Machine Learning


1
2
Data Science

What is Data Science ?

- Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.[1][2] Data science is related to data mining, machine learning and big data
- Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data
- It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science.

Quick overview of the process

data engineers

Data Team

- Make the data available
- Clean the data
  - missing data
  - data type
  - textual to numerical
  - scaling

Sources
- files
- db
- warehouse

No	Year	Profit	...
1	1990	50	
2	1991	70	
3	1992	75	
4	1993	80	

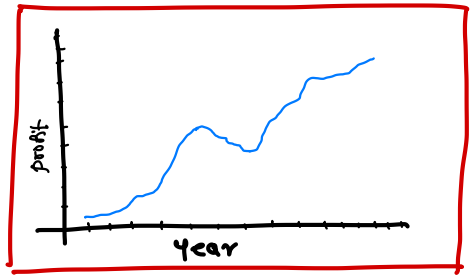
Cleaned / processed data

⇒

business Analysts

Business Intelligence Team

- Find business insights
- Prepares the dashboard

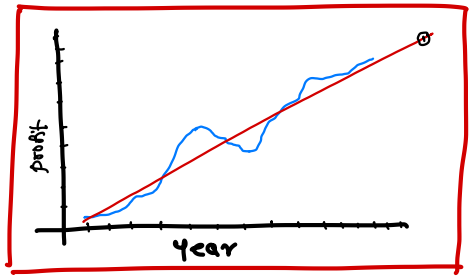


⇒

Statistician / machine learning dev

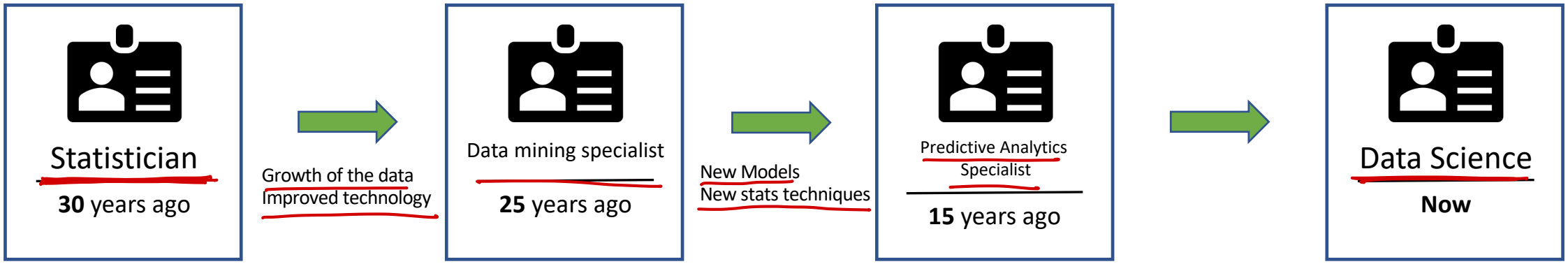
Data Science Team

- Use data analytics tools
- Develop models for various tasks



model = formula
- Future value
  - Regression
  - Classification

How is it evolved ?



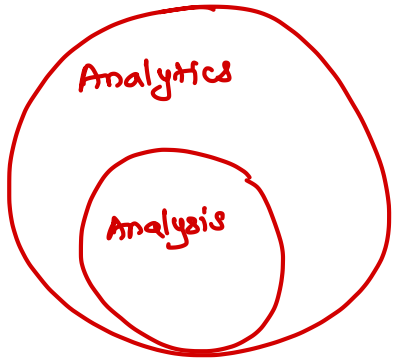
Responsible for
- Gathering data
- Cleaning data
- Applying statistical methods
- Analyzing data

Responsible for
- Extracting patterns from the data

Responsible for
- Perform more accurate forecast

Analysis vs Analytics

- Analysis
  - Analysis is performed on the things that are already happened in the past
  - We do Analysis to explain How and or Why something happened
  - E.g.
    - Analyzing data by separating into chunks



- Analytics
  - Refers to the future after finding patterns
  - We use Analytics to explore potential future events
  - Branches
    - Qualitative ← textual / categorical
      - Intuition and experience
      - Analysis
    - Quantitative
      - Formulas → stats
      - Algorithms

Jargons

- Business Intelligence (BI)
  - Process of analyzing and reporting historical business data
  - Aims to explain past events using business data
  - Can be used for taking strategic decision
- Machine Learning (ML)
  - The ability to predict outcomes without being explicitly programmed
  - It is about creating and implementing algorithms that let machines receive data and use this data to
    - Make predictions
    - Analyze patterns
    - Give recommendations
  - Machine Learning can not be implemented without data
- Artificial Intelligence (AI)
  - Simulating human knowledge and decision making with computers
  - Managed to reach AI through the Machine Learning

Data Science Techniques

Data Collection

- Raw data
  - Can not be analyzed straight away
  - It is untouched data that is accumulated and stored on the server
  - Also known as raw facts or primary data
  - Can be collected by various techniques like
    - Survey
    - Automated tools

Data Pre-Processing

- This process tries to fix the problem that has occurred while data gathering
- Before processing with data analysis, it is important to remove the wrong data
- Techniques
  - Class Labeling
    - Labeling the data to the correct data types
    - e.g. numeric and categorical
  - Data cleansing
    - Deal with inconsistent data
    - Also known as data cleaning or data scrubbing
  - Dealing with missing values
  - Data balancing
  - Data shuffling
    - Prevents unwanted patterns
    - Improves predictive performance

No textual

Analyzing Data

- Once the data is cleaned and formatted, it can be analyzed for various reasons
- It explains past performance
- It can answer simple questions like
  - What happened ?
  - When did it happen?
- Or it can answer complex questions like
  - How did marketing team performed last quarter in terms of revenue
  - How does that compare to the performance in the same quarter last year
- Frequently used terms
- Metric
  - used to gauge the business performance or progress
  - metric = measure + business meaning
- Key Performance Indicators (KPI):
  - Key: related to the business goals
  - Performance: how successfully you have performed within a specified timeframe
  - Indicators: shows values indicates somethings about the business
- dashboards

Predictive Analytics - Traditional

- After the analysis is over, the next logical step is analytics
- It can be performed traditional statistical modelling like
  - Regression
    - A model used for quantifying causal relationships among the different variables included in your analysis
    - Mostly used for predicting future values
  - Clustering
    - Creating different clusters (groups) by understanding data
  - Factor Analysis
  - Time Series analysis

Predictive Analytics – Machine Learning

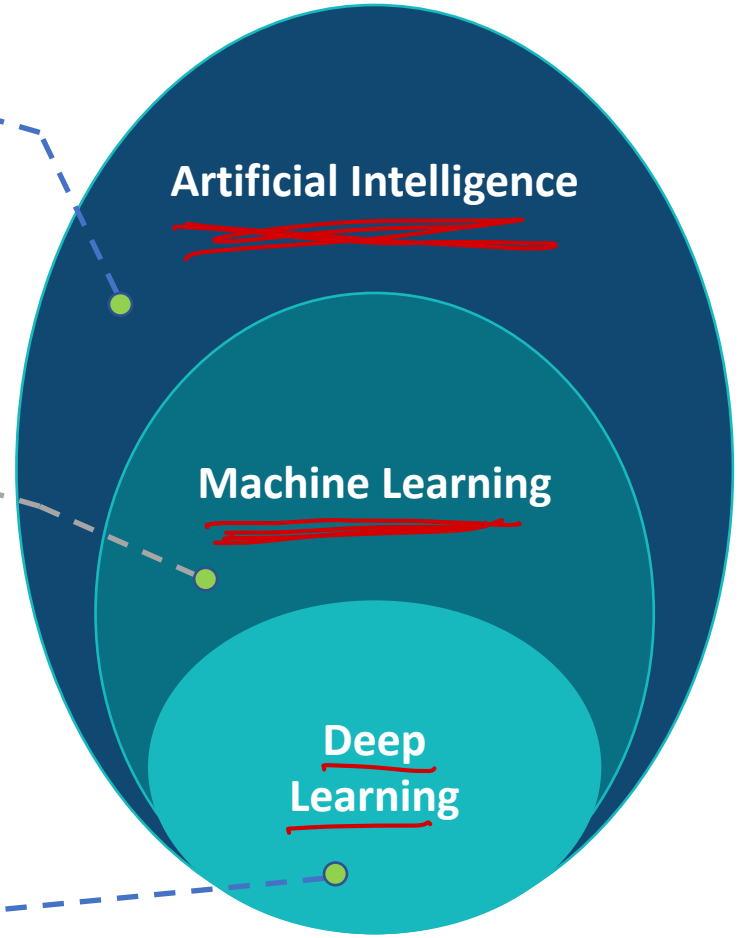
- Utilizes artificial intelligence to predict behavior in unprecedented ways
- There are different techniques
  - Supervised Learning
  - Unsupervised Learning
  - Reinforcement Learning

Biggest confusion

Artificial Intelligence:
- A technique which enables machine to mimic human behavior

Machine Learning:
- Subset of AI which uses statistical methods to enable machines to improve the experience

Deep Learning:
- Subset of ML which makes the computation of multi-layer neural network feasible



Artificial Intelligence

When did it start?

- Greek Mythology – Talos
  - Talos was a giant animated bronze warrior programmed to guard the island of Crete
- 1950 – Alan Turing
  - Alan Turing published a landmark paper in which he speculated about the possibility of creating machines that think
  - What he created is known as Turing Test which is used to determine whether or not the computer can think intelligently like human being
- 1951 – Game AI
  - Christopher Strachey wrote a checkers program and Dietrich Prinz wrote one for chess
- 1956 – The birth of AI
  - John McCarthy first coined the term Artificial Intelligence at Dartmouth Conference
- 1959 – First AI laboratory
  - MIT AI lab was first set up in 1959 and research on AI began

When did it start?

- 1960 – General Motors Robot
  - First robot was introduced to General Motors assembly line
- 1961 – First chatbot
  - The first AI chatbot called ELIZA was introduced in 1961
- 1997 - IBM Deep Blue
  - IBM's Deep Blue beats world champion Garry Kasparov in the game of chess
- 2005 - DARPA Grand Challenge
  - Stanford Racing Team's autonomous robotic car, Stanley wins the 2005 DARPA Grand Challenge
- 2011 – IBM Watson
  - IBM's question answering system, Watson, defeated the two grated Jeopardy champions Brad Ruther and Ken Jennings

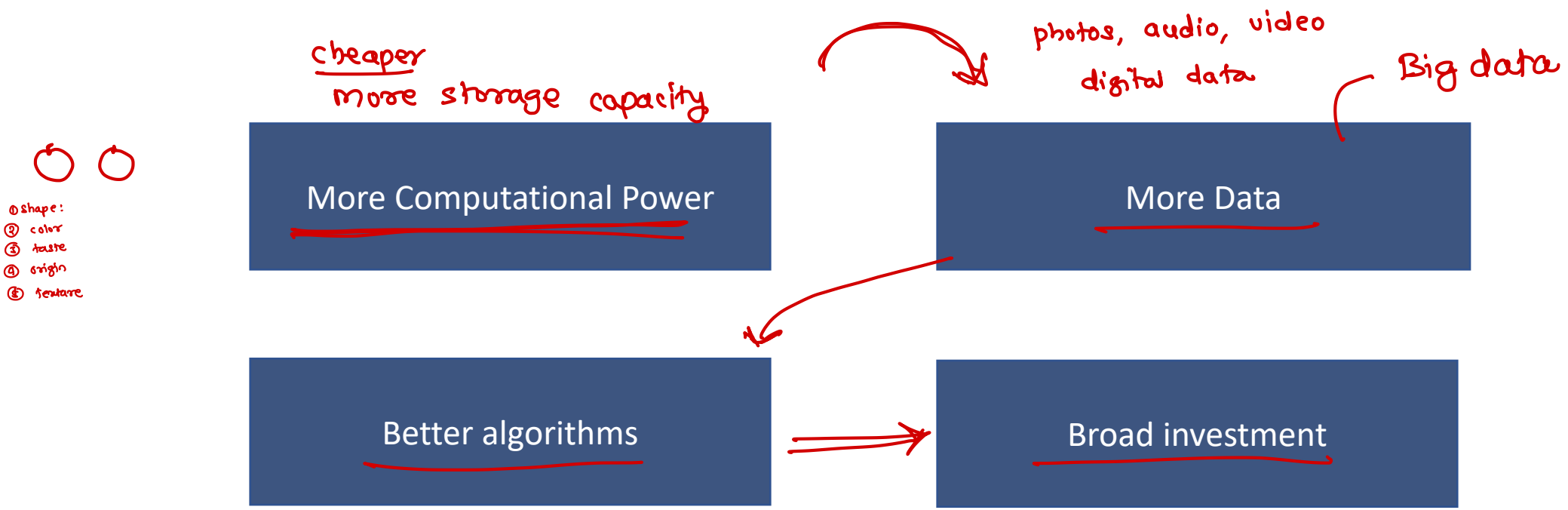
What is AI?

- Artificial Intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans
- Any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals
- The theory and development of computer system able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision making and translation
- Often used to describe machines (or computers) that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving"

Aspects of AI (1955)

- Simulating higher functions of the human brain
- Programming a computer to use general languages
- Arranging hypothetical neurons in a manner so that they can form concepts
- A way to determine and measure problem complexity
- Self-improvement
- Abstraction: defined as the quality of dealing with ideas rather than events
- Randomness and creativity

Why are we talking about it now ?



AI applications

- Google's search engine
- JPMorgan Chase's Contract Intelligence (COiN) platform uses AI, machine learning and image recognition software to analyse legal documents
- IBM Watson: Healthcare organizations use IBM AI (Watson) technology for medial diagnosis
- Google's AI Eye Doctor can examine retina scans and identify a condition called as diabetic retinopathy which can cause blindness
- Facebook uses ML and DL to detect facial features and tag your friends
- Twitter uses AI to identify hate speech and terroristic language in the tweets
- Smart Assistants: Siri, Google Assistant, Alexa, Cortana
- Tesla automated cars
- Netflix uses AI for movie recommendations
- Spam filtering

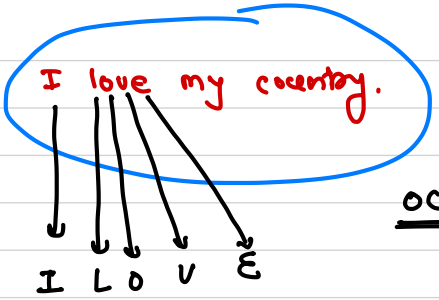
Machine Learning

What is machine learning ?

- A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E
  - Tom Mitchell, 1997

- Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed
  - Arthur Samuel, 1959

- Machine Learning is the science (and art) of programming computers so they can learn from data



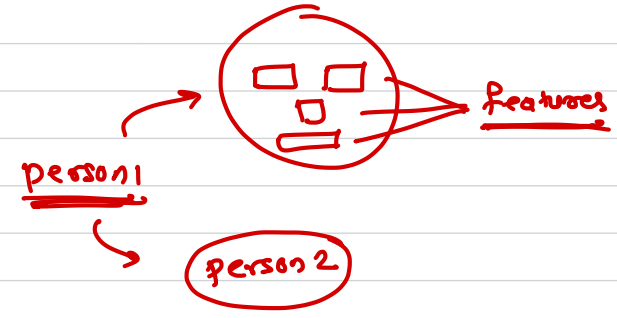
OCR → tensorflow + open CV

CNN

I → letter I

A → letter A

B C D



A

Where to use machine learning ?

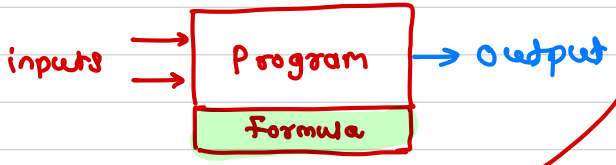
- Problems for which existing solutions require a lot of fine-tuning or long lists of rules:
  - one Machine Learning algorithm can often simplify code and perform better than the traditional approach
- Complex problems for which using a traditional approach yields no good solution:
  - the best Machine Learning techniques can perhaps find a solution
- Fluctuating environments:
  - a Machine Learning system can adapt to new data
- Getting insights about complex problems and large amounts of data

email

① sender ←
② contains words
— lottery.....

```
type = ''
if (email.sender == '100.200.300.400') {
    type = 'spam'
} else if (email contains ['...', '...']) {
    type = 'spam'
} else {
    type = 'ham'
}
```

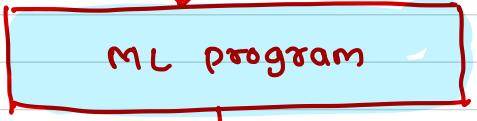
Traditional Approach



ML approach

heartRate	chestPain	heartAttack
78	100	0
90	200	1

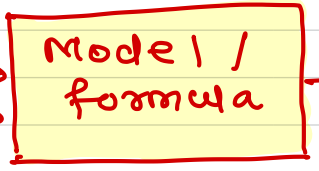
data (answers)



patient:

heart Rate = 80

chest pain = 150



heart Attack

→ 1 ✓

→ 0 ✓

Examples of Applications

- Analyzing images of products on a production line to automatically classify them
  - This is image classification, typically performed using convolutional neural networks (CNN)
- Detecting tumors in brain scans
  - This is semantic segmentation, where each pixel in the image is classified (typically use CNNs)
- Automatically classifying news articles
  - This is natural language processing (NLP), and more specifically text classification (sentiment analysis, email is spam/ham)
- Automatically flagging offensive comments on discussion forums
  - This is also text classification, using the same NLP tools
- Forecasting your company's revenue next year, based on many performance metrics
  - This is a regression task (i.e., predicting values) that may be tackled using any regression model
- Making your app react to voice commands
  - This is speech recognition, which requires processing audio samples: since they are long and complex sequences, they are typically processed using RNNs, CNNs, or Transformers

Examples of Applications

- Detecting credit card fraud
  - This is anomaly detection example
- Segmenting clients based on their purchases so that you can design a different marketing strategy for each segment
  - This is clustering example
- Representing a complex, high-dimensional dataset in a clear and insightful diagram
  - This is data visualization, often involving dimensionality reduction techniques
- Recommending a product that a client may be interested in, based on past purchases
  - This is a recommender system
- Building an intelligent bot for a game
  - This is often tackled using Reinforcement Learning

Types

Types of machine learning

- There are so many different types of Machine Learning systems that it is useful to classify them in broad categories, based on the following criteria
  - Whether or not they are trained with human supervision
    - supervised, unsupervised, and Reinforcement Learning
  - Whether or not they can learn incrementally on the fly
    - online versus batch learning
  - Whether they work by simply comparing new data points to known data points, or instead by detecting patterns in the training data and building a predictive model, much like scientists do
    - instance-based versus model-based learning

Supervised
Unsupervised
Reinforcement Learning

Supervised Learning

- The majority of practical machine learning uses supervised learning
- Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output

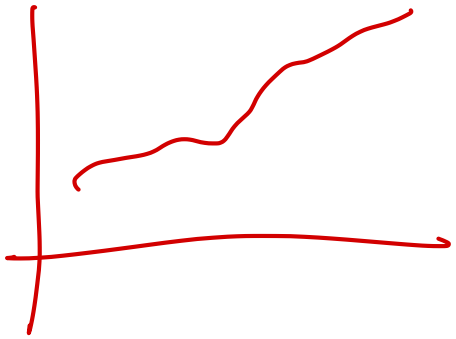
independent variable(s)
dependent variable

$$Y = f(X)$$

- The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data
- It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process
- We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher
- Learning stops when the algorithm achieves an acceptable level of performance

Supervised Learning – Problems

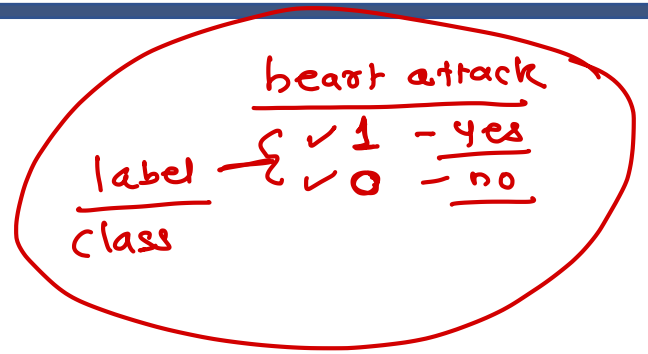
- **Regression**
  - Related to predicting future values
  - E.g.
    - Population growth prediction
    - Expecting life expectancy
    - Market forecasting/prediction
    - Advertising Popularity prediction
    - Stock prediction
  - Algorithms
    - Linear and multi-linear regression
    - Logistic regression
    - Naïve Bayes
    - Support Vector Machine
    - Ridge
    - Lasso



output variable = discrete

Supervised Learning – Problems

- **Classification**
  - Related to classify the records
  - E.g.
    - Find whether an email received is a spam or ham
    - Identify customer segments
    - Find if a bank loan is granted
    - Identify if a kid will pass or fail in an examination
  - Algorithms
    - Logistic Regression
    - Decision Tree
    - Random Forest
    - Support Vector Machine
    - K-nearest neighbor



seg1 / seg2 / seg3

yes / No

Output variable = categorical

Unsupervised Learning

- Unsupervised learning is where you only have input data (X) and no corresponding output variables
- The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data
- These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher
- Algorithms are left to their own devises to discover and present the interesting structure in the data

Unsupervised Learning - Problems

- **Clustering**
  - discover the inherent groupings in the data, such as grouping customers by purchasing behaviour
  - E.g.
    - Batsman vs bolwer
    - Customer spending more money vs less money
  - Algorithms
    - K-means clustering
    - Hierarchical clustering

Unsupervised Learning - Problems

- **Association**
  - An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y
  - E.g.
    - Market basket analysis
  - Algorithms
    - Apriori
    - Eclat

Reinforcement Learning

- It is about taking suitable action to maximize reward in a particular situation
- It is employed by various software and machines to find the best possible behavor or path it should take in a specific situation
- Reinforcement learning differs from the supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task
- In the absence of training dataset, it is bound to learn from its experience

Reinforcement Learning

- Examples
  - Resources management in computer clusters
  - Traffic Light Control
  - Robotics
  - Web system configuration
  - Chemistry
- Algorithms
  - Q-Learning
  - Deep Q-Learning