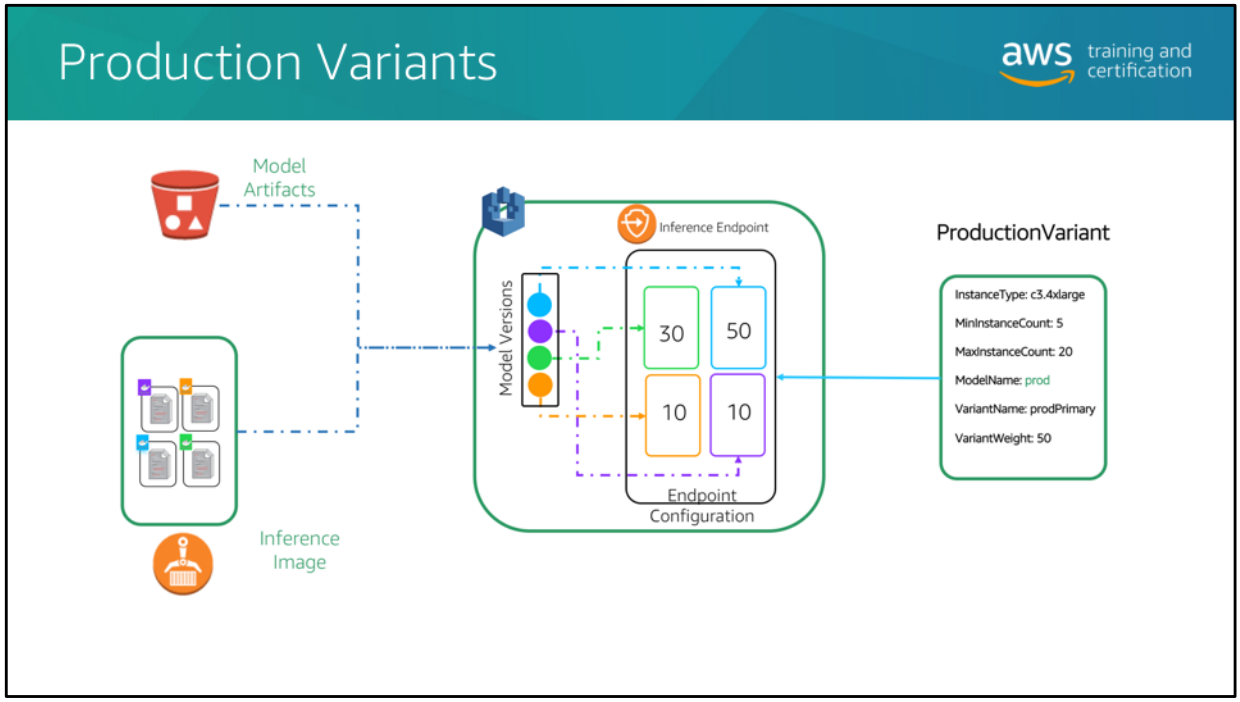
Bayesian Optimization: This is an algorithm that finds a good combination of

hyperparameters without exhaustively searching through all of them. It uses a combination of exploit vs explore heuristics to find the optimal values.

Amazon SageMaker Hyperparameter Tuning: Amazon SageMaker uses a variant of Bayesian Optimization. Amazon SageMaker Hyperparameter Tuning manages the workflow and infrastructure for the optimization for you. It offers the ability to do random search as well. This requires more trials to reach optimal values, but all the trials can be done in parallel. In Bayesian Optimization the trials are not in parallel.


aws training and certification
Demo: Model Tuning with Amazon SageMaker


aws training and certification
Deploying the Best Model to an Endpoint, A/B Testing, and Auto Scaling

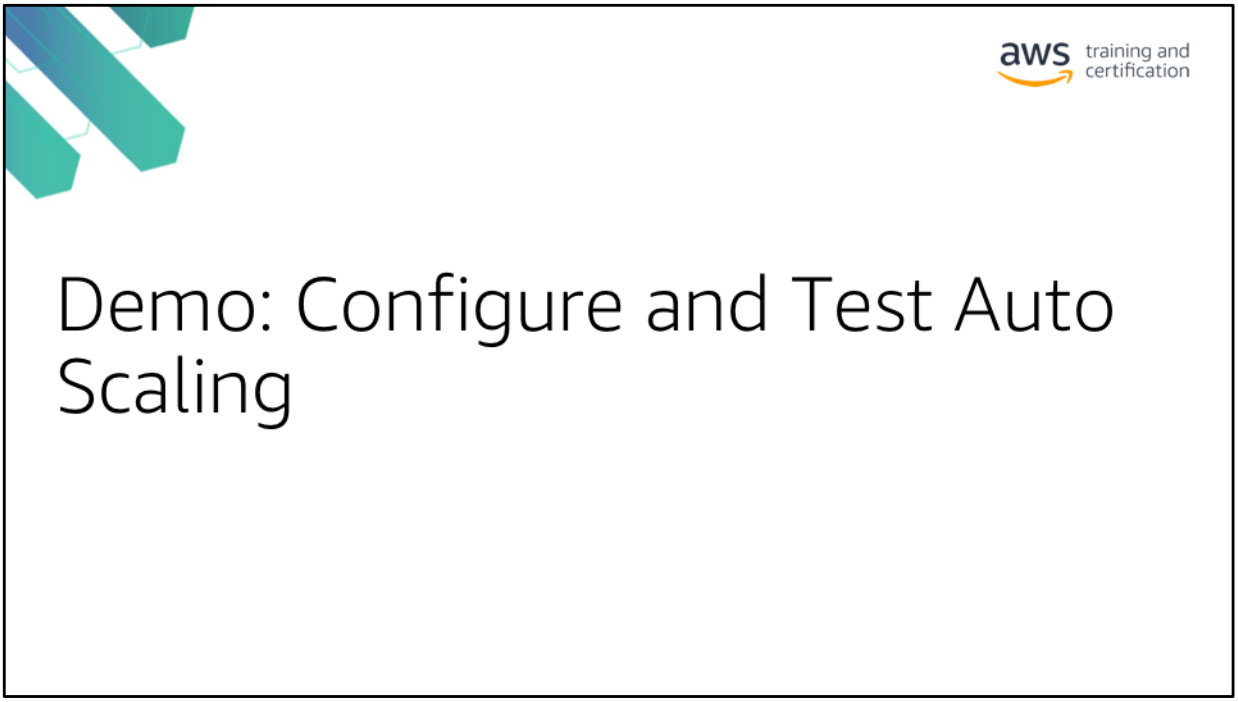


SageMaker conveniently enables data scientists to A/B test model deployments in real time. Below is a diagram from the official SageMaker documentation that outlines the model deployment process. When deploying a model, SageMaker first creates a new EC2 instance, downloads the serialized model data from S3, pulls the docker container from ECR, and deploys the runtime required to run and serve the model. Finally, it mounts a volume with the model data to the runtime.

Amazon SageMaker includes built-in A/B testing capabilities that help you test your model and experiment with different versions to achieve the best results.

After tuning the model using the hyperparameter tuning job, deploy the new model to the previously created endpoint.

Let’s examine the results of the tuning job.


aws training and certification
Demo: Configure and Test Auto Scaling

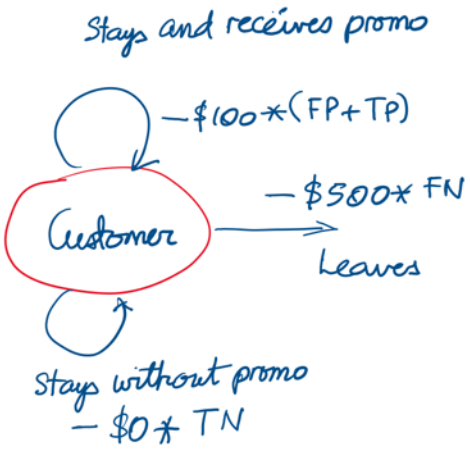
Auto Scaling

aws training and certification

- With AWS Auto Scaling for Amazon SageMaker, instead of having to closely monitor inference volume and change the endpoint configuration in response, you can configure a scaling policy to be used by AWS Auto Scaling.
- AWS Auto Scaling adjusts the number of instances up or down in response to actual workloads, determined by using Amazon CloudWatch metrics and target values defined in the policy.
- Before we put AWS Auto Scaling in place for our last endpoint, let's monitor how the endpoint behaves under load first and understand which metrics to track.


aws training and certification
The Cost of Errors

Cost of Errors



If the customer leaves we lose $500
If we give the customer a promo we lose $100
Assume we only

Relative Cost of Errors

aws training and certification

- A customer that churns is expected to cost the company more than if the company proactively tried to retain a customer who might churn.
- We should consider adjusting this cutoff in an attempt to minimize the costly false negative mistakes.
- This will increase the number of false positives, but it can also increase the number of true positives while reducing the number of false negatives.
- What are the costs for our problem of mobile operator churn?

The costs, of course, depend on the specific actions that the business takes. Let's make some assumptions here.

aws training and certification

Amazon SageMaker Use Case

DigitalGlobe

aws training and certification

- DigitalGlobe provides satellite imagery to Google Maps, Bing Maps and other online map providers.
- They had 100 PB of data and used AWS SnowMobile to transfer it into AWS.

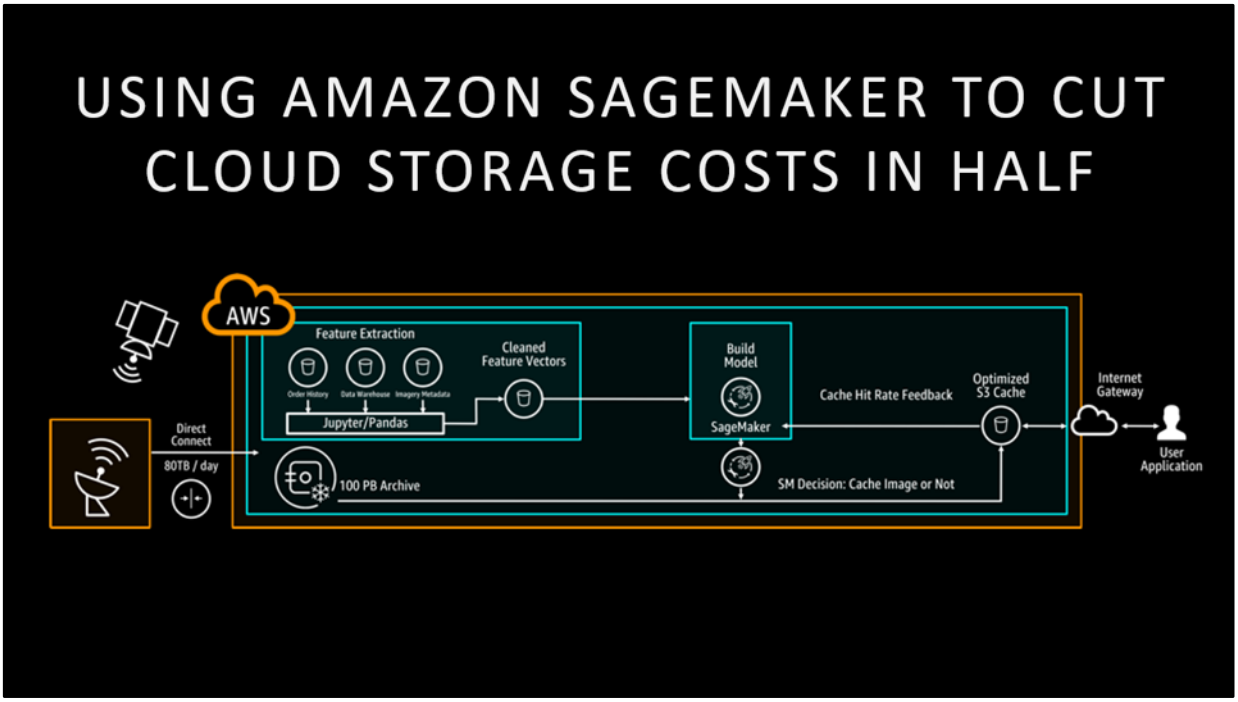
http://blog.digitalglobe.com/industry/digitalglobe-moves-to-the-cloud-with-aws-snowmobile/
http://blog.digitalglobe.com/wp-content/uploads/2018/01/Using-Amazon-SageMaker.png
https://aws.amazon.com/blogs/publicsector/how-digitalglobe-uses-amazon-sagemaker-to-manage-machine-learning-at-scale/
https://www.youtube.com/embed/mkKkSRIxU8M?start=115&end=345&version=3

DigitalGlobe

aws training and certification

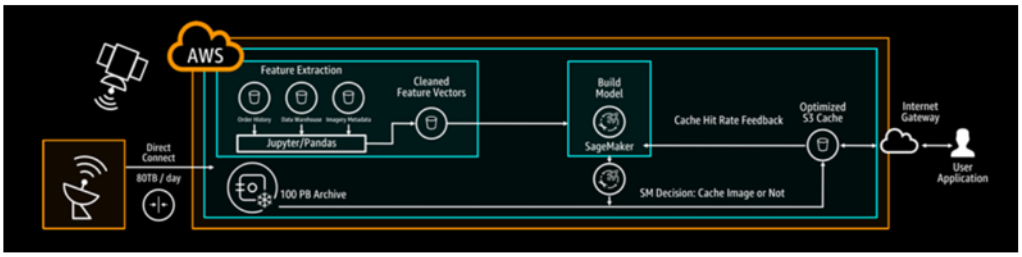
- DigitalGlobe stores its satellite image data in Glacier and uses S3 as a cache.
- DigitalGlobe uses SageMaker to decide what images to precache to optimize access.
- Using SageMaker they were able to improve their hit rate from 40% to 83% trending to 90% in one week.

http://blog.digitalglobe.com/industry/digitalglobe-moves-to-the-cloud-with-aws-snowmobile/
http://blog.digitalglobe.com/wp-content/uploads/2018/01/Using-Amazon-SageMaker.png
https://aws.amazon.com/blogs/publicsector/how-digitalglobe-uses-amazon-sagemaker-to-manage-machine-learning-at-scale/
https://www.youtube.com/embed/mkKkSRIxU8M?start=115&end=345&version=3



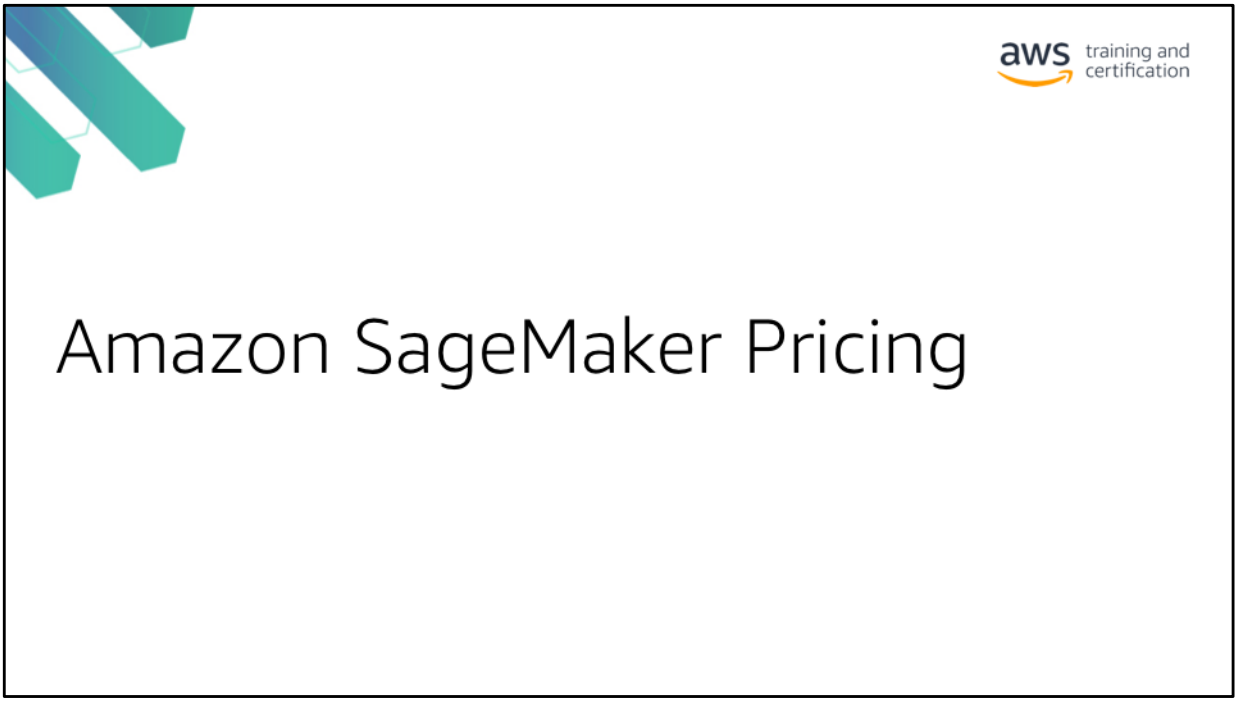
http://blog.digitalglobe.com/industry/digitalglobe-moves-to-the-cloud-with-aws-snowmobile/
http://blog.digitalglobe.com/wp-content/uploads/2018/01/Using-Amazon-SageMaker.png
https://aws.amazon.com/blogs/publicsector/how-digitalglobe-uses-amazon-sagemaker-to-manage-machine-learning-at-scale/
https://www.youtube.com/embed/mkKkSRIxU8M?start=115&end=345&version=3

DigitalGlobe's SageMaker Use Case



Using SageMaker they were able to improve their hit rate from 40% to 83% trending to 90% in one week.

http://blog.digitalglobe.com/industry/digitalglobe-moves-to-the-cloud-with-aws-snowmobile/
http://blog.digitalglobe.com/wp-content/uploads/2018/01/Using-Amazon-SageMaker.png
https://aws.amazon.com/blogs/publicsector/how-digitalglobe-uses-amazon-sagemaker-to-manage-machine-learning-at-scale/
https://www.youtube.com/embed/mkKkSRIxU8M?start=115&end=345&version=3


aws training and certification
Amazon SageMaker Pricing

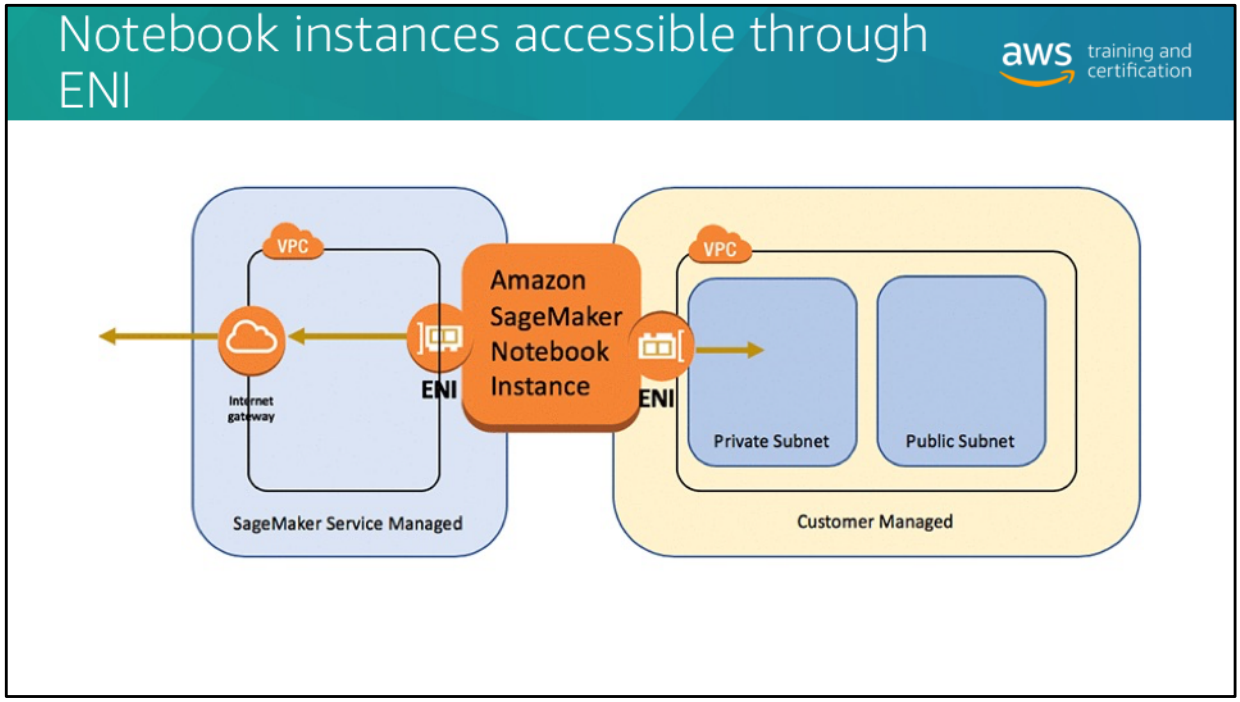
Pricing

aws training and certification

- For notebook, training, and deployment instance prices see https://aws.amazon.com/sagemaker/pricing/
- Training and batch transform instances terminate automatically.
- Stop notebook instances when not in use.
- Terminate or use auto-scaling on real-time endpoint instances.
- SageMaker has a free tier https://aws.amazon.com/free

aws training and certification

Amazon SageMaker Networking



Amazon SageMaker notebook instances can be launched with or without your Virtual Private Cloud (VPC) attached. When launched with your VPC attached, the notebook can either be configured with or without direct internet access.

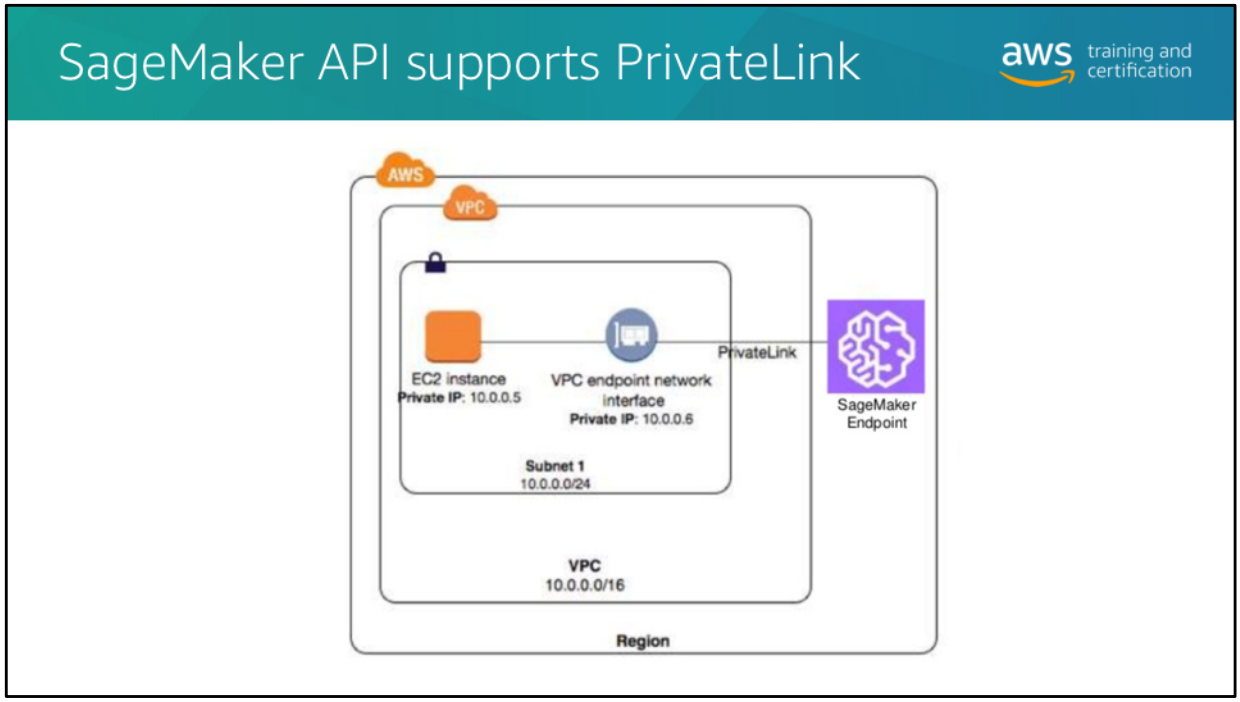
IMPORTANT NOTE: Direct internet access means that the Amazon SageMaker service is providing a network interface that allows for the notebook to talk to the internet through a VPC managed by the service.

Using the Amazon SageMaker console, these are the three options:
1. No customer VPC is attached.
2. Customer VPC is attached with direct internet access.
3. Customer VPC is attached without direct internet access.

Amazon SageMaker notebook instances support Amazon Virtual Private Cloud (Amazon VPC) interface endpoints that are powered by AWS PrivateLink. Each VPC endpoint is represented by one or more Elastic Network Interfaces (ENIs) with private IP addresses in your VPC subnets.

https://aws.amazon.com/blogs/machine-learning/understanding-amazon-

sagemaker-notebook-instance-networking-configurations-and-advanced-routing-options/



All Amazon SageMaker APIs are now fully supported on AWS PrivateLink, which increases the security of data shared with cloud-based applications by reducing the exposure of data to the public Internet. All communication between applications and Amazon SageMaker can be secured inside a Virtual Private Cloud (VPC).

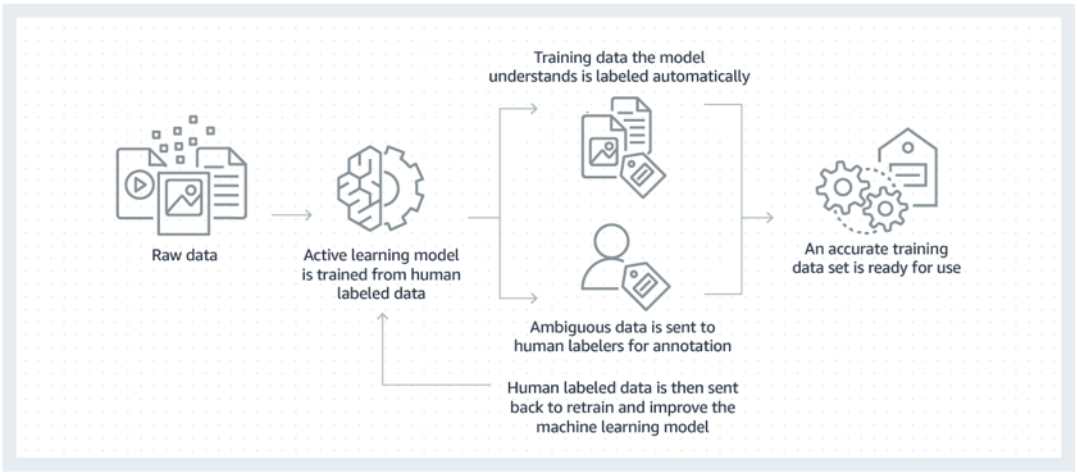
Earlier this year, we announced the support for prediction calls to machine learning models that are hosted on Amazon SageMaker to be secured using AWS PrivateLink. With this new enhancement, every SageMaker API call can be called through an interface endpoint within the VPC. Since all communication is inside the VPC, there is no need for an Internet Gateway, a NAT device, a VPN connection, or AWS Direct Connect.

Amazon SageMaker support on AWS PrivateLink is now available in the US East (N. Virginia), US East (Ohio), US West (Oregon), Europe (Ireland), Europe (Frankfurt), Asia Pacific (Tokyo), Asia Pacific (Seoul), and Asia Pacific (Sydney) AWS Regions. For more information, visit the documentation here.


aws training and certification
Amazon SageMaker Features

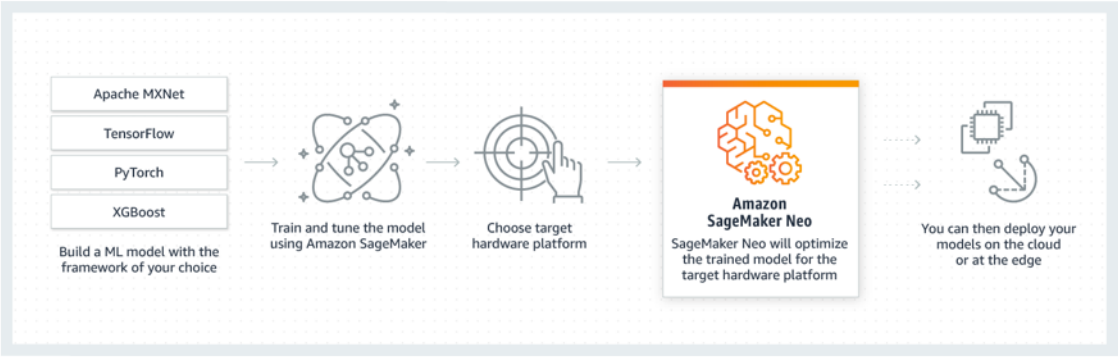
Amazon SageMaker Ground Truth

aws training and certification



Amazon SageMaker Neo

aws training and certification



On day two we will cover an end-to-end example of distributed image classification algorithm in transfer learning mode and the use of Amazon SageMaker's implementation of the XGBoost algorithm to train and host a multiclass classification model.

In the first half, we will use the Amazon sagemaker image classification algorithm in transfer learning mode to fine-tune a pre-trained model (trained on imagenet data) to learn to classify a new dataset. In particular, the pre-trained model will be fine-tuned using caltech-256 dataset.

In the second half, we will use the MNIST dataset for training. It has a training set of 60,000 examples and a test set of 10,000 examples. To illustrate the use of libsvm training data format, we download the dataset and convert it to the libsvm format before training.

Scripps Networks

aws training and certification

Business Model	Use Case
Licenses image and video content for television, internet and other platforms	Use Amazon Rekognition to automatically tag images and video making them searchable

C-SPAN

aws training and certification

Business Model	Use Case
Broadcasts unedited proceedings of the U.S. House of Representatives and the U.S. Senate	Uses Amazon Rekognition to index content by speaker

Daniel Wellington

aws training and certification

Business Model	Use Case
European watch and jewelry manufacturer	Uses Amazon Rekognition to process product returns

K-STAR

aws training and certification

Business Model	Use Case
Sells concert tickets	Uses "Face Ticket" based on Amazon Rekognition to verify tickets instead of paper tickets which has eliminated concert lines

FamilySearch

aws training and certification

Business Model	Use Case
Uses geneological database to connect families across generations	Uses Amazon Rekognition to find relative you most resemble based on family photographs

ARMED Data Fusion System

aws training and certification

Business Model	Use Case
Private intelligence contractor secures large-scale events	Tracks individuals and persons of interest across video feeds in real time

CampSite

aws training and certification

Business Model	Use Case
Software for summer camps	Informs parents when a picture with their kid has been uploaded to the camp photo database

aws training and certification

What are image recognition use cases in your industry?

aws training and certification

What are image recognition options?

Image Recognition Options

aws training and certification

Option	Challenge	Training Cost	Image Type
Rekognition	Easy	N/A	Generic
Build Deep Learning Model	Difficult	$$$	Domain-Specific
Transfer Learning with Existing Model	Moderate	$	Domain-Specific
Humans	Not Scalable	Low	Domain-Specific

aws training and certification

Amazon Rekognition Features

Rekognition Features 1/2

aws training and certification

Feature	Description
Object and Scene Detection	Identifies objects such as vehicles, pets, furniture, and scenes such as sunset, beach.
Facial Recognition	Finds similar faces in large image collections. Enables indexing faces from images.
Facial Analysis	Locates faces. Analyzes attributes, such as face is smiling, eyes are open.

Rekognition Features 2/2

aws training and certification

Feature	Description
Face Comparison	Provides similarity score that measures likelihood that faces in two images are same person.
Unsafe Image Detection	Detects explicit and suggestive content for filtering.
Celebrity Recognition	Detects and recognizes celebrities.
Text In Image	Locates and extract text in images, for example, road signs, license plates, text on t-shirts, mugs, captions on screen. Returns text label, rectangular frame, confidence score.

Transfer Learning

aws training and certification

- Data used in training an algorithm must be properly chosen to be representative
- But suppose we want to apply a algorithm to a new or shifting domain? Retrain!
- Can we somehow use our existing trained algorithm or classifier as a starting point to give us a shortcut?
- This is **Transfer Learning**.

Data used in training a classifier must be properly chosen to be representative
If not? Accuracy will be worse than expected
But suppose we want to apply a classifier to a new or shifting domain? Retrain!
But that's expensive.
Can we somehow use our existing classifier as a starting point to give us a shortcut?
This is **Transfer Learning**.

https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14
What is a classifier: "An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm, that maps input data to a category."

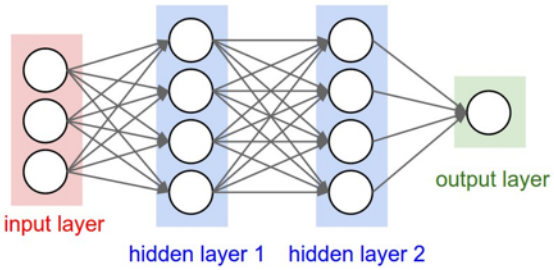
aws training and certification

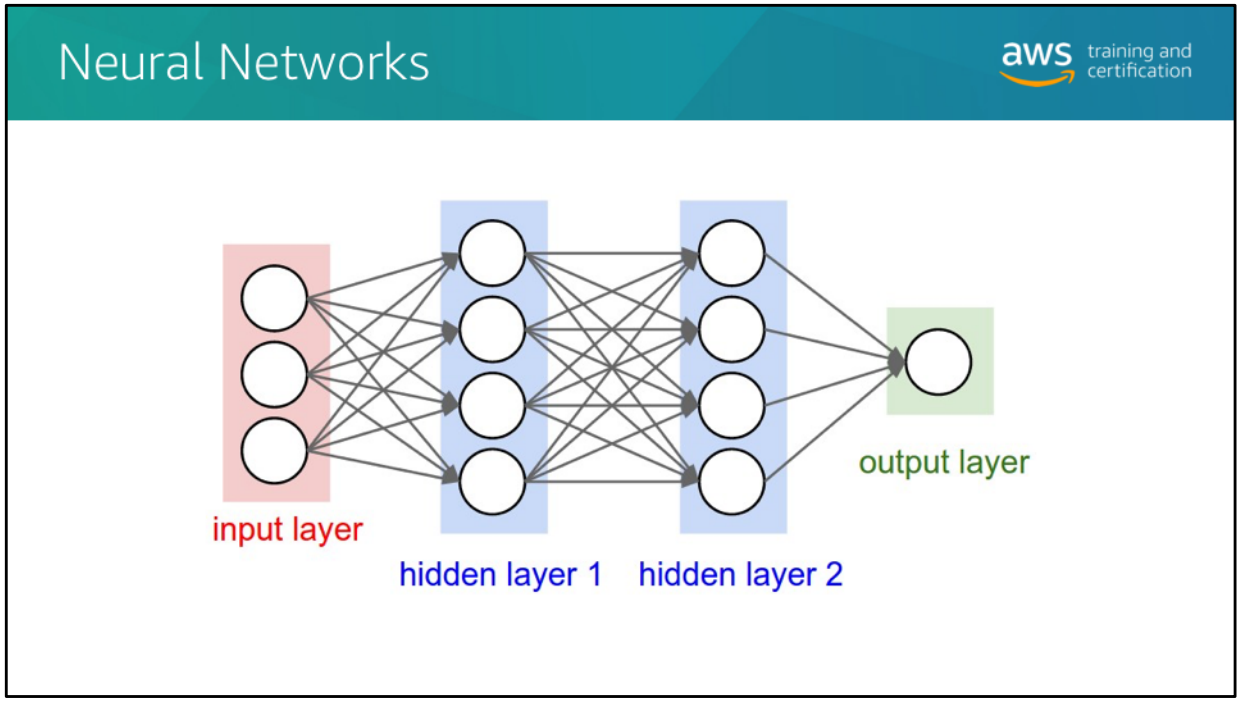
Neural Networks

Image Recognition

aws training and certification

- Modern image recognition systems are built using neural network architectures.
- They use a specialized neural network architecture called the Convolutional Neural Network.
- First let's talk about neural networks. Then we will talk about convolutional neural networks.



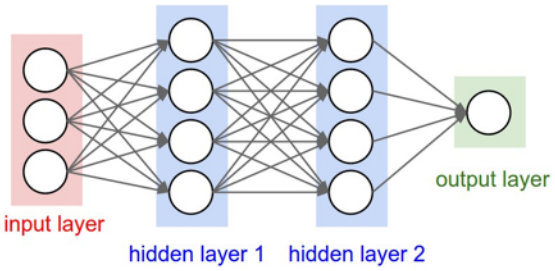


The fundamental algorithm underlying modern image recognition systems is the neural network.
The neural network has an input layer, multiple hidden layers, and an output layer.
The neurons are connected to each other through weights.
Weights indicate how one neuron depends on another in a previous layer for its value.
The training process consists of finetuning the weights so that given inputs create the right output.

Neural Networks

aws training and certification

Neuron Type	Meaning
Neurons	Units of logic
Input Neurons	Inputs to network
Hidden Neurons	Intermediate classification outputs
Output Neurons	Final classification outputs
Weight between two neurons	Influence one neuron has on another

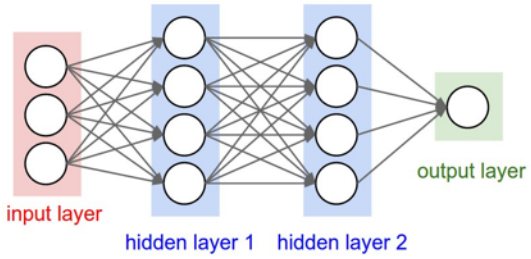


These are the types of objects that make up a neural network

Neural Network Training vs Inference

aws training and certification

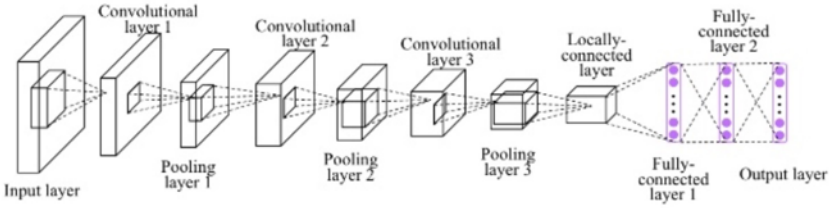
Algorithm	Stage	Description
Forward propagation	Training and Inference	From inputs calculate outputs
Back propagation	Training	Alters weights to reduce error



There are two operations the neural network supports.

- Forward propagation: This is the inference operation. Given a set of inputs the network produces a final output.
- Back propagation: This is the weight adjustment operation in which the error at the final layer is propagated back.
- Training consists of: (1) forward propagating data point, (2) calculating error, (3) back propagating it to adjust the weights. Then repeating this process many times.

Convolutional Neural Networks



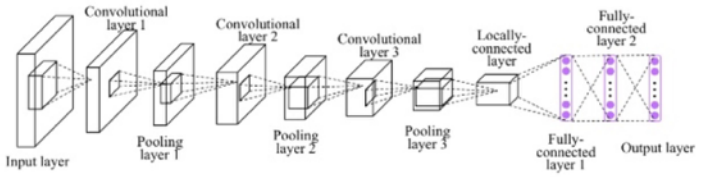
Convolutional Neural Networks are specialized for image detection. The initial part of the network is not fully connected. Instead if is a sandwich of multiple convolutional layers and pooling layers stacked together.

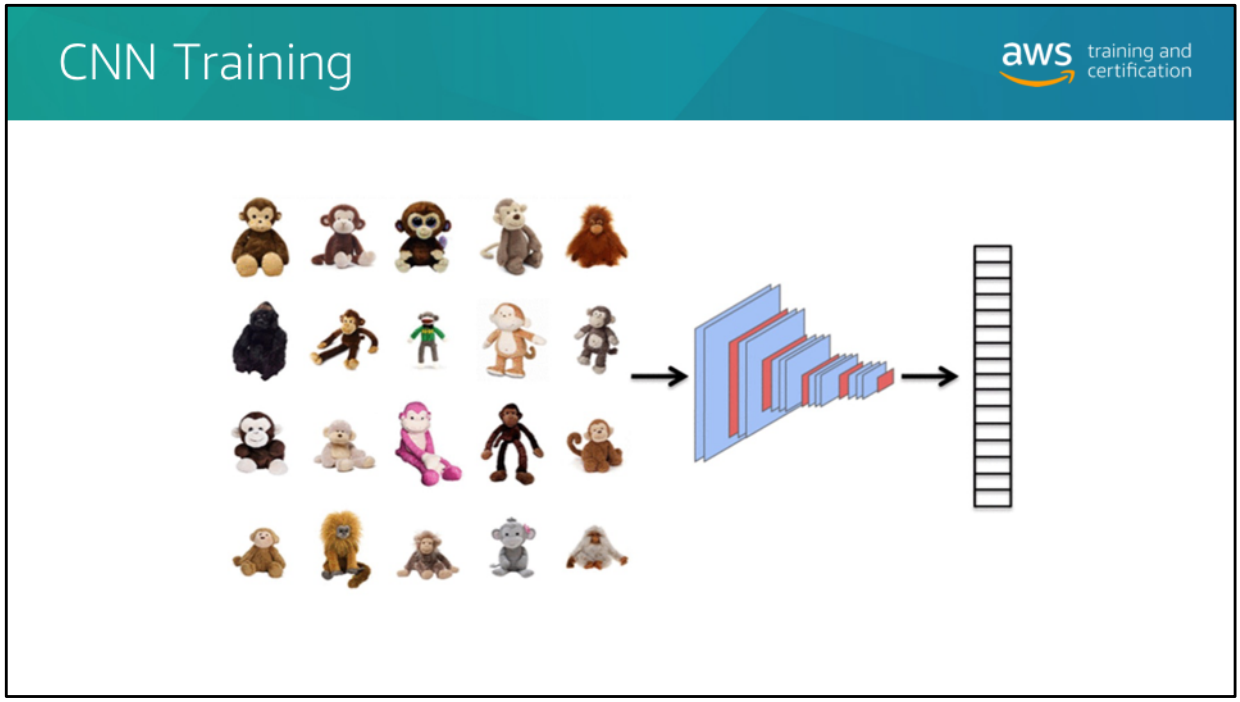
- Convolutional layers: Designed to look for the presence of specific features. What features? The features to look for are selected using back propagation to improve the classification results of the network. The features are encoded in kernels. Back propagation adjusts these kernels iteratively to reduce the error in the results.
- Pooling layers: Reduce resolution of the lower layer and make the images blurry. This reduces the computation required to process the images. It is easier to see objects in blurry images than in crisp images.
- Fully connected layer: Final fully-connected layer is a classifier. Looks at presence and absence of lower-level features to figure out the category of the image.

Convolutional Neural Network

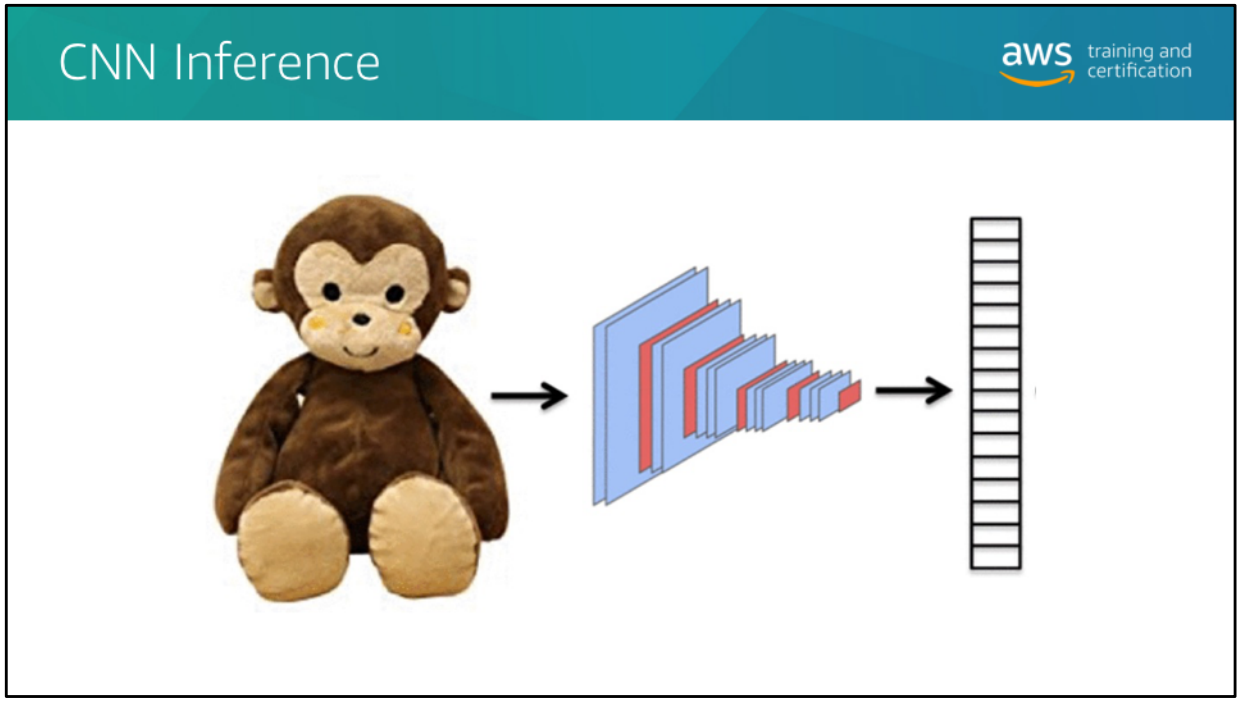
aws training and certification

Layer	Purpose
Convolutional layers	Look for specific features selected using back propagation to improve classification results of network.
Pooling layers	Reduce resolution of lower layer to make images blurry. Easier to see objects in blurry images than in crisp images.
Fully connected layer	Classifier looks at presence/absence of lower-level features to figure out category of image.





Suppose we train our network to recognize objects such as stuffed monkeys. It will learn to extract common patterns across all these images and classify them all as stuffed toy monkeys.

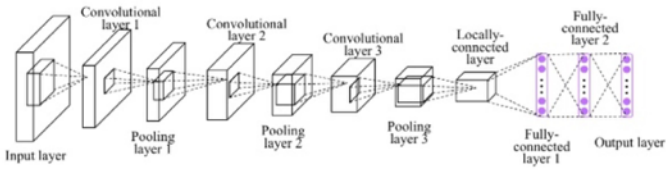


When it sees a new object it has not seen before it will classify that as a monkey.

Transfer Learning

aws training and certification

Layer	Weights	Purpose
Convolutional + Pooling Layers	Fix using pre-training	Detects low-level image features, and combinations of these low-level features
Fully-Connected Layers	Randomize	Detects domain-specific images

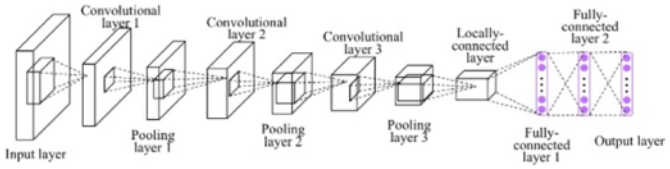


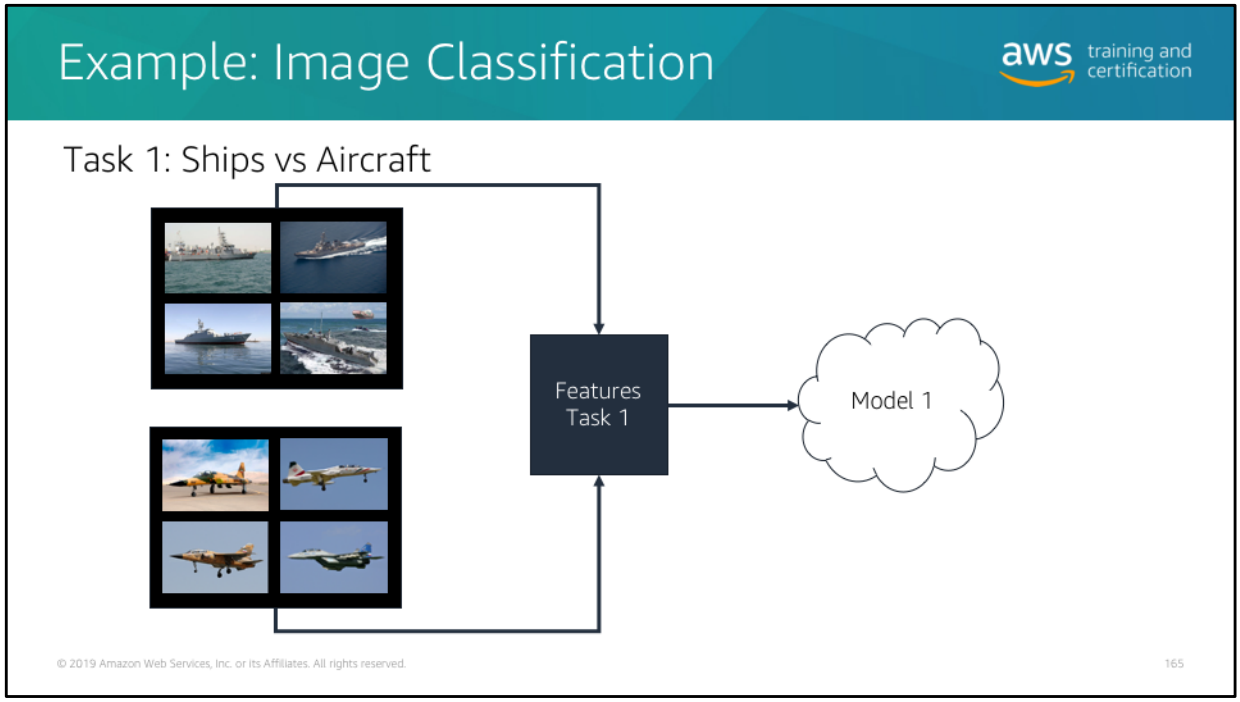
The basic idea of transfer learning is that we fix the kernels and the weights in the convolutional layers. We randomize the weights in the fully connected layer. We train the classifier to recognize our new categories using the old features and the previous neural network was trained to detect.

Transfer Learning Benefits

aws training and certification

Benefit	Details
Customized	Domain specific images
Economical	Saves on training for low-level and more abstract features
Robust	Uses industry-standard lower layers





What to transfer
- Instances?
- Model?
- Features?

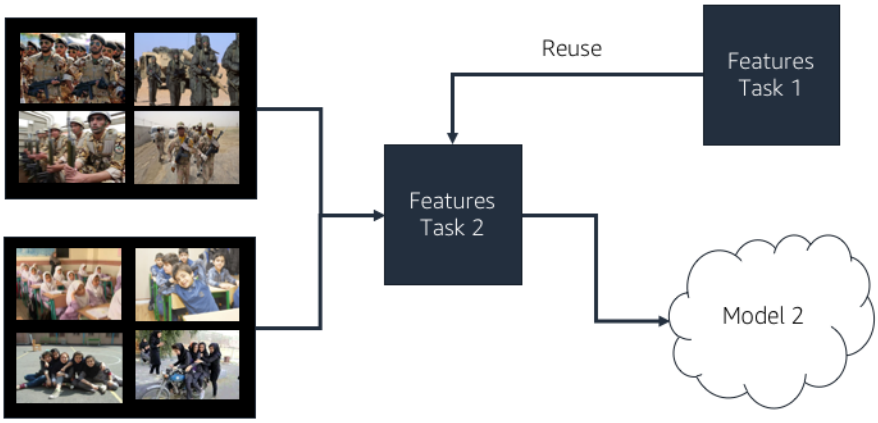
How to transfer
- Weight instances
- Unify features
- Map model

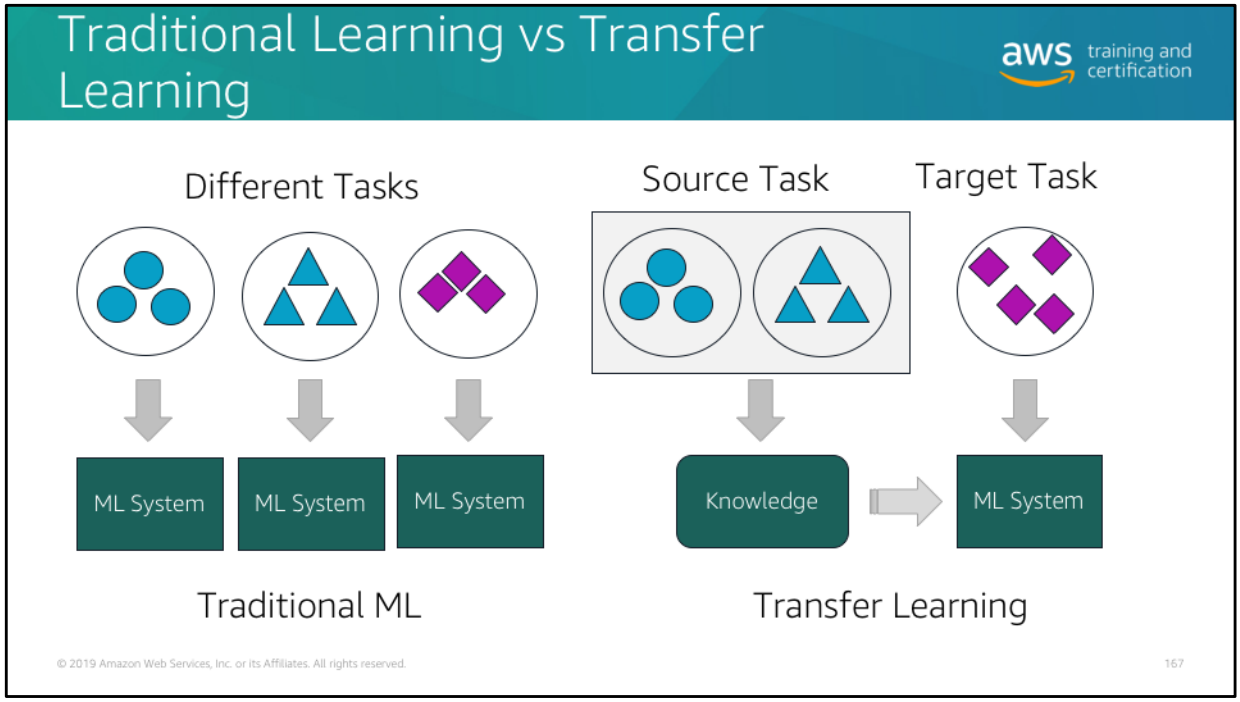
When to transfer In which situations?
- Faster to transfer or to retrain?

Example: Image Classification

aws training and certification

Task 2: Infantry vs Children





Given a source domain and source learning task, a target domain and a target learning task, transfer learning aims to help improve the learning of the target predictive function using the source knowledge, where

- Source domain does not equal target domain
- Source task does not equal target task

Hyperparameters

aws training and certification

Hyperparameter	Meaning
image_shape	Pixel dimensions of images
num_layers	Number of layers in neural network, e.g. 18, 50, 152
num_training_samples	Total number of training samples
num_classes	Number of output classes, e.g. 257 (CalTech), 1000 (Imagenet)
mini_batch_size	Number of training samples in each mini batch, after which error is calculated
epochs	Number of passes through data
learning_rate	How quickly to adjust weights
use_pretrained_model	Use transfer learning vs build new model

Machine Learning Stages

aws training and certification

Stage	Details
Stage Data	Stage training and validation data in S3
Train	Build model
Batch Transform	Perform batch inference
Host Endpoint	Perform real-time inference
Clean up	Delete endpoint (important)

This notebook demonstrates the use of Amazon SageMaker's implementation of the XGBoost algorithm to train and host a multiclass classification model. The MNIST dataset is used for training. It has a training set of 60,000 examples and a test set of 10,000 examples. To illustrate the use of libsvm training data format, we download the dataset and convert it to the libsvm format before training.
To get started, we need to set up the environment with a few prerequisites for permissions and configurations.

What is MNIST?

aws training and certification

- Large database of handwritten digits used as training images in machine learning.
- Created from NIST's original datasets.
- NIST is National Institute of Standards and Technology.

MNIST

What type of classification is this?

aws training and certification

Type	Output
Regression	Single numeric value
Binary Classification	Single binary value
Multi-Class Classification	Multiple classes

Machine Learning Approach

aws training and certification

Feature	Details
Input	MNIST dataset
Output	10 classes: 0, 1, …, 9
Algorithm	XGBoost

XGBoost Concepts

aws training and certification

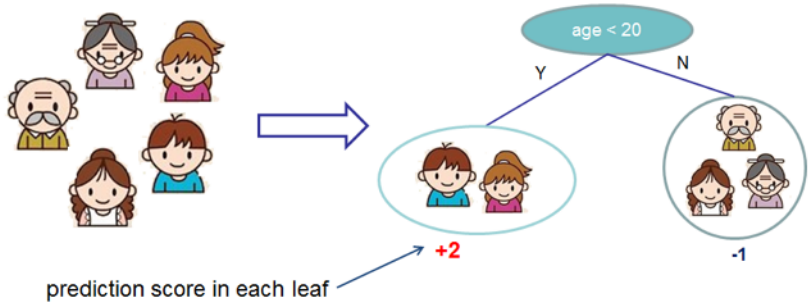
Concept	Details
XGBoost	Fast gradient boosting algorithm
Ensemble	Uses ensemble of simple models (decision trees)
Training	Models are added sequentially and compensate for errors of previous models
CART	Classification and Regression Trees

XGBoost CART

aws training and certification

Input: age, gender, occupation, …

Like the computer game X



prediction score in each leaf

XGBoost Input Formats

aws training and certification

ContentType	Format	Features
text/libsvm	LibSVM	Compact, good for sparse data
text/csv	CSV	Human-readable

Hyperparameters

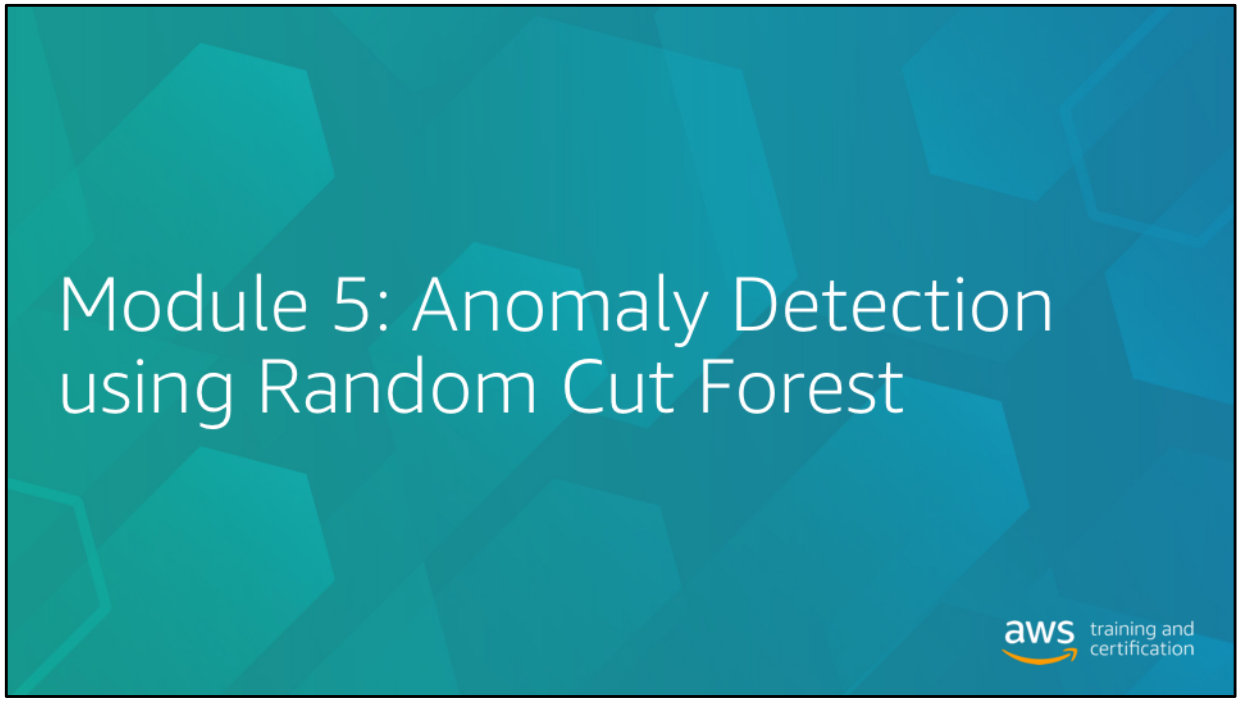
aws training and certification

Hyperparameter	Meaning
max_depth	Max depth of trees in algorithm. Deeper trees can lead to better fit, but can overfit.
eta	Step size shrinkage used in updates to prevent overfitting. Eta shrinks feature weights to make boosting process more conservative.
gamma	Minimum loss reduction required to make further partition on leaf node of tree. Larger values make algorithm more conservative.
min_child_weight	If tree partition step results in leaf node with sum of instance weight less than min_child_weight, building process stops partitioning.

Evaluating Performance

aws training and certification

Measure	Details
Error Rate	Incorrect predictions divided by total predictions
Confusion Matrix	More precise measurement of error per class



This notebook demonstrates the use of Amazon SageMaker's implementation of the XGBoost algorithm to train and host a multiclass classification model. The MNIST dataset is used for training. It has a training set of 60,000 examples and a test set of 10,000 examples. To illustrate the use of libsvm training data format, we download the dataset and convert it to the libsvm format before training.
To get started, we need to set up the environment with a few prerequisites for permissions and configurations.

Random Cut Forest

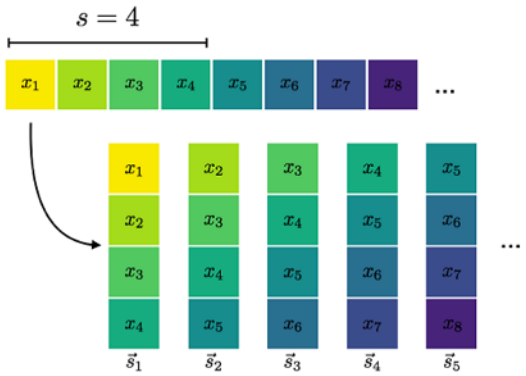
aws training and certification

- Identifies anomalies in data.
- Gives each data point an "anomaly score".
- Can be used for time series data.
- Useful for identifying: potential fraud, cyber attacks, service outages, unexpected scenarios.

Random Cut Forest

aws training and certification

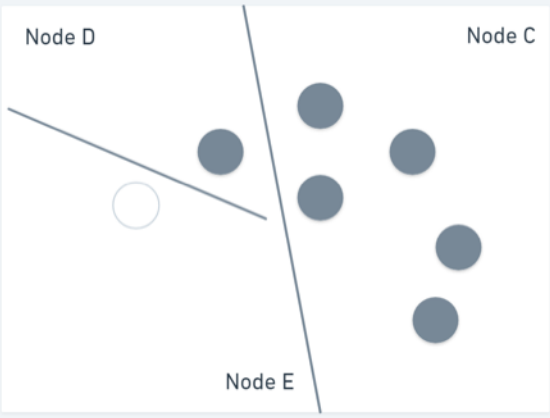
- Shingles are subsequences of length s which help detect breaks in periodicity.
- Small shingles lead to higher sensitivity to small fluctuations.
- Large shingles can miss small scale anomalies.
- Align shingle size with anticipated periodicity in data. E.g. 48 hours for taxi data.



Random Cut Forest

What is the basic idea of Random Cut Forest?

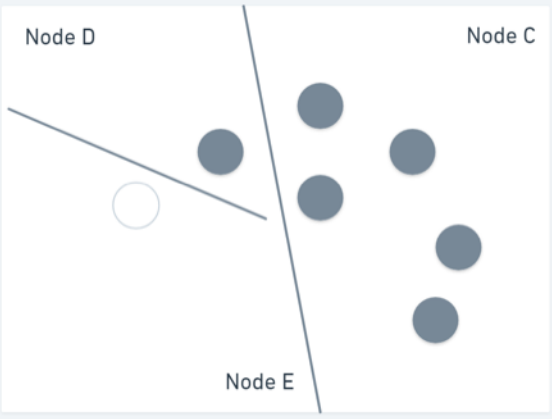
- Insert a random line (cut) to divide points into two groups.
- For each point track how many random cuts isolate it.
- Anomalies will be isolated with a few random cuts.
- Each set of cuts represents a tree.



Random Cut Forest

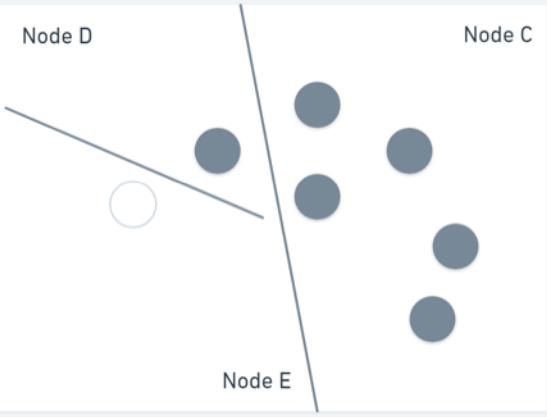
aws training and certification

Hyperparameter	Meaning
num_samples_per_tree	Number of random samples given to each tree from training data set.
num_trees	Number of trees in forest.



Random Cut Forest

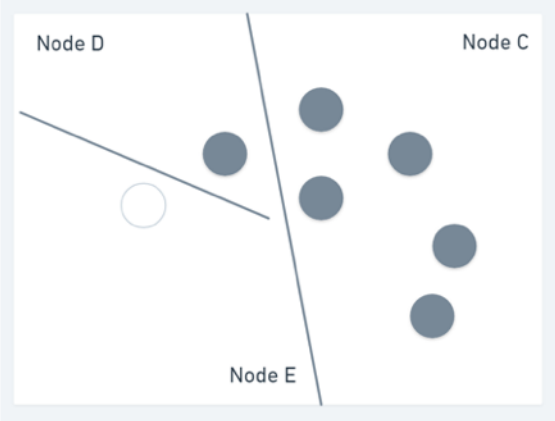
Hyperparameter	Meaning
num_samples_per_tree	Number of random samples given to each tree from training data set.
num_trees	Number of trees in forest.



Random Cut Forest Algorithm

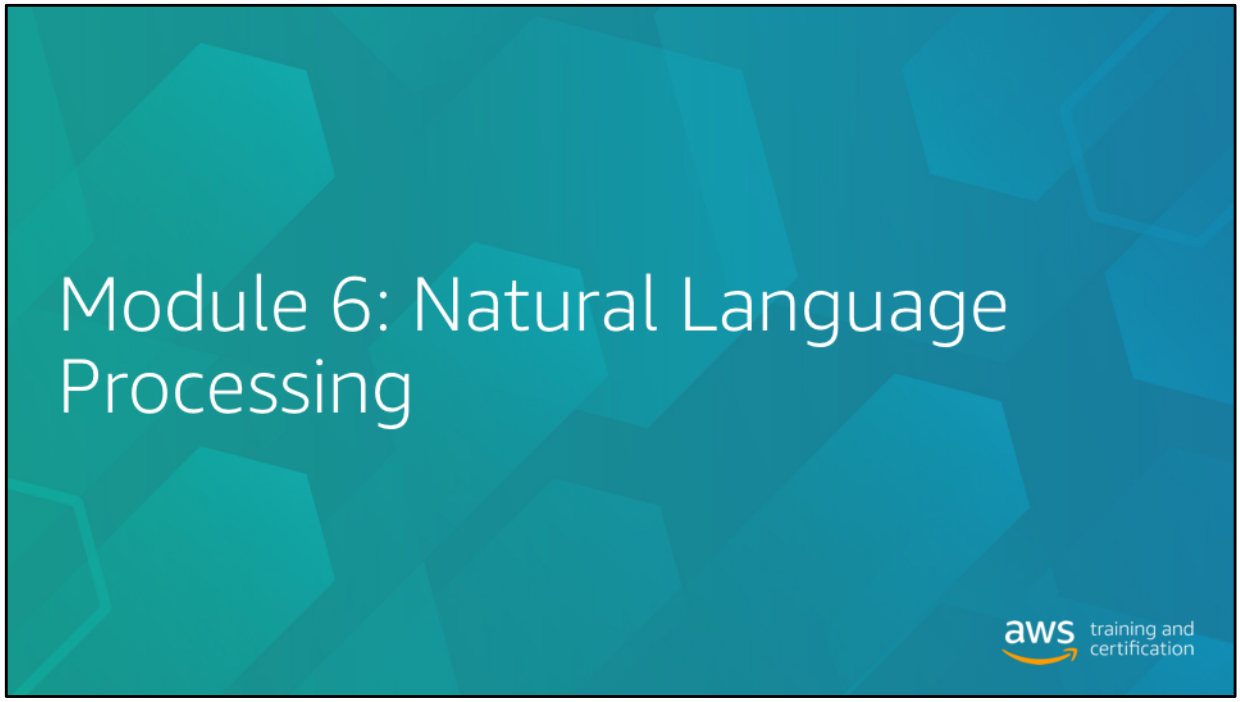
aws training and certification

- Each tree trains on sample of input training data.
- Data point's anomaly score in tree inversely proportional to depth in tree.
- Data point's anomaly score in model is average across trees.
- Set `num_samples_per_tree` so its inverse approximates expected percentage of anomalies in dataset.



aws training and certification

Lab: Random Cut Forest



Amazon SageMaker Neural Topic Model (NTM) is an unsupervised learning algorithm that attempts to describe a set of observations as a mixture of distinct categories. NTM is most commonly used to discover a user-specified number of topics shared by documents within a text corpus. Here each observation is a document, the features are the presence (or occurrence count) of each word, and the categories are the topics. Since the method is unsupervised, the topics are not specified upfront and are not guaranteed to align with how a human may naturally categorize documents. The topics are learned as a probability distribution over the words that occur in each document. Each document, in turn, is described as a mixture of topics.

If you would like to know more please check out the SageMaker Neural Topic Model Documentation (https://docs.aws.amazon.com/sagemaker/latest/dg/ntm.html).

aws training and certification

Amazon Comprehend

Lexis/Nexis

aws training and certification

Business Model	Use Case
Provides legal content and news	Uses Amazon Comprehend for entity recognition to identify judges and attorneys in 200 million documents.

Fred Hutchinson Cancer Research

aws training and certification

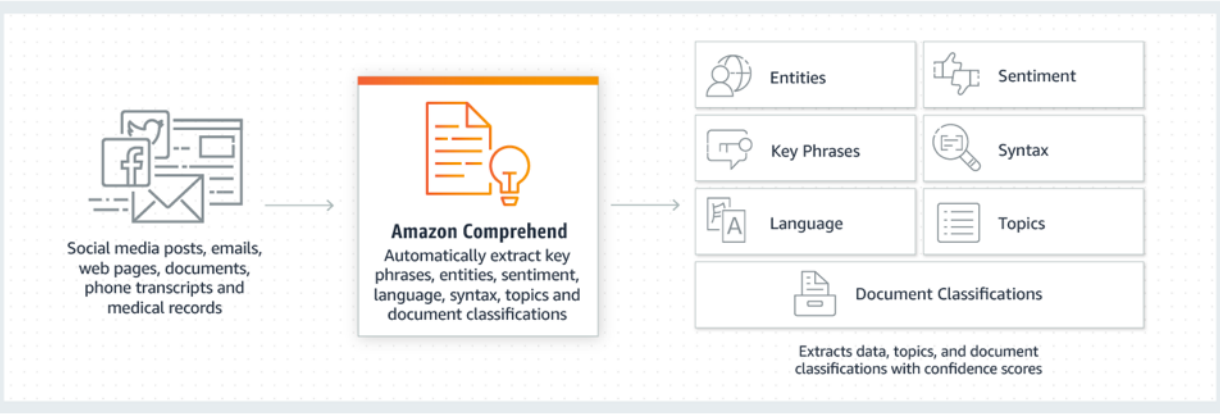
Business Model	Use Case
Cancer research	Uses Amazon Comprehend Medical to analyze unstructured clinical record data to match patients with clinical trials.

FINRA

aws training and certification

Business Model	Use Case
Regulates brokerage industry	Uses Amazon Comprehend to extract individuals and organization, match extracted entities to FINRA records, flag individual of interest, and detect similarities with other documents.

How Comprehend Works



Comprehend Features

aws training and certification

Feature	Details
Keyphrase Extraction	Extracts key phrases with confidence score.
Sentiment Analysis	Determines if text sentiment is positive, negative or neutral.
Syntax Analysis	Identifies parts of speech such as nouns, verbs, adjectives.

Comprehend Features

aws training and certification

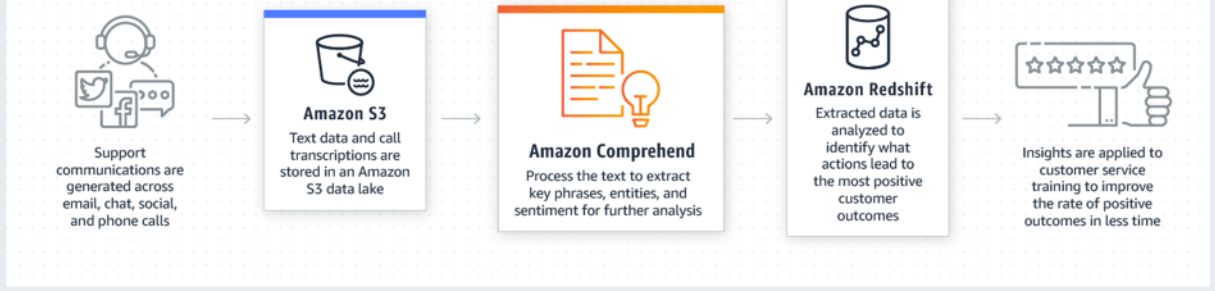
Feature	Details
Entity Recognition	Identifies entities such as organization, person, city.
Medical Named Entity and Relationship Extraction	Extracts medical information such as medication, medical condition, test, treatment, and procedures (TTP), anatomy, Protected Health Information (PHI).
Custom Entities	Identify domain specific terms such as policy numbers by learning from small index of examples.

Comprehend Features

aws training and certification

Feature	Details
Language Detection	Detects language.
Custom Classification	Classifies documents into categories using text examples for each label you want to use.
Topic Modeling	Identifies relevant terms or topics from collection of documents. Identifes most common topics in collection and organizes them in groups by topic.

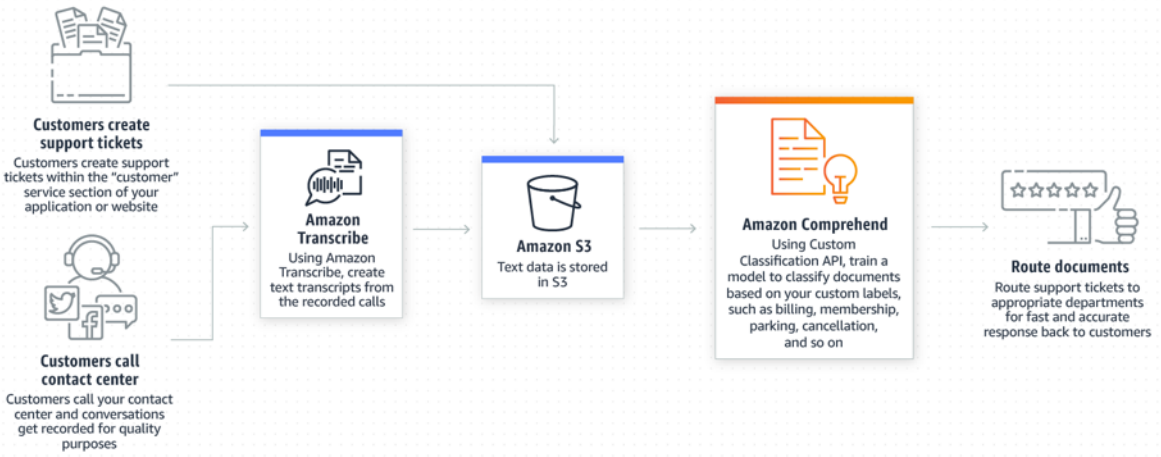
Call Center Analytics



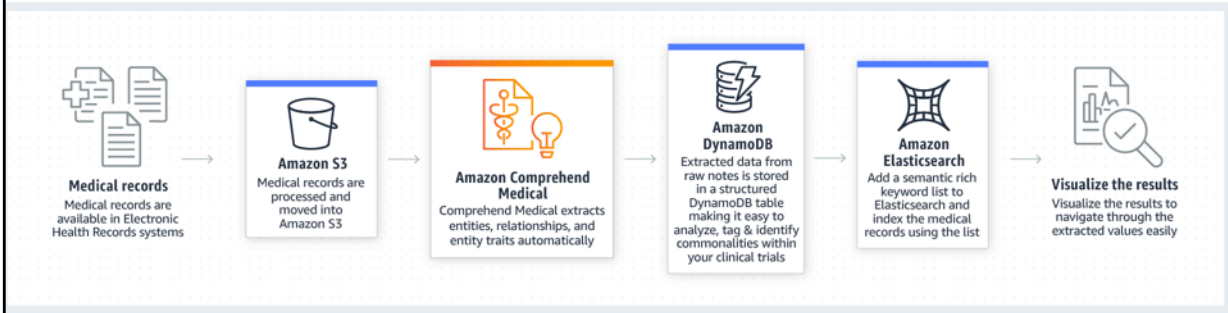
Index and Search Product Reviews

Content Recommendation

Customer Support Ticket Handling



Clinical Trial Recruitment



aws training and certification

What NLP use cases do you see in your industry?

Neural Topic Model

aws training and certification

Aspects	Details
Unsupervised	Does not require labels
Observations	Documents
Features	Occurrence count of words
Categories	Topics (number is prespecified)
Result	Documents described as mixture of identified topics

NTM Uses

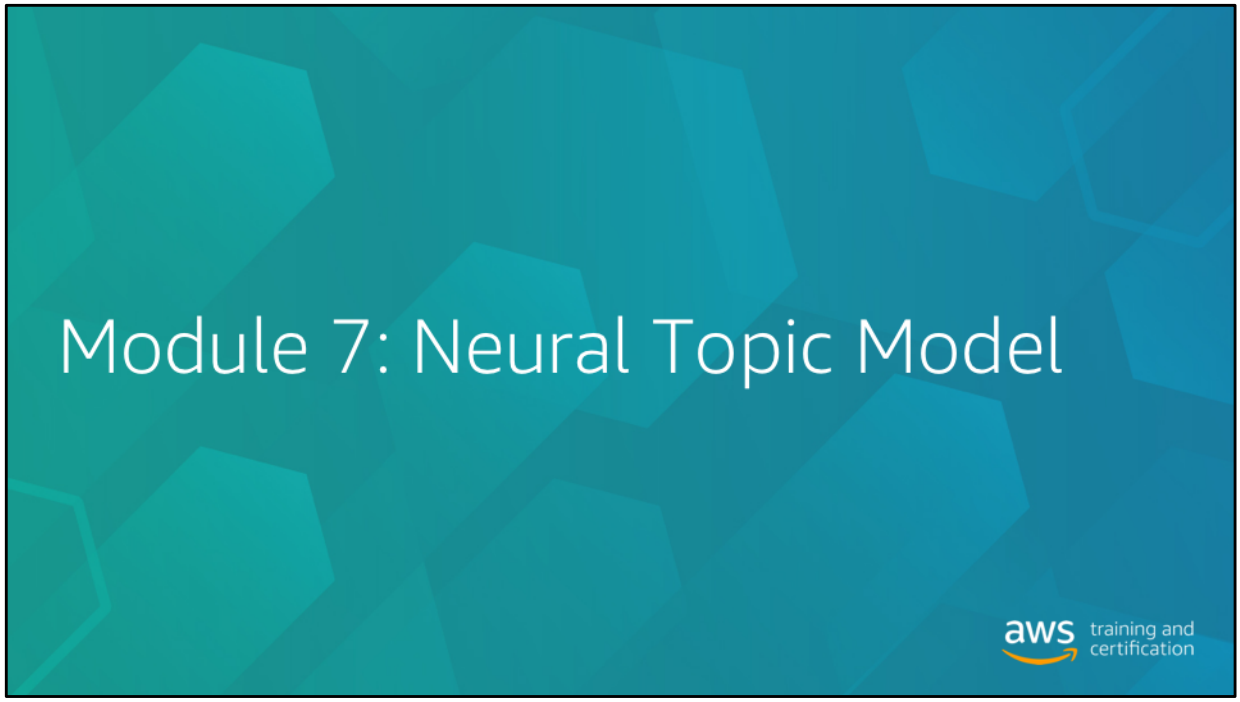
aws training and certification

Domain	Use Case
News	Classify news stories by topic
Support	Classify support issues
Development	Route bugs/issues/tickets to appropriate teams
Healthcare	Match patient records to researchers
Social Media	Analyze messages to detect negative behavior

Topic Model Options

aws training and certification

What are topic model options?



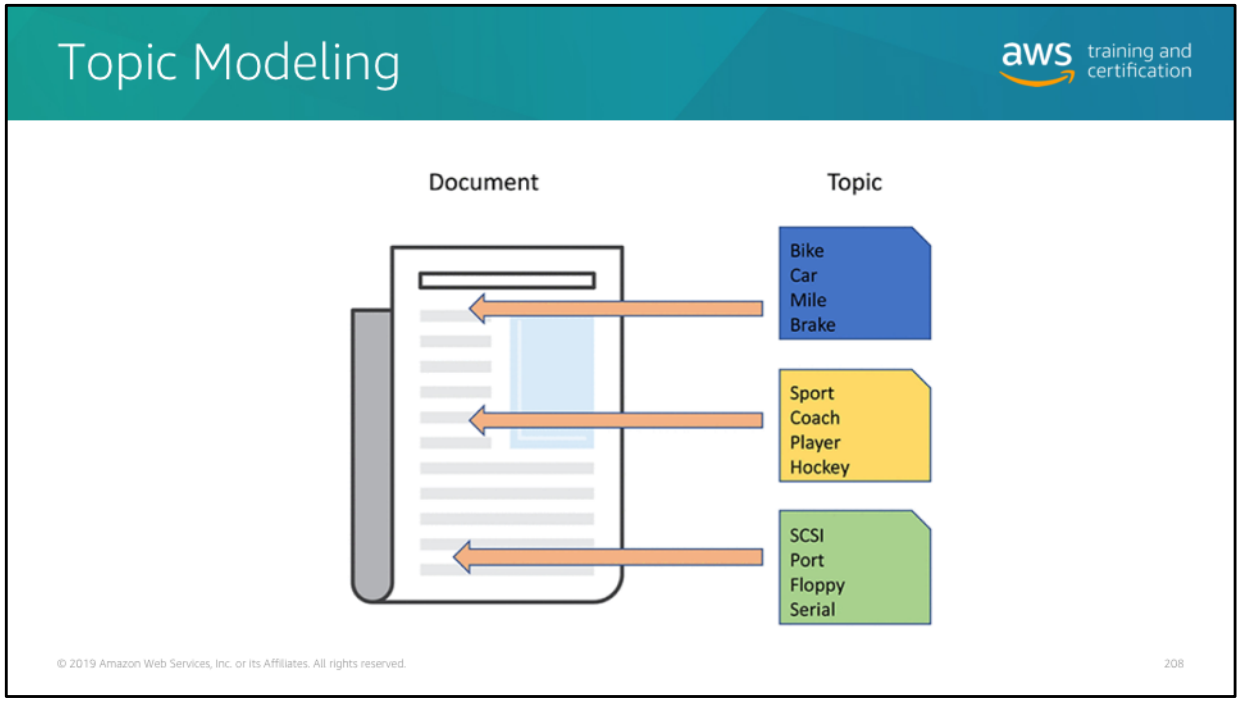
Amazon SageMaker Neural Topic Model (NTM) is an unsupervised learning algorithm that attempts to describe a set of observations as a mixture of distinct categories. NTM is most commonly used to discover a user-specified number of topics shared by documents within a text corpus. Here each observation is a document, the features are the presence (or occurrence count) of each word, and the categories are the topics. Since the method is unsupervised, the topics are not specified upfront and are not guaranteed to align with how a human may naturally categorize documents. The topics are learned as a probability distribution over the words that occur in each document. Each document, in turn, is described as a mixture of topics.

If you would like to know more please check out the SageMaker Neural Topic Model Documentation (https://docs.aws.amazon.com/sagemaker/latest/dg/ntm.html).

Hyperparameters

aws training and certification

Hyperparameter	Required	Description
feature_dim	*	Vocabulary size of dataset
num_topics	*	Number of topics
Other hyperparameters		



The technical definition of topic modeling is that each topic is a distribution of words and each document is a mixture of topics across a set of documents (also referred to as a corpus). For example, a collection of documents that contains frequent occurrences of words such as “bike,” “car,” “mile,” or “brake” are likely to share a topic on “transportation.” If another collection of documents shares words such as “SCSI,” “port,” “floppy,” or “serial” it is likely that they are discussing a topic on “computers.” The process of topic modeling is to infer hidden variables such as word distribution for all topics and topic mixture distribution for each document by observing the entire collection of documents. The figure that follows shows the relationships among words, topics, and documents.

There are many practical use cases for topic modeling, such as document classification based on the topics detected, automatic content tagging using tags mapped to a set of topics, document summarization using the topics found in the document, information retrieval using topics, and content recommendation based on topic similarities. Topic modeling can also be used as a feature engineering step for downstream text-related machine learning tasks. It’s also worth mentioning that, topic modeling is a general algorithm that attempts to describe a set of observations with the underlying themes. Although we focus on text documents here, the

observations can be applied other types of data. For example, topic models can also be used for modeling other discrete-data use cases such as discovering peer-to-peer applications on the network of an internet service provider or corporate network.

Model output

aws training and certification

- In our model, the output format of SageMaker NTM inference endpoint is a Python dictionary with the following format.
- We extract the topic weights, themselves, corresponding to each of the input documents.

```
{
    "predictions": [
        {"topic_weights": [0.02, 0.1, 0,...]},
        {"topic_weights": [0.25, 0.067, 0,...]}
    ]
}
```

Each number represents the probability of a topic

Predictions for 2 separate documents are represented here

The output of the SageMaker NTM model inference looks like the following. (There is a separate prediction output line for each document in the input data. The decimal numbers represent the weights for each of the topics assigned to the document.) The output format of inference: “topic_weights” is a list of non-negative numbers that represent the strength of topics in each document.

The output format of inference: “topic_weights” is a list of non-negative numbers that represent the strength of topics in each document.

Managing Vocabulary

aws training and certification

- Ignoring case
- Ignoring punctuation
- Ignoring frequent words that don't contain much information, called stop words, like "a," "of," etc.
- Fixing misspelled words.
- Reducing words to their stem (e.g. "play" from "playing") using stemming algorithms.

As the vocabulary size increases, so does the vector representation of documents. In this example, the length of the document vector is equal to the number of known words. You can imagine that for a very large corpus, such as thousands of books, that the length of the vector might be thousands or millions of positions. Further, each document may contain very few of the known words in the vocabulary.

This results in a vector with lots of zero scores, called a sparse vector or sparse representation. Sparse vectors require more memory and computational resources when modeling and the vast number of positions or dimensions can make the modeling process very challenging for traditional algorithms.

As such, there is pressure to decrease the size of the vocabulary when using a bag-of-words model. By managing vocabulary, you can use cleaning techniques:

Module 8: BlazingText and Word2Vec

aws training and certification

Problem

aws training and certification

Cluster documents that are similar to each other.

Classify documents into categories.

BOW (Bag of Words)

aws training and certification

the dog is on the table

0	0	1	1	0	1	1	1
are	cat	dog	is	now	on	table	the

BOW (Bag of Words)

aws training and certification

Feature	Pros	Cons
Treat document as bag of words and count frequencies	Easy to compute	Requires large sparse vectors, inefficient memory usage
		Does not notice similarity of words
		Does not notice relevance of words

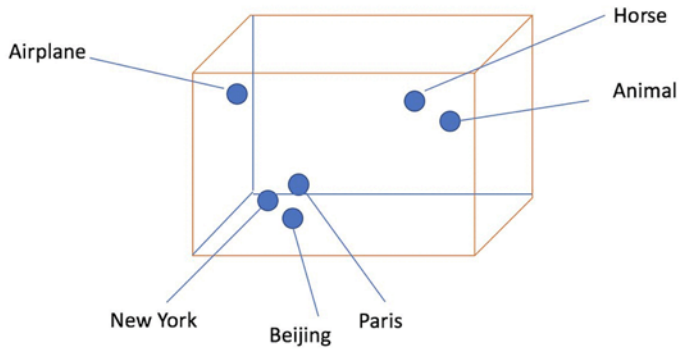
Word2Vec

aws training and certification

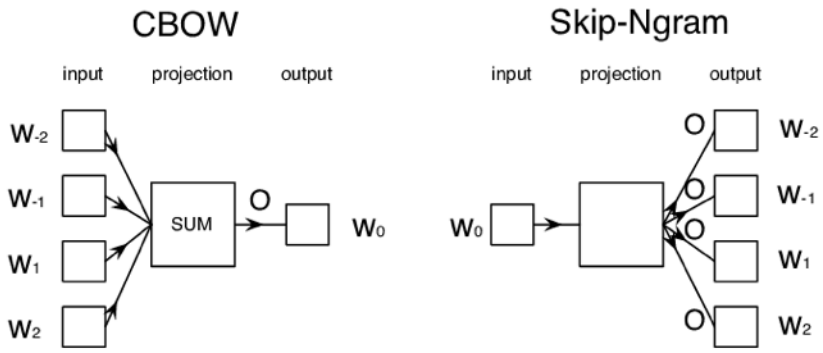
- How can we solve the problem of high dimensionality of words?
- Use Word2Vec to embed words in dense low-dimensional space.
- Word2Vec vectors capture semantic relationships.
- *King - Man + Woman = Queen*
- *Paris - France + Poland = Warsaw*

Embeddings

aws training and certification



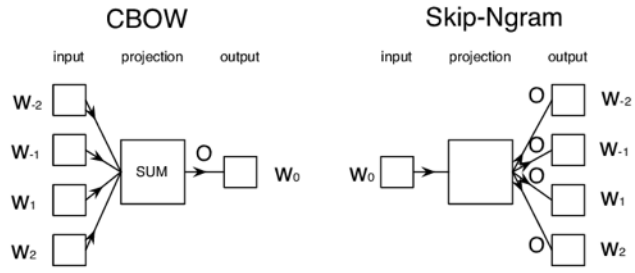
Word2Vec NN Arch



CBOW vs Skip Gram

aws training and certification

Algorithm	CBOW	Skip Gram
Maps	Neighbors to Word	Word to Neighbors
Performance	Fast	Accurate
Output	Predict Next Word	Word2Vec Embedding



Word2Vec Demo

aws training and certification

- Word2Vec Embedding Projector

BlazingText

aws training and certification

What is BlazingText?

- Generates Word2Vec using custom vocabulary.
- Has two modes.
- Unsupervised: Generates Word2Vec.
- Supervised: Generates Word2Vec and uses it to classify documents.

BlazingText vs NTM

aws training and certification

How does BlazingText differ from Neural Topic Model?

- NTM uses unsupervised learning to cluster documents.
- BlazingText/Word2Vec uses supervised learning to classify documents.
- BlazingText requires pre-specified categories for classification.

BlazingText vs Comprehend

aws training and certification

What is the difference between BlazingText and Comprehend?

- Comprehend is the fully managed option.
- Does not require training in many cases.
- Is not customized to domain-specific vocabulary.
- BlazingText and Comprehend represent convenience vs control tradeoff.

BlazingText Comparison

aws training and certification

Comparison	BlazingText	NTM	Comprehend
Training	Supervised	Unsupervised	Limited
Vocabulary	Custom	Custom	Generic
Convenience	Moderate	Moderate	High
Customization	High	High	Low

BlazingText Case Study

aws training and certification

- Mobile coupon and cash-back shopping app.
- Uses BlazingText for real-time search.
- Results include related terms.

BlazingText Data

aws training and certification

- BlazingText can be used in supervised or unsupervised mode.
- For supervised mode, the training/validation file should contain a training sentence per line along with the labels.
- Labels are words prefixed by `__label__`.
- If `subwords` set to true, model can generate vectors for out-of-vocabulary (OOV) words.

BlazingText Hyperparameters

aws training and certification

- BlazingText: Hyperparameters
- BlazingText: Model Tuning

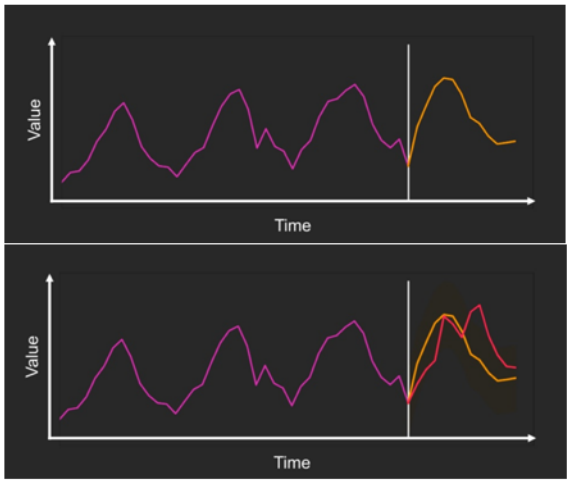
Module 9: Forecasting

aws training and certification

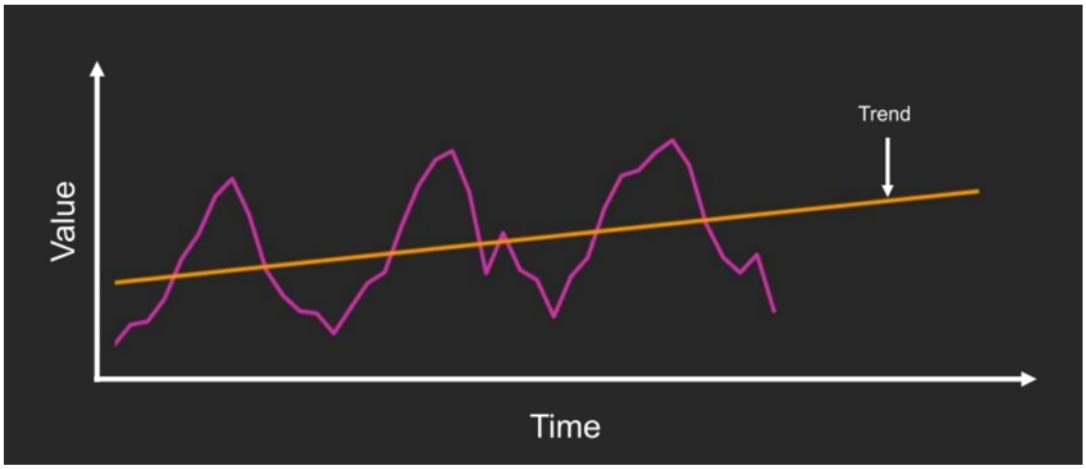
Forecasting

aws training and certification

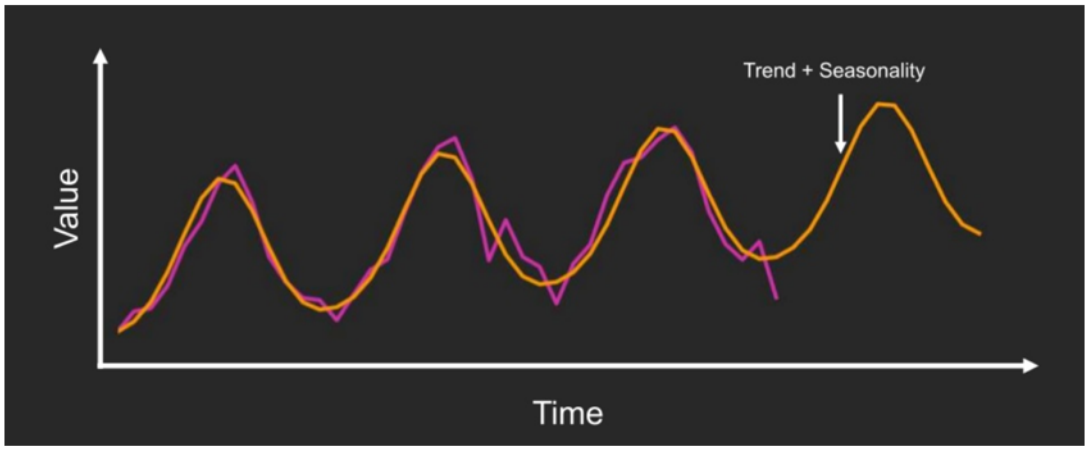
- Analyze past time series values.
- Predict future values.
- Measure accuracy against actual values.
- Probabilistic forecasts include confidence levels.



Time Series Model with Linear Trend



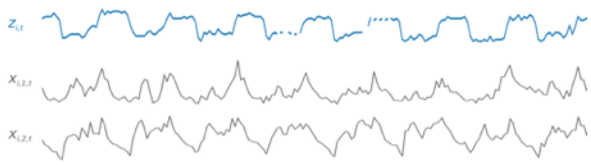
Time Series Model with Seasonality



Deep Learning Time Series Models

aws training and certification

- Can use covariates.
- Can be more accurate than traditional models.
- Can detect seasonal patterns with different periods.
- Support cold-start forecasts for new items.
- Auto-Regressive LSTM = DeepAR



Deep Learning Time Series Models

aws training and certification

- $Z_{i,t}$ = target.
- $X_{1,i,t}$ = feature 1.
- $X_{2,i,t}$ = feature 2.
- Algorithm uses past values of Z, X_1 and X_2 to predict the next value of Z