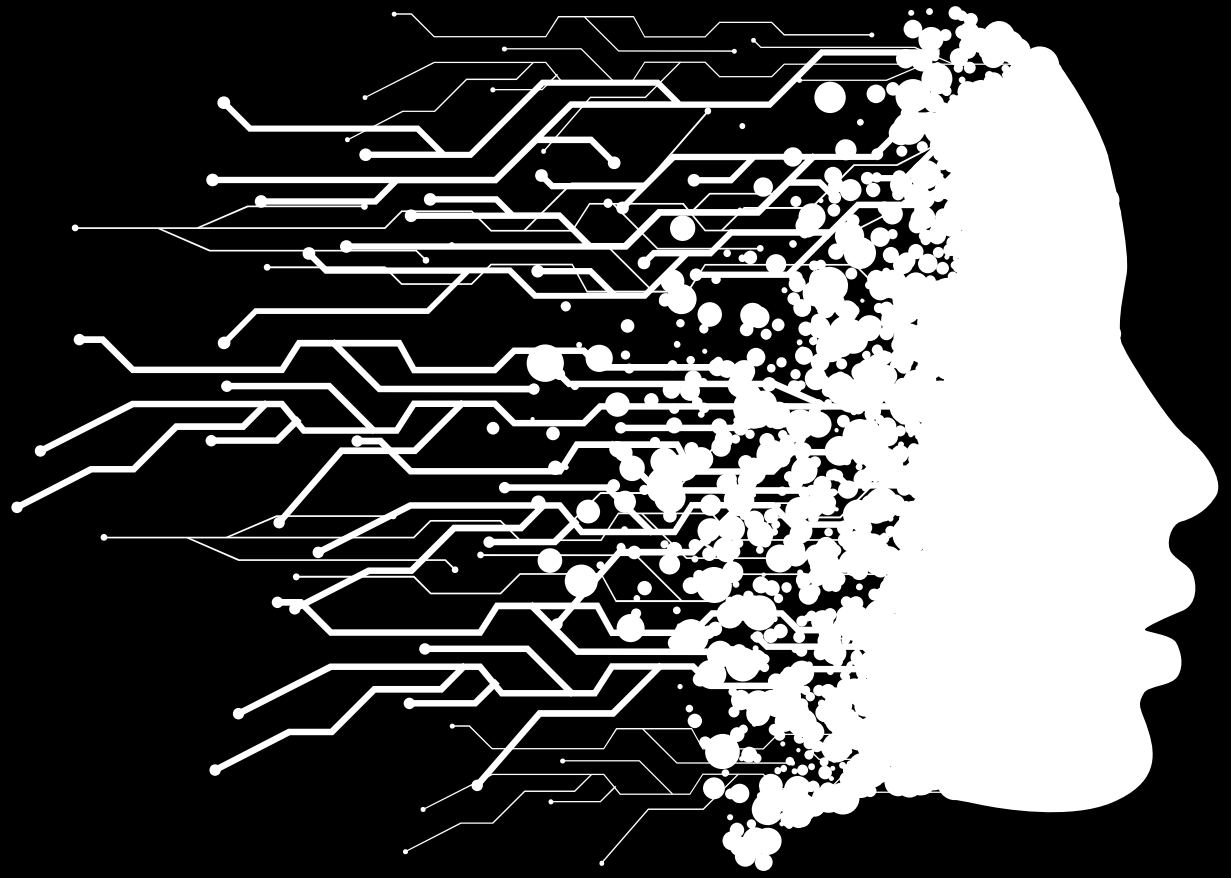


Machine Learning With Spark (MD188)

We will be starting soon

Welcome to Day 1 Spark Machine Learning

Please perform PRE-WORK

1. Access the virtual Lab using link https://html.inspiredvlabs.com Use the username TEKMD190-XX (replace XX with your number) and password **TEKmd189!23**

2. Download material for the Lab:

https://tinyurl.com/bdtSparkML

Last Name	First Name	User-Name
ALAMELU KRISHNAN	SURESH	TEKMD190-01
BRAGG	JAMES	TEKMD190-02
GARIMELLA	HIMABINDU	TEKMD190-03
JAJOO	SANDESH	TEKMD190-04
JOSHI	CECIL	TEKMD190-05
LANNERS	QUINN	TEKMD190-06
MADAS	ANANT	TEKMD190-07
MAXSON	CRAIG	TEKMD190-08
MISRA	SMARAN	TEKMD190-09
PASUMARTHY	VENKATA	TEKMD190-10
POUDYAL	BASHUDEV	TEKMD190-11
RUTHERFORD IV	ROB ROY	TEKMD190-12
SAMALLA	APARNA	TEKMD190-13
SCHLEY	DANIEL	TEKMD190-14
VONGSOUVANH	BOUAPHAYCHIT	TEKMD190-15
ZHENG	LEI	TEKMD190-16

We will be starting soon

Logistics

- Timing – 8:30 AM to 4:30 PM CST
- Lunch – Approx. noon to 1 PM CST
- Periodic Breaks – At logical points

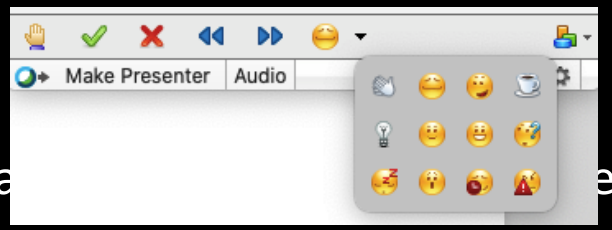
Virtual class tips

Audio

Please mute yourself if not speaking but please please unmute to ask questions.

Notifications

Give us green right check mark in participa



e.

✓

Please put a smiley emoji in participant panel when you are back.

🙂

Please raise hand in participant panel or put on chat or mute & ask question.

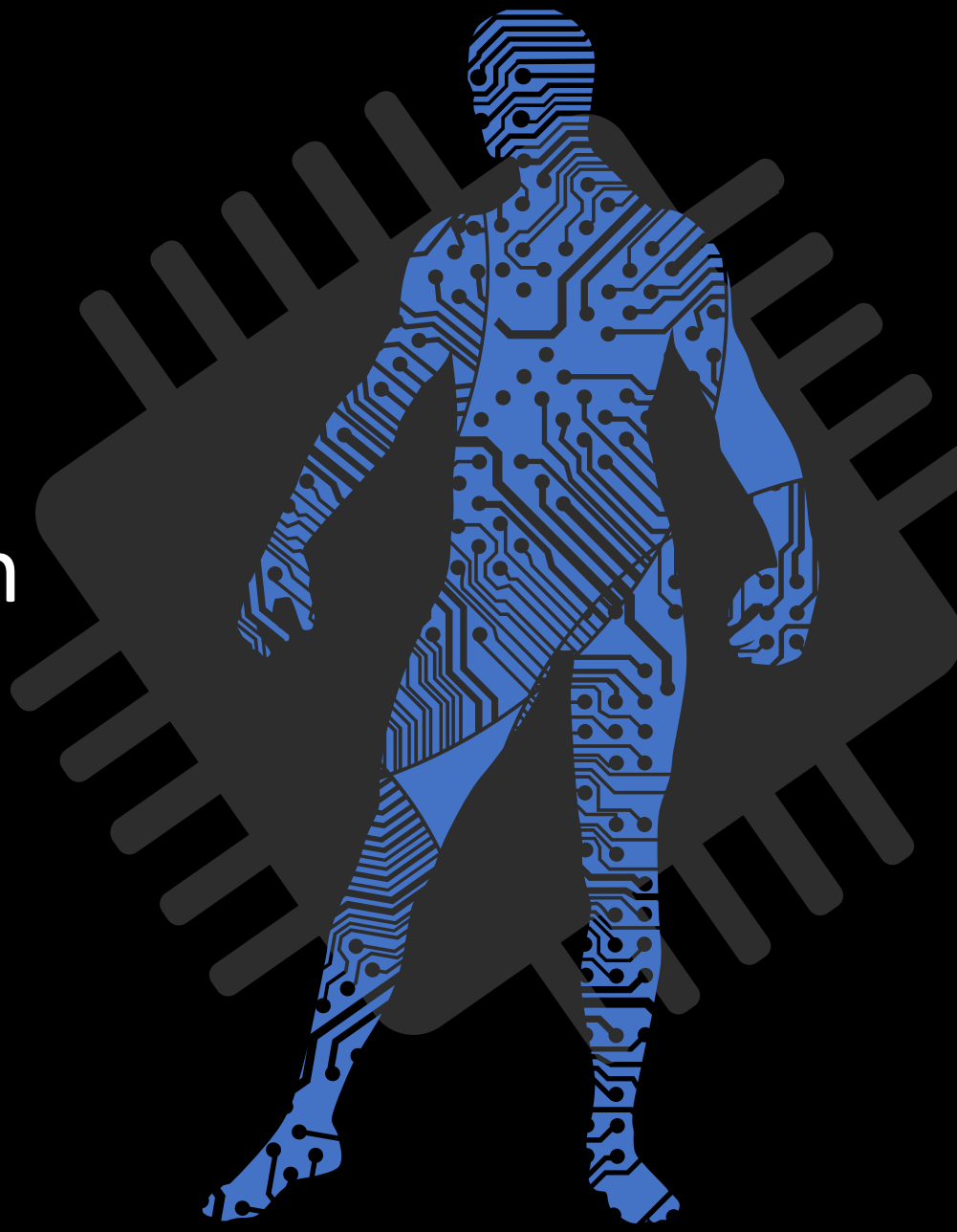
Chat

Please send chat to “All Participants” unless you have some answer or private info to send to “All Panelists”

Let’s make it interactive !

Agenda – Day 1

1. Introductions
2. Machine Learning Introduction
3. Spark Essentials
4. Spark Development (ML)
5. Python Introduction
6. Machine Learning Techniques

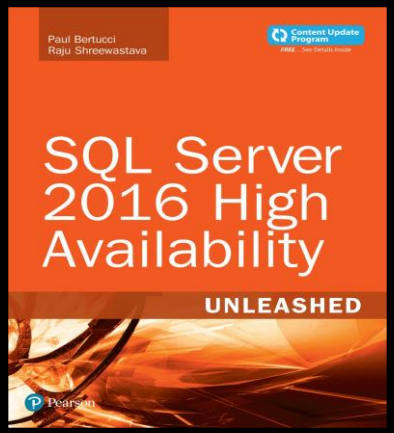


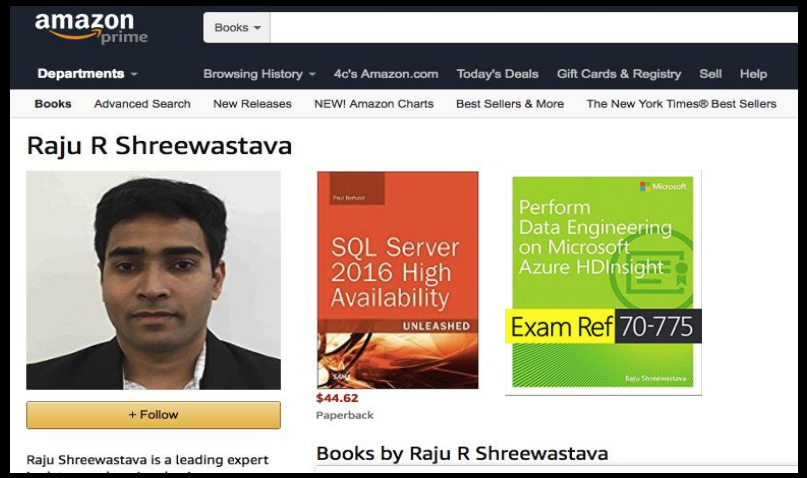
Introduction

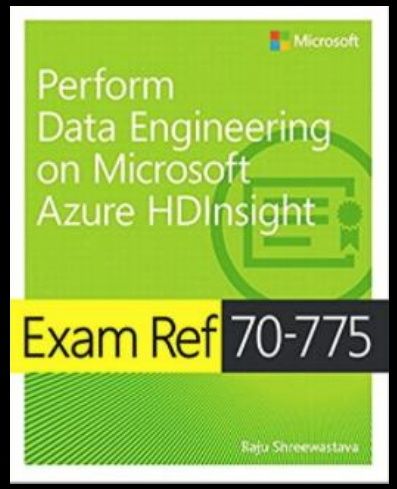
- Raju Shreewastava
- 20+ yrs. in Data & Analytics
- Live in California
- Passion for teaching
- Presenter in several conferences

Author

- Big Data High Availability (Pearson Publication)
- Perform Data Engineering on Microsoft Azure HDInsight (Microsoft Press)







Sanjay Oza

- 25+ years in Software and Data Engineering
- Multiple years of management experience in software and hardware development
- Worked in networking industry
- Like to teach and mentor
- **Patents**
  - Detecting and mitigating a high-rate distributed denial of service - Approved
  - Individually assigned server alias address for contacting a server - Pending

Your turn

Who? Name, Role & Group

What? Overall experience & Main Skill (Any Python or ML experience?)

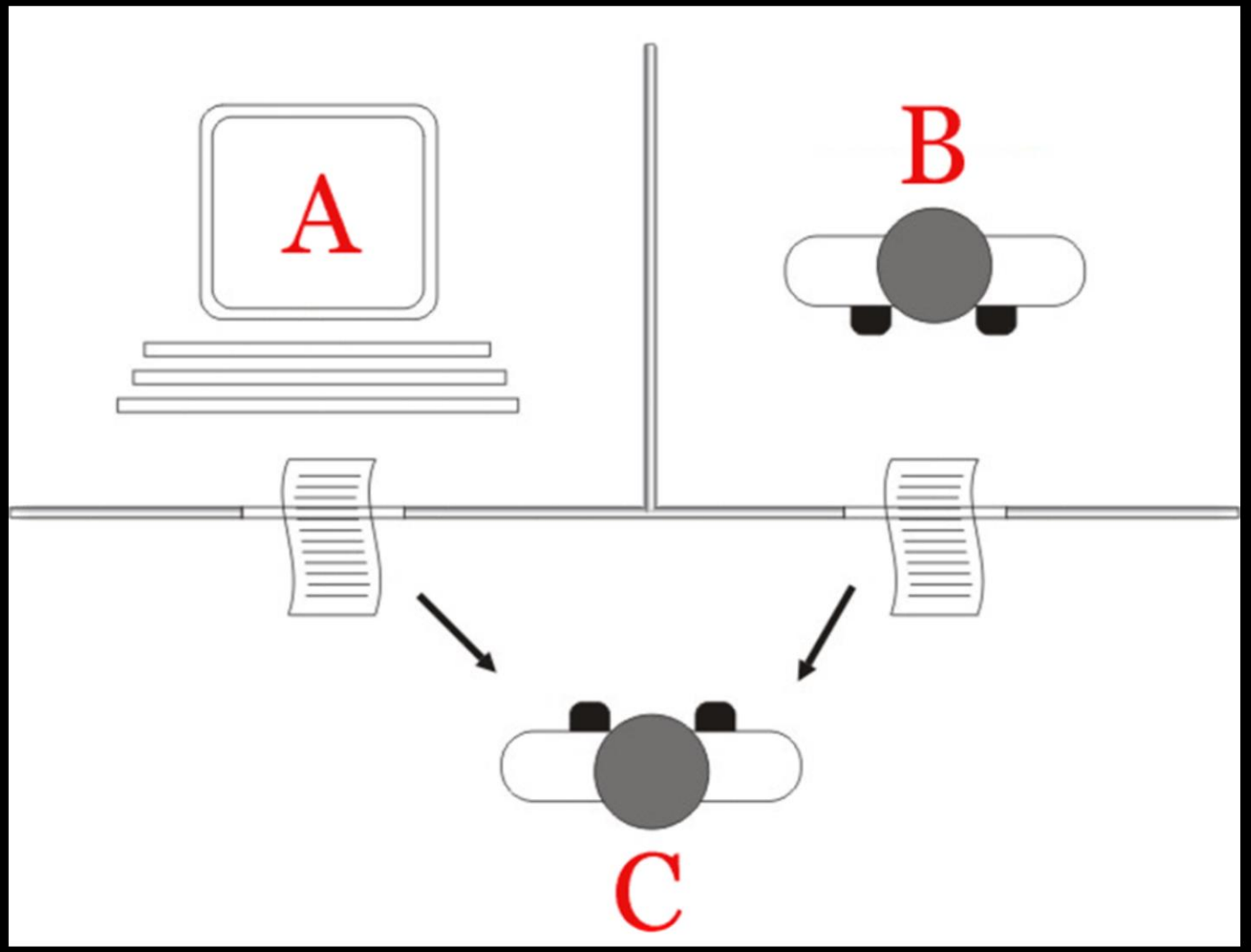
Why? Are you here (Key Motivations)

How? You can plan use this (if already known)

What is Machine Learning?

Your thoughts, perceptions and myths

1950 – Turing Test



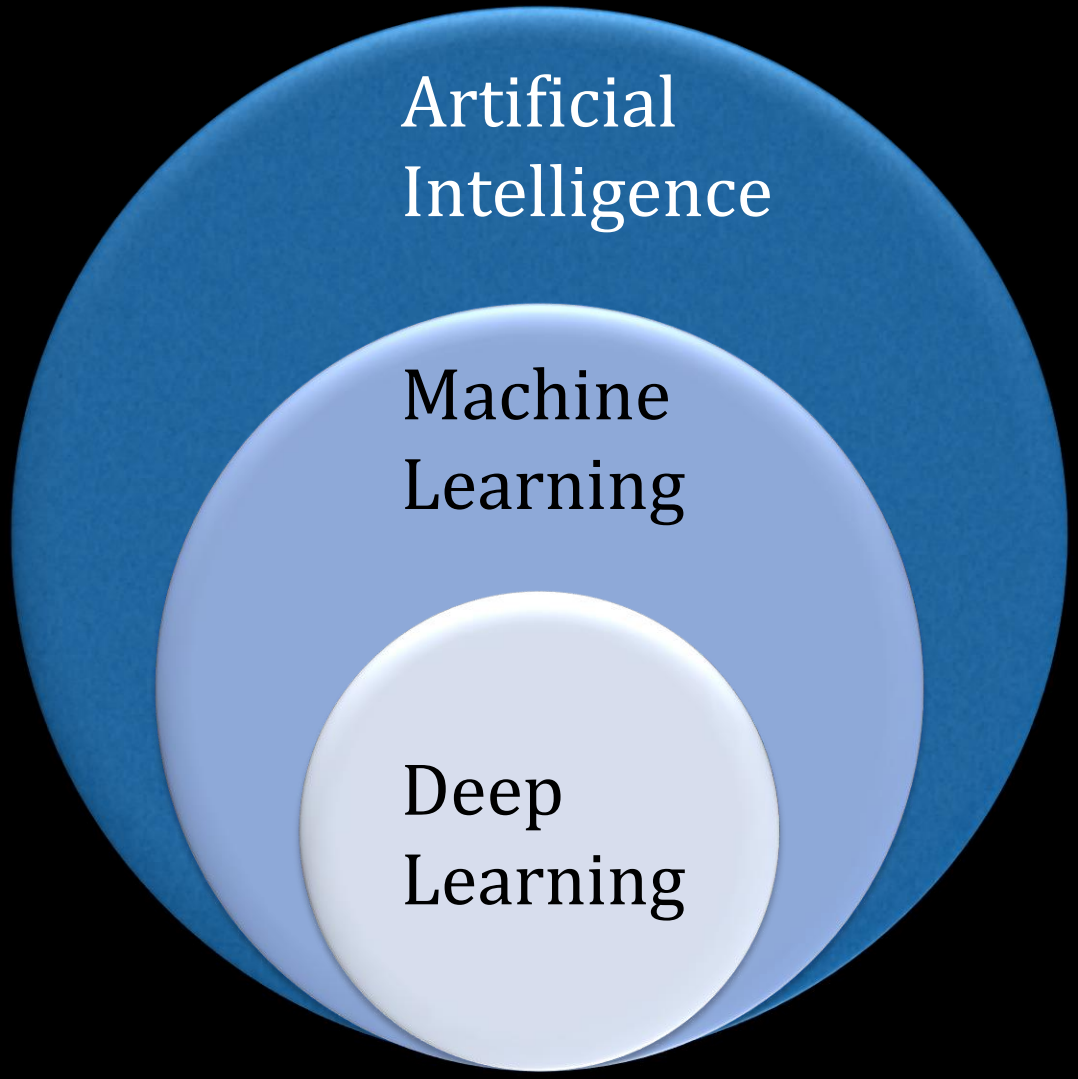
History of Machine Learning

- 1950 – Turing Test
- 1951 – First Neural Network
- 1967 – “Nearest Neighbor” Algorithm is written
- 1974 – First AI winter
- 1979 – Stanford Cart
- 1997 – IBM Deep Blue beats Garry Kasparov

- 2006 - Geoffrey Hinton coins the term “deep learning”
- 2014 – Facebook develops Deep Face & Google Buys DeepMind
- 2015 – Amazon, Google & Microsoft ML Cloud offerings
- 2017 – Number of open source libraries for AI and ML

Source https://www.bbc.com/timelines/zypd97h

Machine Learning is type of Artificial Intelligence



Intelligent Behavior

Learning from data to make predictions and gain insights

Learning using Artificial Neural Networks

What is Artificial Intelligence?

The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making and translation between languages

Source - Wikipedia

Applications include:

- Neural Networks
- Game playing
- Natural language processing
- Robotics
- Speech Recognition

What is Machine Learning?

- Humans learn from our past experiences
- Machine follow instructions given by humans
- What if we humans can train machines to learn from the historical data?

Instead of programming a computer, you give a computer examples and it learns what you want

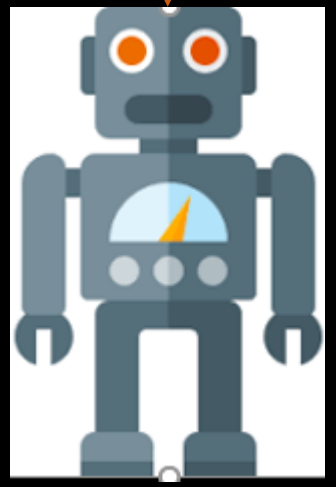
Machines Learn by Example

User ID	Gender	Age	Estimated Salary	Purchased
15624510	Male	19	19000	Yes
15810944	Male	35	20000	No
15668575	Female	26	43000	No
15804002	Male	19	76000	No
15728773	Male	27	58000	No
15598044	Female	27	84000	No
15694829	Female	32	150000	Yes
15600575	Male	25	33000	No

Learns from the data

New Input Data

Gender, Age, Estimated Salary



Purchase? Yes or No

Deep Learning

- How can we extract images of appliances from the picture?
- Deep learning is good at dealing with problems such as high dimension data
- Programming the models requires little guidance from the programmer

What is Deep Learning?

What is Deep Learning?

Wikipedia:

Deep learning is a set of algorithms in machine learning that attempt to model high-level abstractions in data by using architectures composed of multiple non-linear transformations.

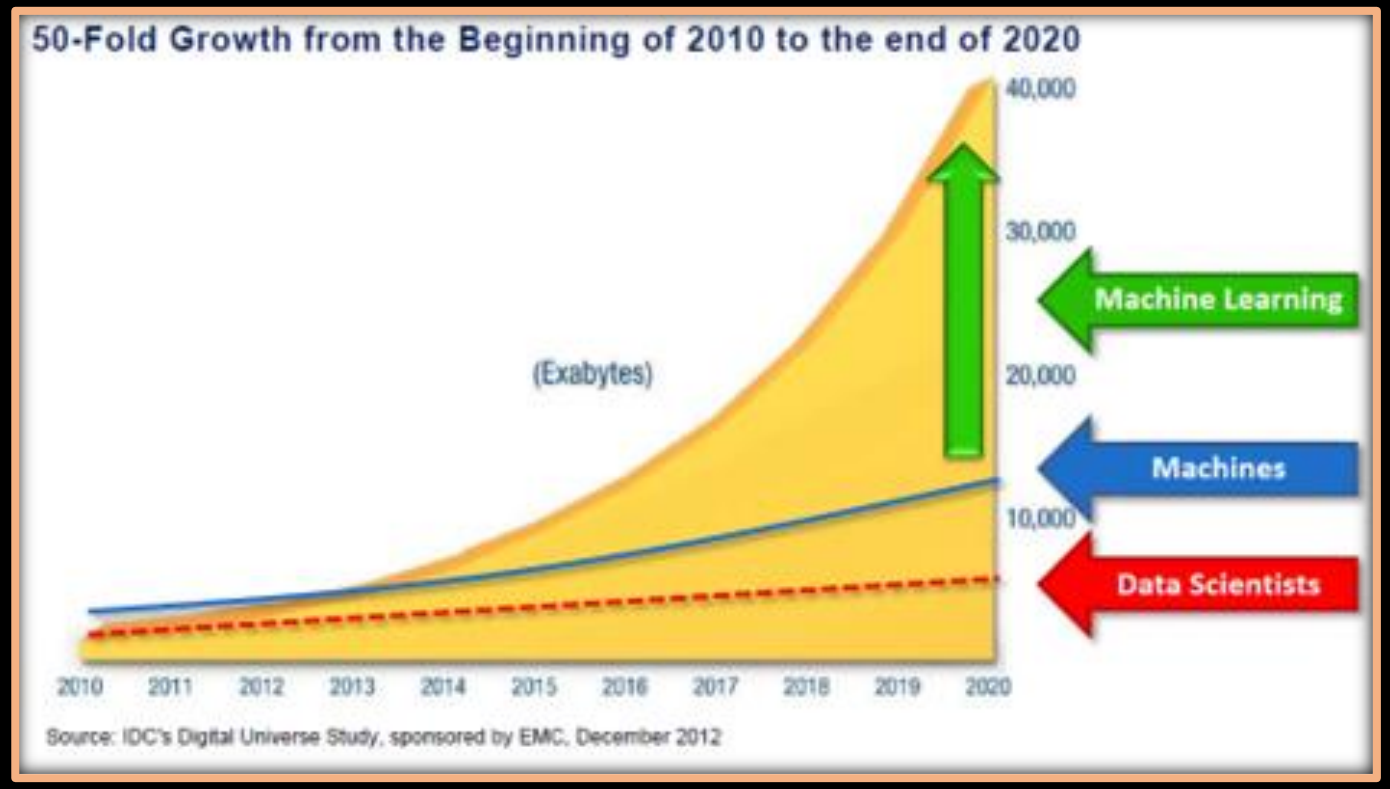
Example: Input data (image)

DeepFace rotating a celebrity's face to use for facial verification. Image credit: Facebook

Prediction (who is it?)

Facebook's DeepFace (Yann LeCun) recognises faces as well as humans

Machine Learning – Why is it popular now?



1 terabyte = 1,000 GB
1 petabyte = 1,000 terabytes
1 exabyte = 1,000 petabytes

- Data Exhaust – as we move around, we leave data markers everywhere
- Sophistication of Machine Learning algorithms
- Increasing computing power and availability of computing hardware and software

Machine Learning Applications

Marketing and Sales

The ability to capture data, analyze it and extract information to provide better shopping experience or implement marketing campaign

Health care

Wearable devices and sensors that access patient's health information in real time use machine learning

Machine Learning Applications

Financial Services

Uses machine learning to detect fraud, glean insights into customer income and expense patterns

Transportation

Identifying traffic patterns and trends helps transportation industry provide more efficient routes.

Traditional Software Development v/s Machine Learning

Traditional Software Development

Convert inches to centimeters (cm)

Input: **inches**

Relationship: **cm** = **inches * 2.54**

Output: **cm**

Traditional Software Development

Recognize Cat or Dog?

Traditional Software Development

Cat or Dog?

Input:

Rules:

Rule 1
Rule 2
Rule 3
… **thousands of lines of code**

Output:

“cat”

Traditional Software Development

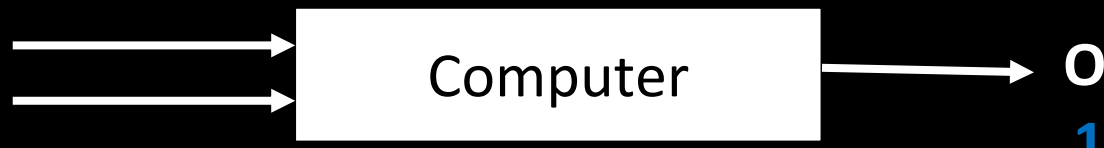
Programming v/s Machine Learning

Input Data

X = 1 , 2 , 3 , 4

Program

Square (X)



Output

1 , 4 , 9 , 16

Input Data

X = 1 , 2 , 3 , 4

Output

1 , 8 , 27, 64

Model

y = Cube (X)

Machine Learning

Input:

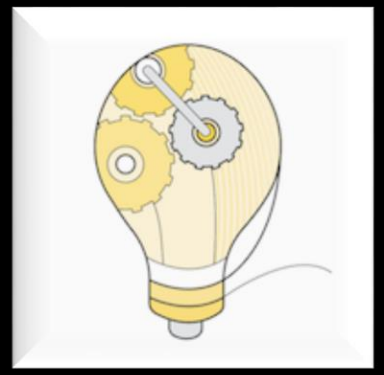


Model:

Output: **[“cat”, “dog”, “cat”]**

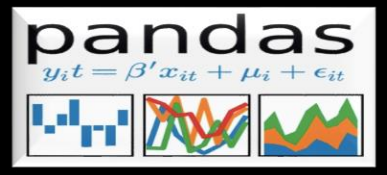
Machine Learning Offerings

Machine Learning Offerings



Machine Learning Using

Python Based Libraries



Machine Learning

Deep Learning

Python and Jupyter Notebook

- Introduction to Jupyter Notebook
- Spark Examples/Python Introduction

Spark Overview

Big Data Before Spark

Hadoop Components

Storage

HDFS

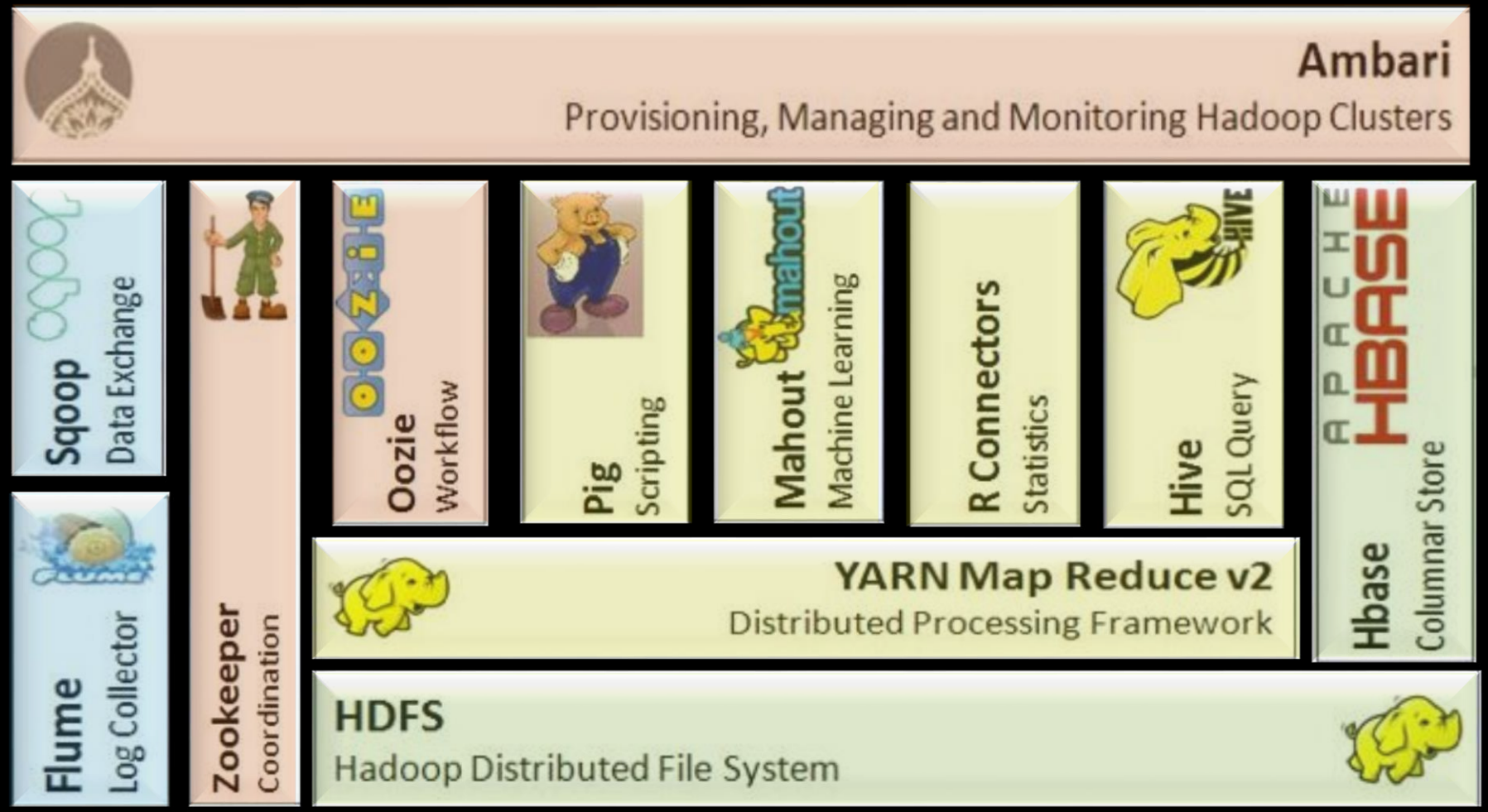
Hadoop Cluster

Processing

MapReduce

Before Spark Ecosystem

- Data Ingestion
- Administration
- Storage
- Processing



Apache Spark

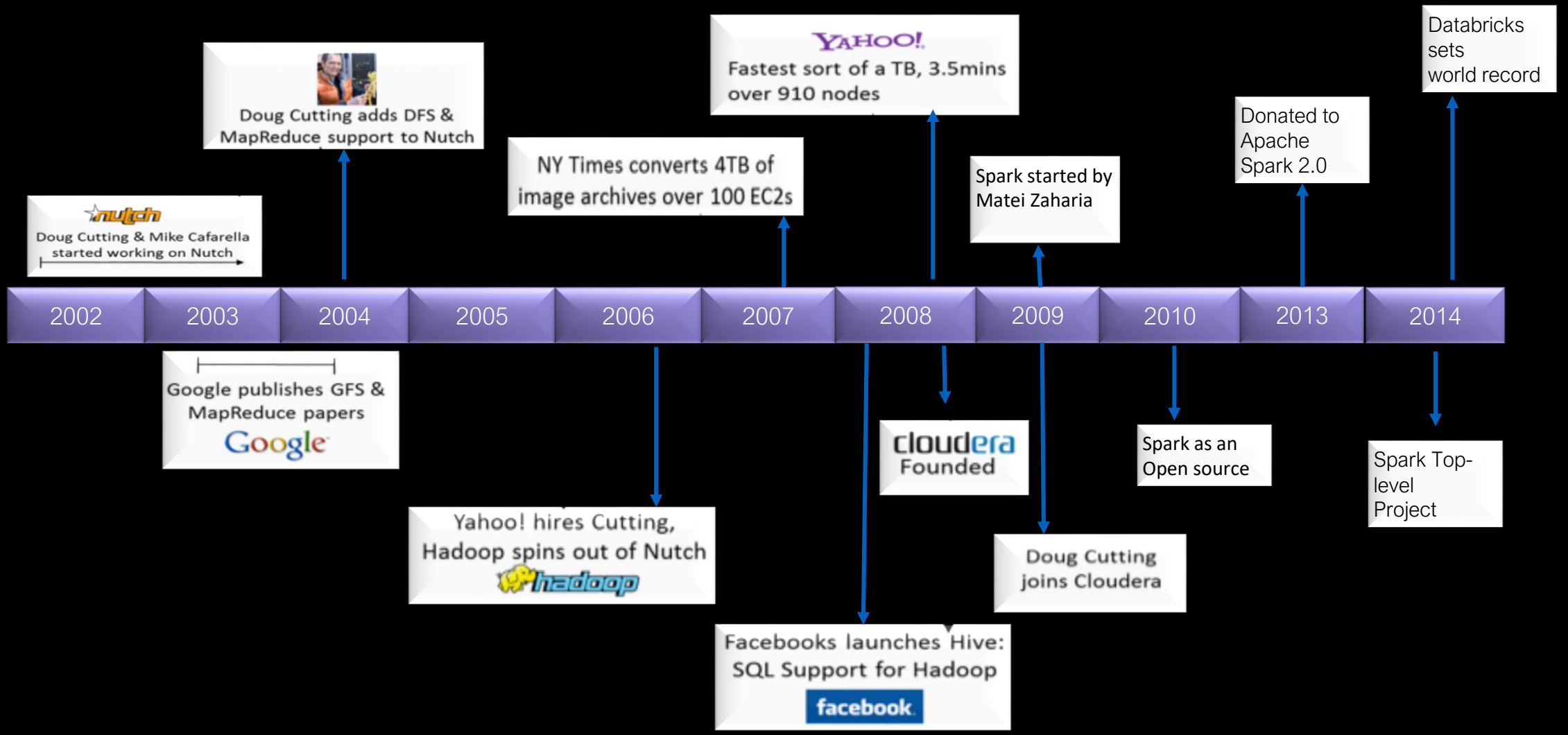
Apache Spark is an open source cluster computing framework for both Batch and Real Time processing

Developed by AMP-Lab (UC Berkley) in 2009, open sourced in 2010 and moved to Apache Foundation in 2013

A unified analytics engine for large scale real- time processing

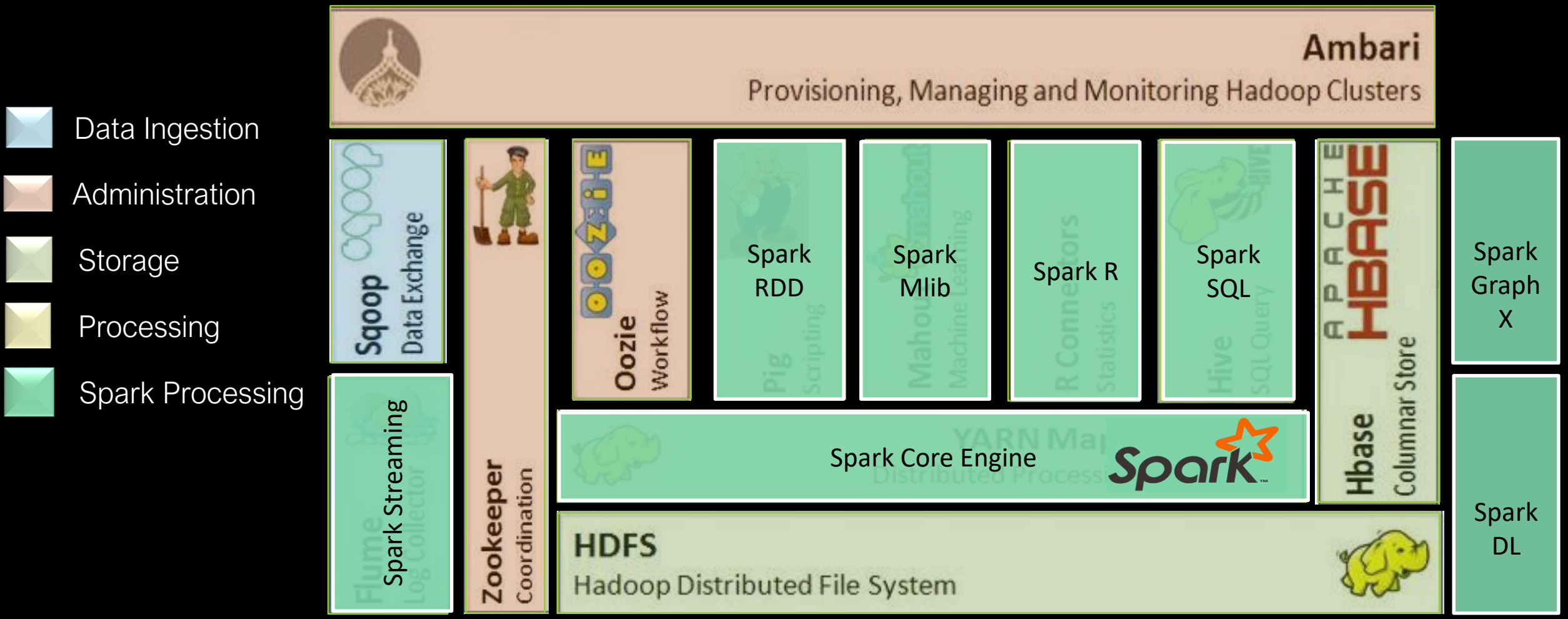
Provides interface for programming entire clusters with data parallelism and fault-tolerance

Spark History



Big Data After Spark

After Spark - Ecosystem



Spark v/s MapReduce

	MapReduce	Spark
Storage	Disk Only	In memory or on disk
Operations	Map and Reduce	Map, Reduce, Join, Sample, etc.
Execution Model	Batch	Batch, interactive, streaming
Programming Environments	Java	Scala, Java, Python, and R. Can also use SQL

Spark Essentials

Spark Features

Generality – Combine SQL, streaming, and complex data analytics

Runs Everywhere – Runs on Hadoop, Apache Mesos, Kubernetes, Standalone, Cloud

Ease of Use – Write applications in Java, Python, Scala, R and SQL

Real Time – Use of in-memory computation helps low latency

Speed – Runs work loads 100 times faster

Powerful Caching – Both disk and memory caching

Spark Components

Structured data processing. Can run queries from existing Hadoop Deployment	Enables scalable, high throughput stream processing of live data	Machine learning libraries to run on top of Spark	Support for Graph and Graph parallel computing	Support for R language so that capabilities of Spark can be used with R
Spark SQL	Spark Streaming	MLlib (Machine Learning)	GraphX Processing	SparkR (R on Spark)

Spark Core Engine

Spark Environment

Data Frame

Spark SQL

Spark Streaming

MLlib (Machine Learning)

GraphX Processing

SparkR (R on Spark)

RDD API

Spark Core Engine

Spark ML Development

Spark Machine Learning Libraries

Support for 2 packages:

- **Spark.mllib** – used with original RDD API
- **Spark.ml** – used with Spark Dataframes

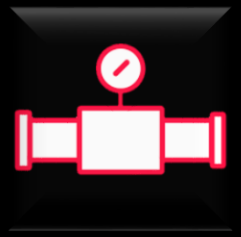
Spark ML Lib Components

ML Algorithms

Support for common machine learning algorithms for classification, clustering, regression and collaborative filtering

Featurization

Includes feature extraction, transformation, dimensionality reduction and feature selection



Pipelines

Has tools for constructing, evaluating and tuning machine learning pipelines

Spark ML Lib Components

Persistence

Save and load algorithms, models and pipelines

Utilities

Tools for data handling, statistics, and linear algebra

Why Spark's Lazy Evaluation is Smart?

```
dataframe = sparkContext.read('log.txt')
##  ( Assume millions of lines in log.txt file)
errors = dataframe.filter(lamda x: "error" in x)
## only say 5000 lines have errors
print "Here are 10 errors:"
for line in errors .take(10):
        print line
## only 10 error lines used, so being lazy avoid million lines read
```

Quiz

1. **Which execution model is supported in Spark?**
   a. Batch only
   b. Batch, Interactive, streaming
   c. None of the above

2. **Programming languages supported in Spark?**
   a. C, C++, JAVA, Python
   b. JAVA, Python, R
   c. Python, R, Scala, JAVA

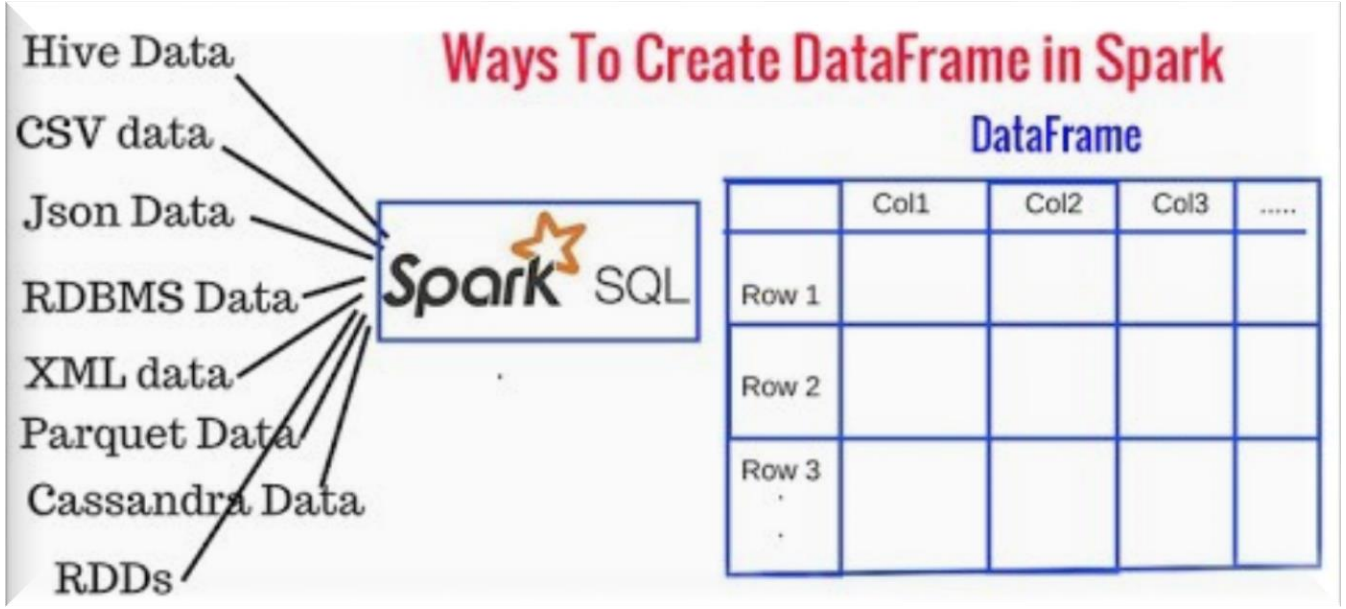
Spark DataFrame

Spark DataFrame

- DataFrame is a distributed of rows under named columns, i.e. it is like a table in relational database

- Some of the characteristics are:
  - **Immutable** – once created cannot be changed, it can be transformed into different DataFrame
  - **Lazy Evaluations** – means task is not executed until an action is applied
  - **Distributed** – DataFrames are distributed by nature

Creating Spark DataFrame



Spark DataFrame Demo

DataFrame Example

- Open file ‘DataFrame Example’ using Jupyter
- Example of reading a csv file into a dataframe
- Some of operations that can be performed on the dataframe
- Also example of converting Spark Dataframe into Pandas Dataframe
- Some examples of using Pandas Dataframe

Apache Spark Reference

http://spark.apache.org/

Machine Learning Techniques

Machine Learning Techniques

Supervised Learning

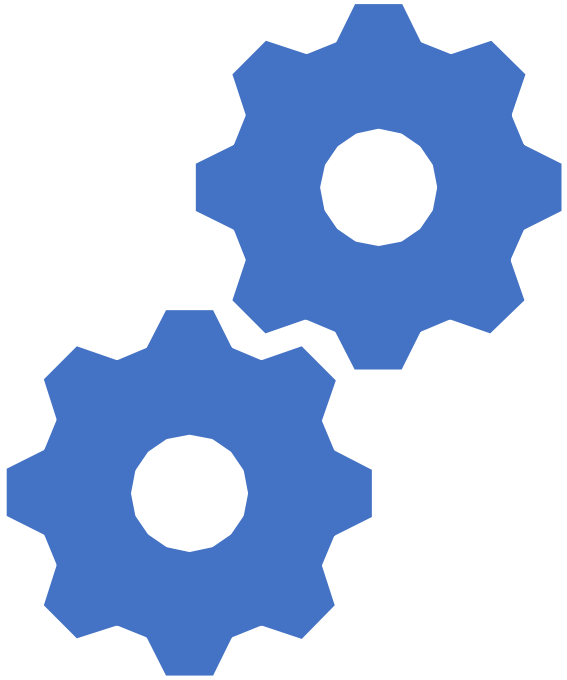
Unsupervised Learning

Supervised Learning

Diameter, Thickness → Features

Currency → Label

Diameter	Thickness	Currency
1.0430 inches	0.0790 inches	US dollar coin
1.0433 inches	0.07680 inches	Canadian dollar coin (Loonie)
0.9154 inches	0.09173 inches	One Euro coin



Supervised Learning

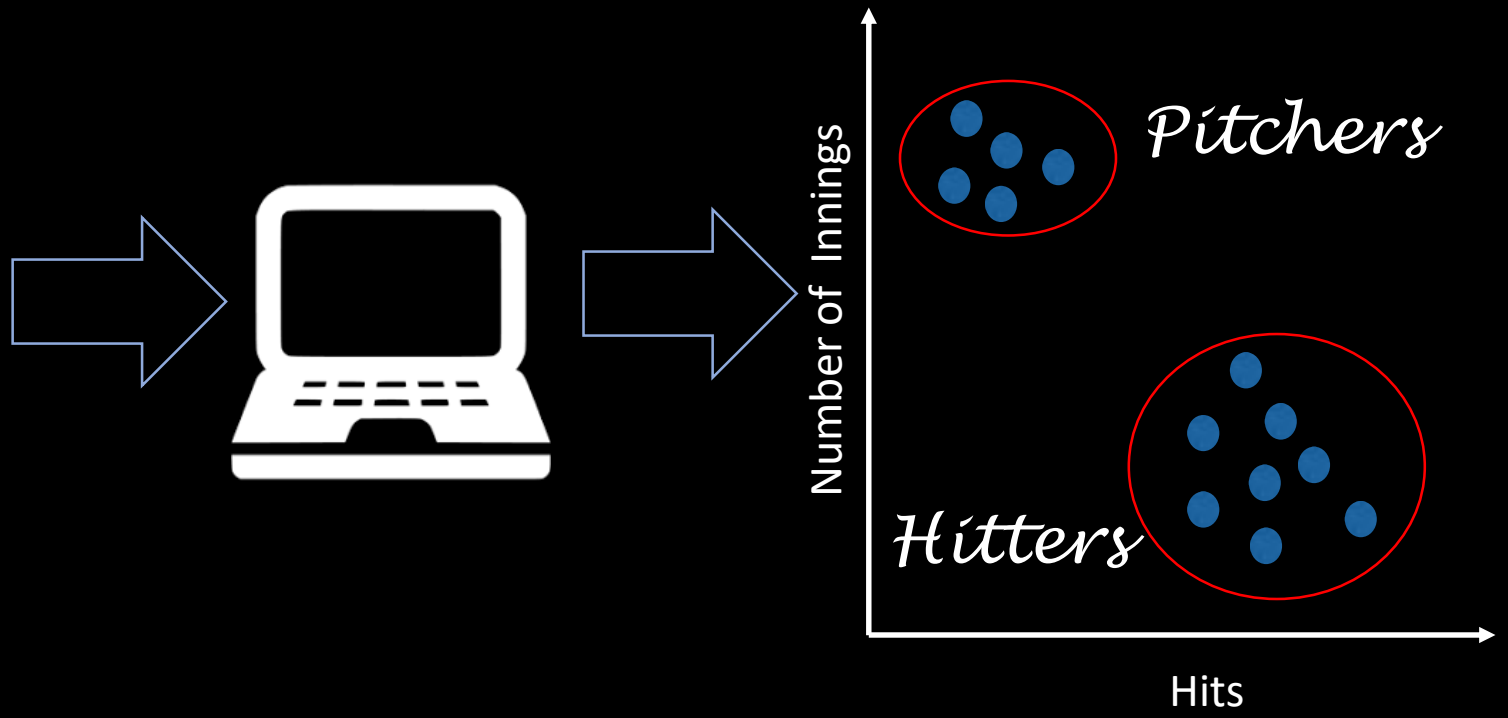
- Supervised Learning technique is where algorithms are trained using ***labelled*** inputs, i.e. a relation between input and labelled output
- The learning algorithm learns from a set of inputs and the corresponding labelled outputs to create a model
- It then modifies the model accordingly to reduce errors in its learning
- Some of the methods applied are classification, prediction and gradient boosting

Unsupervised Learning

ALL PLAYERS | Compute | Sort | Format | Minimum | Close

Batting Statistics
2009 Central High Raiders

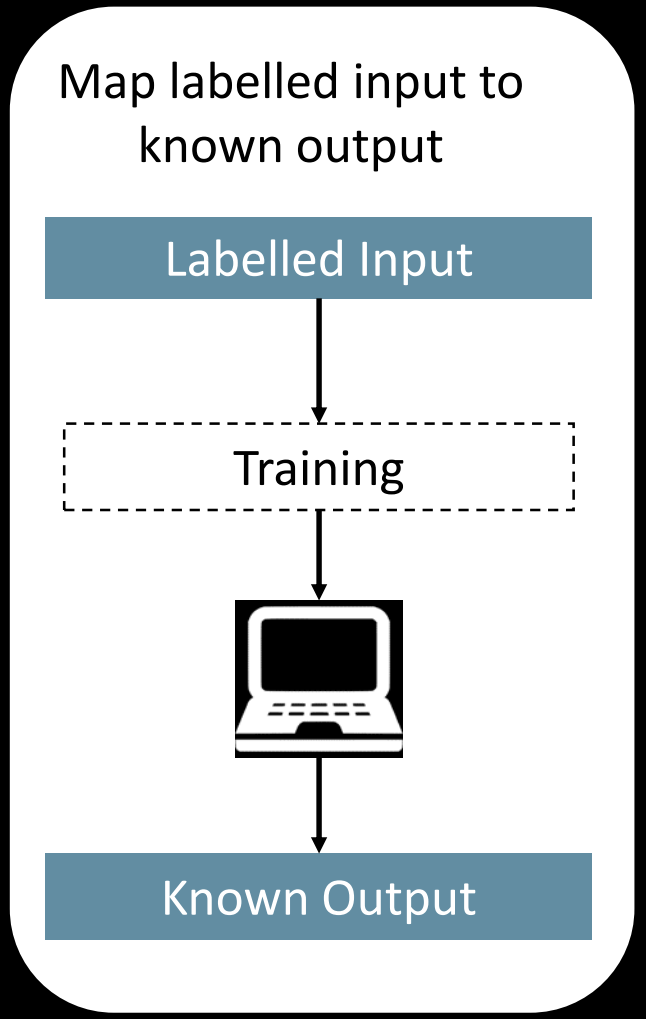
Player	G	AB	R	H	2B	3B	HR	TB	BB	SO	SH	SF	SB	RBI	OBP	Slg	BA
Joe Anderson	6	20	4	8	1	0	0	9	7	7	0	0	0	4	.556	.450	.400
Jake Johnson	6	22	6	8	1	0	4	21	7	4	0	0	4	6	.517	.955	.364
Tony Valenza	6	29	5	10	0	1	0	12	1	7	0	0	3	2	.367	.414	.345
John Lex	6	16	3	5	1	1	1	11	1	2	0	0	0	2	.353	.688	.313
Art VanDelay	6	19	1	5	0	1	0	7	0	2	0	0	0	0	.263	.368	.263
Frank Hansen	6	23	5	6	1	0	0	7	6	5	0	0	0	2	.414	.304	.261
Bob Guyer	6	26	2	6	0	0	1	9	2	4	0	0	0	7	.286	.346	.231
Ron Nowakowski	5	9	0	2	0	0	0	2	0	4	0	1	0	1	.200	.222	.222
Frank Butler	6	23	3	5	2	0	1	10	4	5	0	0	0	2	.357	.435	.217
Tim Bradley	6	19	1	4	1	0	0	5	5	5	0	0	0	1	.375	.263	.211
TOTALS	6	206	30	59	7	3	7	93	33	45	0	1	7	27	.386	.451	.286



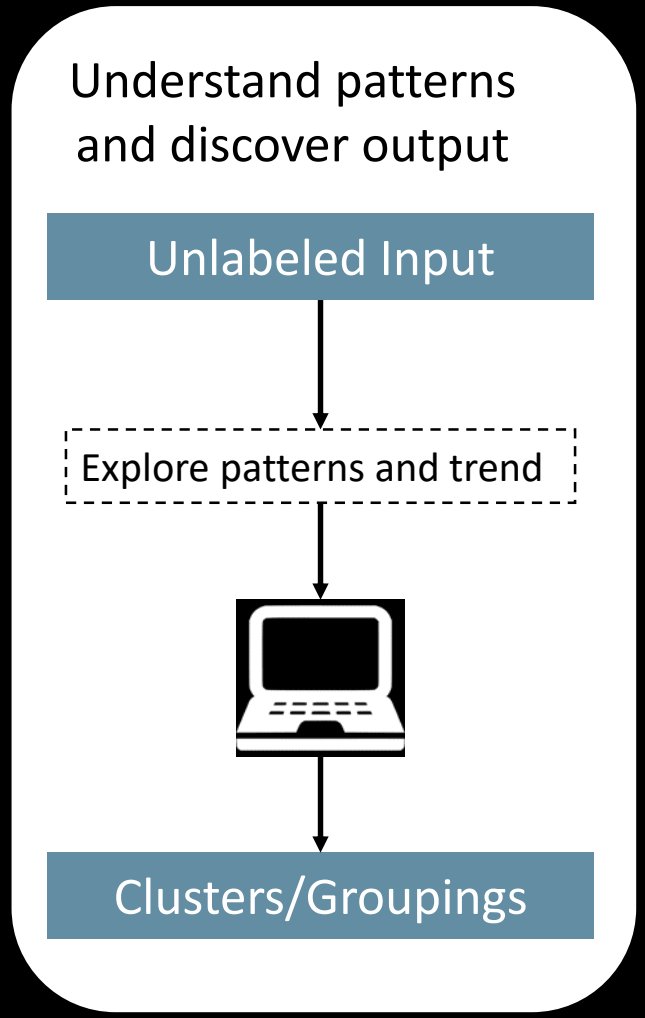
Unsupervised Learning

- Unsupervised learning is where there is no labelled data, the algorithm creates clusters/groupings
- The objective is to explore the data and discover underlying patterns and capture useful insights
- Popular methods include nearest neighbors, K-Means clustering, etc.
- Used in recommendation systems and anomaly detection

Learning Approaches

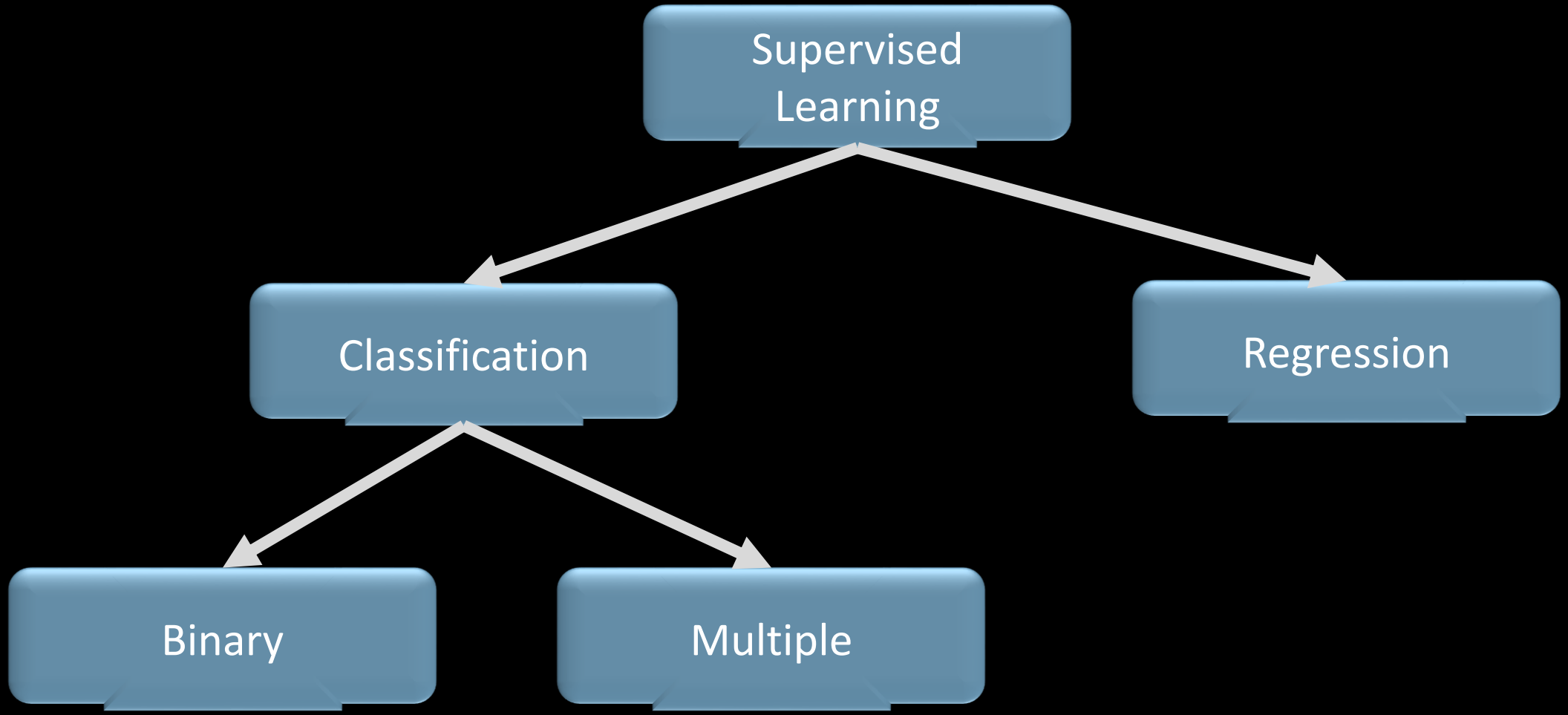


Supervised

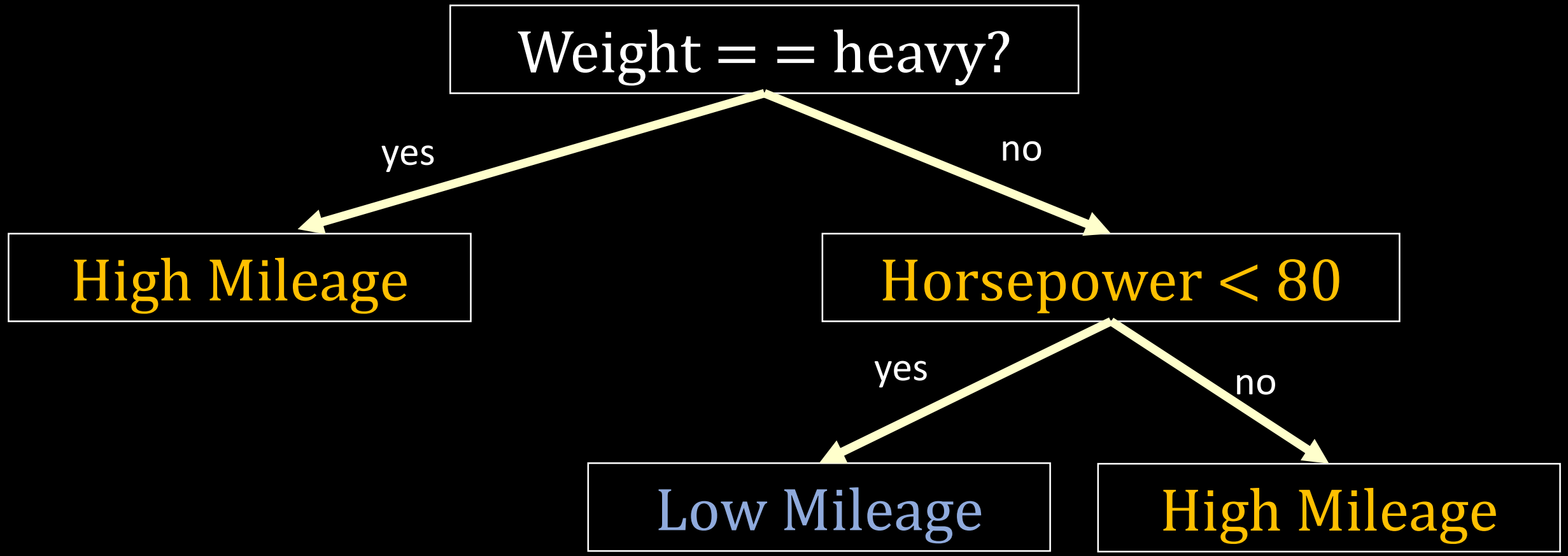


Unsupervised

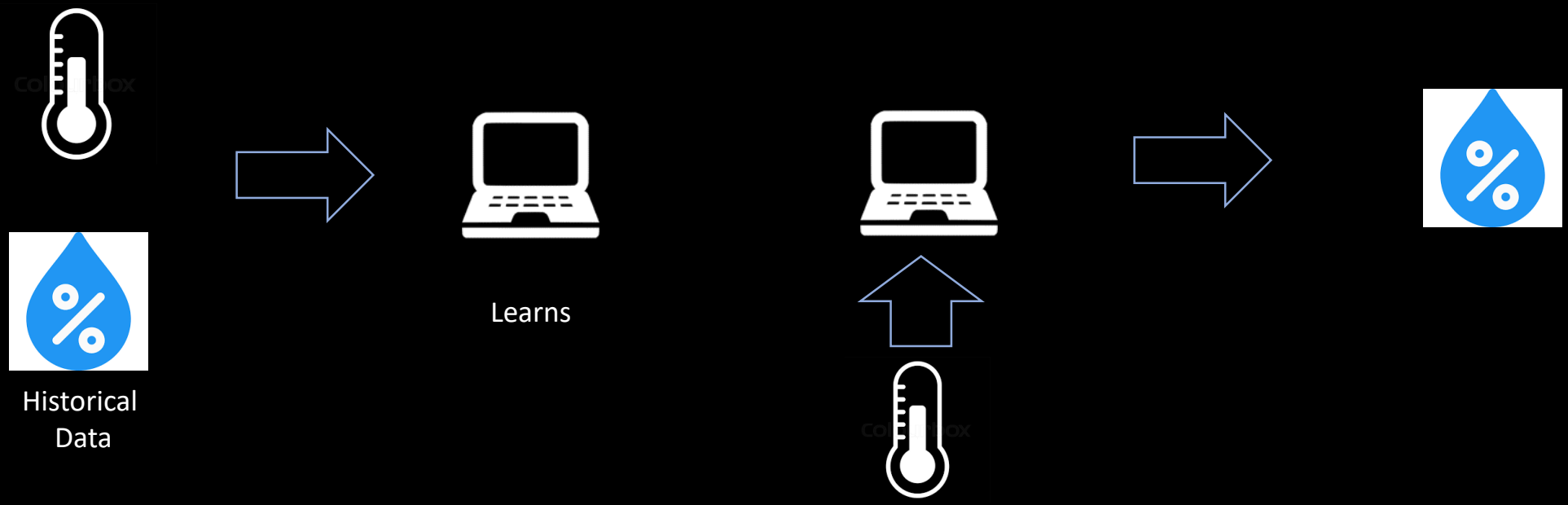
Supervised Learning



Classification - Binary

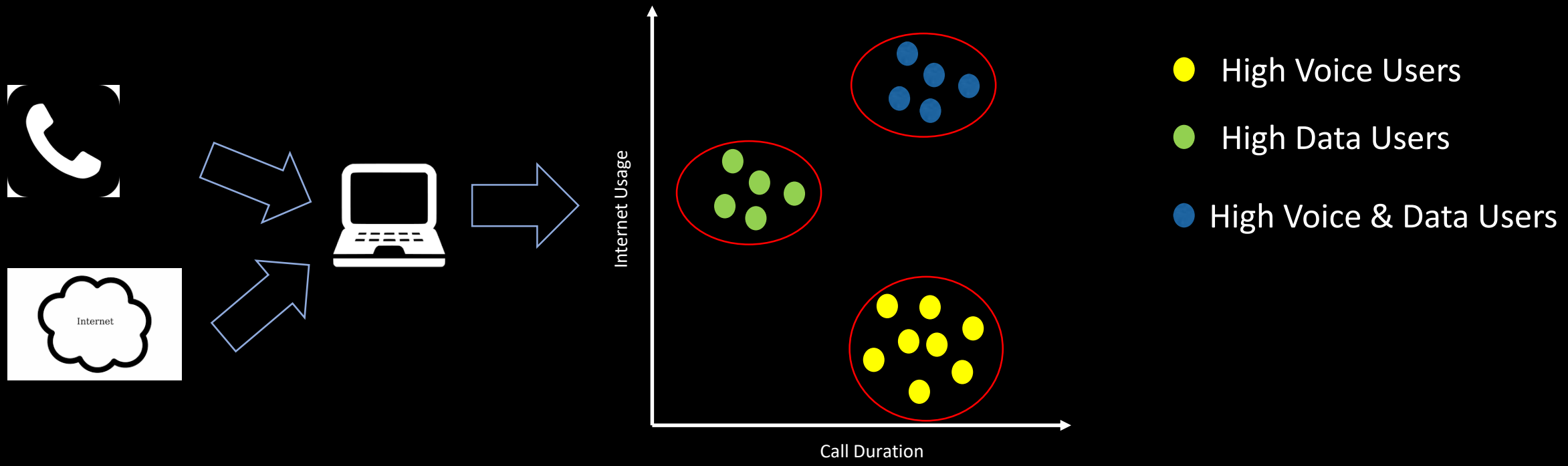


Regression



- Regression models output a continuous variable
- Continuous variable is one that has precise value, e.g. home prices, stock price, etc.

Clustering – Unsupervised Learning



Clustering is where model groups data based on similarities – creating a cluster

Association – Unsupervised Learning

Bread
Fruits
Eggs
Butter
Meat

Bread
Jam
Milk
Butter

Bread
Fruits
Butter
Meat

Bread
Butter

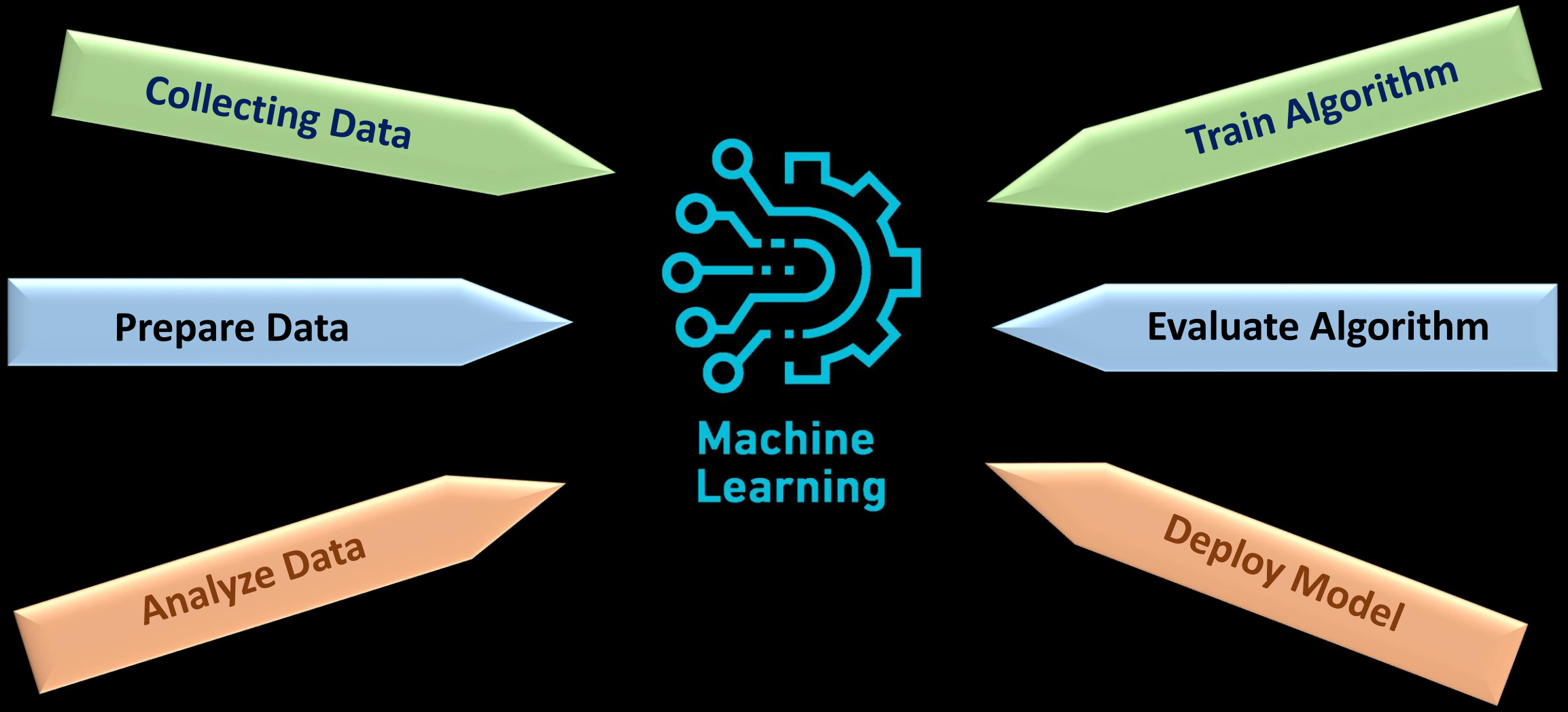
- Association is finding co-relations in data or rules that describe large amounts of similar data
- Classic example is people who buy item X will by item Y

Quiz

1. Which of the following technique is Unsupervised learning?
   a) Clustering
   b) Regression
   c) Classification
2. Which technique would take images and predict the age of person?
   a) Clustering
   b) Regression
   c) Classification
3. Unsupervised Learning has output labels?
   a) True
   b) False

Machine Learning Development

Machine Learning Process



Machine Learning Process

1. Collecting Data

2. Prepare Data

3. Analyze Data

4. Train Algorithm

5. Evaluate Algorithm

6. Deploy Model

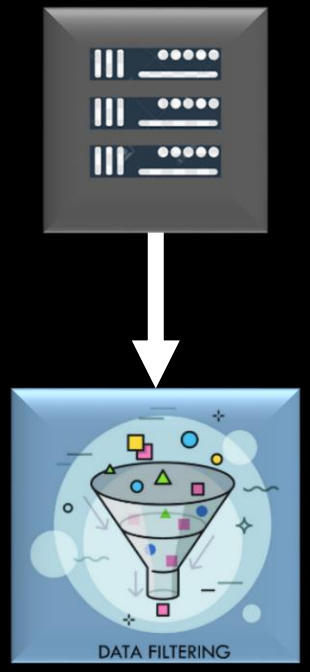
Collect relevant data from various sources

Store in servers

Machine Learning Process

1. Collecting Data
2. **Prepare Data**
3. Analyze Data
4. Train Algorithm
5. Evaluate Algorithm
6. Deploy Model

Start to clean and convert the collected data into format that can be used by algorithms



Clean Data

Machine Learning Process

1. Collecting Data
2. Prepare Data
3. **Analyze Data**
4. Train Algorithm
5. Evaluate Algorithm
6. Deploy Model

Data is analyzed, visualized and is prepared for model

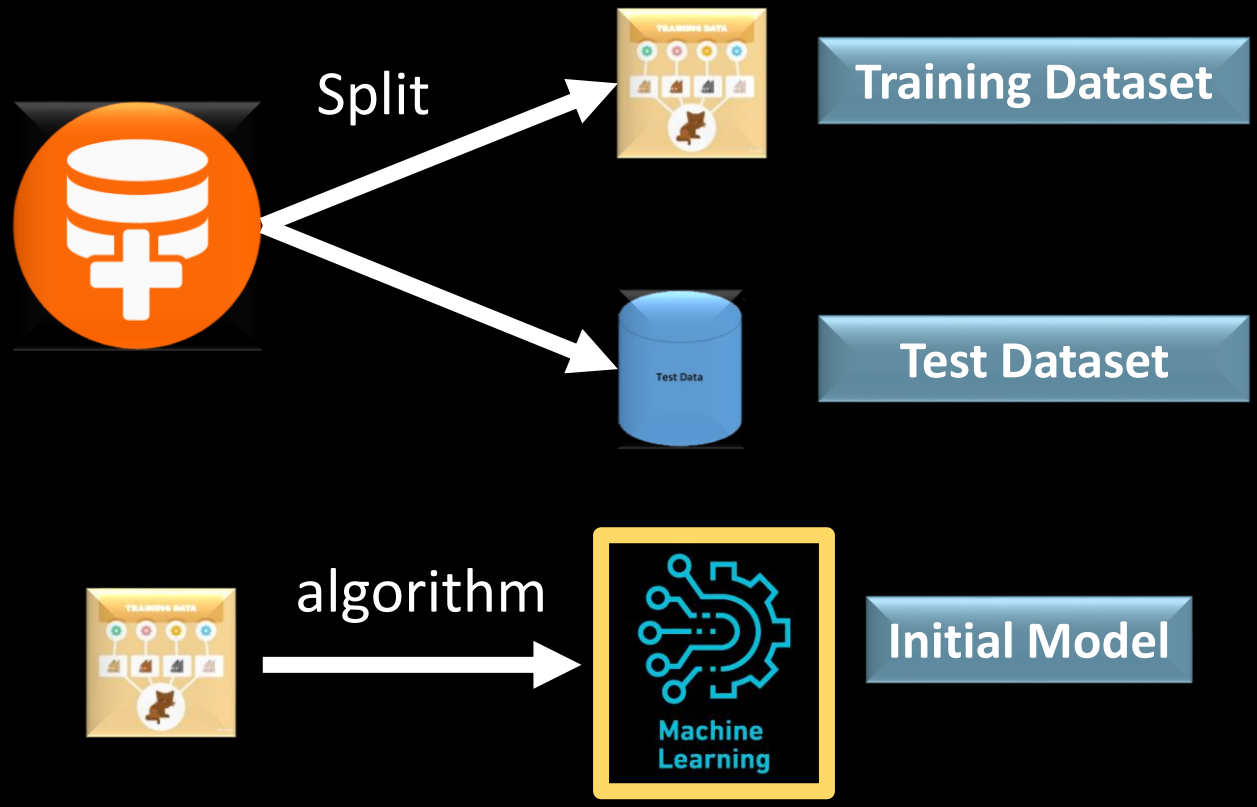
Feature Selection

Data ready
for algorithm

Machine Learning Process

1. Collecting Data
2. Prepare Data
3. Analyze Data
4. **Train Algorithm**
5. Evaluate Algorithm
6. Deploy Model

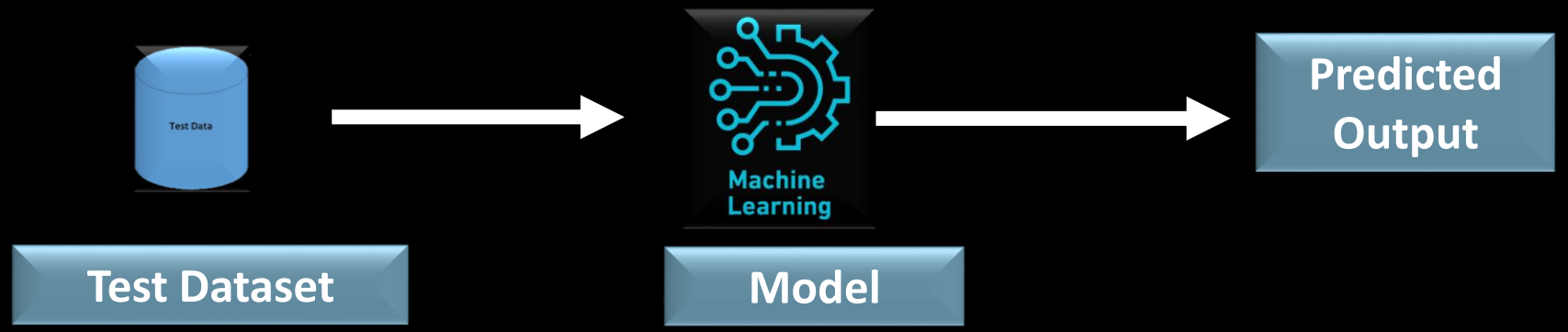
Algorithm is trained on the training dataset, where the algorithm understands the pattern and data relationship



Machine Learning Process

1. Collecting Data
2. Prepare Data
3. Analyze Data
4. Train Algorithm
5. **Evaluate Algorithm**
6. Deploy Model

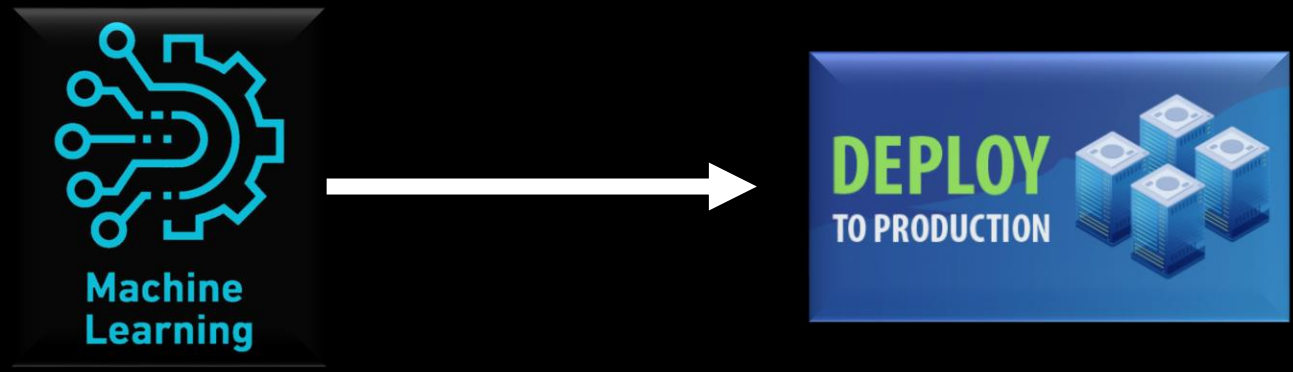
Test dataset is used to determine the accuracy of the model



Machine Learning Process

1. Collecting Data
2. Prepare Data
3. Analyze Data
4. Train Algorithm
5. Evaluate Algorithm
6. **Deploy Model**

- If the accuracy and performance of the model is acceptable then the model is deployed in production
- In production environment if the performance degrades then the model is pulled and additional data is use to improve performance



Machine Learning Process (Code)

Training

```
model.fit(X_train, y_train)
```

Evaluate with Test Data

```
model.predict(X_test)
```

Score the Model

```
model.score(X_test, y_test)
```

The Data Science Process (DIAPERS)

1. Define Problem Statement
2. Ingest Data
3. Analyze Data
4. Prepare Data for ML
5. Evaluate Models
6. Refine model
7. Ship it

Data

Structured Data

User ID	Gender	Age	Estimated Salary	Purchased
15624510	Male	19	19000	Yes
15810944	Male	35	20000	No
15668575	Female	26	43000	No
15603246	Female	27	57000	Yes
15804002	Male	19	76000	No
15728773	Male	27	58000	No
15598044	Female	27	84000	No
15694829	Female	32	150000	Yes
15600575	Male	25	33000	No
15727311	Female	35	65000	No
15570769	Female	26	80000	No
15606274	Female	26	52000	No

Structured Data Types

Numerical

- Exact numbers as data points
- Two types
  - ***Continuous***
    - Can assume any value within a range
    - Height, weight, salary, etc. e.g. 6.3, 40.32, 32.5
  - ***Discrete***
    - Data has distinct values
    - Number of students, units sold, etc. e.g. 15, 30, 10

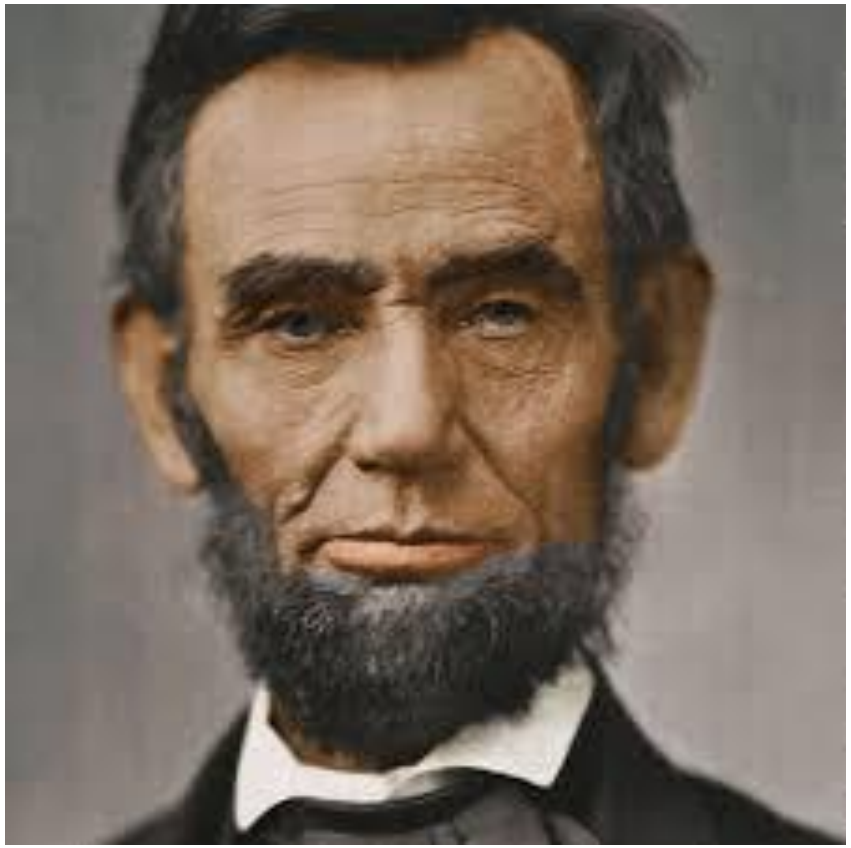
Categorical

- Qualitative data that has no inherent mathematical meaning
- Can assign numbers to categories in order to represent them, however the numbers have no mathematical meaning
- 1 for red, 2 for blue, 3 for green

Ordinal

- A mixture of numerical and categorical data
- Categorical data has the mathematical meaning
- Movie ratings in scale 1 to 5
  - Ratings can be 1, 2, 3, 4 or 5
  - 1 means worst rating, 5 being the best

Unstructured Data



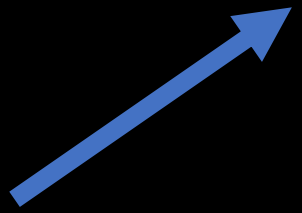
```
[ [216, 203, 125,  10,  84, 241, 149, 159, 212, 118, 135, 158,  11,  91,  36, 177, 176, 253, 132, 210, 159,  20, 153, 131, 132,  55, 16, 132],
  [184,  34,  95, 225,  60, 218,  49, 193,  93, 119,  68, 133, 195, 104, 248,  18,  18, 136,  90,  71,  81,  41, 233,  53,  46,  87, 86, 243],
  [ 85,  61, 220, 170, 206,  34, 141,  97,  66, 217, 124, 143, 241, 205,  76, 123,  66,  72, 231, 116, 244,  74, 155, 144,  47, 230, 171, 165],
  [156,  87, 181,  90, 160,   2, 184, 112, 108,  62, 223, 153,  93, 244,  83, 187,  83,  18, 134,  28, 121, 244, 202, 176, 228, 233,  76,  13],
  [ 76, 238, 128, 183, 119, 130,  34,  12, 112, 254,  90, 167,  64,  89, 170, 221, 196,  69,  82,  11,  65,  86, 254, 111, 134,   0, 148, 246],
  [105, 178, 254,  31,  32, 133,  57,  40,   6,  85, 115,  56, 132,  84,  35, 119, 158, 182, 106,  77,  84, 106, 164, 230,  54,  42,  55, 130],
  [ 25,  86, 222,  59, 242, 111,  59, 183, 236, 214, 251,   7, 142,  90, 179,  80, 163, 159,  26, 143, 108, 109, 229, 223, 220, 196,  21,  18],
  [ 21,  42, 109, 188,  91,  93, 246, 236, 125,  48, 151,  12, 178,  26, 118, 135,  77,  84, 179, 208, 114, 224,  99, 246,  68,  21,  69,  39],
  [253,  66,  78,  55,  39, 107, 248,  90, 124, 107,  51,  92, 150, 234,  91, 177, 146,  80,   8, 179, 148, 229, 233,  59, 164, 199, 252,  43],
  [ 79,  60,   5,  70,  37, 218,  19,   9,  90,  74, 198, 129,  61, 160, 206,  11,  37, 171,  44, 241, 228, 190, 232,  99,   7, 100,  83, 225],
  [211,  38,  52, 167, 206, 139, 215, 209, 202, 102, 122,  77,  86, 117, 134,  22, 176,  94,  22, 201,   6,  73, 156, 226,  36,   0,  50, 119],
  [159,  24, 197, 215,  16, 243, 177,  13, 108, 211,   6,  97,  75, 214, 121,  92, 154, 109, 213, 163, 123,  20, 190, 174,  89,   6, 136, 164],
  [183, 136, 245, 175, 233,  62, 141, 117, 150,  74, 182, 175,  36, 230,  93, 109, 212,  43,  10,  75, 234, 124,  70, 244, 161,  76, 241, 223],
  [150,   7, 184,  20, 133,  22, 112, 212,  48,  30, 156, 113, 127, 207, 219, 173, 223, 127, 202, 172,  39,  98, 134, 124, 130,  34, 210, 101],
  [101,  77,  87,  37, 152, 112,  34, 106,  30,  23,  79, 214, 245, 152, 129, 243, 109, 213, 170, 190, 220,  25,  76, 205, 135, 227, 225, 165],
  [108, 184, 172, 121,   8,  83, 106, 116, 235,  55,  73, 204,  50,  40, 124, 153, 225, 157,  13,  28, 105,  62, 242, 214,  56, 159, 137,  67],
  [ 14,  75,  26,  47,  74, 205,  45, 219,  27,  18,  79,  28,  49, 224,  85, 214, 180, 105, 183,  87,  18,  64,   7,  61, 125,  87,  38,  98],
  [122, 146,   4,  72, 150, 249,  77,  90,   6, 132, 134, 151, 164,  29,  94, 188, 251, 177,   0, 206, 193, 182, 231,  43,  32,  32,  80, 147],
  [ 26,  39,  76,  12,  35,  61, 103, 233, 204, 136,  82,  28,   5,  68, 229, 197,  52, 215, 224, 117, 101,   4, 154,   4, 205,  50, 251, 114],
  [ 68, 176,  23, 246,  11,  57,  62,  25,  38,  17, 136, 106, 113, 140, 254,  43, 231, 150,  12, 114,  77,   8, 214, 187,  92,  66, 195,  70],
  [ 20, 241, 148, 151,  37,   4,  14, 231, 225,  53, 232, 240, 223,  59, 234, 134, 247, 242, 212,  63, 201,  38,  63, 200, 128, 139, 167, 173],
  [ 60, 244,  33, 111, 143, 127, 168, 237, 189,  63, 125, 181,  92,  91,  14, 211,  21,  26, 253, 109, 174, 100, 138, 138, 221, 204,  29, 230],
  [ 81, 174, 217,  93,  65, 134,   7,  36, 176, 122, 226,  23, 223,  28, 202,   5,  54, 205, 169,  14,  88, 178,  84, 198,  96, 201, 230, 193],
  [215, 168, 125,  92,  70, 151, 183, 210,  36,  32,  19,  51,  42,  64,  19, 146, 183, 246,   0, 184, 236,   7, 226, 118, 113, 241,  85,  89],
  [ 31, 158, 210,  16, 199,  58, 224,   7, 203,  86, 103,  45,  28,  54,  92, 204, 243, 117,  75, 208, 248, 223,  87, 250,  14,  43, 102,  66],
  [ 13, 236, 138,  67, 236, 109, 113,  46, 115,  19, 214, 154, 199, 248,  55, 172, 214, 249, 125, 154, 139, 141, 188,  78, 107, 200, 196,  16],
  [ 65, 150, 158, 254, 114, 177, 120,  15,  65,  58,  79, 171, 118,  32, 250,  81,  27,  85, 128, 146, 144, 234, 139,  26,   6,  68, 133, 205],
  [123,  68, 216,  34, 139,  34,  34, 175, 213,  72,  76,  19,  32, 138, 132, 111, 242, 249, 177,  89,  61,  72, 252,  79,  20, 171, 174, 177] ]
```

Data

Social Network Ads

User ID	Gender	Age	Estimated Salary	Purchased
15624510	Male	19	19000	Yes
15810944	Male	35	20000	No
15668575	Female	26	43000	No
15603246	Female	27	57000	Yes
15804002	Male	19	76000	No
15728773	Male	27	58000	No
15598044	Female	27	84000	No
15694829	Female	32	150000	Yes
15600575	Male	25	33000	No
15727311	Female	35	65000	No
15570769	Female	26	80000	No
15606274	Female	26	52000	No

Data



Features (X)

Independent Variables

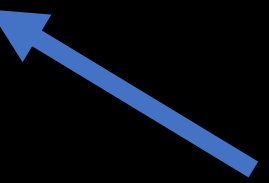
Social Network Ads

User ID	Gender	Age	Estimated Salary	Purchased
15624510	Male	19	19000	Yes
15810944	Male	35	20000	No
15668575	Female	26	43000	No
15603246	Female	27	57000	Yes
15804002	Male	19	76000	No
15728773	Male	27	58000	No
15598044	Female	27	84000	No
15694829	Female	32	150000	Yes
15600575	Male	25	33000	No
15727311	Female	35	65000	No
15570769	Female	26	80000	No
15606274	Female	26	52000	No

Data

Social Network Ads

User ID	Gender	Age	Estimated Salary	Purchased
15624510	Male	19	19000	Yes
15810944	Male	35	20000	No
15668575	Female	26	43000	No
15603246	Female	27	57000	Yes
15804002	Male	19	76000	No
15728773	Male	27	58000	No
15598044	Female	27	84000	No
15694829	Female	32	150000	Yes
15600575	Male	25	33000	No
15727311	Female	35	65000	No
15570769	Female	26	80000	No
15606274	Female	26	52000	No



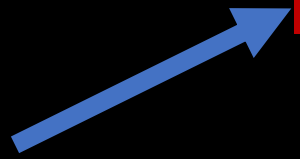
Target or Output (y)

Dependent Variable

Data

Social Network Ads

User ID	Gender	Age	Estimated Salary	Purchased
15624510	Male	19	19000	Yes
15810944	Male	35	20000	No
15668575	Female	26	43000	No
15603246	Female	27	57000	Yes
15804002	Male	19	76000	No
15728773	Male	27	58000	No
15598044	Female	27	84000	No
15694829	Female	32	150000	Yes
15600575	Male	25	33000	No
15727311	Female	35	65000	No
15570769	Female	26	80000	No
15606274	Female	26	52000	No



Sample

Data for algorithm

$$X = \begin{bmatrix} 15624510 & 1 & 19 & 19000 \\ 15810944 & 1 & 35 & 20000 \\ 15668575 & 0 & 26 & 43000 \\ 15603246 & 0 & 27 & 57000 \\ 15804002 & 1 & 19 & 76000 \\ 15728773 & 1 & 27 & 58000 \\ 15598044 & 0 & 27 & 84000 \\ 15694829 & 0 & 32 & 150000 \\ 15600575 & 1 & 25 & 33000 \\ 15727311 & 0 & 35 & 65000 \\ 15570769 & 0 & 26 & 80000 \\ 15606274 & 0 & 26 & 52000 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Data Quality

Best Data Qualities

Clean

Coverage

Complete

Otherwise Garbage in Garbage out

Best Data Qualities

15603246	Female	27	57000	1
15804002	Male	19	76000	3
15728773	F	27	58000	2
15598044	Female	27	84000	13
15694829	Female	32	150000	2
15600575	Mal		33000	
15727311	Female	35	65000	1
15570769	Female	26		1
15606274	Female	26	52000	3

Best Data Qualities

15603246	Female	27	57000	1
15804002	Male	19	76000	3
15728773	F	27	58000	2
15598044	Female	27	84000	13
15694829	Female	32	150000	2
15600575	Mal		33000	
15727311	Female	35	65000	1
15570769	Female	26		1
15606274	Female	26	52000	3

Best Data Qualities

Clean

15603246	0	27	57000	1
15804002	1	19	76000	3
15728773	0	27	58000	2
15598044	0	27	84000	1
15694829	0	32	150000	2
15600575	1	28	33000	1
15727311	0	35	65000	1
15570769	0	26	48000	1
15606274	0	26	52000	3

Best Data Qualities

Coverage

15603246	Female	27	57000	1
15804002	Male	19	76000	3
15728773	F	27	58000	2
15598044	Female	27	84000	13
15694829	Female	32	150000	2
15600575	Mal		33000	
15727311	Female	35	65000	1
15570769	Female	26		1
15606274	Female	26	52000	3

Depth

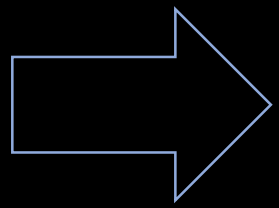
Best Data Qualities

Complete

15603246	Female	27	57000	1
15804002	Male	19	76000	3
15728773	F	27	58000	2
15598044	Female	27	84000	13
15694829	Female	32	150000	2
15600575	Mal		33000	
15727311	Female	35	65000	1
15570769	Female	26		1
15606274	Female	26	52000	3

Breadth

Data Split



Training Data

70 %

Test Data

30 %

- On training data set, use fit() method to train the model
- On the test data set, use predict() method to make predictions on the trained model
- After fit and predict, use score() on the model to see its accuracy

Statistics Brush Up

Mean, Median, Mode

Mean

Sample									Total	Num items	Mean
35	45	67	12	24	48	55	58	30	430	10	43

Median

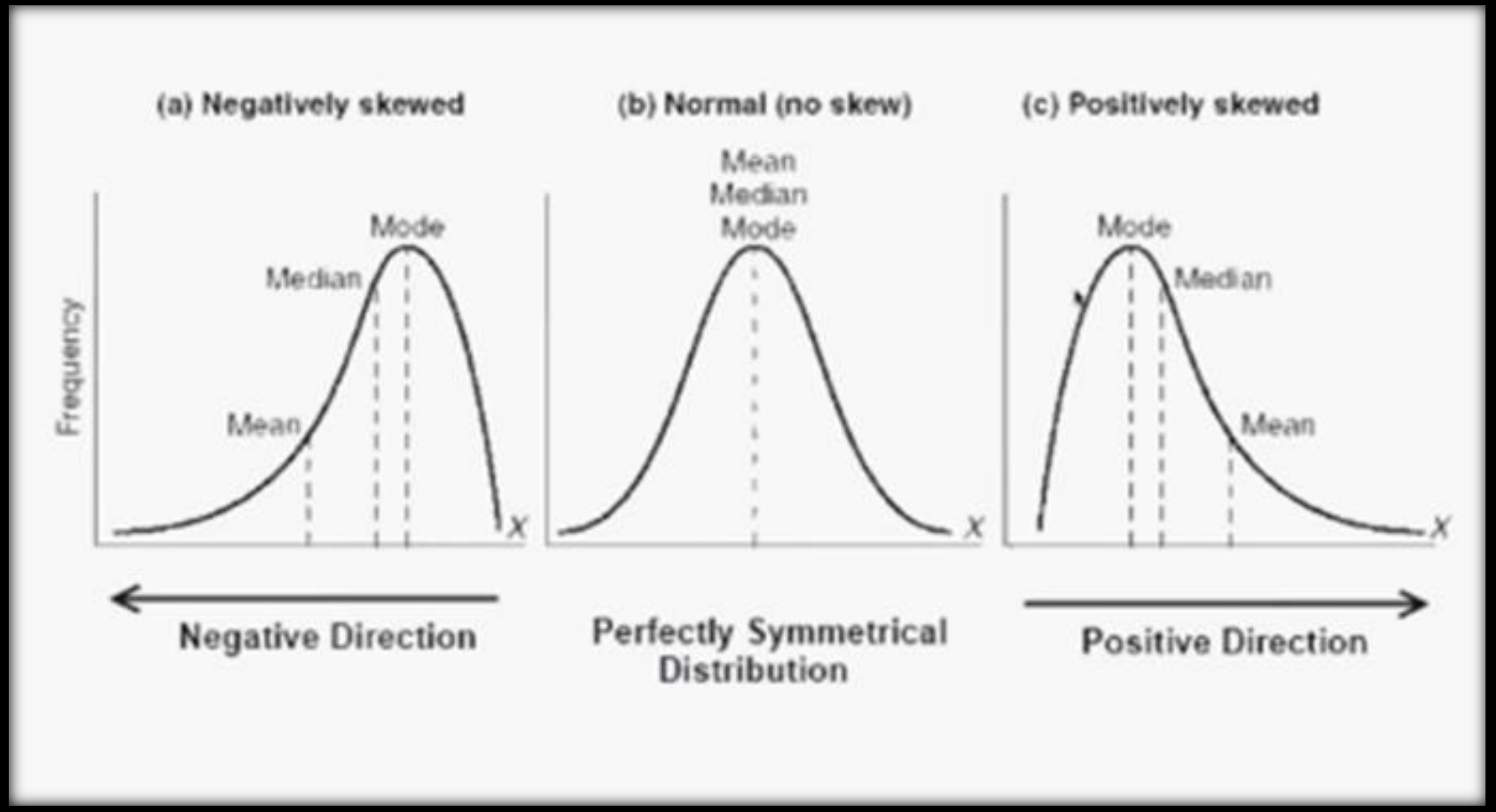
Sample							Median
1	2	3	4	5	6	7	4

Sample							Median
1	2	3	4	5	6	7	8	(4 + 5) / 2 = 4.5

Mode

Sample									Mode
2	3	4	2	2	4	5	2	2	2

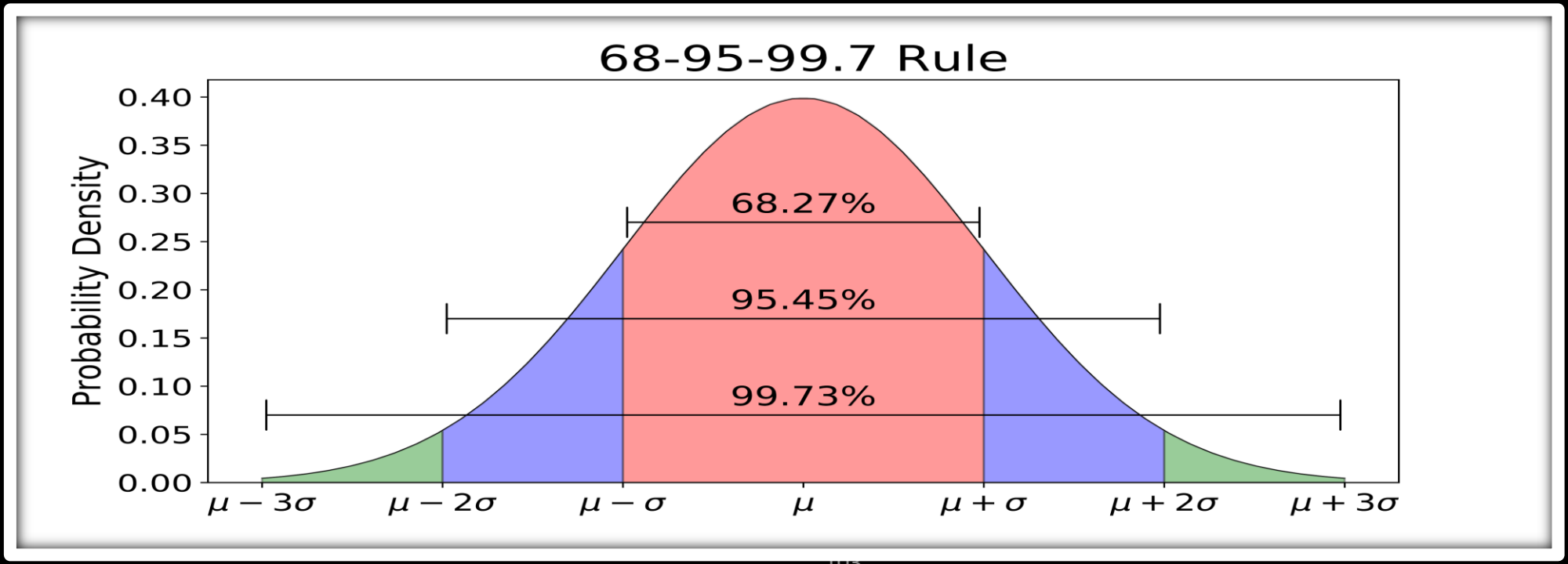
Distribution



Variance and Standard Deviation

Variance is how spread out the data is

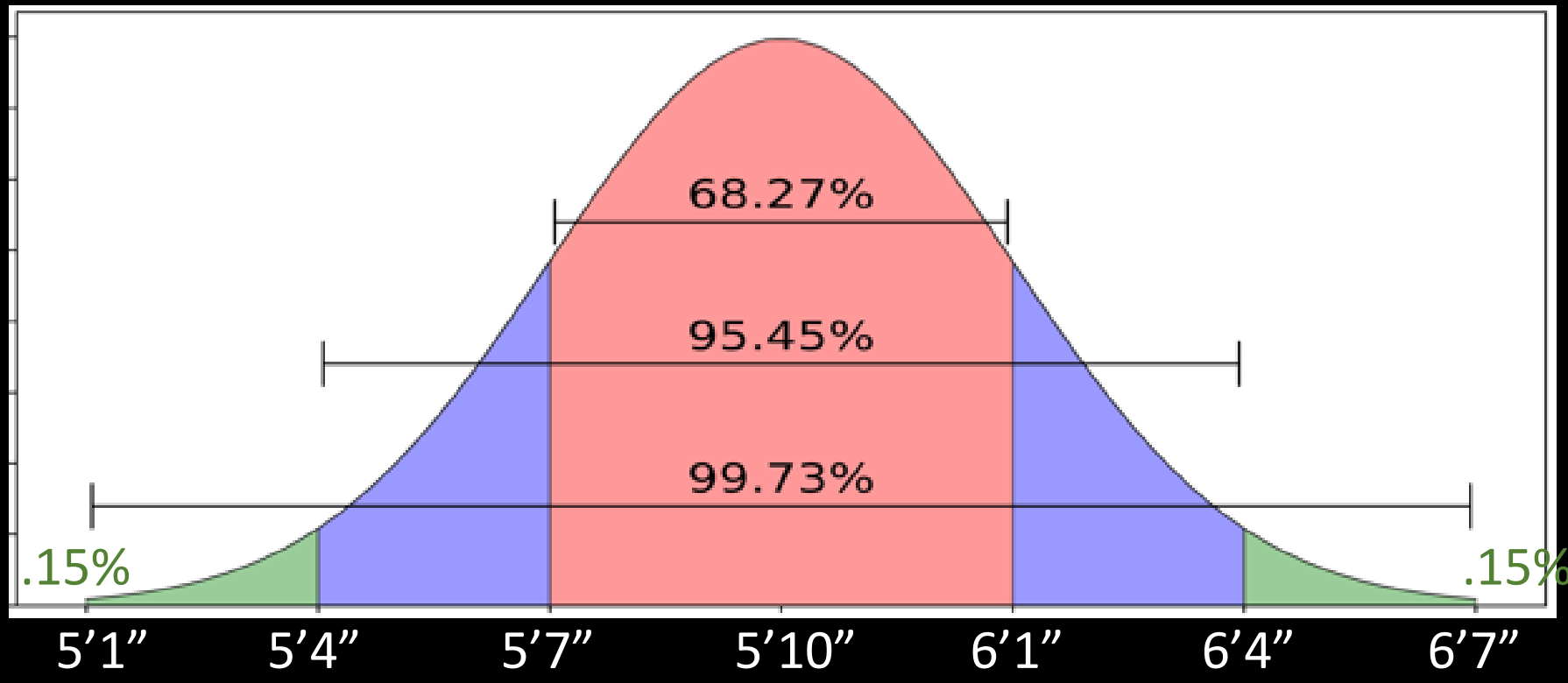
Standard Deviation is a measure of the amount of variation or dispersion of a set of values



Standard Deviation

Average American Men Height is 5’10’’

Standard Deviation is 3”



Formulas

$$mean(x) = \frac{\sum_{i=1} x_i}{count(x)}$$

$$variance = \sum_{i=1}^{n} (x_i - mean(x))^2$$

Covariance of two groups of numbers describes how those numbers change together

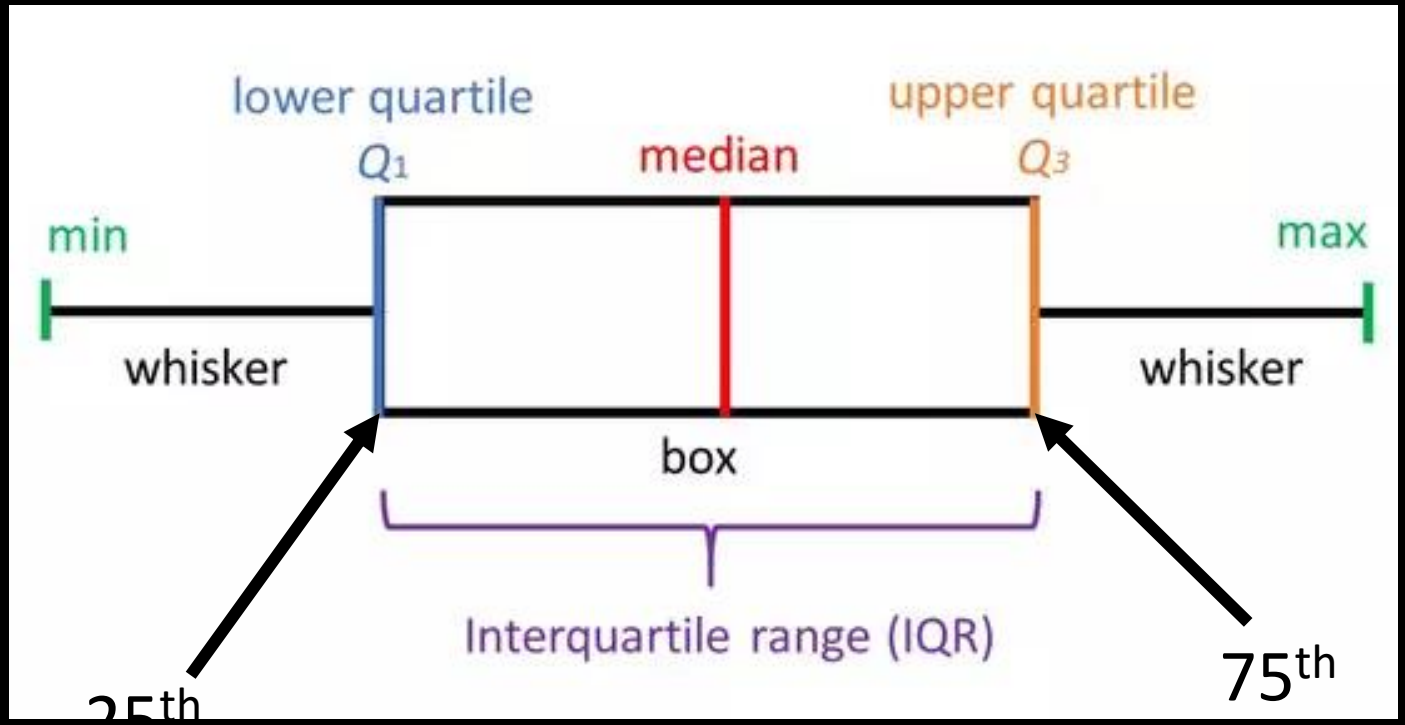
$$covariance = \sum_{i=1}^{n} (x_i - mean(x)) \times (y_i - mean(y))$$

$$standard\ deviation = \sqrt{covariance}$$

Correlation describes relationship between two groups of numbers. Some algorithms might exhibit poor performance if the data is highly correlated

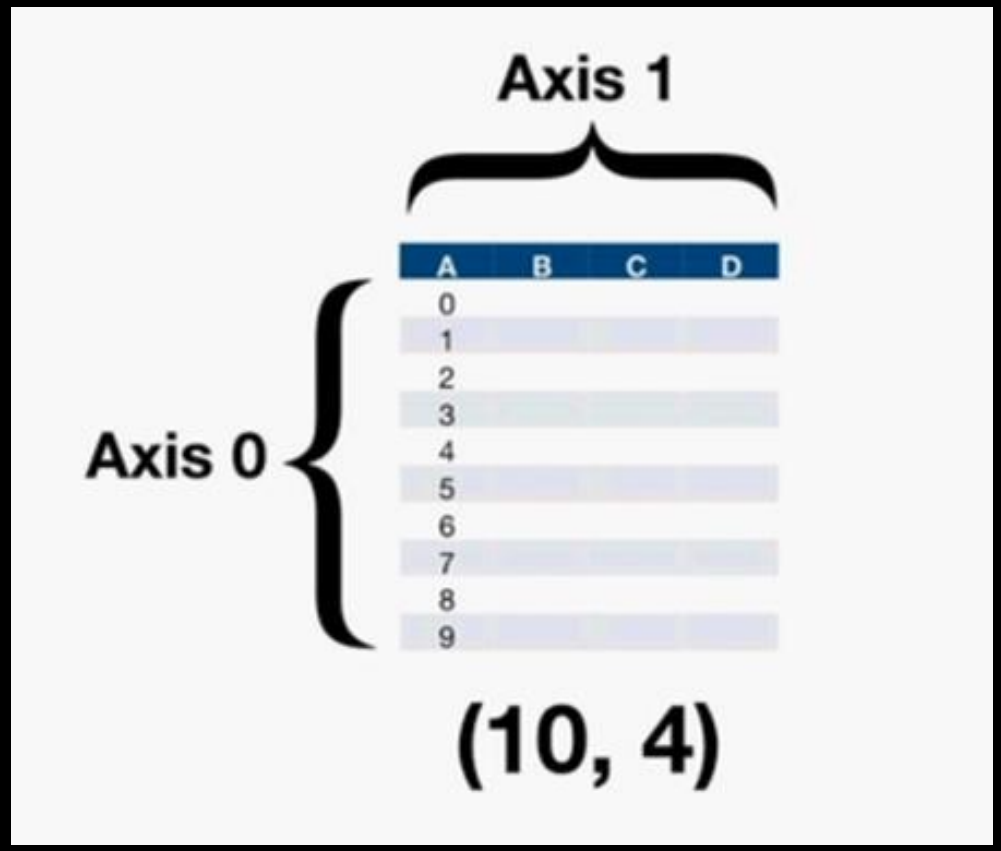
Percentile and Quartiles

- Percentile is a point at which *x*% of the values are less than the value at that point
- Interquartile range is the area in the middle of the distribution that contains 50% of the values (used for finding outliers)

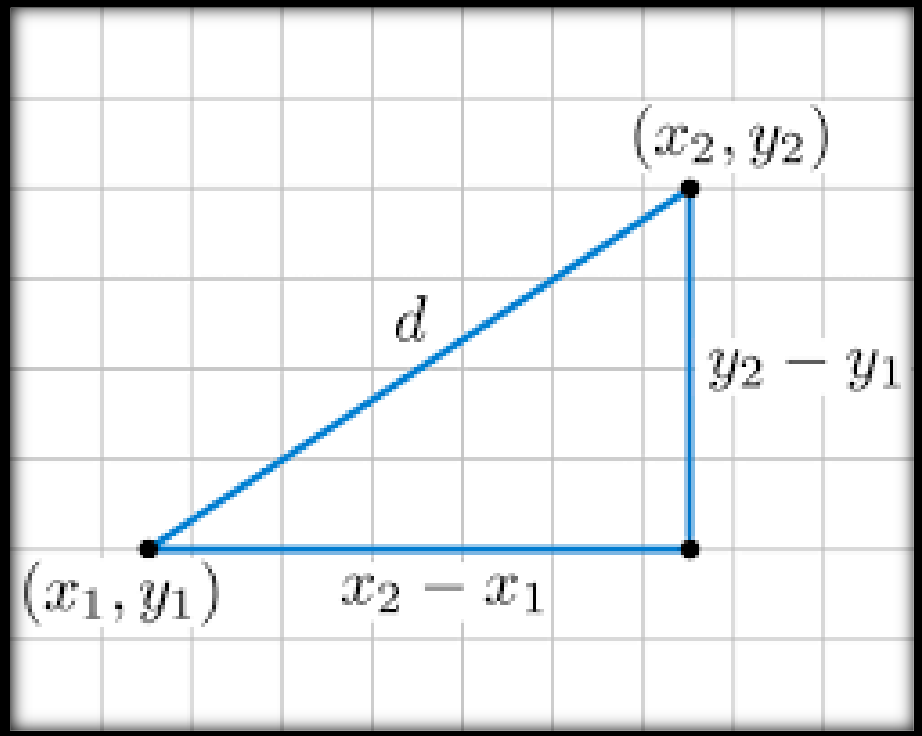


Axis and Shape

- Axis 0 represents rows, where as Axis 1 represents columns
- Shape is defined as (Axis 0, Axis 1). Used on a data set find the number of rows and columns



Euclidean Distance



$$x = [3,7] \qquad y = [6,11]$$

$$d = \sqrt{(x2 - x1)^2 + (y2 - y1)^2}$$

$$d = \sqrt{(6 - 3)^2 + (11 - 7)^2}$$

$$d = \sqrt{(3)^2 + (4)^2}$$

$$d = 5$$

Lab: Hands on

Spark Descriptive Statistics Lab

- Open notebook – “Spark Lab/Descriptive Statistics”
- Use “pima_diabetes.csv” file – dataset about Pima Indians diabetes. Whether a person has diabetes on or not based on number of features
- Contains examples using Pandas Dataframe
- Write code for Spark Dataframes (Marked in Red)

Machine Learning Algorithms

Machine Learning - Predictions

$y = f(X)$

$\hat{y} = \hat{f}(X)$ Prediction

$\hat{y} = 22.5$ Regression – continuous value

$\hat{y} = 0.89$ Classification – probability

$\hat{y} = 1$ Classification – prediction

Popular Algorithms

Supervised

- Linear Regression
- Logistic Regression
- Super Vector Machine (SVM)
- K Nearest Neighbors (KNN)
- Decision Trees
- Random Forest

Unsupervised

- K-Means
- Apriori
- C-Means

Linear Regression

Simple Linear Regression (Univariate)

Constant

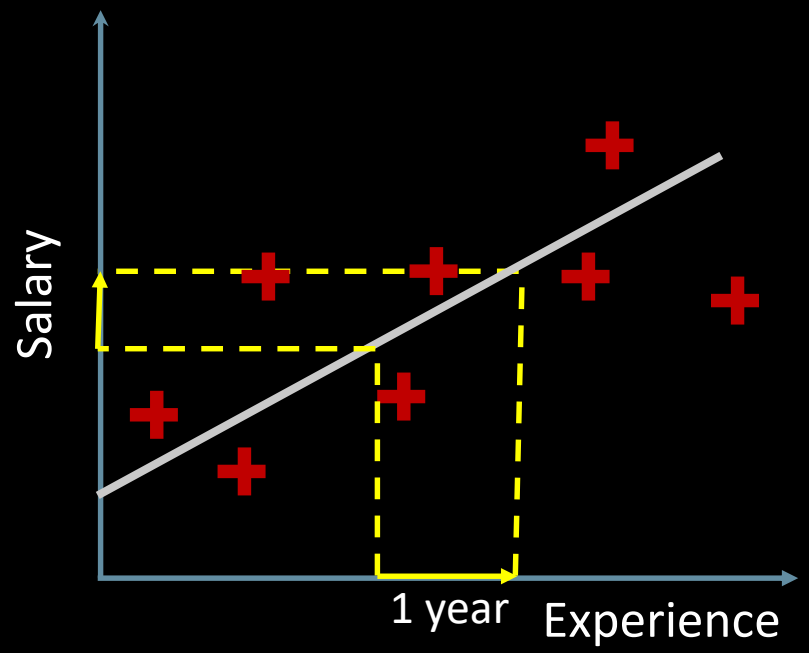
Coefficient

$$y = b_0 + b_1 * x_1$$

Dependent Variable

Independent Variable

$$y = b + mx$$



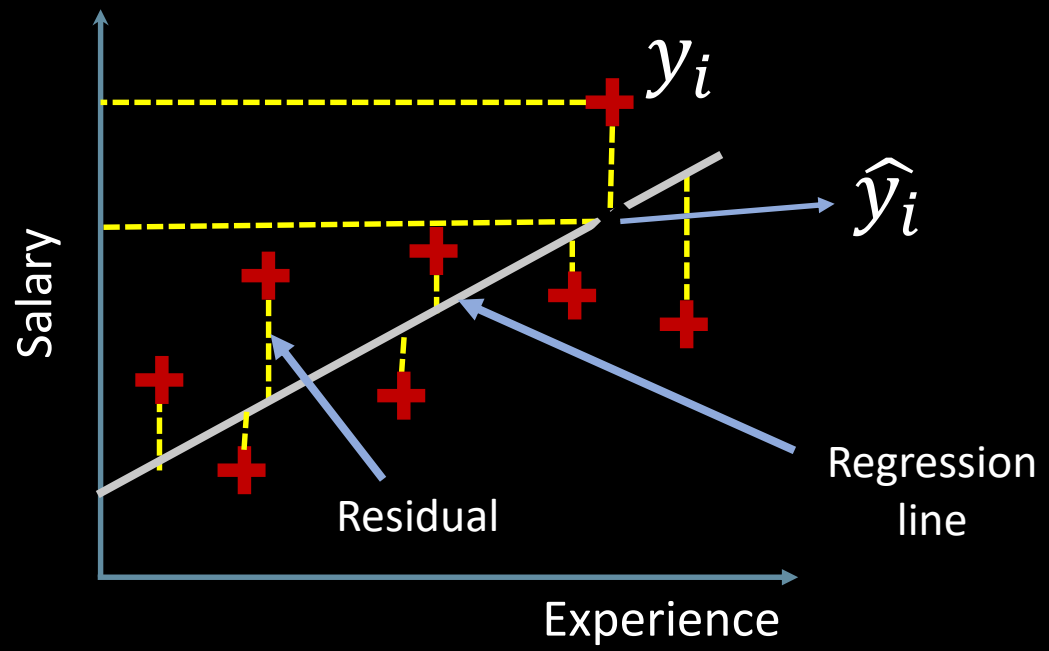
Salary = b_0 + b_1 * Experience

b_0 is the y intercept

b_1 is the slope of the line

Not just drawing a line, but trying to find the best fit line based on the observations

Find the Best Fitting Line



- The regression line is the line that best fits the data (in terms of having the smallest overall distance from the line to the points)
- This technique is used for finding the best fitting line is called “least squares method”

y_i Observed value

$\hat{y}_i$ Modeled value

SUM$(y_i - \hat{y}_i)^2$ → minimum

Multiple Linear Regression - Multivariate

$$y = b_0 + b_1x_1 + b_2x_2 \ \ldots\ldots bnxn$$

- y is the dependent variable
- x_i are the independent variables

Multiple Linear Regression

R&D Spend	Administration	Marketing Spend	State	Profit
165349.2	136897.8	471784.1	0	192261.83
162597.7	151377.59	443898.53	1	191792.06
153441.51	101145.55	407934.54	0	191050.39
144372.41	118671.85	383199.62	0	182901.99
142107.34	91391.77	366168.42	0	166187.94
131876.9	99814.71	362861.36	0	156991.12
134615.46	147198.87	127716.82	1	156122.51
130298.13	145530.06	323876.68	0	155752.6

$$profit = 1.63 + 7.91x_1 + 3.8x_2 + 6.88x_3 - 0.12x_4$$

x_1 = R&D Spend
x_2 = Administration
x_3 = Marketing Spending
x_4 = State (0 = California, 1 = New York)

Model Evaluations - Loss Function

- Loss Function
  - A method to evaluate how well our algorithm is modeling our dataset
  - If predictions are way off then the loss function will have higher value
  - Observing changes to the loss function helps understand algorithm changes
  - Absolute value of (prediction – actual value)
  - Also referred to as Cost Function

Loss Function – Mean Squared Error

- Mean Squared Error (MSE)
  - How close the regression line is to the set of points
  - Smaller the MSE, the closer the best fit line

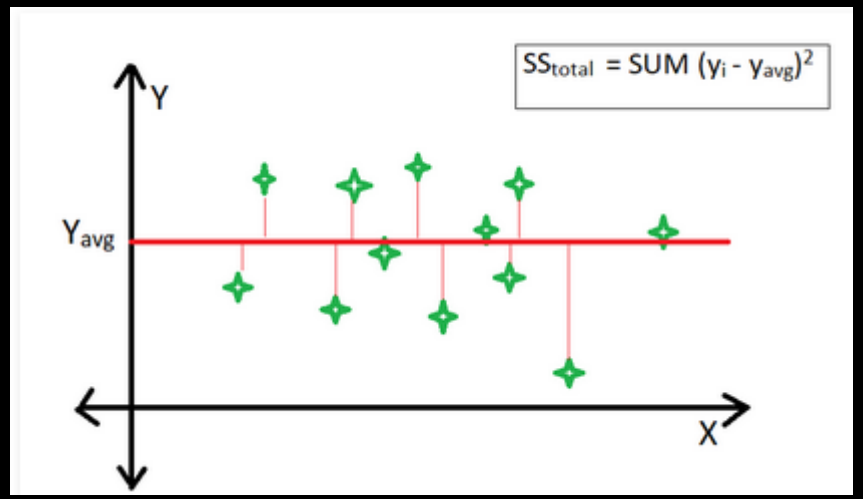
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y)^2$$

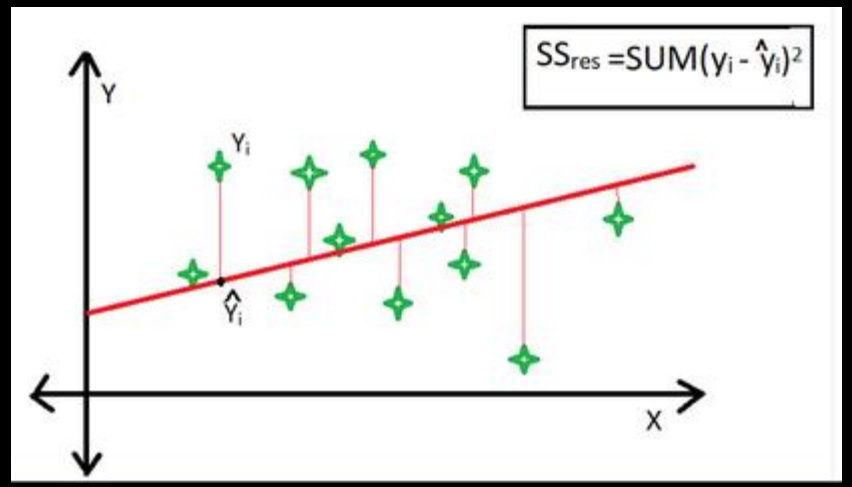
- Root Mean Squared Error (RMSE)
  - How spread out the residuals are from the regression line
  - Indicates how concentrated are the data points around the line

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y)^2}$$

R-Squared

- R^2
  - Statistical measure of how close the data are to the fitted line
  - Usually values between 0% and 100%
  - In general higher the R^2 value, the better the model fits the data.





$$R^2 = 1 - \frac{SSres}{SStot}$$

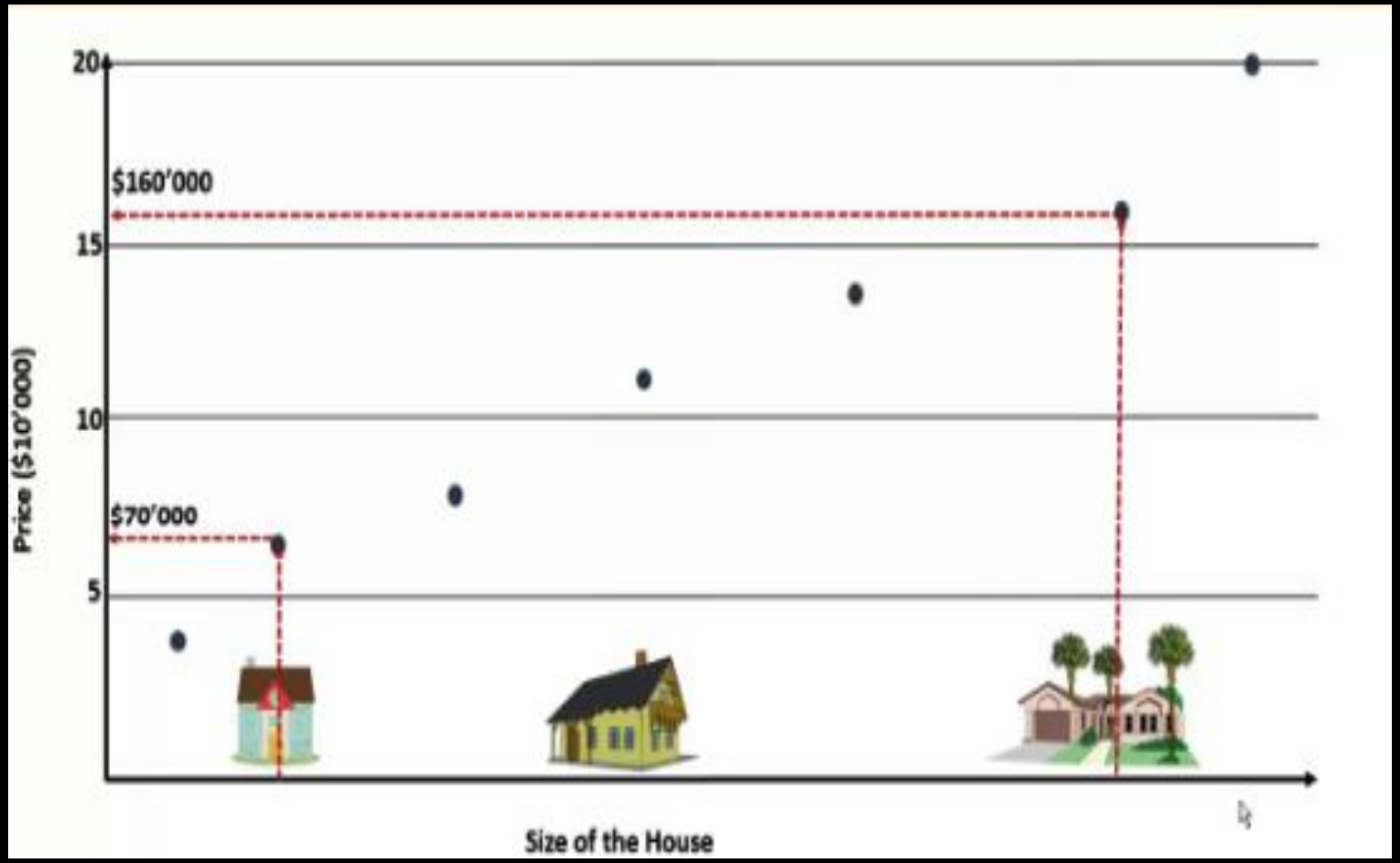
Linear Regression Demo

Linear Regression Hands-on

Real Use Case : House Price estimates

REDFIN

Zillow



Use case: Boston Real Estate Data

- Dataset – from scikit-learn datasets
- Linear Regression example with following features:

CRI: Per capita crime rate by town
ZN: Proportion of residential land zoned for lots over 25,000 sq. ft
INDUS: Proportion of non-retail business acres per town
CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX: Nitric oxide concentration (parts per 10 million)
RM: Average number of rooms per dwelling
AGE: Proportion of owner-occupied units built prior to 1940
DIS: Weighted distances to five Boston employment centers
RAD: Index of accessibility to radial highways
TAX: Full-value property tax rate per $10,000
PTRATIO: Pupil-teacher ratio by town
B: 1000(Bk — 0.63)², where Bk is the proportion of [people of African American descent] by town
LSTAT: Percentage of lower status of the population
PRICE: Median value of owner-occupied homes in $1000s

Spark Lab/Spark Linear Regression Real Estate