Fundamentals of Machine Learning with AWS

Asli Bilgin
@asli
asli@noktaconsulting.com

About this course

General housekeeping

- Questions encouraged: Q&A panel
- Commentary or tech support: Group Chat
- Breaks – after each segment 10 minutes – 15 minutes at the halfway point (about 2.5 hours)
- Polling

Where are you located?

- North America
- South America
- Asia Pacific
- Southeast Asia
- Europe
- Africa
- In the Clouds

Segment 1: High Level Overview of AI & ML

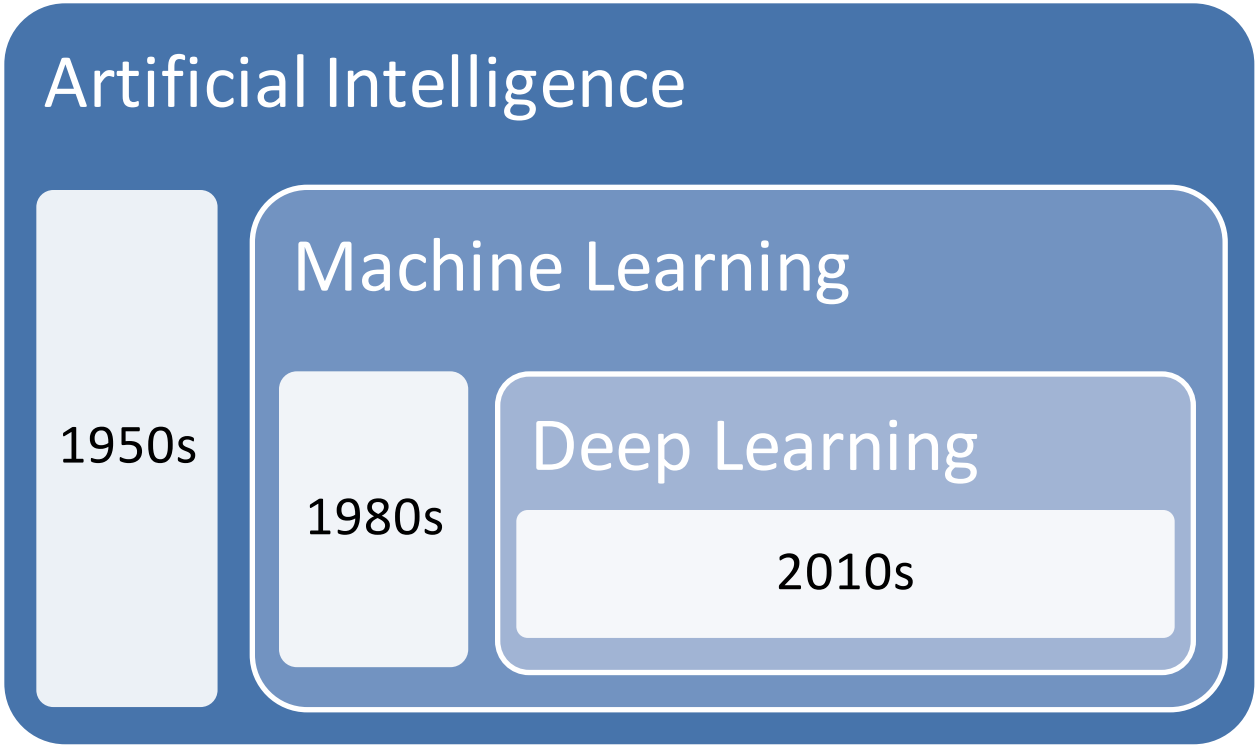
What is Artificial Intelligence?

What is Machine Learning?

What is Deep Learning?

Evolution of Artificial Intelligence, Machine Learning and Deep Learning

Evolution of AI & ML



What is Artificial Intelligence?

Behavior by machines that mimics human behavior in an "intelligent" manner.

Artificial Intelligence: A Modern Approach, authors Stuart Russell and Peter Norvig define AI as

Thinking humanly	Thinking rationally	Acting humanly	Acting rationally

The Surge of AI & ML

90% of all the data in the world was created in the last 2 years (attunity.com)

270% was the growth of AI in the enterprise in past 4 years (Gartner)

73% of company data is not analyzed (inc.com)

What is Machine Learning?

- Machine Learning is about using data to teach computers to “think”
- Begins with data
  - Ideally lots of it
  - Learnable data
  - Historical data
- Data is used to make predictions
- Data can be seen as experiences which can be generalized
- Avoids “hard coding” and rules

Data is the Heart of ML

"There's the joke that 80% of data science is cleaning the data and 20% complaining about cleaning the data"

– Kaggle founder & CEO Anthony Goldbloom

How Does ML Differ from AI?

Used when hard code or rules based programming is impractical

How does it work?

ML leverages **data** to	recognize patterns in data via **algorithms**	interpret those patterns in order to make **predictions**	which are then used to take **action** or make decisions

What is Deep Learning (DL)?

like the human brain, uses neural networks to identify features directly from data

"Deep" refers to the "layers" inside these neural nets

Layers contain hierarchy and relationships

How AI, ML & DL Differ in Application

8 legs
Spins webs
Round body

Spider

Lots of legs
Long body

Caterpillar

6 legs
Big eyes
Wings

Housefly

Artificial Intelligence

Machine Learning

Deep Learning

Segment 2: Which Use Cases Can Machine Learning Address?

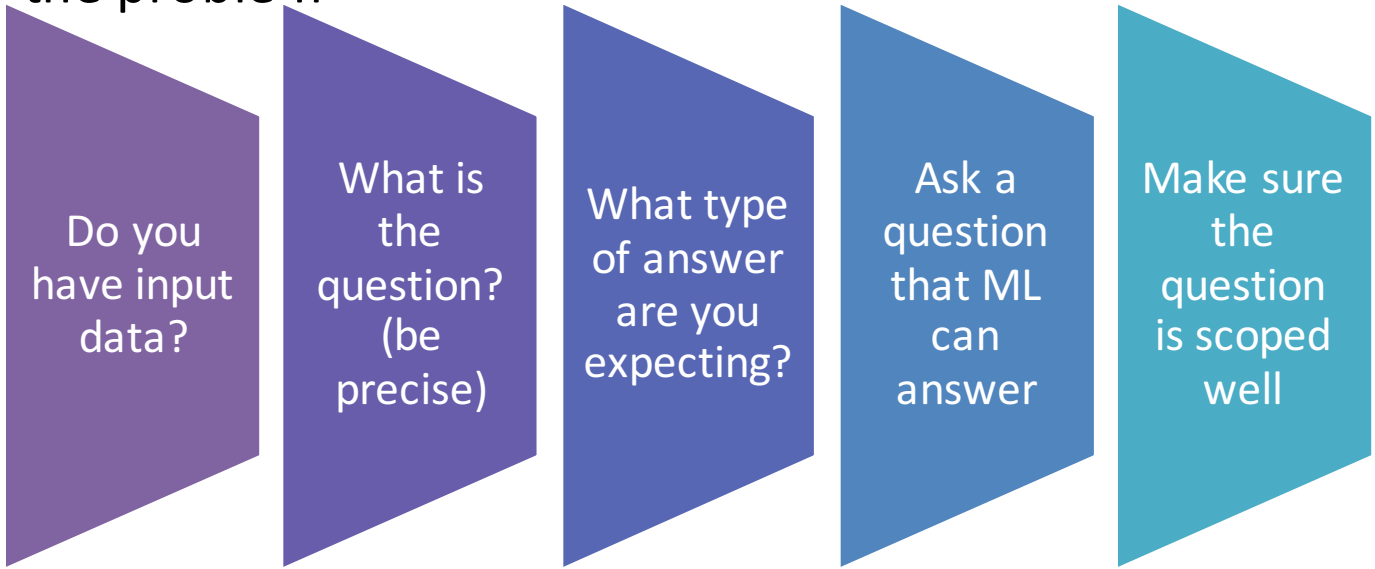
Personalization
Search
Marketing
Finance
Personal Productivity
Product Management

What type of system do you plan to build using Machine Learning on AWS?

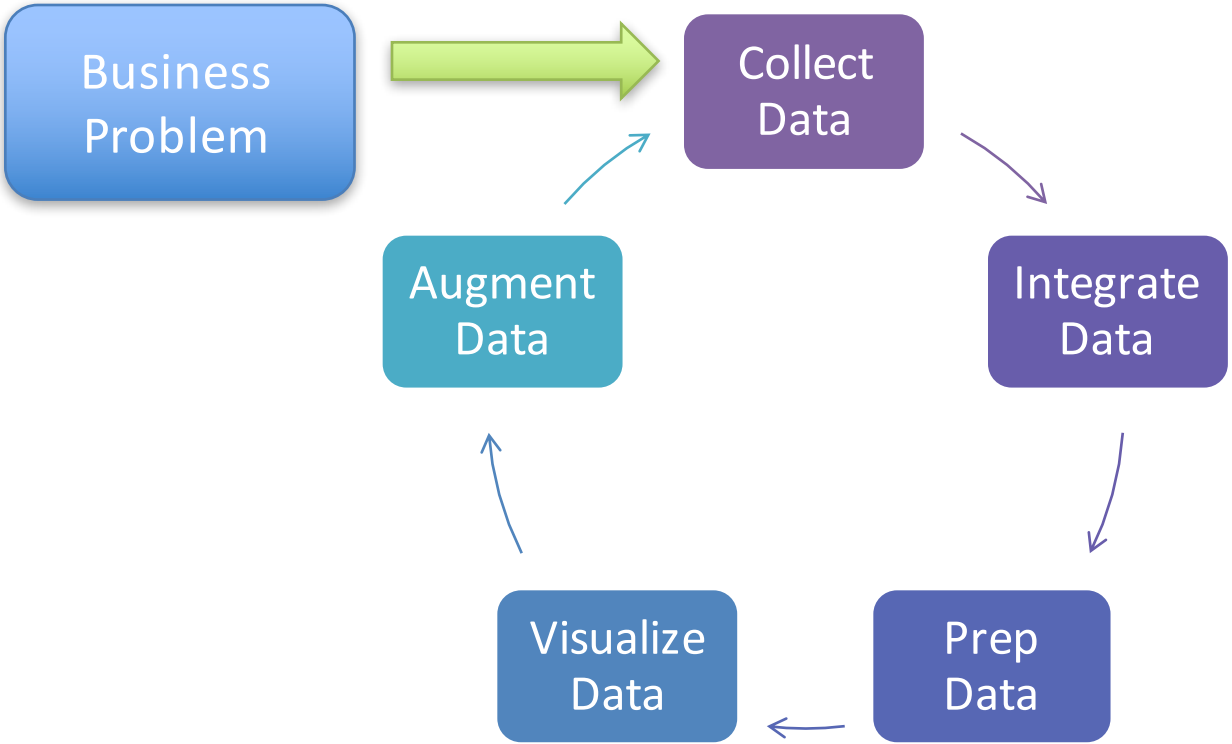
- No idea
- Banking
- Personal Productivity
- Marketing
- Retail
- Products / ISV
- None of the above? Comment in Group Chat

How to Frame a Suitable Problem

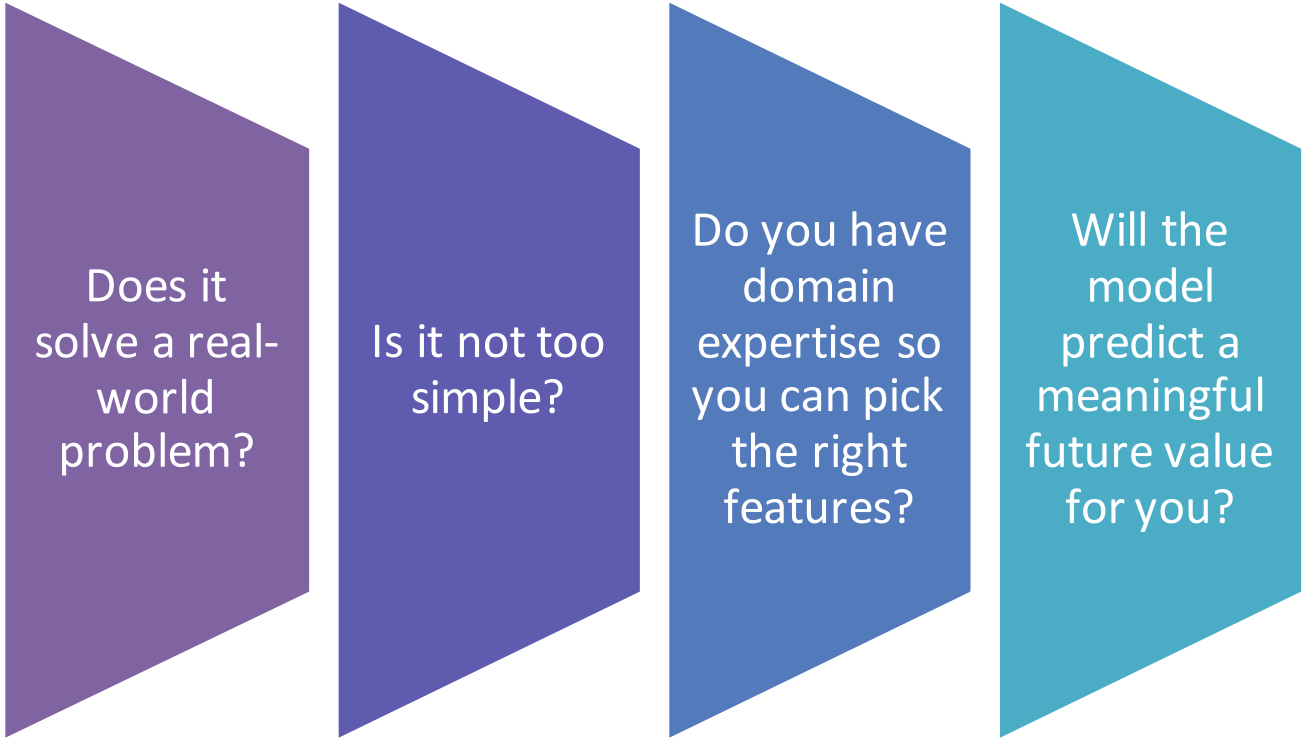
In order to determine whether machine learning is suitable for your scenario, you first need to frame the problem



Once a Problem is Framed… Address the Data



How to Choose a Good Business Problem?



Industry Adoption of AI

High	Medium	Low
• High Tech • Telecom • Automotive • Financial Services	• Retail • Media & Entertainment • Consumer Packaged Goods	• Education • Healthcare • Travel / Tourism

Source: McKinsey, How AI Can Deliver Real Value (2017)

Personalization

- Customer care
- Customer churn
- Predictive support
- Recommendations

Search

- Ranking
- Quality control
- Image and video search

Marketing

- Targeted marketing
- Outreach campaigns
- Promotions

Finance

- Fraud detection
- Transaction classification
- Market anomaly detection
- Customer account closure risks
- Chatbots

Personal Productivity

- Spam detection
- Natural language email response
- Time and goal tracking
- Personal assistant
- Social robots

Product Management

- Sales prediction
- Product improvement recommendations
- Demand forecasting
- Customer churn prediction
- Customer support

Amazon.com Product Management

Discovery (Search)

Fulfillment

Feature Enhancement

New Products (e.g. Alexa)

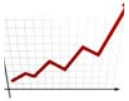
Segment 3: How does Machine Learning Work?

Workflow- Build, Train Deploy
Supervised Learning
Unsupervised Learning

Workflow for Getting Started

Ask Question

- What are you solving for?
- What type of problem is it?



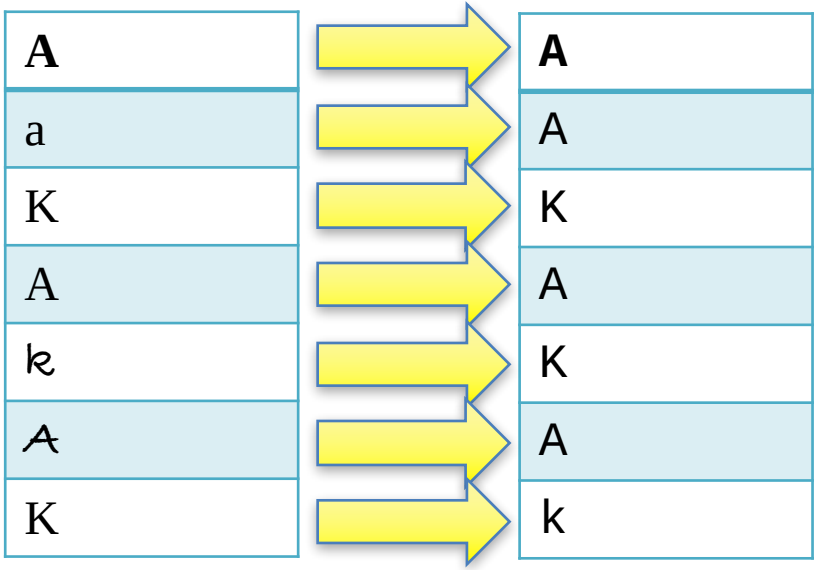
Identify Algorithms

- Identify 2-3 algorithms appropriate for job

Parameters

- Select short list of features

Supervised Machine Learning



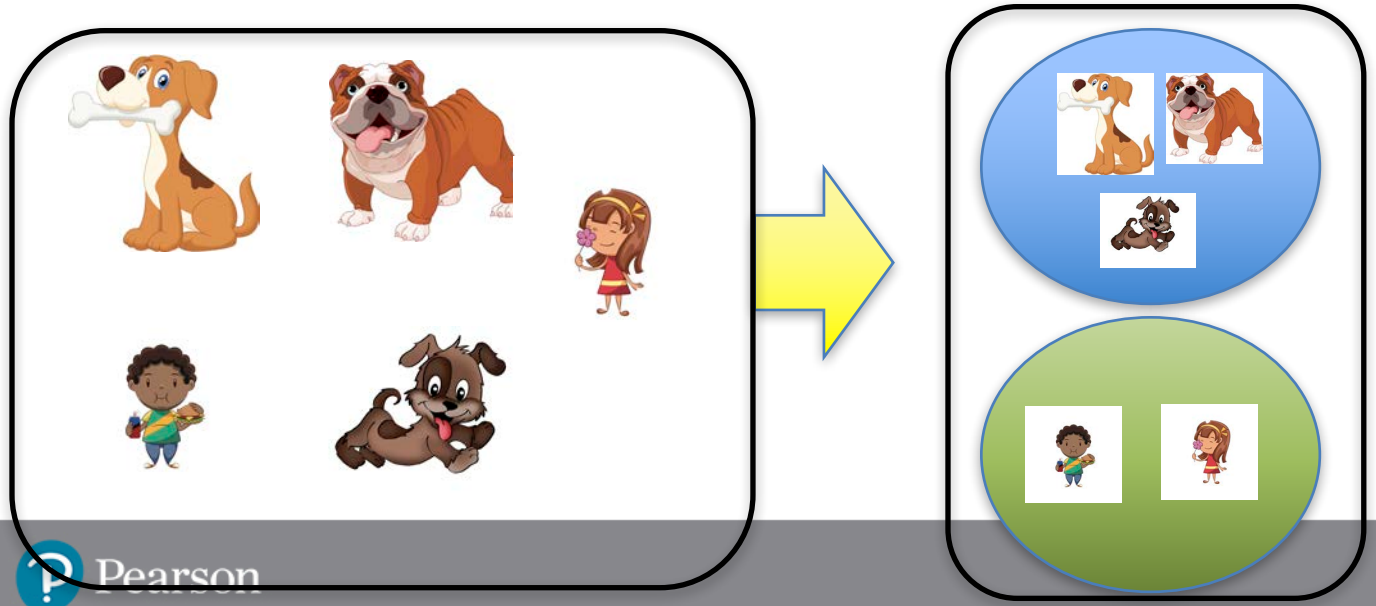
Labeled Training Data HISTORICAL INPUT (Handwriting Images)

Labeled Training Data HISTORICAL OUTCOME (Digital Text)

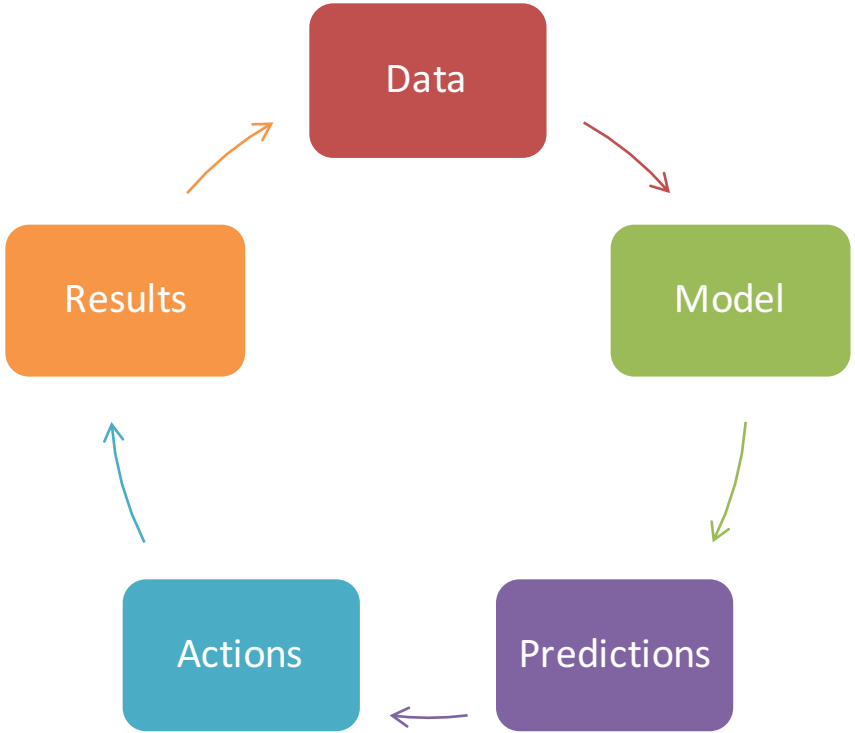
- Data is the teacher
- Establish relationship between labeled data points
- Leverage those relationship patterns to make predictions on new incoming data

Unsupervised Machine Learning

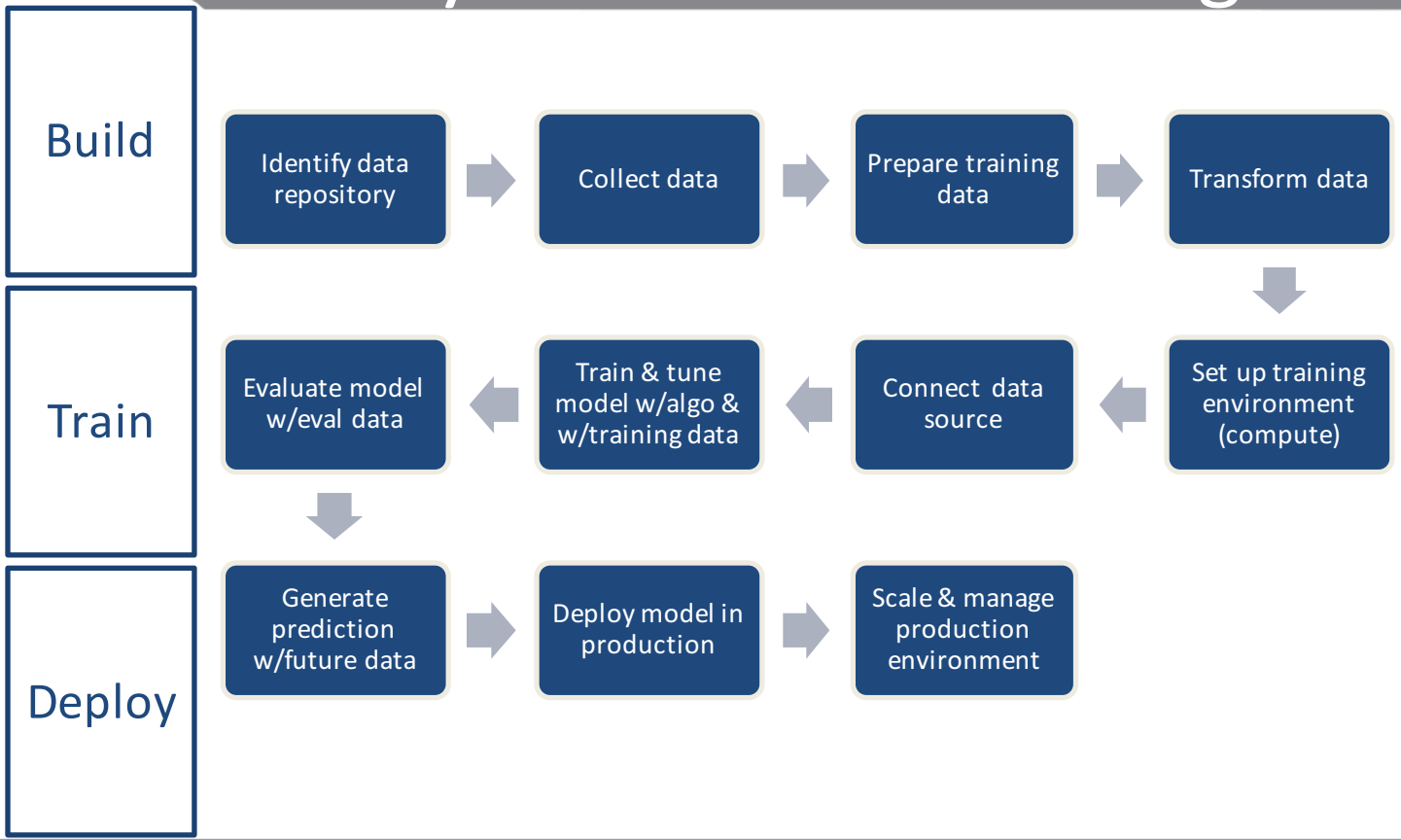
- There are no predefined labels
- The machine will distinguish patterns and classify the objects or provide an answer of some sort



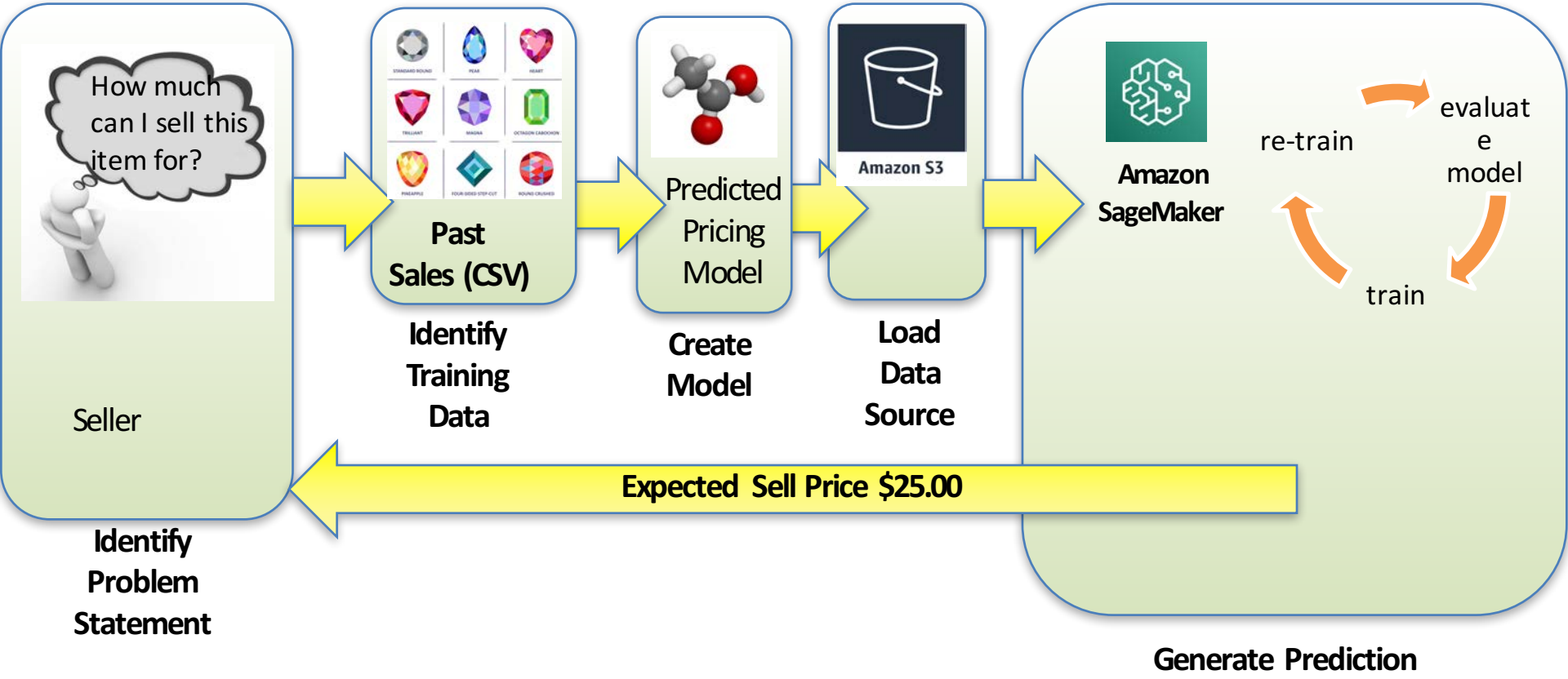
How Does ML Learn Over Time?



Life Cycle of ML Processing



Curator Project Workflow

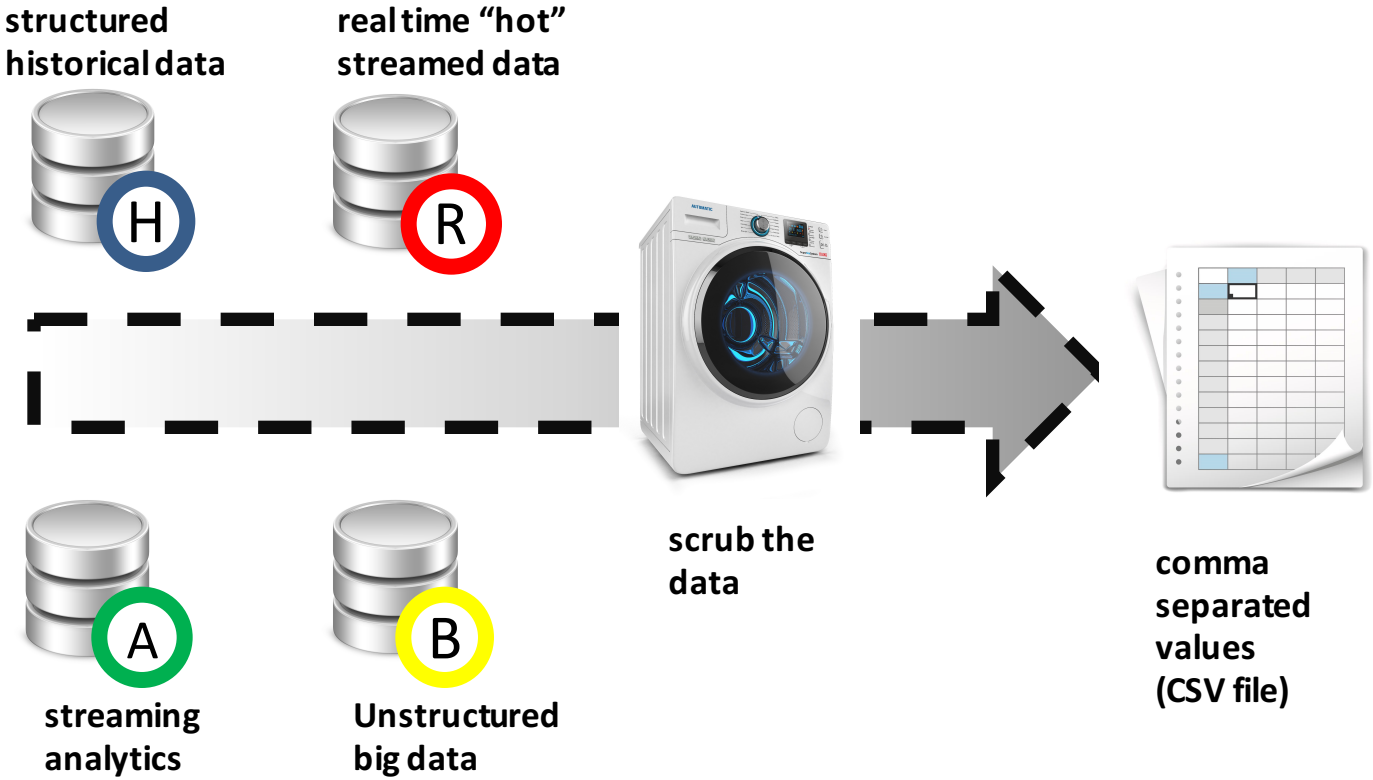


Curator Project: Sample Business Example

A curator would like to know the value of his collection. Machine Learning can predict prices for the curated collection as well as the sum of the prices for the items

What will be the selling price of a group containing those items? Does it make sense to break up the collection?

Many Types of Data



What type of data to you plan to use with Machine Learning? (Multi-select)

- No idea
- Relational database on premise
- Relational database in cloud
- Unstructured data on premise
- Unstructured data in cloud
- Real time data
- Other? Use Group Chat to comment

Mapping scenarios to models

Scenario 1 – Identifying Bots

An Instagram user wants to know whether comments on her photos are posted by a bot

- Do we have sample input data?
  - Yes – the historical comments on previous photos
- What are we solving for?
  - Whether future photos have comments posted by a bot
  - Yes or No – Is it a bot?
- Which model type is appropriate?
  - Binary Supervised Learning

Scenario 2 – Predicting Pricing

A franchise owner wants to sell his store and has asked the real estate agent: “At which price would the property sell?”

- Do we have sample input data?
  - Yes – comparative sales data
- What are we solving for?
  - At which price could we expect to sell?
  - Dollar numeric
- Which model type is appropriate?
  - Regression Supervised Learning

Scenario 3 – Predicting Choice

In order to hit quarterly numbers, your supervisor wants to create a marketing campaign for the financial services asset that will grow the most

- Do we have sample input data?
  - Yes – past customer choices
- What are we solving for?
  - Which asset is most appealing to a new customer?
  - Financial services asset class (category)
- Which model type is appropriate?
  - Multiclass Supervised Learning

Segment 4: How does Deep Learning Work?

High Level Overview
Key Concepts and Taxonomy
Neural Networks

High Level Overview of Deep Learning

- Is considered to be it a subset of Machine Learning
- Is based on neural networks - Convolutional neural network (CNN)
- Takes advantage of performance and scale that is provided by cloud computing
- Is inspired by neural science – how the brain works
- Doesn’t necessarily have to provide explicit features (e.g. eyes, nose, mouth) for it to “work”

Key Concepts and Taxonomy

- Convolutional neural network (CNN)
- Layers
- Results get better with more data, more models

How Does DL Differ from ML?

DL

- is used when features are not identified in advance
- "learns" the features
- can be more accurate
- is larger training dataset
- has a longer training time

ML

- conducts feature extraction manually
- enables you to choose your own features
- often has a smaller training dataset
- has a shorter training time

Segment 5: Amazon Artificial Intelligence and Machine Learning Overview

High Level Overview of AWS
AWS ML & AI: Platform Services
AWS ML & AI: Application Services
AWS ML & AI: Foundational Services

Did you attend re:Invent 2019?

- Yes – I absorbed every session on Machine Learning
- Yes – But I am not up to speed on all the developments around Machine Learning
- No – But I watched a few sessions online, including machine learning
- No – But I watched the keynote and a few sessions
- No – But I plan to
- No – What is re:Invent?

Amazon's Innovation with AI/ML

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

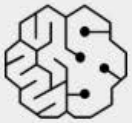
aws

Contact Sales | Support | English | My Account | Sign In to the Console

Products | Solutions | Pricing | Documentation | Learn | Partner Network | AWS Marketplace | Explore More

Machine Learning | Overview | AI Services | ML Services | Frameworks | Infrastructure | Learn ML | Blog | Partners

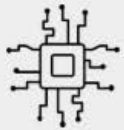
Machine Learning on AWS



ML Services
Build, train, and deploy ML fast

AI Services
Easily add intelligence to your applications

Frameworks
Choice and flexibility with broadest framework support



Compute
Fastest and lowest-cost compute options

Analytics and Security
Comprehensive capabilities, no compromise

Learning Tools
Get deep on ML with AWS DeepRacer and DeepLens

Machine Learning & AI Services

Platform Services

Application Services

Foundational (Frameworks / Interfaces)

Platform Services: SageMaker

Platform Services	 Amazon SageMaker	 Amazon SageMaker Ground Truth
What does it do?	Platform for end-to-end machine learning to build, train, and deploy models at scale	Enables you to leverage half a million Amazon Mechanical Turk human labelers and machine learning models to label your data
What's an example of it in use?	Map handwriting to digital data	Label ambiguous data for processing

AI Services : Vision

	 Amazon Rekognition	 Amazon Textract
What does it do?	Two services to conduct video & image analysis	Extracts text, tables and fields in forms from scanned documents
What's an example of it in use?	Facial recognition in the airport security line	Converting archives of paper documents into digital form

AI Services : Speech

	 Amazon Polly	 Amazon Transcribe
What does it do?	Automated text to speech service	Service for Automatic Speech Recognition
What's an example of it in use?	Language teaching courseware	Receive audio dictation or call center recordings and convert into digital text transcript

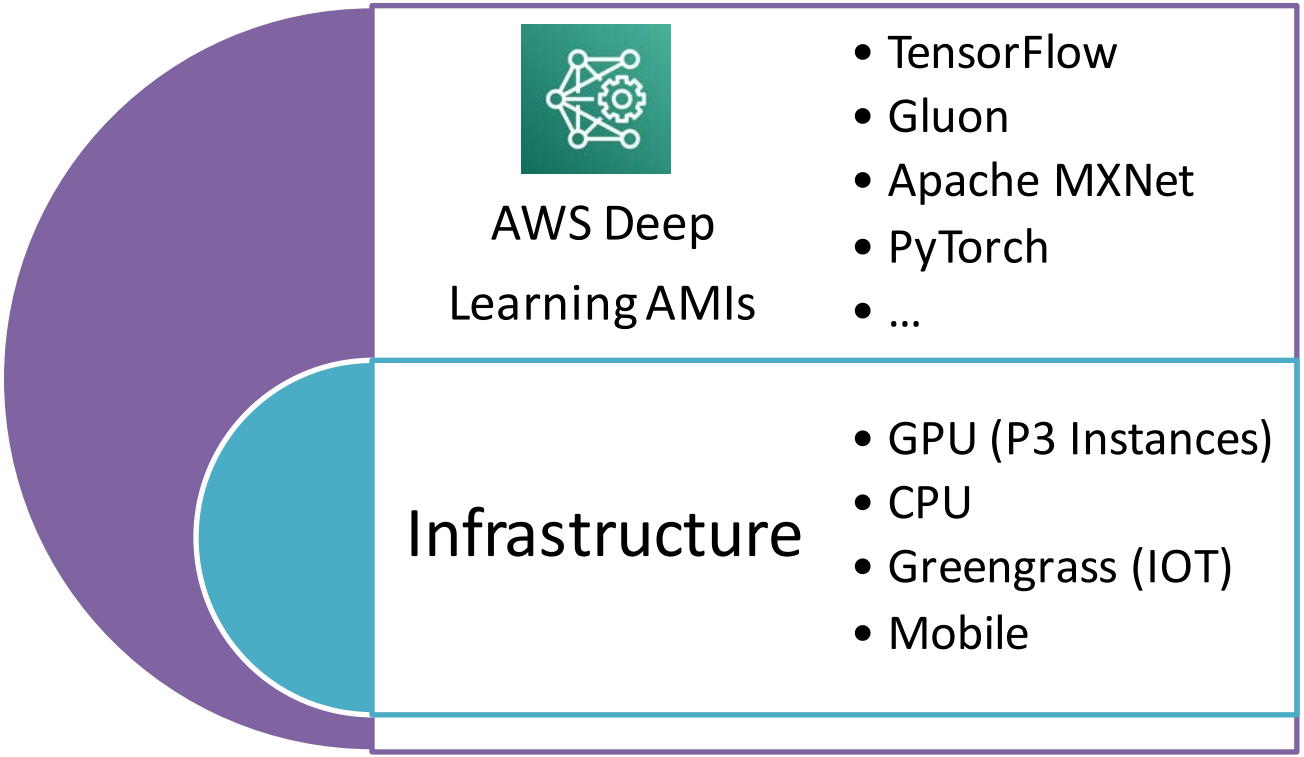
AI Services : Language

	 Amazon Translate	 Amazon Comprehend
What does it do?	Service to translate languages using deep learning via neural machine translation	Natural Language Processing service
What's an example of it in use?	Localize content with natural sounding translation quickly	Ability to gauge customer satisfaction based on negative and positive phrases in a review

AI Services : Recommendations, Forecasting, Chatbot

	 Amazon Personalize	 Amazon Forecast	 Amazon Lex
What does it do?	Machine learning service that creates sophisticated recommendation engines	Creates time series forecasts to predict trends	Service to build conversational elements via automatic speech recognition & Natural Language Understanding
What's an example of it in use?	Creates marketing promotions that are tailored to a specific demographic	Optimizing inventory supply by predicting customer purchases and demand	Human voice dialogue to Amazon Alexa or Chatbot to triage a customer support issue

AWS ML & AI: Foundational Services



New AI Services – re:Invent 2019

- Amazon Fraud Detector _ ADDD ICON!!!
- Amazon CodeGuru (preview)
- ContactLens Amazon Connect
- Amazon Kendra

Sample Case Studies from AWS

- Capital One
  - Conversational chatbot using Natural Language Processing
- Intuit
  - 2013 All In AWS
  - Sagemaker / Quickbooks
  - 90% faster ML model deployment
  - Tax season for Mint, Quickbooks & Turbotax
- NFL
  - Play prediction

Sample Case Studies from AWS

- Cerner
  - Health data - 250 M people
  - Make healthcare data actionable
  - Eliminating inefficacy, variation, cost and waste
- GE Healthcare
  - Improving patient's lives
  - Radiology efficiency

Formula One

"We are the most data rich sport in the world," MD of motorsports, Ross Brawn.

- Formula One
  - Deep Learning
  - 65 years of historical race data used to make race predictions with Amazon SageMaker
  - Streams real time race data to SageMaker to analyze driver performance, and share with fans
  - Data rich – 120 sensors per car, 1.1M telemetry data points per second, thousands of streams

Summary of AWS AI/ML Innovation

- Turn into a smart art
- Personalization
- Forecasting
- Image and video analysis
- Automatic speech recognition
- Natural language processing
- Fraud detection

What's your level of experience with AWS?

- Don't have an account : Never logged into the AWS Console
- Hobbyist : Played around for fun
- Basic : Used a few services for personal use or demos
- Intermediate : Conducted training courses and knowledgeable but never wrote production systems on AWS
- Professional : Earn my living using AWS
- Advanced : Wrote complex systems for large enterprises on AWS

Which title primarily describes your role?

- Developer
- Data scientist
- Application Developer
- DBA
- Student
- DevOps
- Web Developer
- General technologist
- Other? Use Group Chat to comment

Segment 6: Machine Learning Concepts and Taxonomy

The Big Picture

What is Input Data?

What are Features?

What is a Target?

What are Observations?

What is Labeled Data?

What is Unlabeled Data?

What is Ground Truth?

What are Hyperparameters?

What is Input Data?

Training data

- With answers / ground truth
- Used to train a model

Test data

- Evaluation data
- With answers, but answers disregarded until end of evaluation to compare quality

Validation data

- Used to verify training accuracy
- Helps optimize parameters

Inputs		Output
A	B	
0	0	0
0	1	1
1	0	1
1	1	1

Using Curator Input Data

itemname	winningBid	itemMaterialNam	typeName	isitMovable	hallmarkName
Butterfly - open	0.99	Sterling Silver	Foliage	NULL	None
San Francisco Trolley, marked	3	Sterling Silver	US	NULL	STER
anniverary charm	19.38	Silver tone	Personal Engraveme	1	unmarked
elf on mushroom	15.5	Sterling Silver	Magic	NULL	None
Dangling pointy fob with chain	9.99	NULL	NULL	NULL	NULL
Consett shield	9.99	Sterling Silver	Shield - Germany - P	NULL	NULL
Firenze coin with duomo on it	9.99	NULL	NULL	NULL	NULL
Puerto Rico country shield	9.99	Sterling Silver	Shield - Germany - P	NULL	Sterling Silver

What are Features?

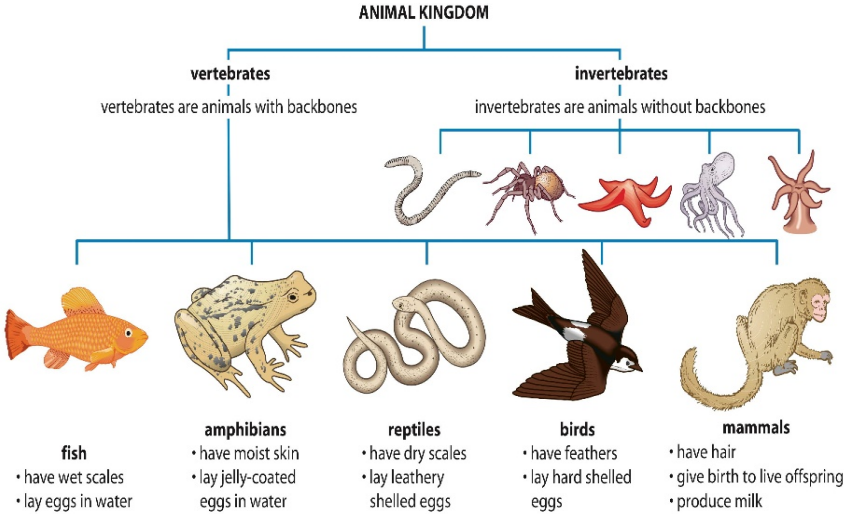
Serve as foundational bricks for a dataset

Measure characteristics of your data set

Are also known as attributes, properties, or variables

Analogy: Imagine as columns in the dataset

Will contain a direct correlation between feature quality and insight quality



Example of Features with the Curator

itemname	winningBid	itemMaterialNam	typeName	isitMovable	hallmarkName
Butterfly - open	0.99	Sterling Silver	Foliage	NULL	None
San Francisco Trolley, marked	3	Sterling Silver	US	NULL	STER
anniverary charm	19.38	Silver tone	Personal Engraveme	1	unmarked
elf on mushroom	15.5	Sterling Silver	Magic	NULL	None
Dangling pointy fob with chain	9.99	NULL	NULL	NULL	NULL
Consett shield	9.99	Sterling Silver	Shield - Germany - P	NULL	NULL
Firenze coin with duomo on it	9.99	NULL	NULL	NULL	NULL
Puerto Rico country shield	9.99	Sterling Silver	Shield - Germany - P	NULL	Sterling Silver

winningBid	typeName	makerName
12.99	Spiritual	TC Sterling
9.5	Spiritual	Monet
7.99	Ladies silver	Monet
7	Spiritual	Griffith (RL Griffith & Son)
9	Spiritual	Wells

Curator Feature Examples

typeName: Luck
isItMoveable: No
MaterialName: Sterling Silver

typeName: Shield
isItMoveable: No
MaterialName: European Silver

What is a Target?

Serves as the "answer" in your dataset

Measures the outcome

Is a special attribute or "feature" in your dataset

Is used for predictions by connecting the target to the other attributes via patterns

Example of a Target with the Curator

itemname	winningBid	itemMaterialName	typeName	isitMovable	hallmarkName
Butterfly - open	0.99	Sterling Silver	Foliage	NULL	None
San Francisco Trolley, marked	3	Sterling Silver	US	NULL	STER
anniverary charm	19.38	Silver tone	Personal Engraveme	1	unmarked
elf on mushroom	15.5	Sterling Silver	Magic	NULL	None
Dangling pointy fob with chain	9.99	NULL		NULL	NULL
Consett shield	9.99	Sterling Silver		NULL	NULL
Firenze coin with duomo on it	9.99	NULL		NULL	NULL
Puerto Rico country shield	9.99	Sterling Silver		NULL	Sterling Silver

contains the answer

What are Observations?

Serve data points in a dataset

Measure characteristics

Also referred to as examples or instances

Analogy: Occur as rows in the dataset

Contain feature(s) and a target

Observation Examples with the Curator

itemname	winningBid	itemMaterialName	typeName	isitMovable	hallmarkName
Butterfly - open	0.99	Sterling Silver	Foliage	NULL	None
San Francisco Trolley, marked	3	Sterling Silver	US	NULL	STER
anniverary charm	19.38	Silver tone	Personal Engraveme	1	unmarked
elf on mushroom	15.5	Sterling Silver	Magic	NULL	None
Dangling pointy fob with chain	9.99	NULL	NULL	NULL	NULL
Consett shield	9.99	Sterling Silver	Shield - Germany - F	NULL	NULL
Firenze coin with duomo on it	9.99	NULL	NULL	NULL	NULL
Puerto Rico country shield	9.99	Sterling Silver	Shield - Germany - F	NULL	Sterling Silver

observations / "experience"

What is Labeled Data?

- Labeled data is a way of classifying the answer, when you know the answer
- Labels would be predicted output, and you can consider a feature to be input
- Labeled data is a means of tagging
- Labeling is analogous to learning insects in school as a child

What is Labeled Data?

8 legs
Spins webs
Round body

Lots of legs
Long body

6 legs
Big eyes
Wings

Features

Spider

Caterpillar

Housefly

Label

What is Unlabeled Data?

- Does not have a defined "answer"
- Is data without definition– for example: a diary, random Post It notes, photographs
- Can be labeled by machines (color, object detection, etc.)
- Can be labeled by humans– for example By Date, Topic, CreatedBy

What is Ground Truth?

- Contains the final answer that you are trying to predict
  Refers to the classification of training data
- Consists of the schema that you need to identify for the data
- Determines whether a model predicts accurately or not
- Serves the real answers compared to the predicted answers
- Can be wrong
- Is found in existing or incoming data
- Is used to predict future outcomes

What are hyperparameters?

- A model parameter is required by a model to make predictions whose value can be “learned” by the data. It’s internal.
- A hyperparameter, on the other hand, is external to the model whose value cannot be estimated by the data. Generally they are “set”
- Often these terms are used interchangeably – For further explanations see: StackOverflow, Quora

What are Predictions or Inferences?

- Delivered as results
- Contains the answer you are trying to predict
- Forms the response using future data
- Are generated by the model

How do Predictions Work?

- Predictions use the model to deliver results
- You feed in observations, predictions return targets with scores
- Scores serve as the probability

Segment 7: How to Use Machine Learning?

Preparing Data

Model Training

Refining Models

Conducting Predictions

Data Pre-Processing

Refers to manipulations done to your data before sending it into algorithm

For example, my regression algorithm would like numbers for my categories.

Used a function that creates dummy columns that are each binary in nature

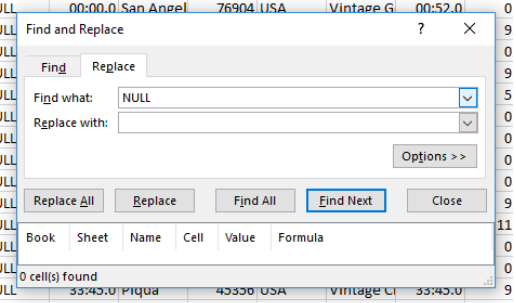
Specific algorithms, such as Linear Learner & XGBoost, require dummy encoded data

Improve Your Model: Options for Massaging the Data

- Create new observations
- Remove anomalies
- Remove outliers
- Remove attributes with poor correlation to target
- Remove Missing Values
- Create Dummy Variables

Scrubbing the Data

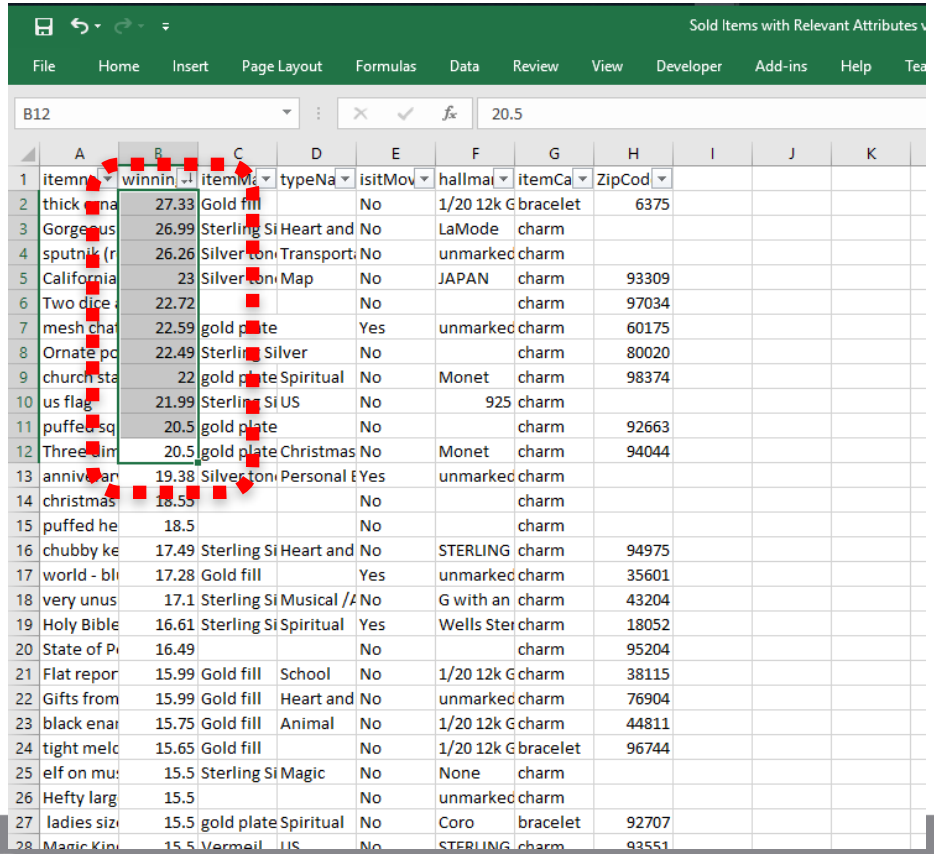
itemNam	itemdescr	winningBi	typeName	isit3D	isitMovab	soldDate	location	ZipCode	CountryNa	title	auctionEn	priceSe
Nebraska	in beautifu	3	Map	NULL	NULL	27:23.0	Springville	84663	USA	Vintage U	06:56.0	8
University	from UNC	14.99	School	NULL	NULL	28:01.0	Richmond	23226	USA	Vintage St	28:01.0	14
Gifts from		15.99	Heart and	NULL	NULL	00:00.0	San Angel	76904	USA	Vintage G	00:52.0	0
New Jersey -	pink sta	9.99	State - US	NULL	NULL							9
snake head		0.99	Ladies silv	1	NULL							0
cougar head,	growli	5.99	Animal	NULL	NULL							9
Concave pig -	runnin	5.99	Animal	NULL	NULL							5
tiny wooden	shoe/ c	0.99	Travel	NULL	NULL							0
Liberty bell,	flat 177(	0.99	US	NULL	NULL							0
New Mexi	marked or	0.99	Map	NULL	NULL							0
WISCONSIN	State Ma	0.99	Map	NULL	NULL							0
Colonial H	tricorn hat	9.99	US	1	NULL							9
Oregon state	map in	11.99	Map	NULL	NULL							11
Zwickau -	attempted	0.99	Shield - G	NULL	NULL							0
Zwickau -	attempted	0.99	Shield - G	NULL	NULL							0
Wisconsin	navy blue	5.5	Map	NULL	NULL	33:45.0	Piqua	45356	USA	Vintage Cl	33:45.0	9
A TOTEM	POLE, with	7.25	Travel	1	NULL	26:49.0	Atlanta	30305	USA	Vintage g	26:49.0	9
Large Bud	Gold tone	7.99	Religious	NULL	NULL	35:35.0	Chalmers	47929	USA	Vintage g	35:35.0	7
VIP small	gothic lett	0.99	Personal I	NULL	NULL	46:59.0	San Anton	78216	USA	Vintage F	46:59.0	0



Cleaning data

- Conduct a Find for NULLs and replace with Empty string
- Replace missing values

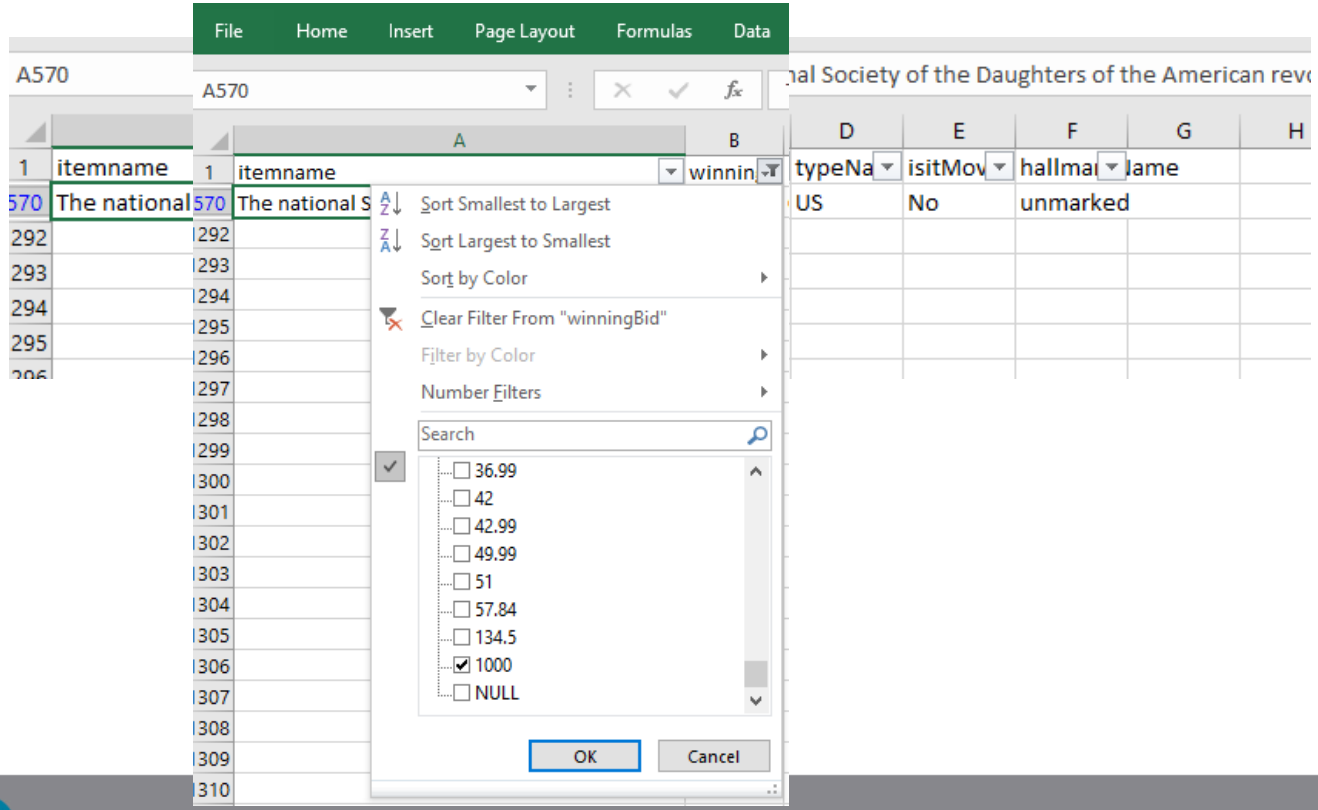
Eliminating Noise



	A	B	C	D	E	F	G	H
1	itemN	winnin	itemM	typeNa	isitMov	hallmar	itemCa	ZipCod
2	thick ena	27.33	Gold fill		No	1/20 12k G	bracelet	6375
3	Gorgeous	26.99	Sterling Si	Heart and	No	LaMode	charm	
4	sputnik (r	26.26	Silver ton	Transport	No	unmarked	charm	
5	California	23	Silver ton	Map	No	JAPAN	charm	93309
6	Two dice	22.72			No		charm	97034
7	mesh chat	22.59	gold plate		Yes	unmarked	charm	60175
8	Ornate pc	22.49	Sterling	Silver	No		charm	80020
9	church sta	22	gold plate	Spiritual	No	Monet	charm	98374
10	us flag	21.99	Sterling Si	US	No	925	charm	
11	puffed sq	20.5	gold plate		No		charm	92663
12	Three dim	20.5	gold plate	Christmas	No	Monet	charm	94044
13	anniv ar	19.38	Silver ton	Personal E	Yes	unmarked	charm	
14	christmas	18.55			No		charm	
15	puffed he	18.5			No		charm	
16	chubby ke	17.49	Sterling Si	Heart and	No	STERLING	charm	94975
17	world - bl	17.28	Gold fill		Yes	unmarked	charm	35601
18	very unus	17.1	Sterling Si	Musical /	No	G with an	charm	43204
19	Holy Bible	16.61	Sterling Si	Spiritual	Yes	Wells Ster	charm	18052
20	State of P	16.49			No		charm	95204
21	Flat repor	15.99	Gold fill	School	No	1/20 12k G	charm	38115
22	Gifts from	15.99	Gold fill	Heart and	No	unmarked	charm	76904
23	black enar	15.75	Gold fill	Animal	No	1/20 12k G	charm	44811
24	tight melc	15.65	Gold fill		No	1/20 12k G	bracelet	96744
25	elf on mus	15.5	Sterling Si	Magic	No	None	charm	
26	Hefty larg	15.5			No	unmarked	charm	
27	ladies size	15.5	gold plate	Spiritual	No	Coro	bracelet	92707
28	Magic Kin	15.5	Vermeil	US	No	STERLING	charm	93551

- Most of the items sold for under 25
- When we look at the raw data, there are only 11 records that are over $20; this may be generating noise
- Can add a filter to raw data

Eliminating Anomalies



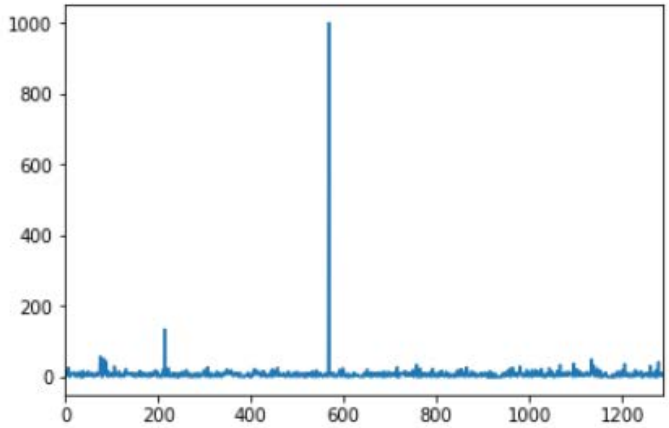
Using Plotting to Find Anomalies

Plotting Now let's plot the prices visually - using a simple line graph

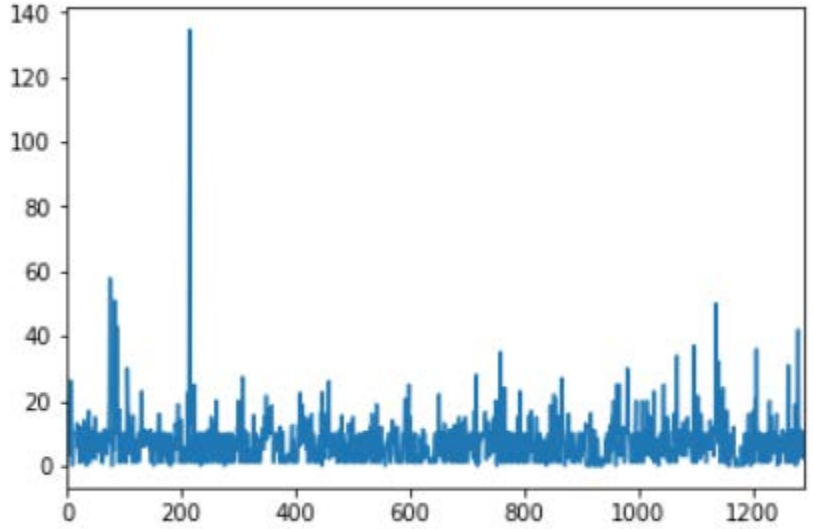
In [52]:
```
import matplotlib # this library to generate plots (in other words, graphs)
%matplotlib inline
```

In [54]:
```
curator_sales['winningBid'].plot()
```

Out[54]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe64470d198>
```

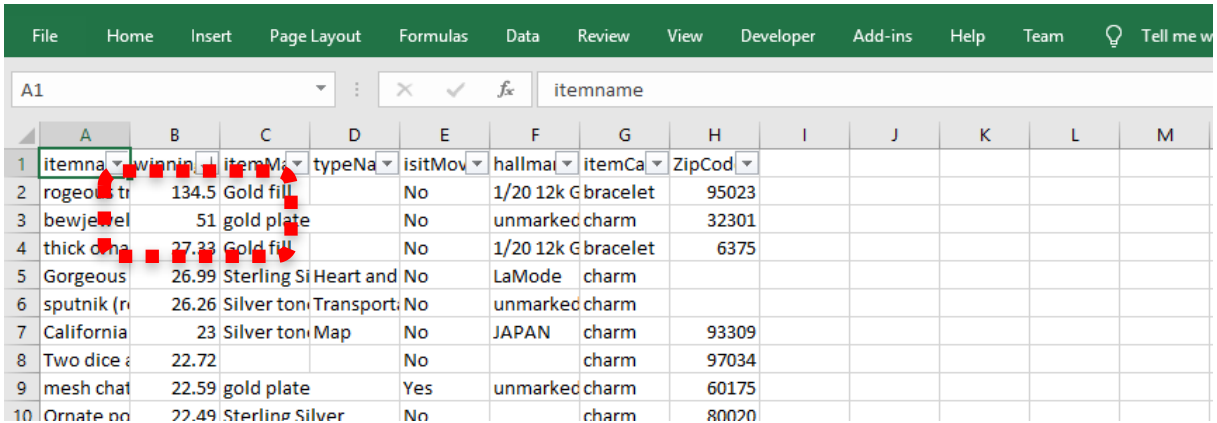


More normalized pricing results



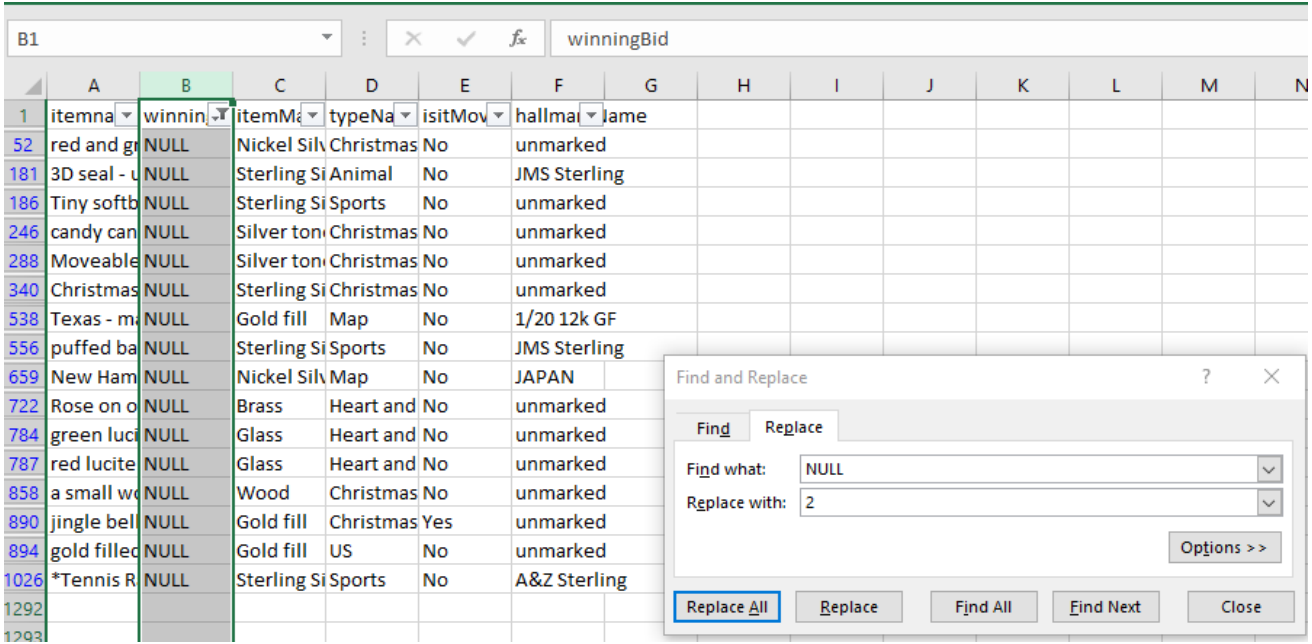
Remove Outliers: Sample

- Sort the source data to identify outliers
- There are two records that are outliers
- Adjust the SQL query to eliminate these
- WHERE winning bid <50



	A	B	C	D	E	F	G	H
1	itemna	winnin	itemM	typeNa	isitMov	hallmar	itemCa	ZipCod
2	rogeou	134.5	Gold fil		No	1/20 12k G	bracelet	95023
3	bewjewel	51	gold plate		No	unmarked	charm	32301
4	thick d	27.33	Gold fil		No	1/20 12k G	bracelet	6375
5	Gorgeous	26.99	Sterling Si	Heart and	No	LaMode	charm	
6	sputnik (r	26.26	Silver ton	Transport	No	unmarked	charm	
7	California	23	Silver ton	Map	No	JAPAN	charm	93309
8	Two dice	22.72			No		charm	97034
9	mesh chat	22.59	gold plate		Yes	unmarked	charm	60175
10	Ornate po	22.49	Sterling Silver		No		charm	80020

Eliminating Missing Values



	A	B	C	D	E	F	G
1	itemna	winnin	itemMa	typeNa	isitMov	hallmai	lame
52	red and gr	NULL	Nickel Silv	Christmas	No	unmarked	
181	3D seal - u	NULL	Sterling Si	Animal	No	JMS Sterling	
186	Tiny softb	NULL	Sterling Si	Sports	No	unmarked	
246	candy can	NULL	Silver ton	Christmas	No	unmarked	
288	Moveable	NULL	Silver ton	Christmas	No	unmarked	
340	Christmas	NULL	Sterling Si	Christmas	No	unmarked	
538	Texas - ma	NULL	Gold fill	Map	No	1/20 12k GF	
556	puffed ba	NULL	Sterling Si	Sports	No	JMS Sterling	
659	New Ham	NULL	Nickel Silv	Map	No	JAPAN	
722	Rose on o	NULL	Brass	Heart and	No	unmarked	
784	green luci	NULL	Glass	Heart and	No	unmarked	
787	red lucite	NULL	Glass	Heart and	No	unmarked	
858	a small wo	NULL	Wood	Christmas	No	unmarked	
890	jingle bel	NULL	Gold fill	Christmas	Yes	unmarked	
894	gold filled	NULL	Gold fill	US	No	unmarked	
1026	*Tennis R	NULL	Sterling Si	Sports	No	A&Z Sterling	
1292							
1293							

Eliminating missing values

dataframe.dropna()

- removes rows of the data frame that has missing info
- dataframe.dropna(how='all')
- Set axis to eliminate rows or columns

sklearn.impute.SimpleImputer

- Can replace missing values with means

pandas.notnull()

- Can drop rows with null values

Dropping null values: Sample

df = dataframe
pd = pandas library

```
df = df[pd.notnull(df['winningBid'])]
df = df[pd.notnull(df['isitMovable'])]
df = df[pd.notnull(df['itemMaterialName'])]
df = df[pd.notnull(df['typeName'])]
df = df[pd.notnull(df['hallmarkName'])]
```

Missing Value Handling Depends on Data Type

numeric

- Use an average or sample mean
- Use a default value that is most often used (e.g. $9.99)
- Use random numbers within the bell curve range

categorical

- Fill using common or default value
- Fill based on frequency, most frequent fields randomly filled into rows
- Random values within frequency ranges

Pre-processing Missing Values with sklearn

sklearn.impute.SimpleImputer - Settings

“mean”

- swap missing values using the mean for each column.
- numeric data only

“most_frequent”

- swap with most frequent value along each column.
- strings or numeric data

“constant”

- Swap w/fill_value.
- strings or numeric data

Sample Scenario – most frequent

- For example, I would like use most frequent for my item material category and mean for missing price values in the target
  - Nickel Silver
  - Sterling Silver
- Based on simple calculation to look for the frequently used fields in the item name field.

Modeling Training & Refinement

Start with Application Services

 Amazon Rekognition	 Amazon Lex	 Amazon Personalize	 Amazon Forecast
Two services to conduct video & image analysis	Service to build conversational elements via automatic speech recognition & Natural Language Understanding	Machine learning service that creates sophisticated recommendation engines	Creates time series forecasts to predict trends

1. First consider the high level application services
2. Designed for general purpose problems

Where to Begin?

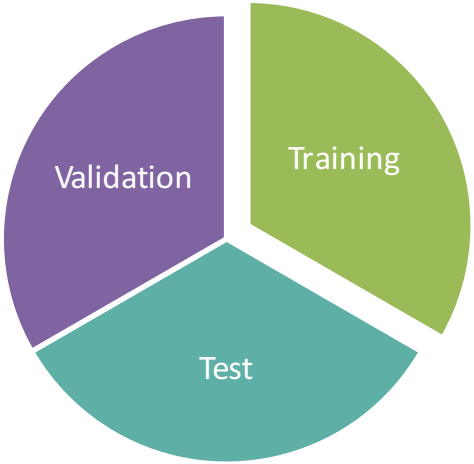
- Begin with the question
- From there, once you know the question, the dataset, and parameters, you can then choose an algorithm
- Keep it simple (e.g. linear, factorization machine)
- Start with basic models as we reviewed in the previous lesson

Correlation after Refinement

Target: Winning Bid

Attribute	Correlation to Target (%)	Attribute Type
isitMovable	1%	Binary
ZipCode	33%	Categorical
hallmarkName	5%	Categorical
itemCategoryName	<1%	Categorical
itemMaterialName	7%	Categorical
typeName	5%	Categorical
itemname	50%	Text

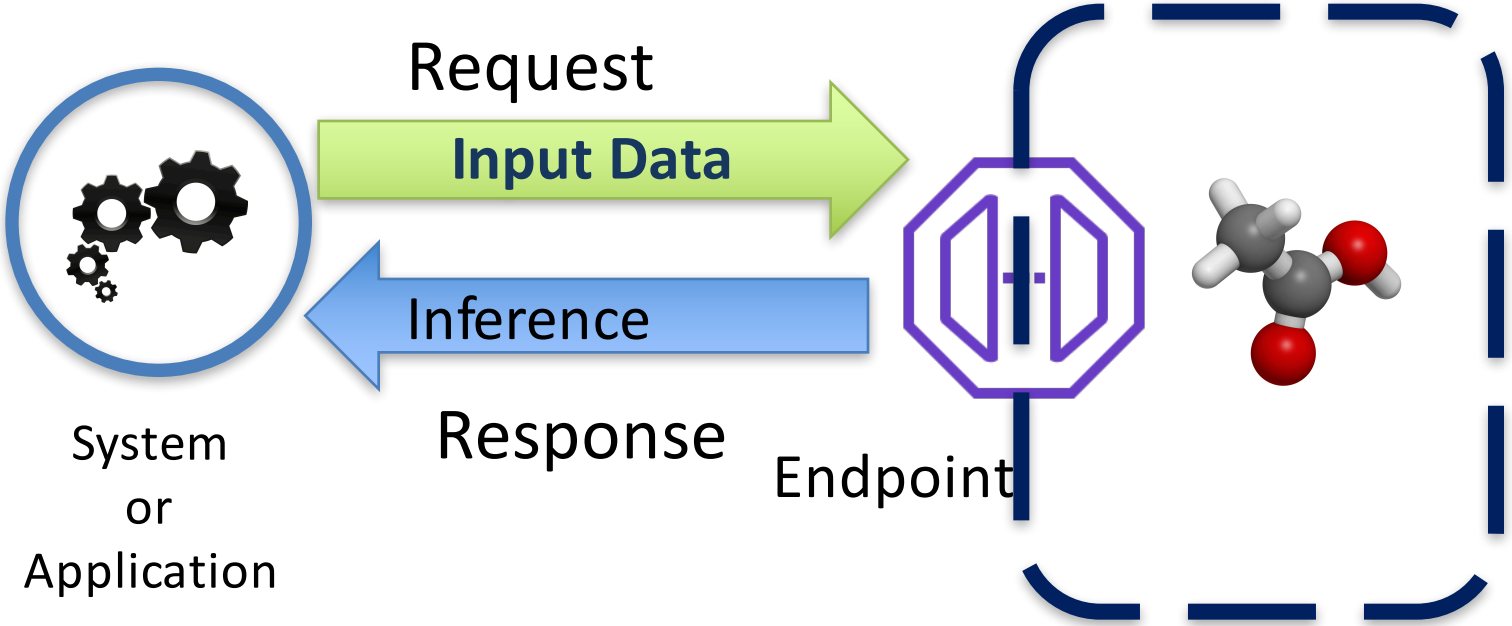
3 Types of Learning Data



- If you don't provide validation data, the algorithm uses a sample of the training data to calibrate
- If you provide test data, the algorithm logs include the test score for the final mode

Predictions

Prediction Overview



What are the Types of Predictions?

- Amazon SageMaker hosting services
- Persistent Endpoint
- One at a time
- Online
- Synchronous
- Low latency: Response in milliseconds

Real Time Inference

- Amazon SageMaker batch transform
- Many at once (entire dataset)
- Asynchronous
- Response minutes-hours
- Can be used to pre-process data

Batch Predictions

Segment 8: Selecting the Appropriate Data

What is the Best Kind of Data?

Academic Sources for Data

Commercial Sources for Data

Data Sources Supported by AWS

What is the Best Kind of Data?

- Generalizable data
  - Data that is applicable to a group or class of things
  - E.g. day of week, month
  - Should not be specific – a person's signature, exact time occurrence
- Complete data
  - Missing or invalid data is not helpful
- Properly formatted data
  - Follow formatting rules when feeding in data
  - Formatting rules depend on algorithm type

How Much Data is Needed?

- It depends
- Determined by similar problems
  - Examine similar situations and the amount of data that collected
  - Examine number of observations or rows
- Can be determined if you are an expert or if you consult an expert on the space
- Depends on how the problem is framed

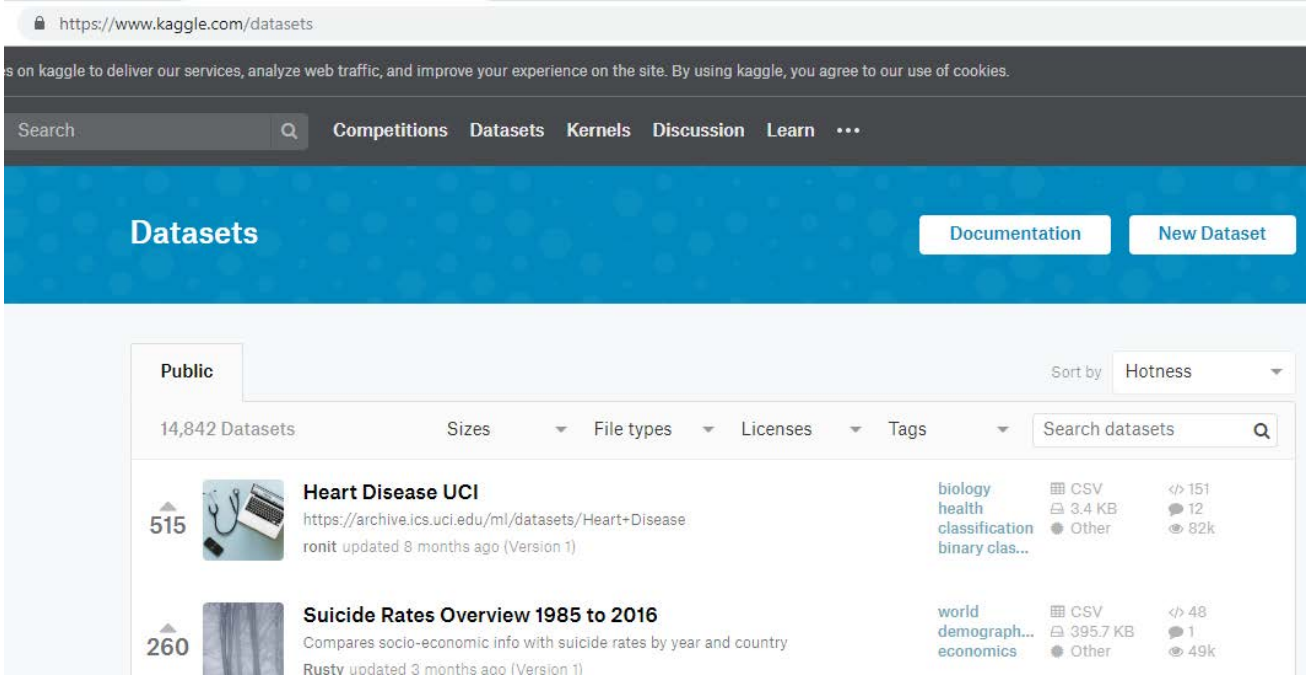
How to Determine Which Features to Select

- Try to select a variety of different “views” on the features
- You could consult an expert
- You may choose to create additional features by integrating or aggregating with sum, counts, etc.

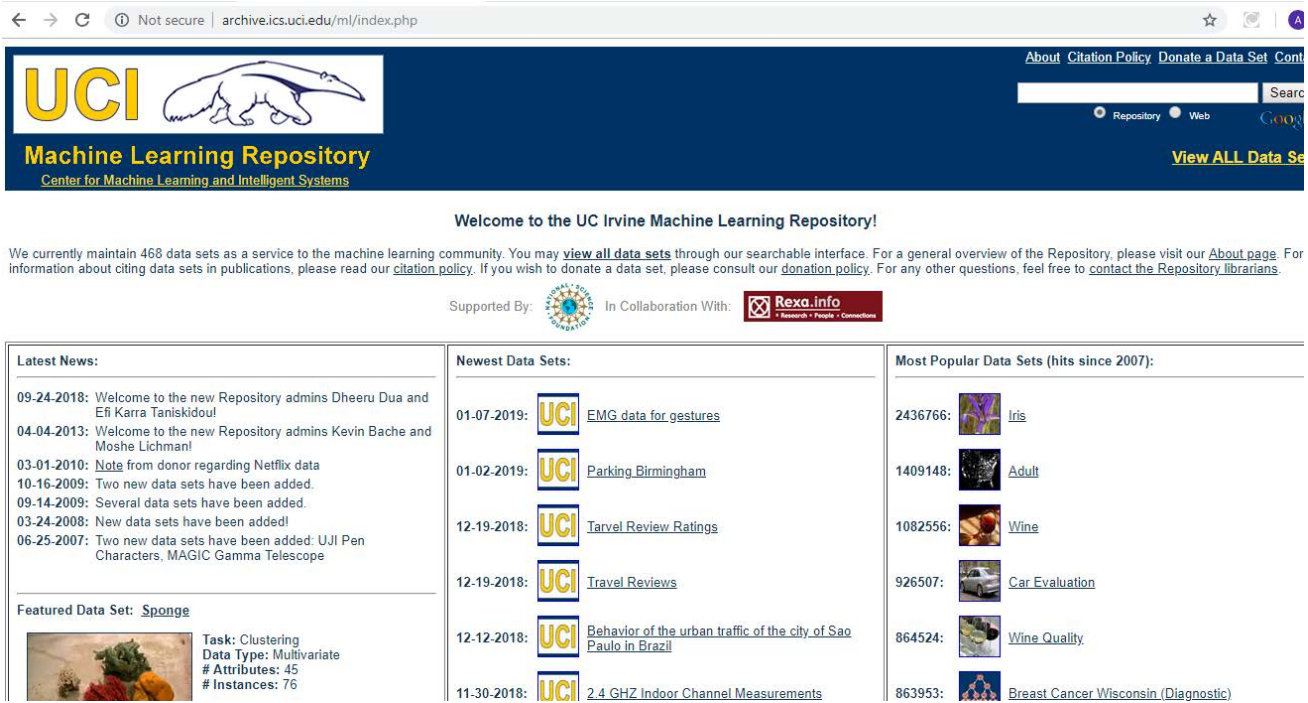
Where can you get sample data?

Online Repositories

Kaggle



UCI



Wine Quality

UCI

About Citation Policy Donate a Data Set Contact

Search

Repository Web

Google

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

View ALL Data Sets

Wine Quality Data Set

Download: Data Folder, Data Set Description

Abstract: Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests (see [Cortez et al., 2009], [Web Link]).

Data Set Characteristics:	Multivariate	**Number of Instances:**	4898	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	12	**Date Donated**	2009-10-07
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	969006

Features – Wine Quality

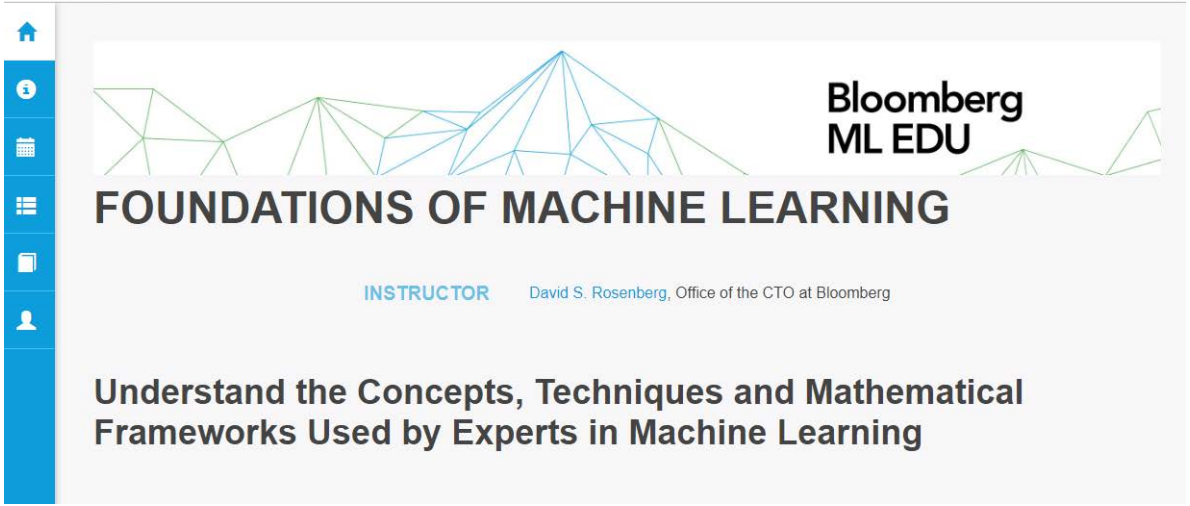
Data Set Information:
The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult: [Web Link] or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are munch more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

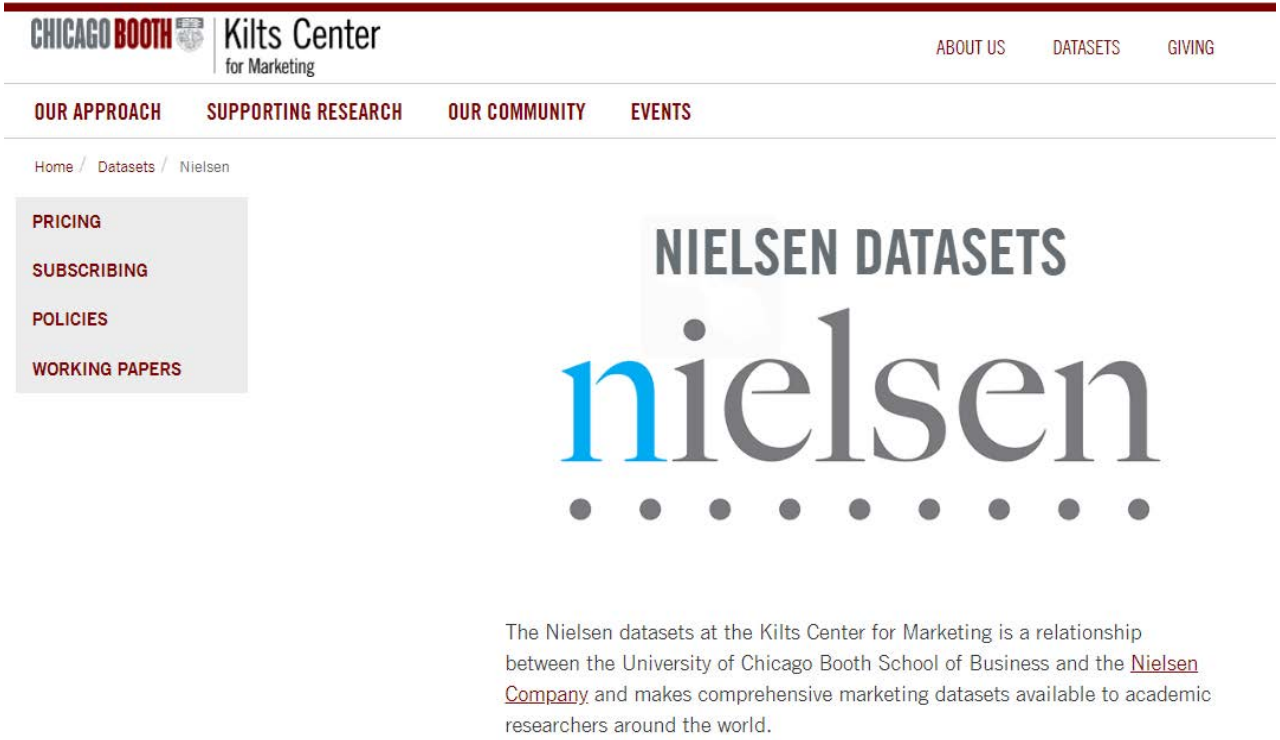
Attribute Information:

For more information, read [Cortez et al., 2009].
Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

Bloomberg Resources



Nielsen Academic DataSets



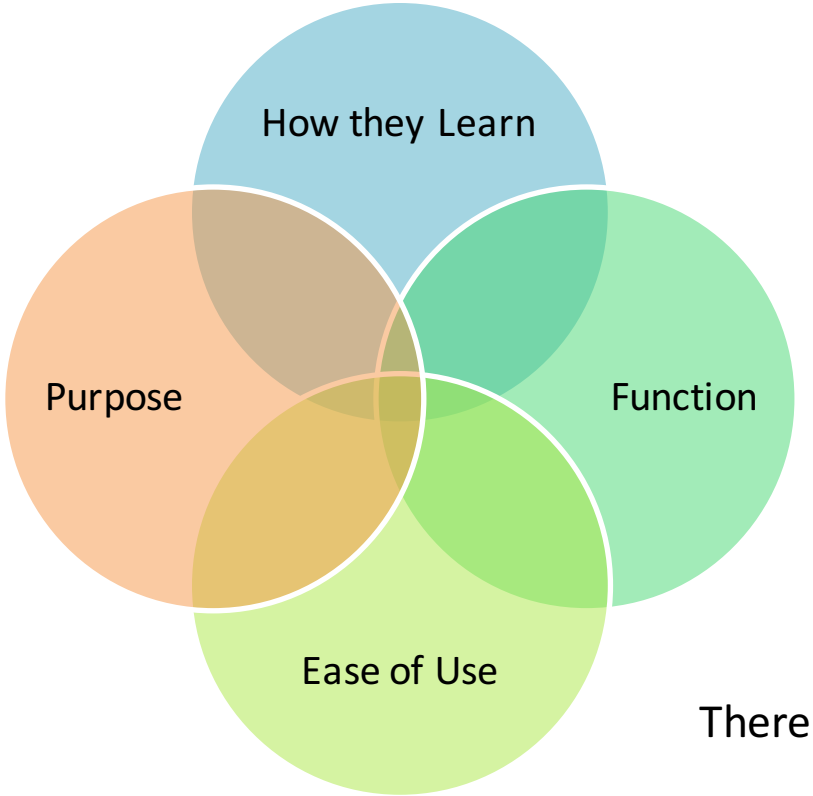
Other commercial sources

Dun and Bradstreet	Benefits: detect fraud, enhanced customer insight, identify growth opportunities. The Dun & Bradstreet Data Cloud.
Planet.com	Image tracking of earth, any point on earth, 72 milimeters

Segment 9: An Introduction to Algorithms

- Machine Learning Algorithm Families
- Common Algorithms
- Use Cases for Popular Algorithms
- Built in AWS Algorithms Provided by AWS

Machine Learning Algorithm Families



There can be overlaps due to similarities

Algorithms by How They Learn

Supervised

- Classification – predicts discrete answers (groups)
- Regression - predicts continuous value outputs
- Ensembling - combines predictions of multiple ML models that aren't strong enough on their own

Unsupervised

- Association
- Clustering
- Dimensionality Reduction
  - Feature selection – subset
  - Feature extraction – data transformation

Built in Algorithm Types by Ease of Use

Identifying an algorithm based on how easy it is to learn

Easy

- Linear Regression
- Classification
- AWS High Level Services

Advanced

- Sequence to Sequence
- LGA

Built in Algorithm Types by Purpose: Samples

Identifying an algorithm based on its purpose

- Image Classification
- Sequencing
- Topics in a document
- Neuro topic model

Common Algorithms by Function

Function	Sample Popular Algorithm
Regression	• Linear
Dimensionality Reduction	• Principal Component Analysis (PCA)
Classification	• Logistic Regression
Clustering	• K-Means

Linear Regression

Purpose

- predict continuous value outputs

Sample predictions

- Price
- Age

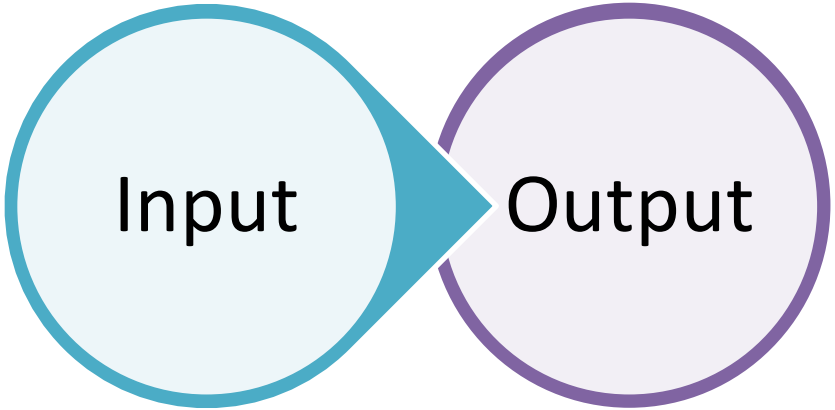
Technique

- Shows relationship between variables
- Correlation between target and input features
- Measures impact of one variable on another

Etymology

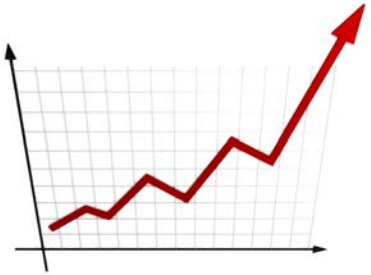
- Linear – line
- Chart x, y axis
- Change is constant

Linear Regression with Curator



- Material: Gold Plate
- Theme: Vacation
- Maker: Monet

- Price: $25



Logistic Regression

Purpose

- Can predict binary outcomes
- Multinomial logistic regression for multi-class

Sample predictions

- Is this spam?
- Will I pass the test?

Technique

- Decision Trees
- Regression b/c it finds relationships b/w variables

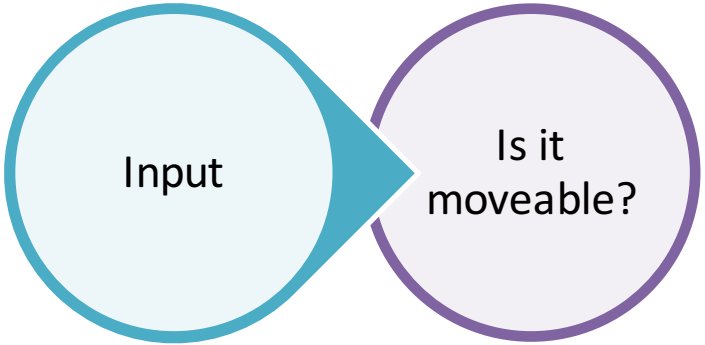
Etymology

- Logistic – runs a logistic function
- Can't chart x, y axis

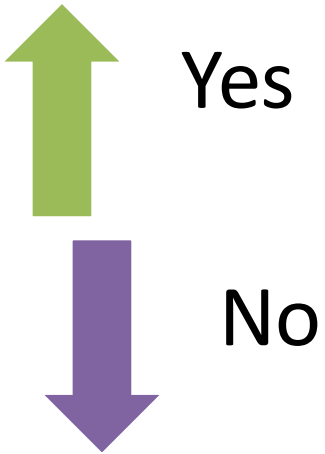
Learning Type

- Supervised

Logistic Regression with Curator



- Material: Gold Plate
- Theme: Vacation
- Maker: Monet



Random Forest

Purpose

- Predicts classification as primary use case and also can do regression
- Trains quickly

Sample predictions

- Image classification or object detection
- Identifying loyal customers
- Likelihood for a customer to like a product

Random Forest

Technique

- Ensemble approach – divide and conquer – weak learners on their own combined to become strong
- Forest is multiple decision trees – each classifier is a “learner”
- Each tree gets a classification
- Calculates votes by tree for particular class
- Information Gain
- Decision Tree Algos

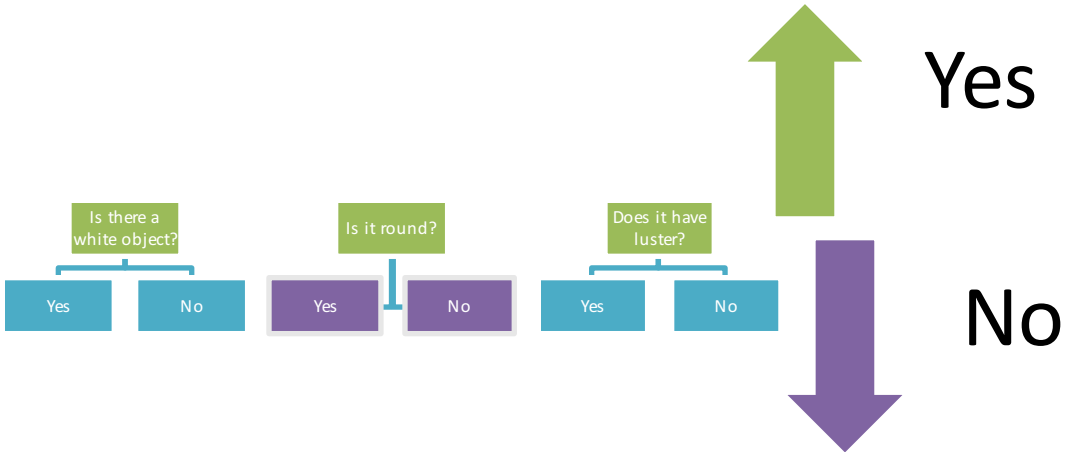
Etymology

- Averages multiple trees

Learning Type

- Usually supervised

Random Forest with Curator



Does this charm image contain a pearl?

Segment 10: Predictions

What are Predictions or Inferences?

How do Predictions Work?

What are the Types of Predictions?

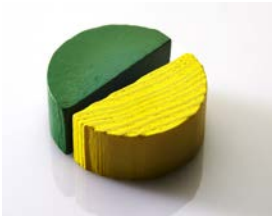
Regression Batch Predictions

- Single score in one column
- Contains raw number
- One row for each observation
- Very simple

Multiclass Batch Predictions

- Multiple columns
- One column per class
- One column as best answer
- Provides scores for each class
- Evaluates each class as a binary

Binary Predictions



- Results in two columns
- Best answer and score
- Score serves as probability
- Best answer is the True/False

Key Tenets for Better Predictions

- Know which features to include
  - There are LOTS of features available in our database but we will only select a subset
  - Experiment with correlation to target
  - Know which are meaningless – e.g. Primary Key
- Correct the data
  - Invalid data, missing data, outliers
- Transform those features
  - Break apart features
  - Group features
- Examine correlation and adjust

Segment 11: A Closer Look at AWS ML Services

- Amazon SageMaker High Level Overview
- Key Components of Amazon SageMaker
- How to Get Started with Amazon SageMaker

What is Amazon SageMaker?

- AWS' goal: "put machine learning in the hands of every developer"
- Readily train, build & deploy machine learning models with scale
- Democratize ML so that you can GTM faster
- Enable multiple roles (e.g. dev, data scientist) to work the same project at the same time
- No need to worry about the infrastructure or manage servers

What is Amazon SageMaker?

- Enables data scientists and developers to "quickly and easily build and train machine learning models, and then directly deploy them into a production-ready hosted environment" (AWS Documentation)
- Leverages data to make better business decisions and applies algorithms to this data
- Contains out of the box machine learning algorithms, including the most popular ones

Who Should use Amazon SageMaker?

- Developers
- Data Scientists
- Technologists
- Those who cannot answer their question using regular analytics or math

What are the Benefits of Amazon SageMaker?

- Easier maintenance versus rules-based engines– rules are typically hard-coded whereas machine learning is dynamic
- Pattern recognition
- Scalable
- Pay as you go – can start small and grow

re:Invent 2019 Machine Learning Updates

New Features re:Invent 2019

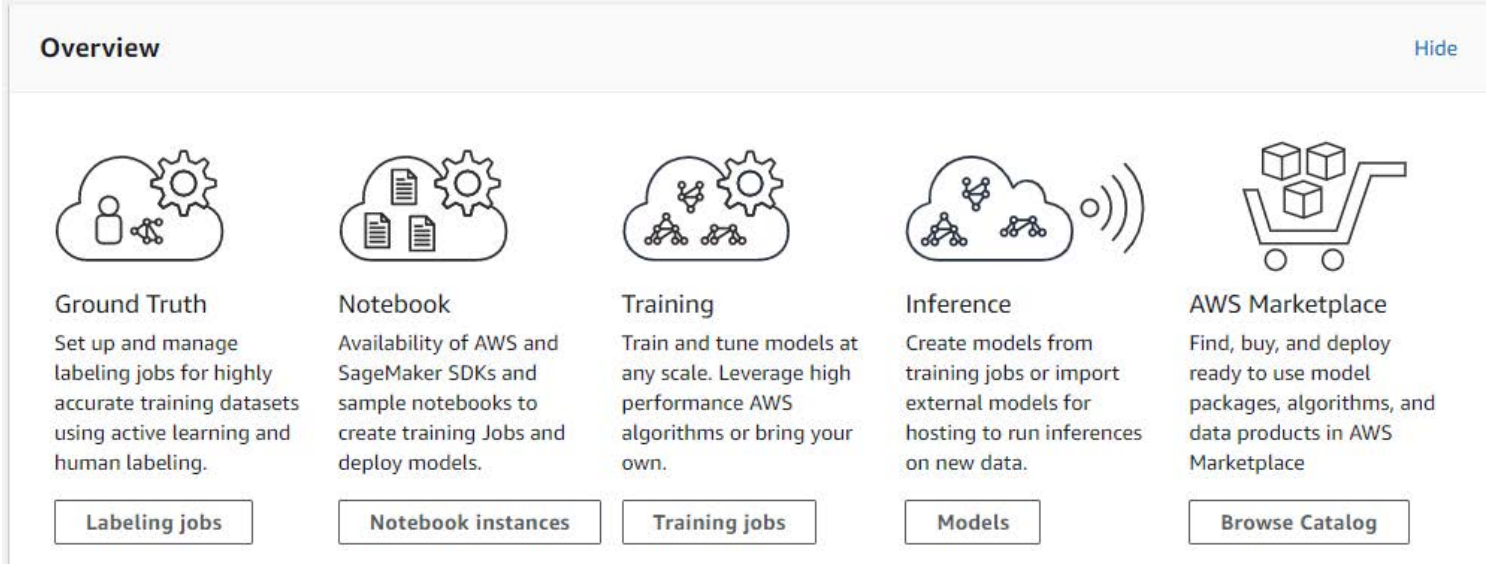
- Ground Truth Data Labeling
- ML Marketplace
- SageMaker Studio (NEW)
- SageMaker Neo

SageMaker Studio - IDE

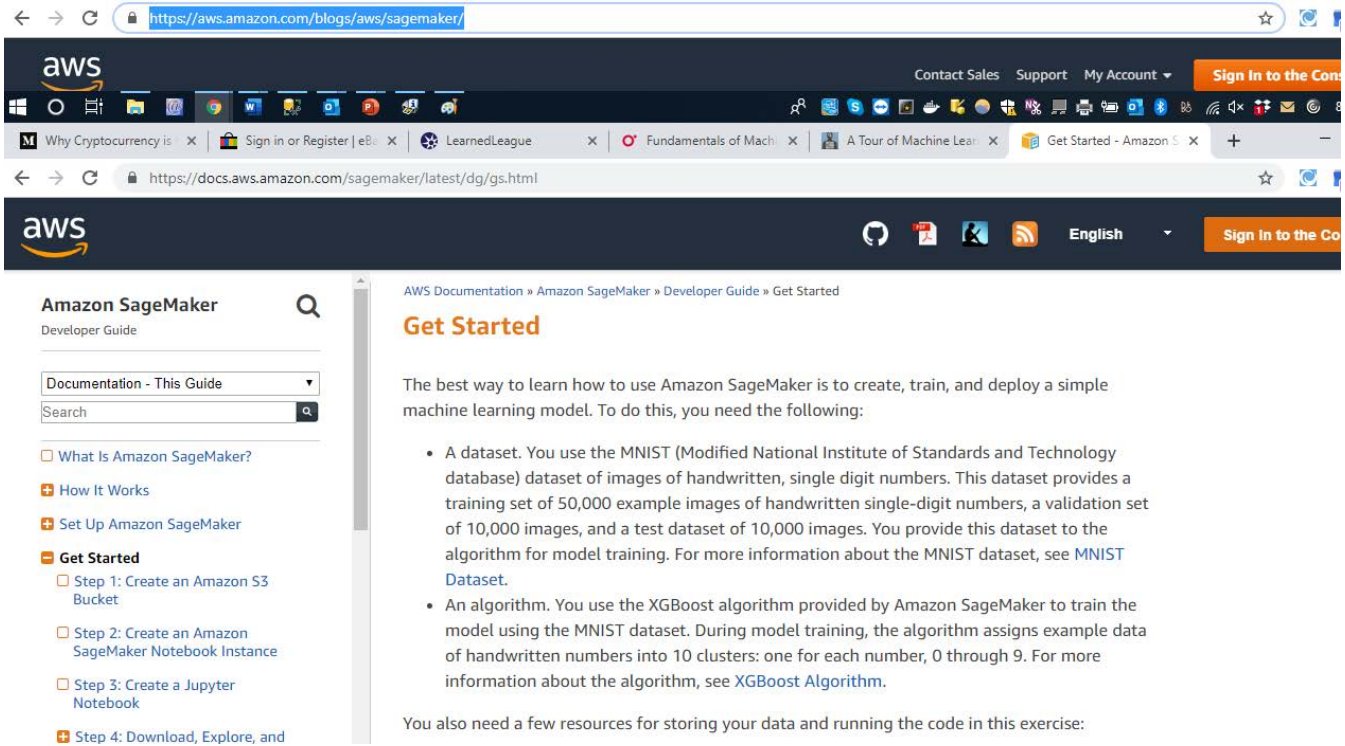
- Built in Algorithms
- SageMaker Notebooks (NEW)
- SageMaker Experiments (NEW)
- Model Training
- SageMaker Debugger (NEW)
- Model Tuning
- SageMaker Autopilot INEW)
- Model Hosting
- SageMaker Model Monitor (NEW)

Amazon SageMaker High Level Overview

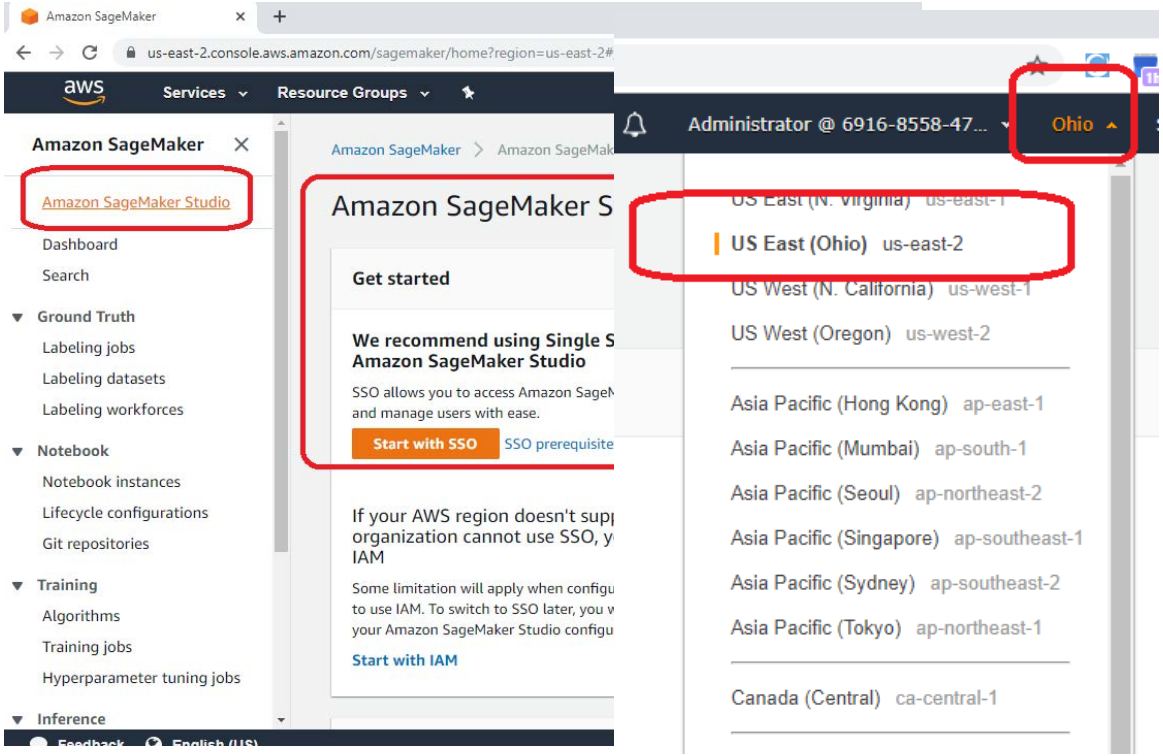
Reference: SageMaker Console



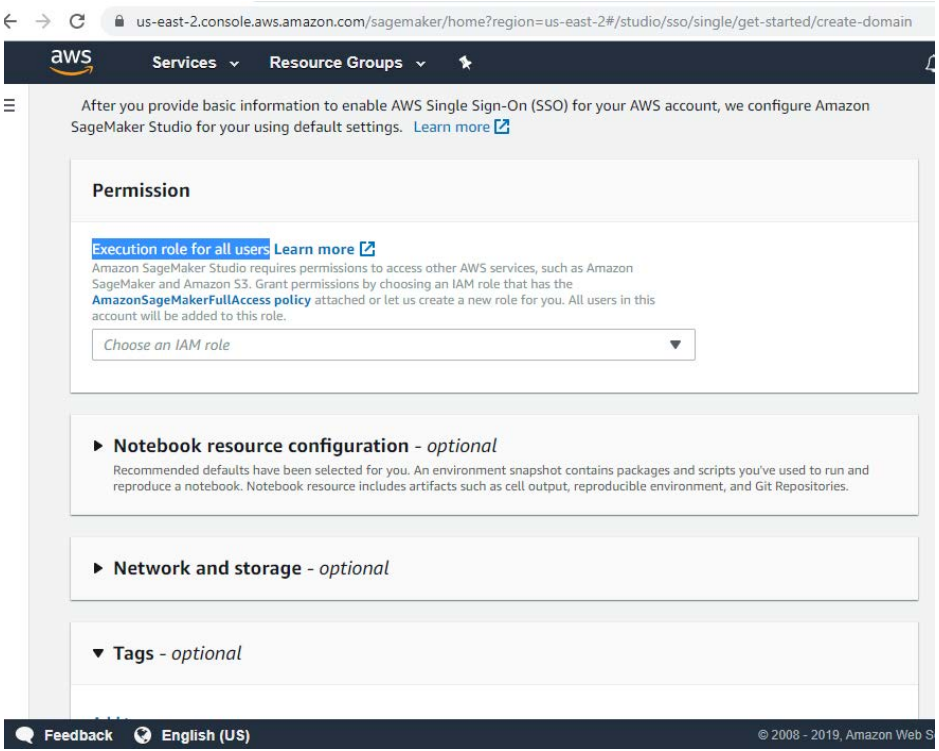
How to Get Started



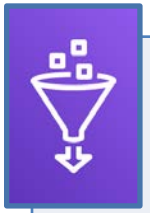
NEW: Amazon SageMaker Studio



NEW SLIDE: SSO with Studio



Options for Data Sources



AWS Glue

Amazon Athena

Amazon EMR

Amazon Redshift

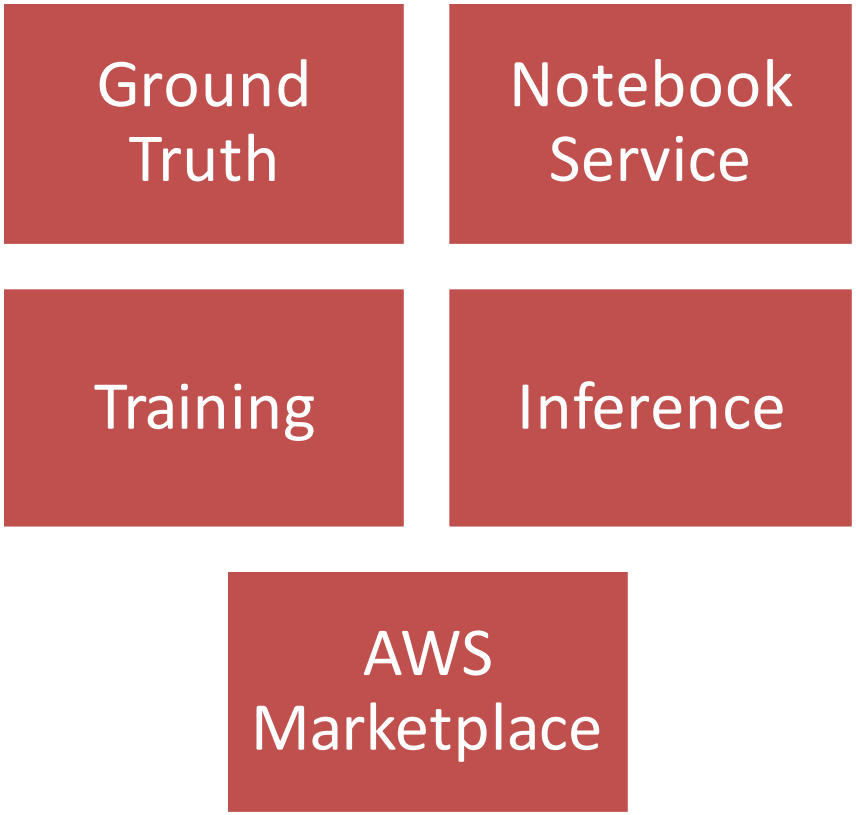
Amazon S3 (easiest way to begin)

Amazon Kinesis (e.g. Formula One streaming data)

When to Use Amazon SageMaker

- Try the high level services first
- They are easy to use, most are simply an API call away
- If not enough, click down one level to SageMaker to build something more customized

Key Components in the SageMaker Console



Amazon SageMaker Notebooks Service

With a Jupyter notebook on an Amazon SageMaker notebook instance, you can….

- Set up S3 and IAM
- Pre-process data
- Transform data
- Create training jobs
- Deploy models
- Evaluate models
- Generate Predictions
- Kill completed services and assets

Amazon SageMaker Training Service

The training service enables models to be "taught" to make predictions via….

- Training jobs
- Input data – training data
- Compute resources
- Model artifacts
- Training code

Amazon SageMaker Hosting Service

The Hosting services enables you to deploy our model so that you can….

- Generate inferences or predictions
- Push out predictions (inferences) via HTTPs endpoints
- Integrate endpoints into your app
- Optimize for subsecond latency

Benefits

- No minimum fees
- No upfront commitments
- Pay by the second

Manage your Costs Wisely

- Training jobs automatically terminate when done
- * Stop your Jupyter Notebooks when complete
- Delete unneeded S3 Storage

Amazon Free Tier Benefits

Free SageMaker hours

First 2 months

- 250 hours of medium instance size
- 50 hours of large
- 125 hours XL

Begins when you create first notebook

Segment 12: Call to Action & Conclusion

Summary

Next Steps

References

Summary

- What is machine learning
- Key taxonomy
- Data sources
- Workflow
- Models and algorithms
- Making predictions

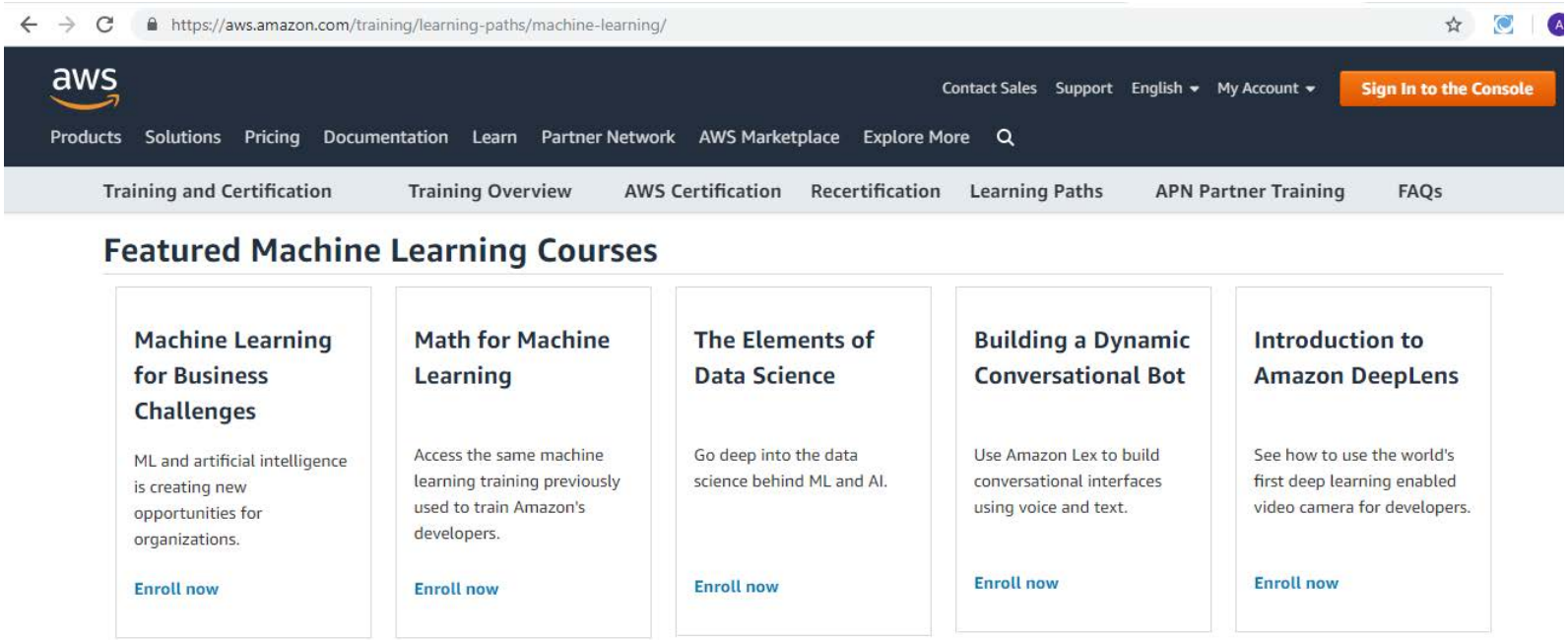
SageMaker Developer Guide

https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf

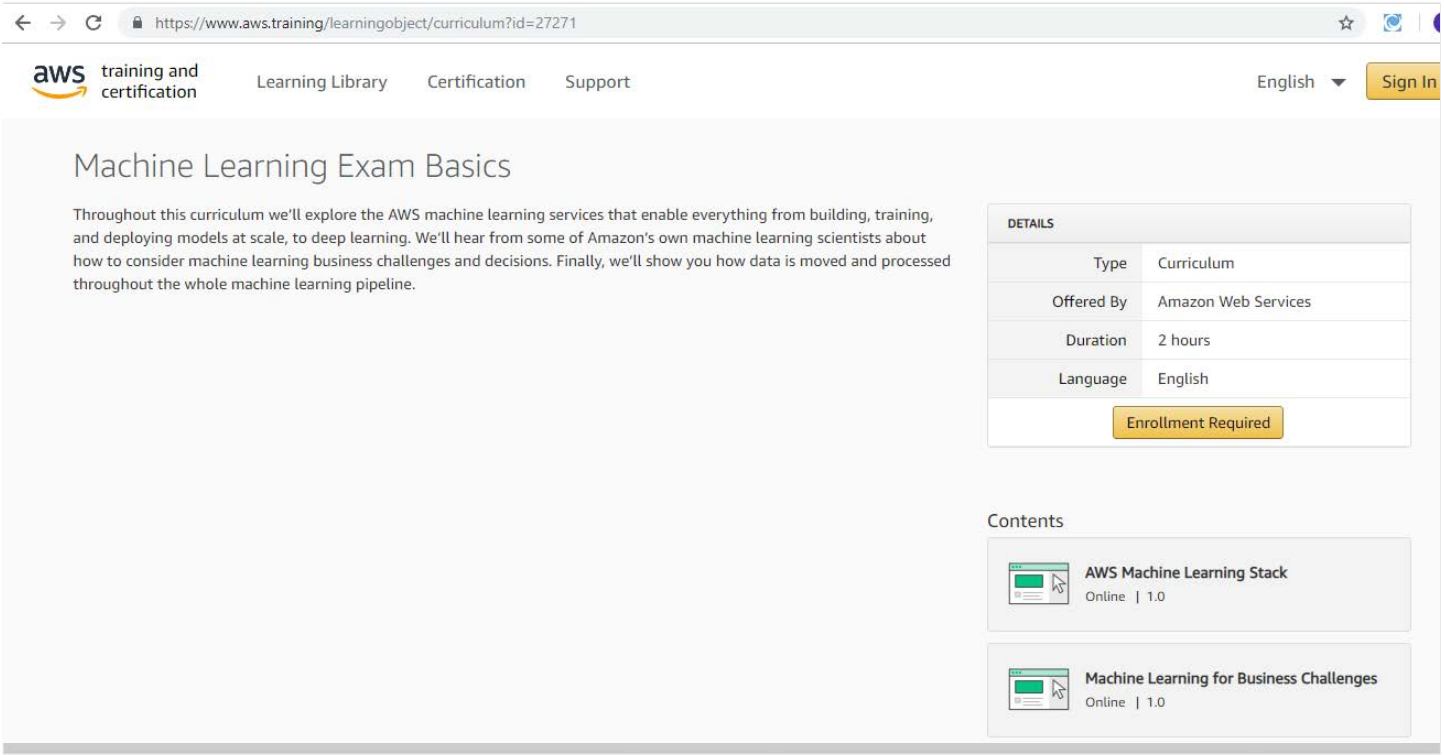
Table of Contents

AWS DeepRacer League

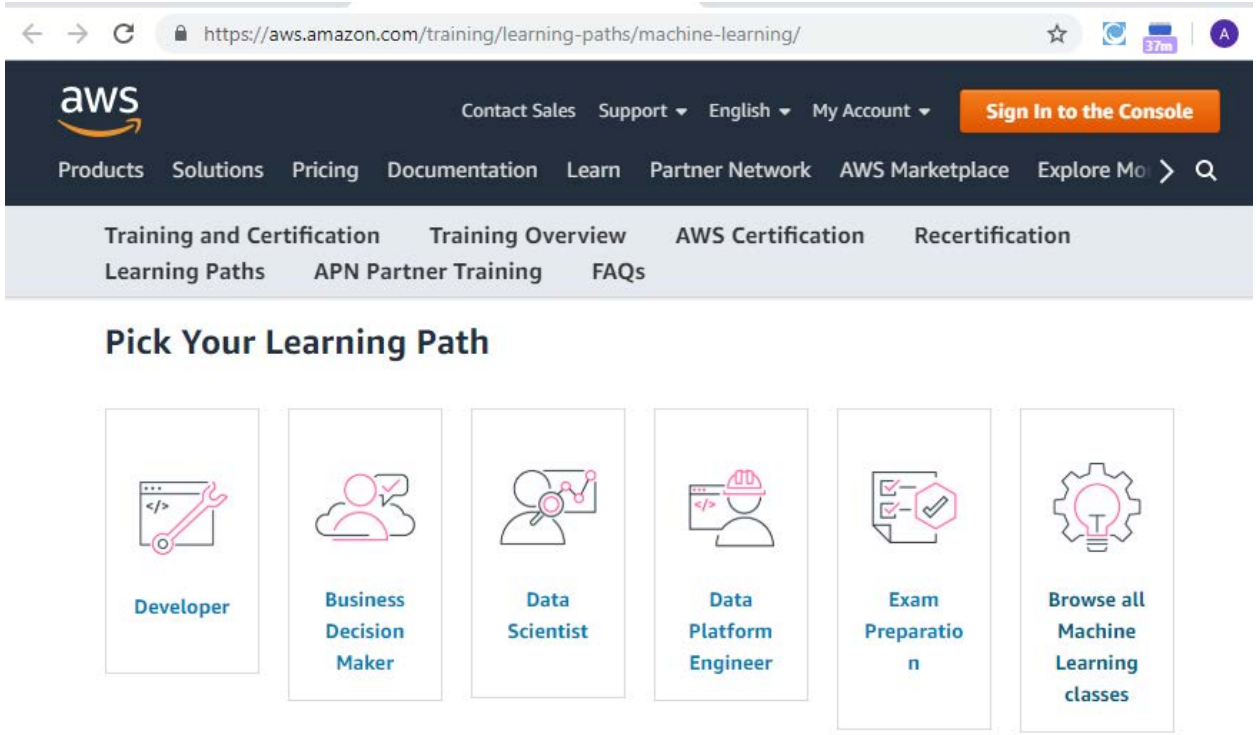
Courses



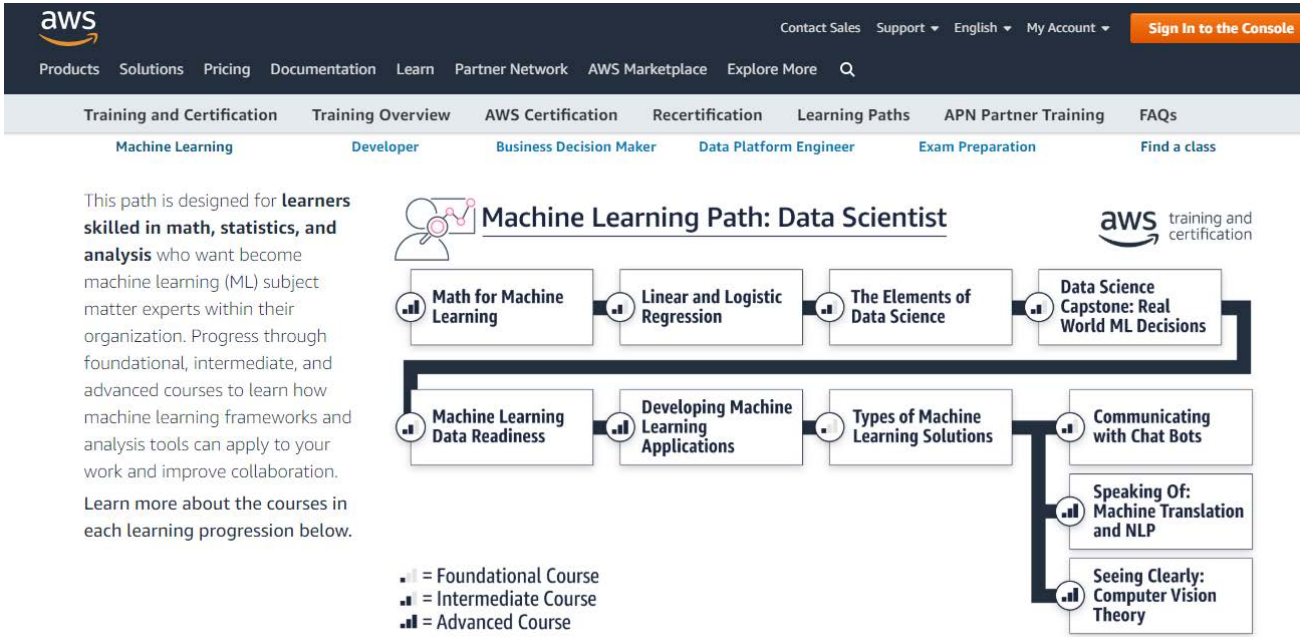
Certifications



Amazon Machine Learning Paths



Example: Data Scientist



Thank you!

Questions?