Data Analytics Workshop

Madhurima Panja
Center for Data Science,
IIIT Bangalore

Recent Trends and Buzzwords

Recent Trends and Buzzwords

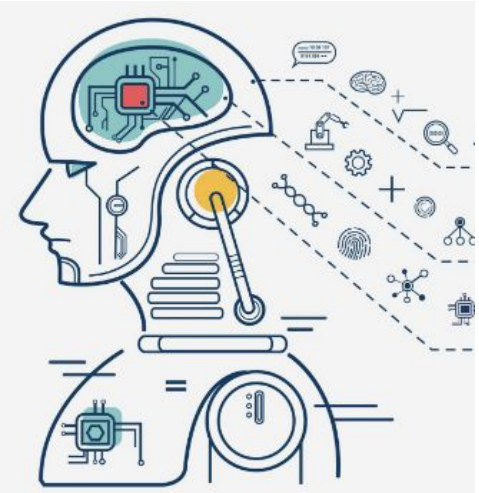
1. Statistics is the study of the collection, analysis, interpretation, presentation and organization of data.

2. Data science is the study of the generalizable extraction of knowledge from data, yet the key word is science.

3. Machine learning is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed.

4. Artificial Intelligence research is defined as the study of intelligent agents: any device that perceives its environment and takes actions that maximize its chance of success at some goal.

5. Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.



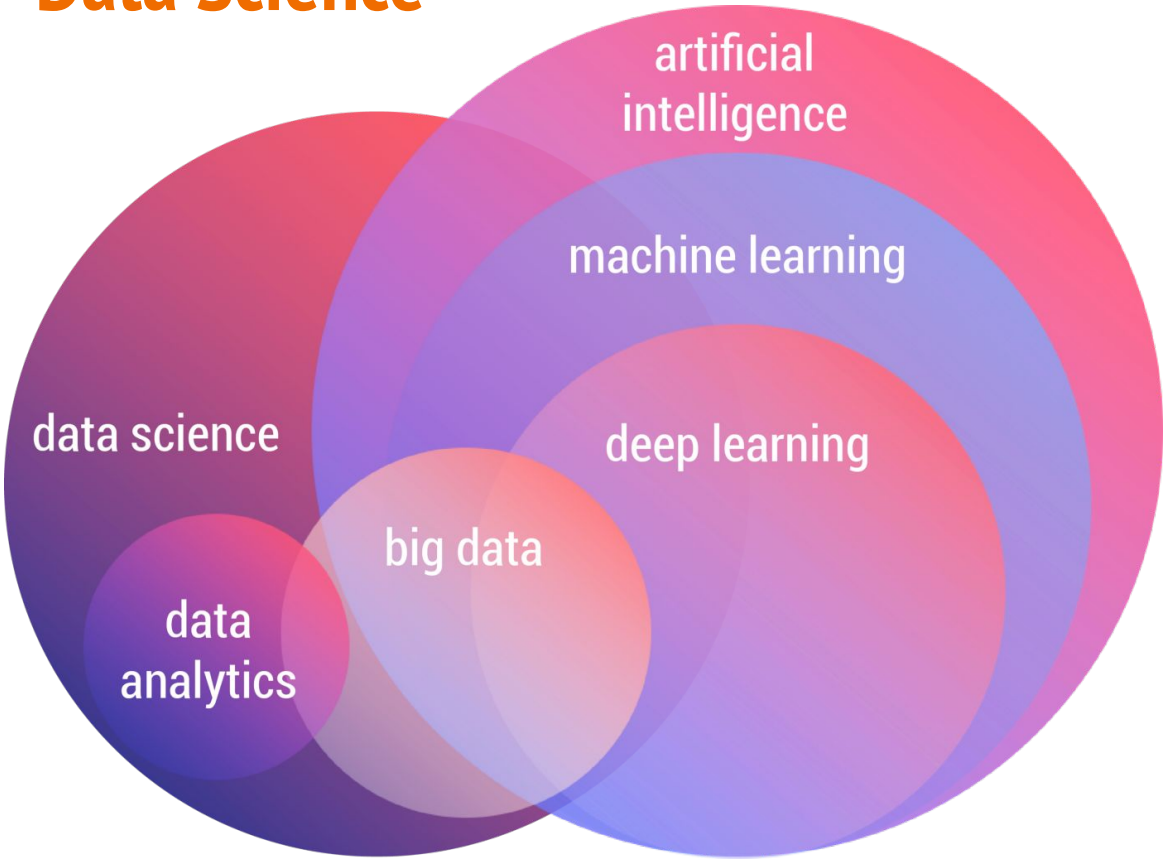
Topic 1: STATISTICS

Statistics

- Data : Large bodies of data with complex data structures are generated from computers, sensors, manufacturing industries, etc.
- Models : Non/Semiparametric models but in complex probability spaces / high-dimensional functional spaces (e.g., deep neural net, reinforcement learning, decision trees, etc.).
- Emphases : Making predictions, causation, algorithmic convergence.
- Data are necessary and at the core of Statistical Learning, Data Science & Machine Learning.
- Statistics : Not only has strong interactions with Probability but also other parts of Data Science (Machine Learning, Artificial Intelligence, etc.).

Topic 2: DATA SCIENCE

Data Science



"***Data is the new oil.*** It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."
- Clive Humby, UK Mathematician and Architect of Tesco's Clubcard.

Role of data: Present

The World is Data Rich

Astronomy

Social Networks

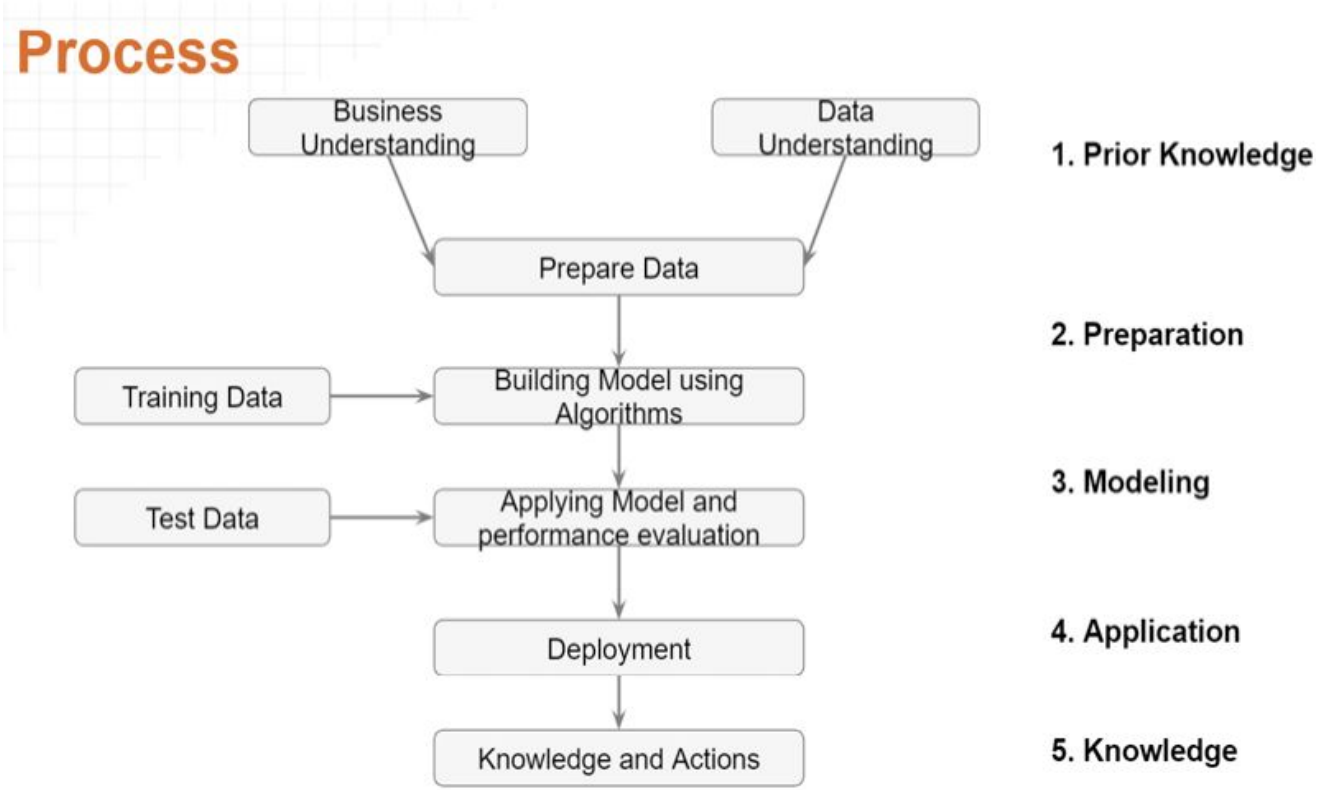
Healthcare

Banking

Genomics

Weather measurements

Data Mining Process



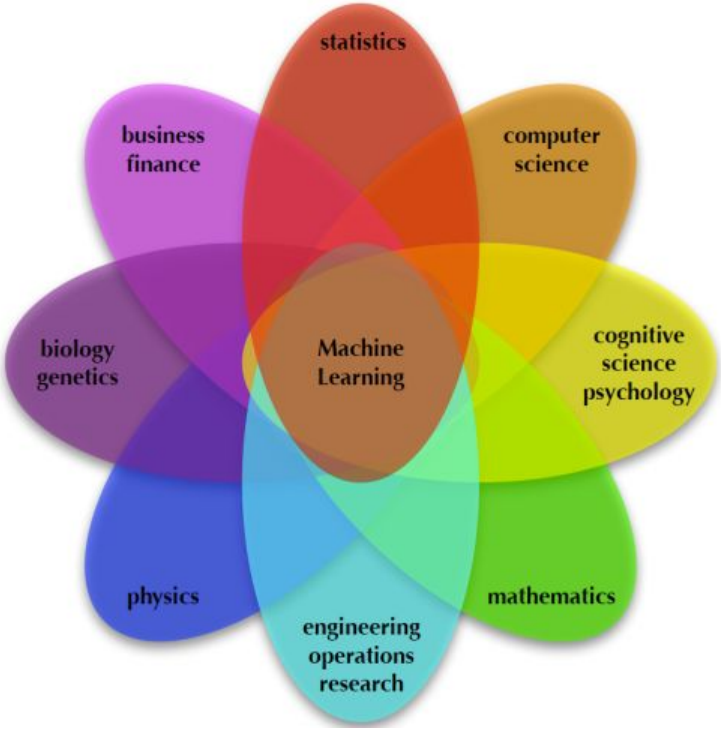
Overview of data science tools

Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of predefined classes. The prediction will be based on learning from known data set.	Decision Trees, Neural networks, Bayesian models, Induction rules, K nearest neighbors	Assigning voters into known buckets by political parties eg: soccer moms. Bucketing new customers into one of known customer groups.
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from known data set.	Linear regression, Logistic regression	Predicting unemployment rate for next year. Estimating insurance premium.
Anomaly detection	Predict if a data point is an outlier compared to other data points in the data set.	Distance based, Density based, LOF	Fraud transaction detection in credit cards. Network intrusion detection.
Time series	Predict if the value of the target variable for future time frame based on history values.	Exponential smoothing, ARIMA, regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the data set based on inherit properties within the data set.	K means, density based clustering - DBSCAN	Finding customer segments in a company based on transaction, web and customer call data.
Association analysis	Identify relationships within an itemset based on transaction data.	FP Growth, Apriori	Find cross selling opportunities for a retailor based on transaction purchase history.

Topic 3: MACHINE LEARNING

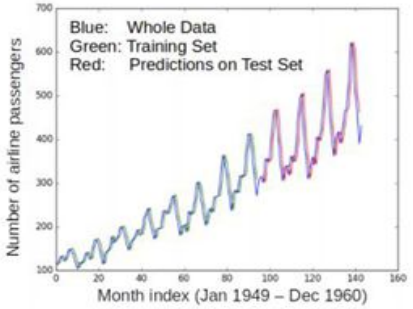
Machine Learning

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.



Introduction to Machine Learning

- Designing algorithms that ingest data and learn a model of the data.
- The learned model can be used to
  1. Detect patterns/structures/themes/trends etc. in the data
  2. Make predictions about future data and make decisions



- Modern ML algorithms are heavily "data-driven".
- Optimize a performance criterion using example data or past experience.

Machine Learning Is Impacting Our Life

8:00 am

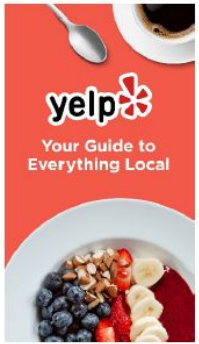
10:00 am

6:00 pm

8:00 pm

8:30 am

4:00 pm



What Machine Learning can do?



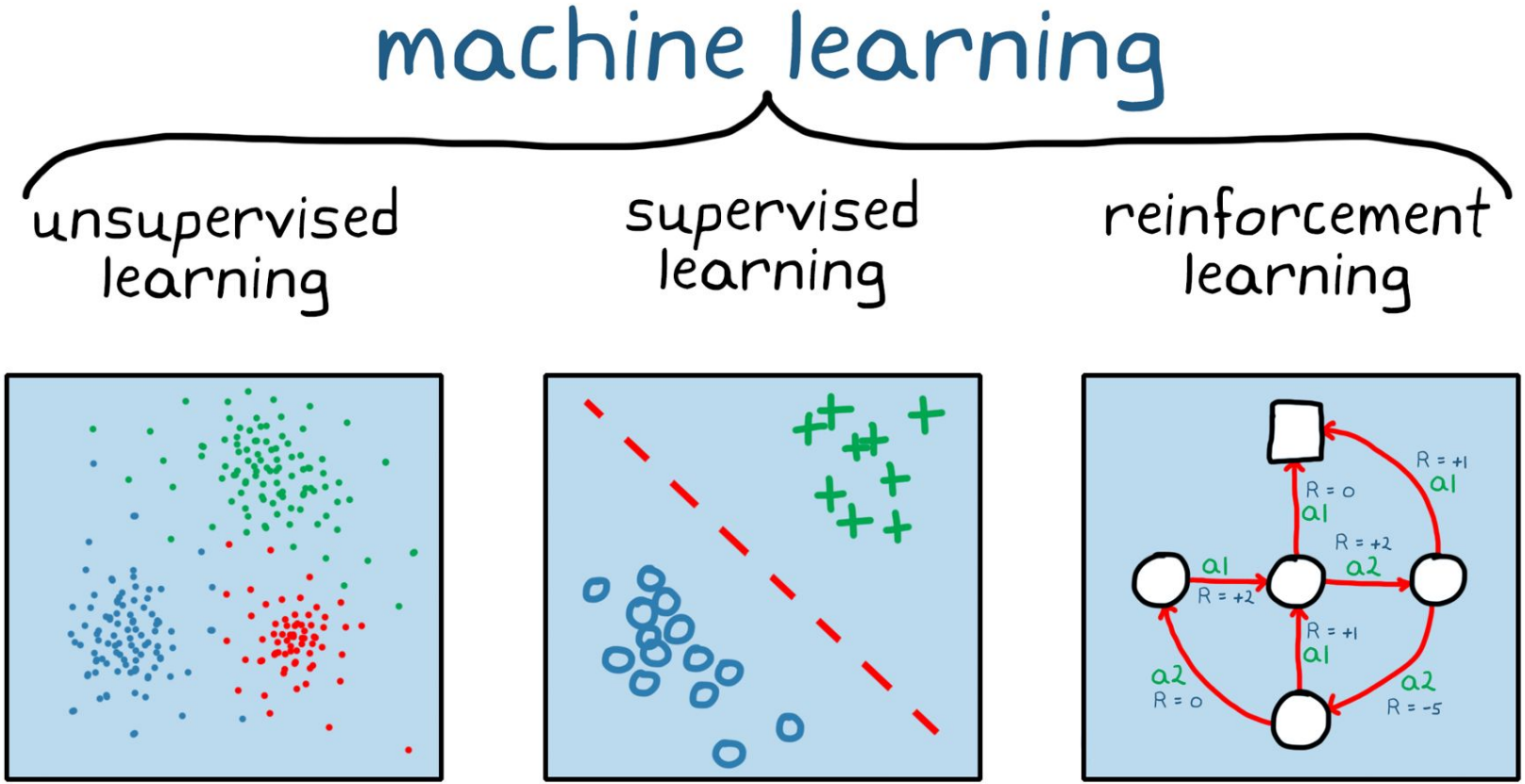



Yes


Maybe

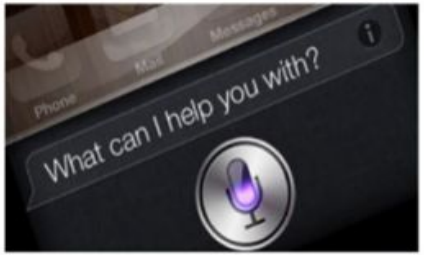
No

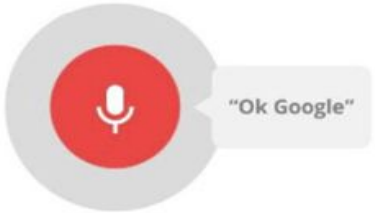
Types of Machine Learning



Applications of Machine Learning In Real World

Broadly applicable in many domains (e.g., internet, robotics, healthcare and biology, computer vision, NLP, databases, computer systems, finance, etc.).





Predictive Policing

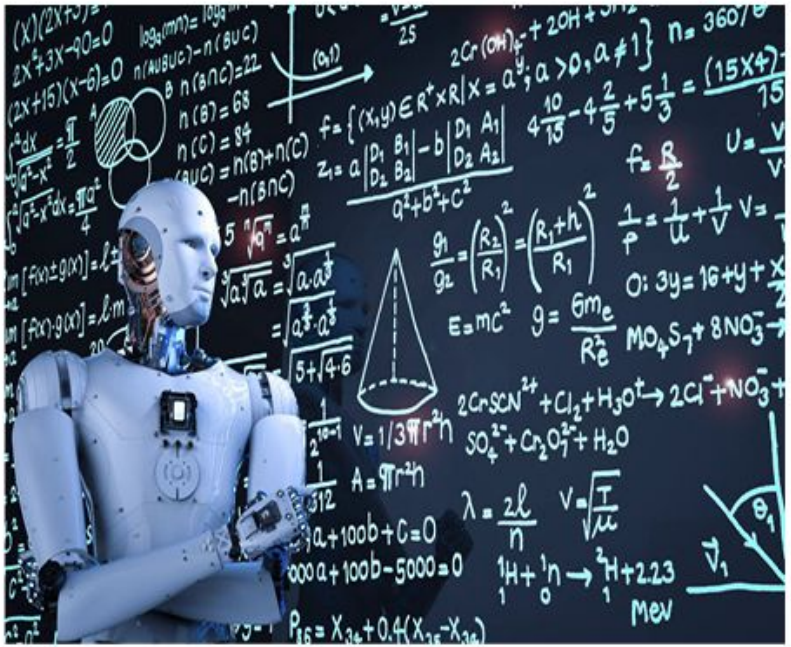
Online Fraud Detection

Topic 4: ARTIFICIAL INTELLIGENCE

A First Look at Artificial Intelligence

- What is Artificial Intelligence?
- What are the main challenges?
- What are the applications of AI?
- What are the issues raised by AI?

- On September 1955, a project was proposed by McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon introducing formally for the first time the term "Artificial Intelligence".



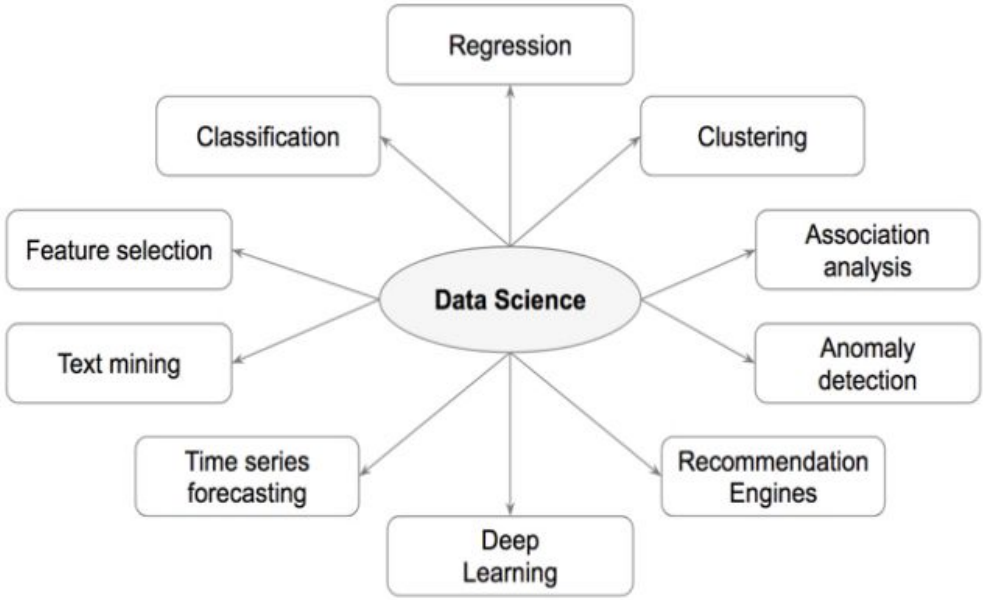
AI Technology: Autonomous Cars

- Originates from 1920 (NY)
- First use of neural networks to control autonomous cars (1989)
- Four US states allow self-driving cars (2013)
- First known fatal accident (May 2016)
- Singapore launched the first self-driving taxi service (Aug. 2016)
- A Arizona pedestrian was killed by an Uber self-driving car (March 2018).

AI Technology: Virtual Assistant / Chatbot

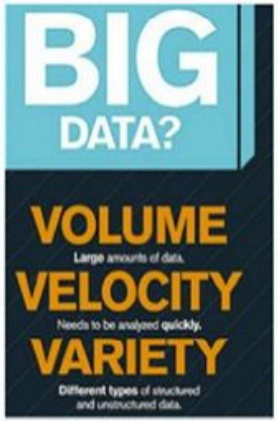
- Voice recognition tool "Harpy" masters about 1000 words (1970s, CMU, US Defense).
- System capable of analyzing entire word sequences (1980).
- Siri was the first modern digital virtual assistant installed on a smartphone (2011).
- Watson won the TV show Jeopardy! (2011).

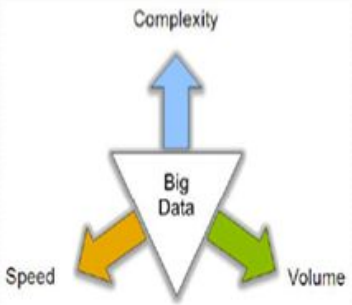
"*When you're fundraising,* it's AI.
When you're hiring, it's ML.
When you're implementing, it's Linear Regression.
When you're debugging, it's printf()."
- Baron Schwartz, Founder and CEO of VividCortex, 2017.



Topic 5: BIG DATA

Characteristics of Big Data: V3





Big Data: V3

Volume:

- Volume of data that needs to be processed is increasing rapidly.
- Need more storage capacity.
- Need more computation facility.
- Need more tools and techniques.

Variety:

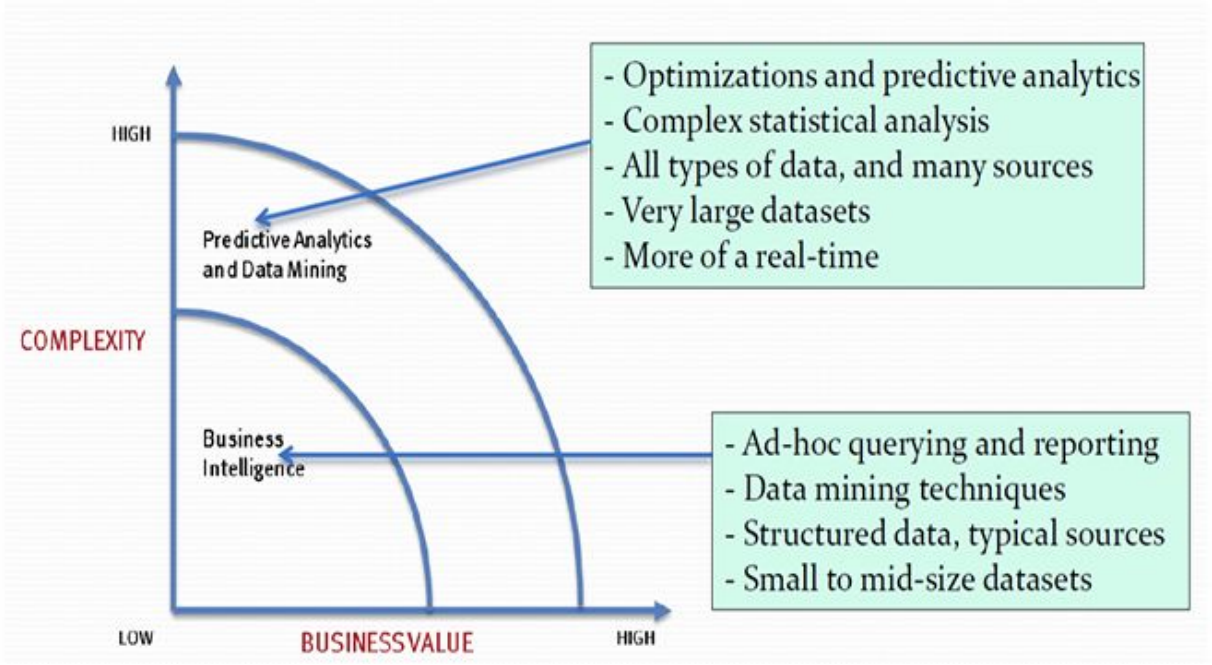
- Various formats, types, and structures.
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dimensional arrays, etc.
- A single application can be generating/collecting many types of data.

Velocity:

- Data is being generated fast and need to be processed fast.
- For time sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.
- Analyze 500 million daily call detail records in real-time to predict customer churn faster.

Big Data vs Small Data

Big data is more real time in nature than traditional applications...



Far From Terminator

- Stephen Hawking BBC, Dec 2 2014
  The development of full artificial intelligence could spell the end of the human race. We cannot quite know what will happen if a machine exceeds our own intelligence, so we can't know if we'll be infinitely helped by it, or ignored by it and sidelined, or conceivably destroyed by it.

R Statistical Software

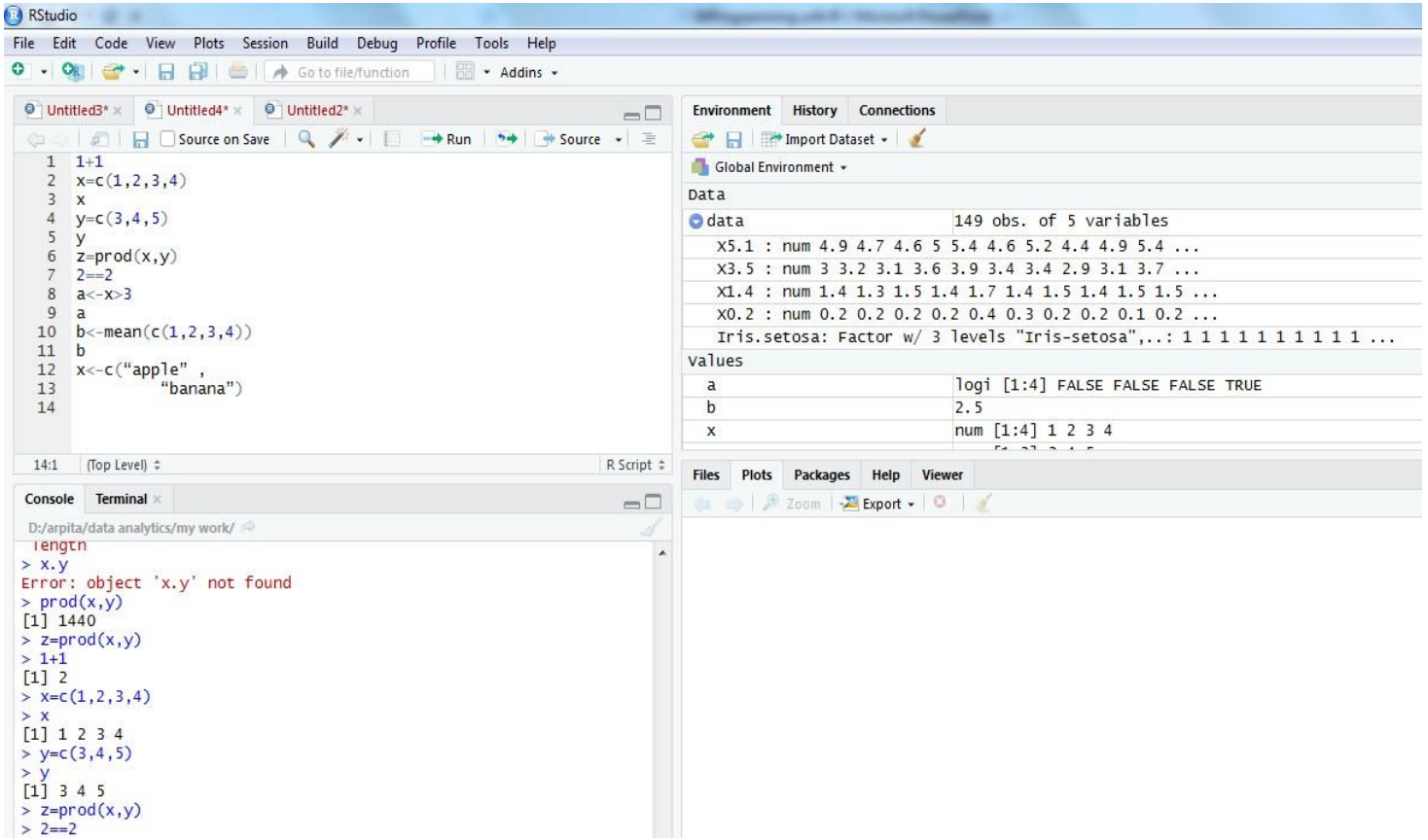
History of R...

- R is an open source programming language and software environment for statistical computing and graphics.
- The R language is widely used among statisticians and data scientists for developing statistical and data analytics tools.
- Modelled after S & S-plus, developed at AT&T labs in late 1980s.
- R project was started by Robert Gentleman and Ross Ihaka Department of Statistics, University of Auckland (1995).
- Currently maintained by R core development team – an international team of volunteer developers (since 1997).

Download R & RStudio

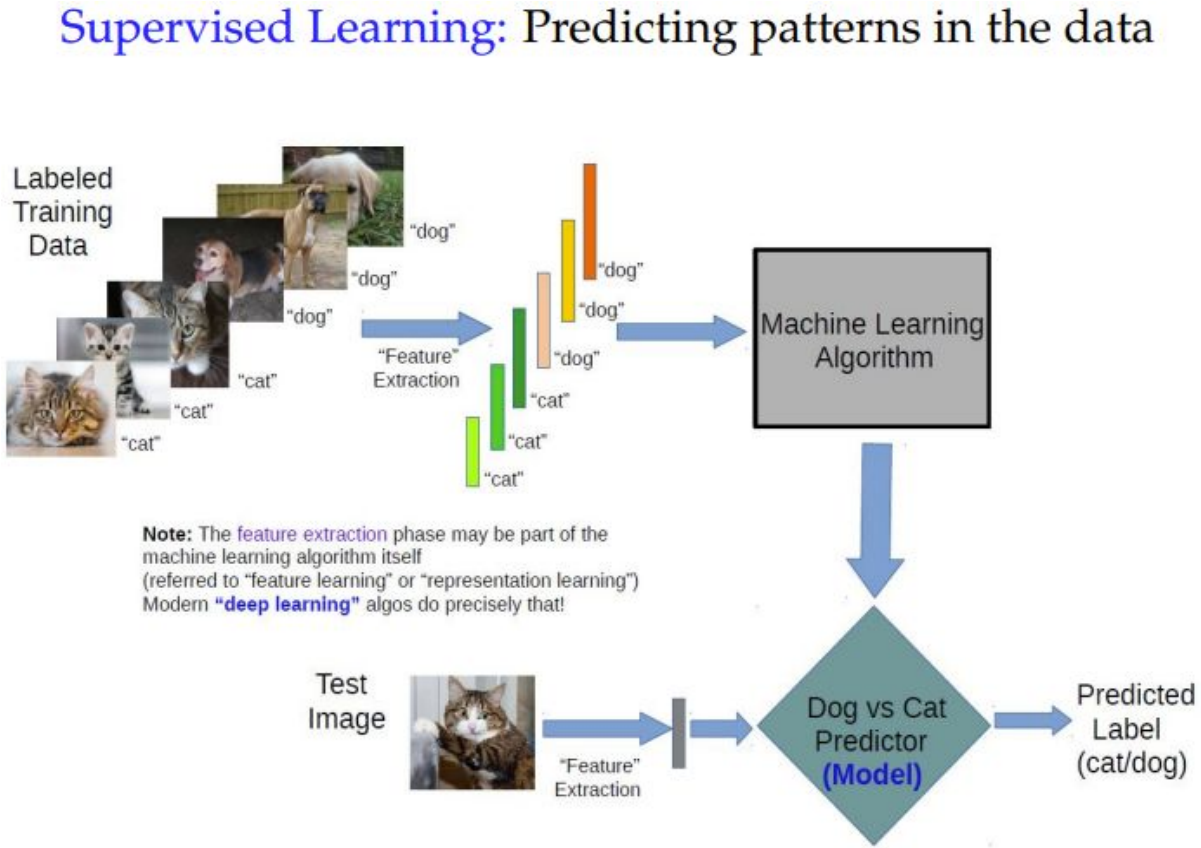
- http://www.r-project.org/
- http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf
- Download R: http://cran.r-project.org/bin/
- Download RStudio: http://www.rstudio.com/ide/download/desktop

R Studio Environment



Supervised Learning

Supervised Learning



CLASSIFICATION VS REGRESSION

Student Profile

CLASSIFICATION

Predicting Student

Pass Or Fail

Student Profile

REGRESSION

Predicting Student Marks

Percentage

Regression Task

- Regression finds correlations between dependent and independent variables.
- Regression algorithms help predict continuous variables such as house prices, market trends, weather patterns, oil and gas prices (a critical task these days!), etc.
- The Regression algorithm’s task is finding the mapping function so we can map the input variable of “x” to the continuous output variable of “y.”

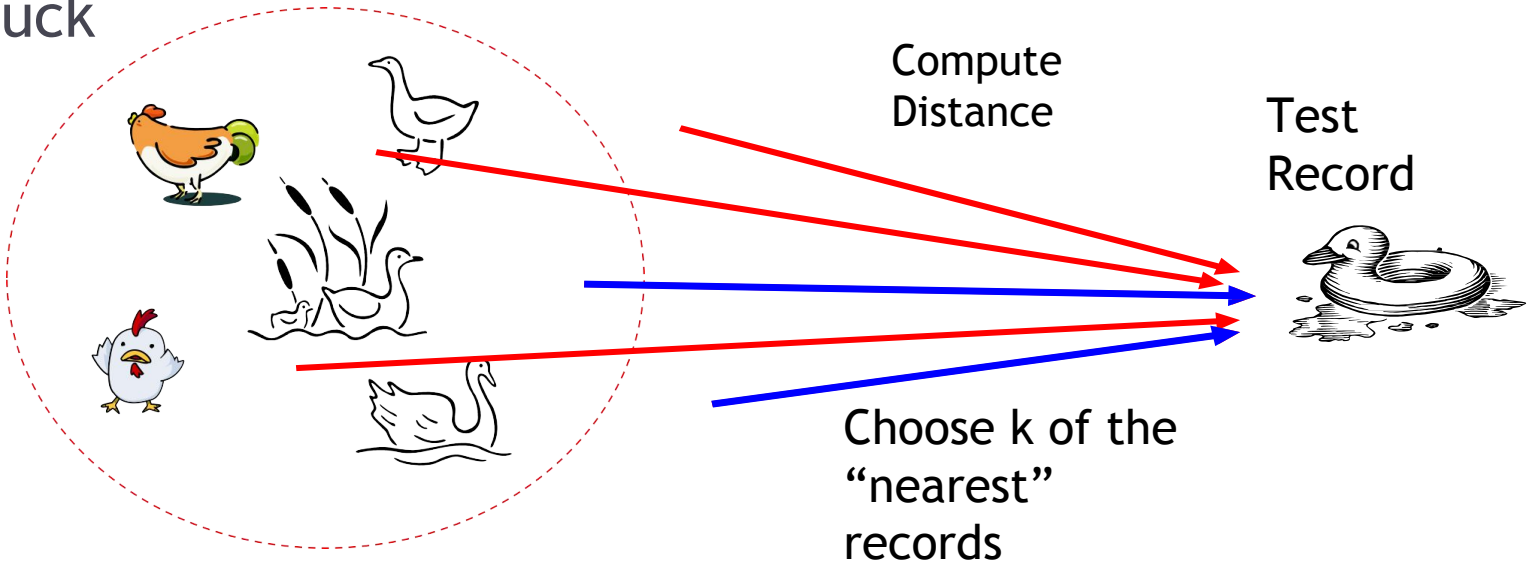
Classification Task

- Classification is an algorithm that finds functions that help divide the dataset into classes based on various parameters.
- Classification algorithms find the mapping function to map the “x” input to “y” discrete output.
- Classification algorithms are used for things like email and spam classification, predicting the willingness of bank customers to pay their loans, and identifying cancer tumor cells.

K Nearest Neighbours

Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck



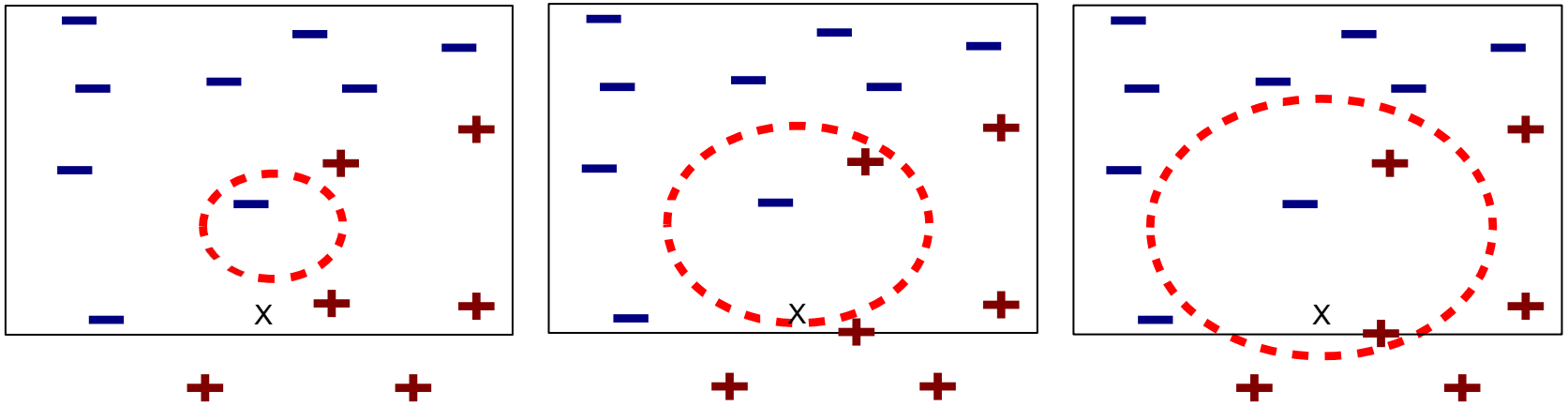
Basic Idea

- *k*-NN classification rule is to assign to a test sample the majority category label of its *k* nearest training samples
- In practice, *k* is usually chosen to be odd, so as to avoid ties
- The *k* = 1 rule is generally called the nearest-neighbor classification rule

Basic Idea

- kNN does not build model from the training data.
- To classify a test instance *d*, define *k*-neighborhood *P* as *k* nearest neighbors of *d*
- Count number *n* of training instances in *P* that belong to class c_j
- Estimate $\Pr(c_j|d)$ as n/k
- No training is needed. Classification time is linear in training set size for each test case.

Definition of Nearest Neighbor



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

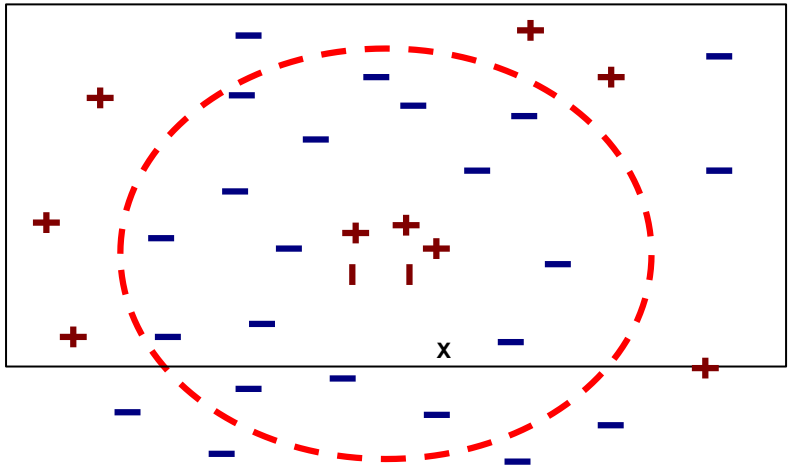
Nearest-Neighbor Classifiers: Issues

- The value of *k*, the number of nearest neighbors to retrieve
- Choice of Distance Metric to compute distance between records
- Computational complexity
  - Size of training set
  - Dimension of data

Value of K

- Choosing the value of k:
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classes

Rule of thumb:
K = sqrt(N)
N: number of training points



Distance Metrics

Minkowsky:

$$D(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_{i=1}^{m} |x_i - y_i|^r \right)^{1/r}$$

Euclidean:

$$D(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

Manhattan / city-block:

$$D(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m} |x_i - y_i|$$

Camberra:

$$D(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m} \frac{|x_i - y_i|}{|x_i + y_i|}$$

Chebychev:

$$D(\boldsymbol{x},\boldsymbol{y}) = \max_{i=1}^{m} |x_i - y_i|$$

Quadratic:

$$D(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^T Q(\boldsymbol{x} - \boldsymbol{y}) = \sum_{j=1}^{m} \left(\sum_{i=1}^{m} (x_i - y_i) q_{ji} \right) (x_j - y_j)$$

Q is a problem-specific positive definite $m \times m$ weight matrix

Mahalanobis:

$$D(\boldsymbol{x},\boldsymbol{y}) = [\det V]^{1/m} (\boldsymbol{x} - \boldsymbol{y})^{\mathrm{T}} V^{-1} (\boldsymbol{x} - \boldsymbol{y})$$

V is the covariance matrix of $A_1..A_m$, and A_j is the vector of values for attribute j occuring in the training set instances $1..n$.

Correlation:

$$D(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{m} (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_{i=1}^{m} (x_i - \overline{x_i})^2 \sum_{i=1}^{m} (y_i - \overline{y_i})^2}}$$

$\overline{x_i} = \overline{y_i}$ and is the average value for attribute i occuring in the training set.

Chi-square:

$$D(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m} \frac{1}{sum_i} \left(\frac{x_i}{size_{\boldsymbol{x}}} - \frac{y_i}{size_{\boldsymbol{y}}} \right)^2$$

sum_i is the sum of all values for attribute i occuring in the training set, and $size_{\boldsymbol{x}}$ is the sum of all values in the vector $\boldsymbol{x}$.

Kendall's Rank Correlation:

$$D(\boldsymbol{x},\boldsymbol{y}) = 1 - \frac{2}{n(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{i-1} \mathrm{sign}(x_i - x_j)\,\mathrm{sign}(y_i - y_j)$$

sign(x)=-1, 0 or 1 if $x < 0$, $x = 0$, or $x > 0$, respectively.

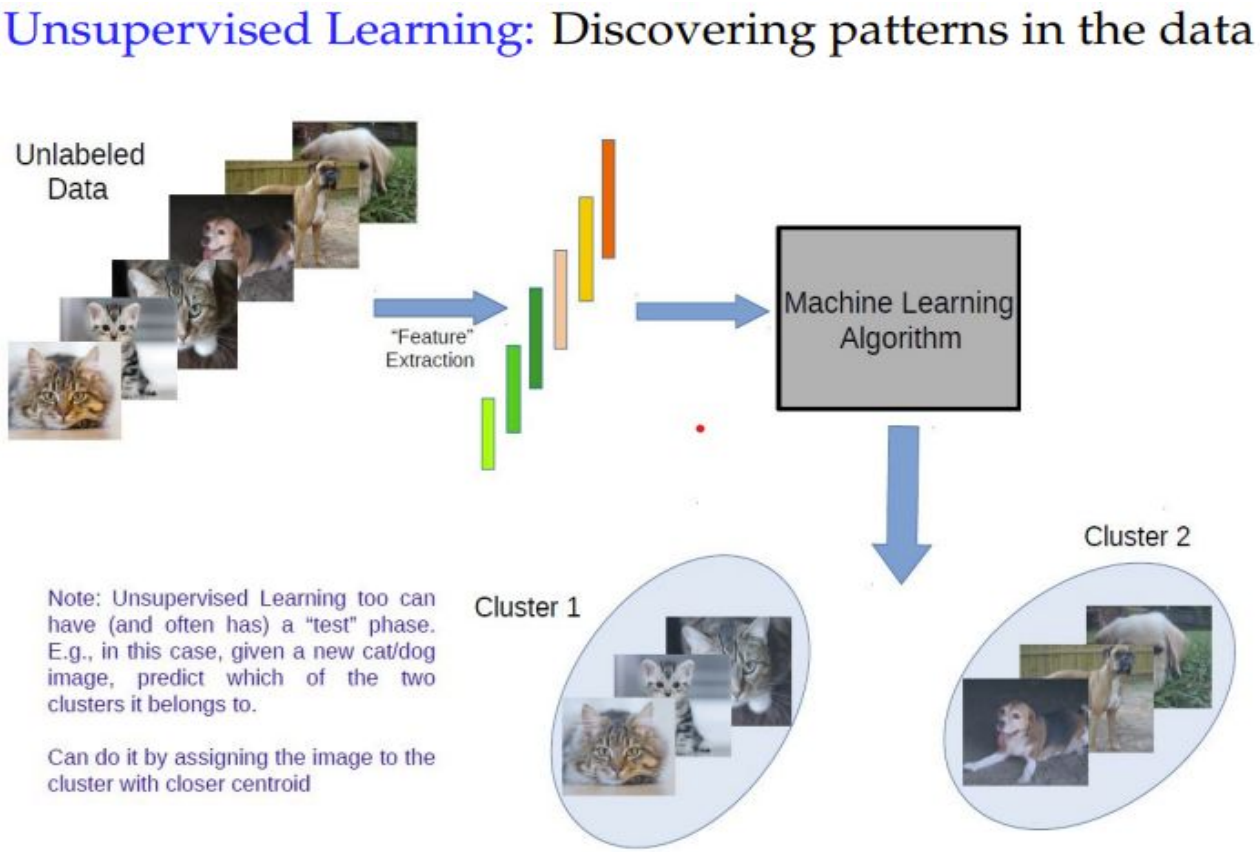
Figure 1. Equations of selected distance functions.
($\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors of m attribute values).

R Lab Session...

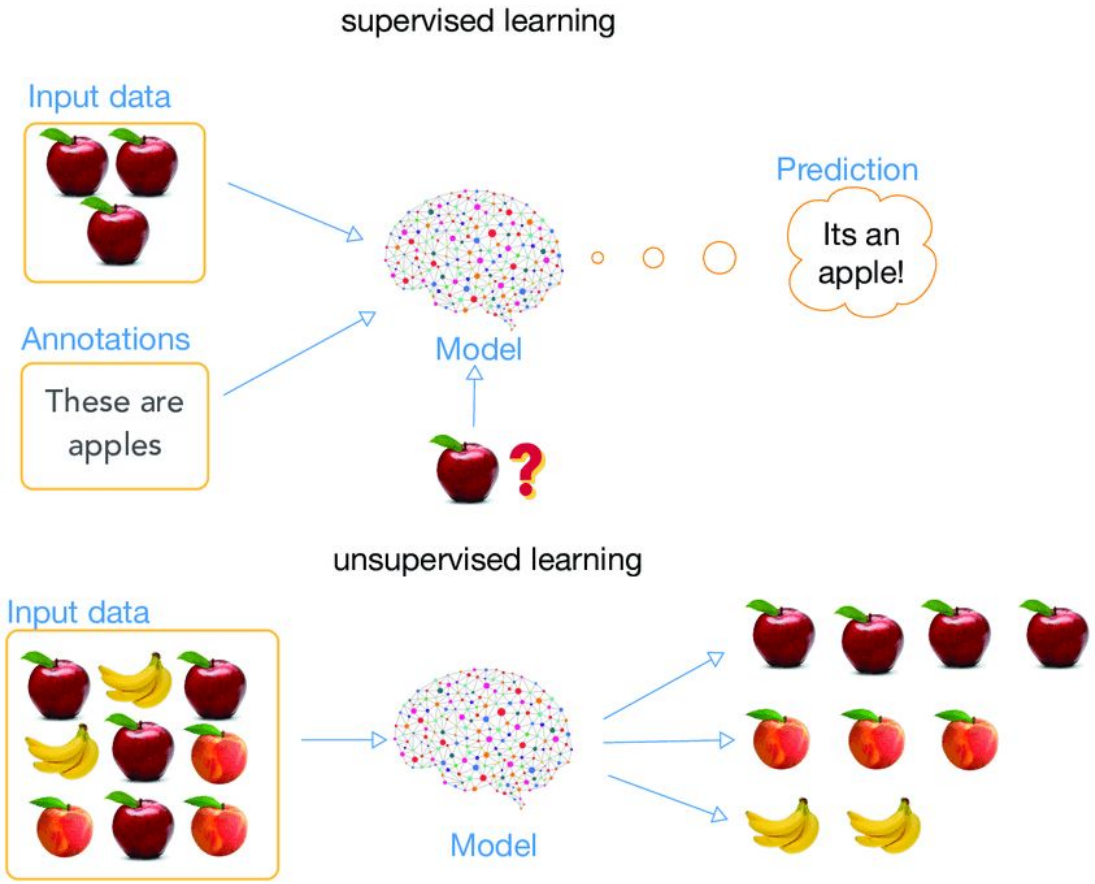
Utilize the Iris dataset from R statistical software to classify the flowers into particular species using KNN algorithm

Unsupervised Learning

Unsupervised Learning



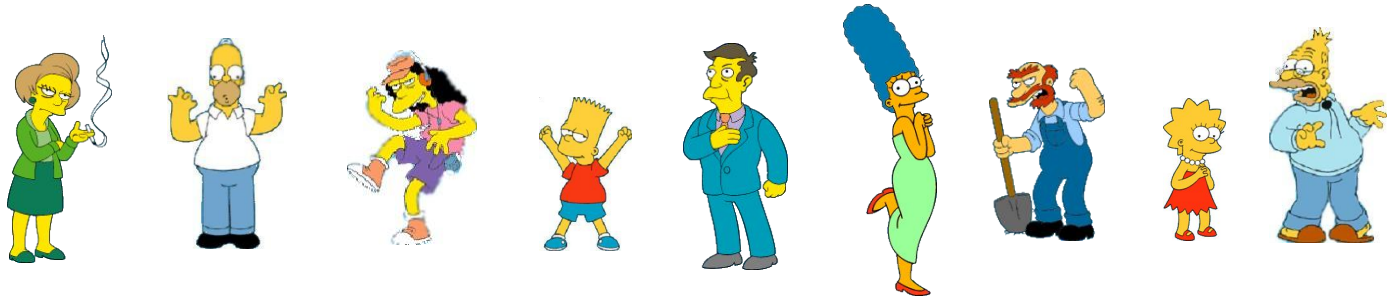
Supervised Learning vs Unsupervised Learning



What is Clustering?

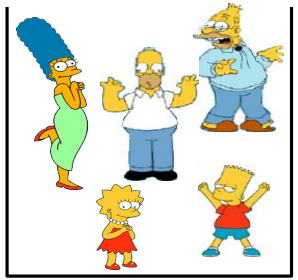
- The organization of unlabeled data into similarity groups called clusters.
- A cluster is a collection of data items which are "similar" between them, and "dissimilar" to data items in other clusters.
  - high intra-class similarity
  - low inter-class similarity
- More informally, finding natural groupings among objects.

What is a natural grouping among these objects?



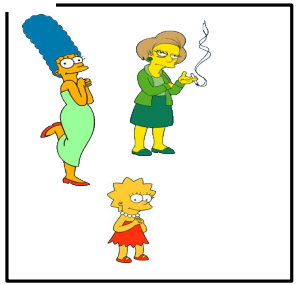
Clustering is subjective

Simpson's Family

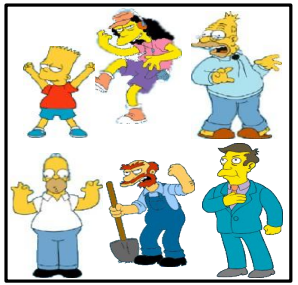


School Employees

Females



Males



Introduction to Clustering

Example 1 : The task of clustering

In order to elaborate the clustering task, consider the following dataset.

Table 2: Life Insurance database

Martial Status	Age	Income	Education	Number of children
Single	35	25000	Under Graduate	3
Married	25	15000	Graduate	1
Single	40	20000	Under Graduate	0
Divorced	20	30000	Post-Graduate	0
Divorced	25	20000	Under Graduate	3
Married	60	70000	Graduate	0
Married	30	90000	Post-Graduate	0
Married	45	60000	Graduate	5
Divorced	50	80000	Under Graduate	2

With certain similarity or likeliness defined, we can classify the records to one or group of more attributes (and thus mapping being non-trivial).

Introduction to Clustering

- Clustering has been used in many application domains:
  - Image analysis
  - Document retrieval
  - Machine learning, etc.

- When clustering is applied to real-world database, many problems may arise.
  - The (best) number of cluster is not known.
  - There is not correct answer to a clustering problem.
  - In fact, many answers may be found.
  - The exact number of cluster required is not easy to determine.

Introduction to Clustering

1. There may not be any a priori knowledge concerning the clusters.

- This is an issue that what data should be used for clustering.
- Unlike classification, in clustering, we have not supervisory learning to aid the process.
- Clustering can be viewed as similar to unsupervised learning.

2. Interpreting the semantic meaning of each cluster may be difficult.

- With classification, the labeling of classes is known ahead of time. In contrast, with clustering, this may not be the case.
- Thus, when the clustering process is finished yielding a set of clusters, the exact meaning of each cluster may not be obvious.

Definition of Clustering Problem

Definition 1: Clustering

Given a database $D = \{t_1, t_2, \ldots, t_n\}$ of n tuples, the clustering problem is to define a mapping $f : D \rightarrow C$, where each $t_i \in D$ is assigned to one cluster $c_i \in C$.

Here, $C = \{c_1, c_2, \ldots, c_k\}$ denotes a set of clusters.

- ★ Solution to a clustering problem is devising a mapping formulation.
- ★ The formulation behind such a mapping is to establish that a tuple within one cluster is more like tuples within that cluster and not similar to tuples outside it.

What do we need for clustering?

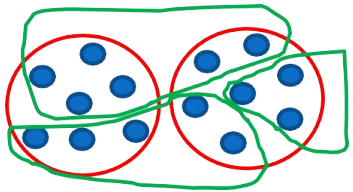
Proximity measure, either

Similarity measure $s(X_i, X_k)$: large if X_i, X_k are similar

Dissimilarity (or distance) measure $d(X_i, X_k)$: small if X_i, X_k are similar

Large d, small s

Large s, small d



Criterion function to evaluate clustering

Algorithm to evaluate clustering

For example by optimizing criterion function

Cluster evaluation (a hard problem)

- **Intra-cluster cohesion** (compactness):
  - Cohesion measures how near the data points in a cluster are to the cluster centroid.
  - Sum of squared error (SSE) is a commonly used measure.
- **Inter-cluster separation** (isolation):
  - Separation means that different cluster centroids should be far away from one another.
- In most applications, expert judgments are still the key

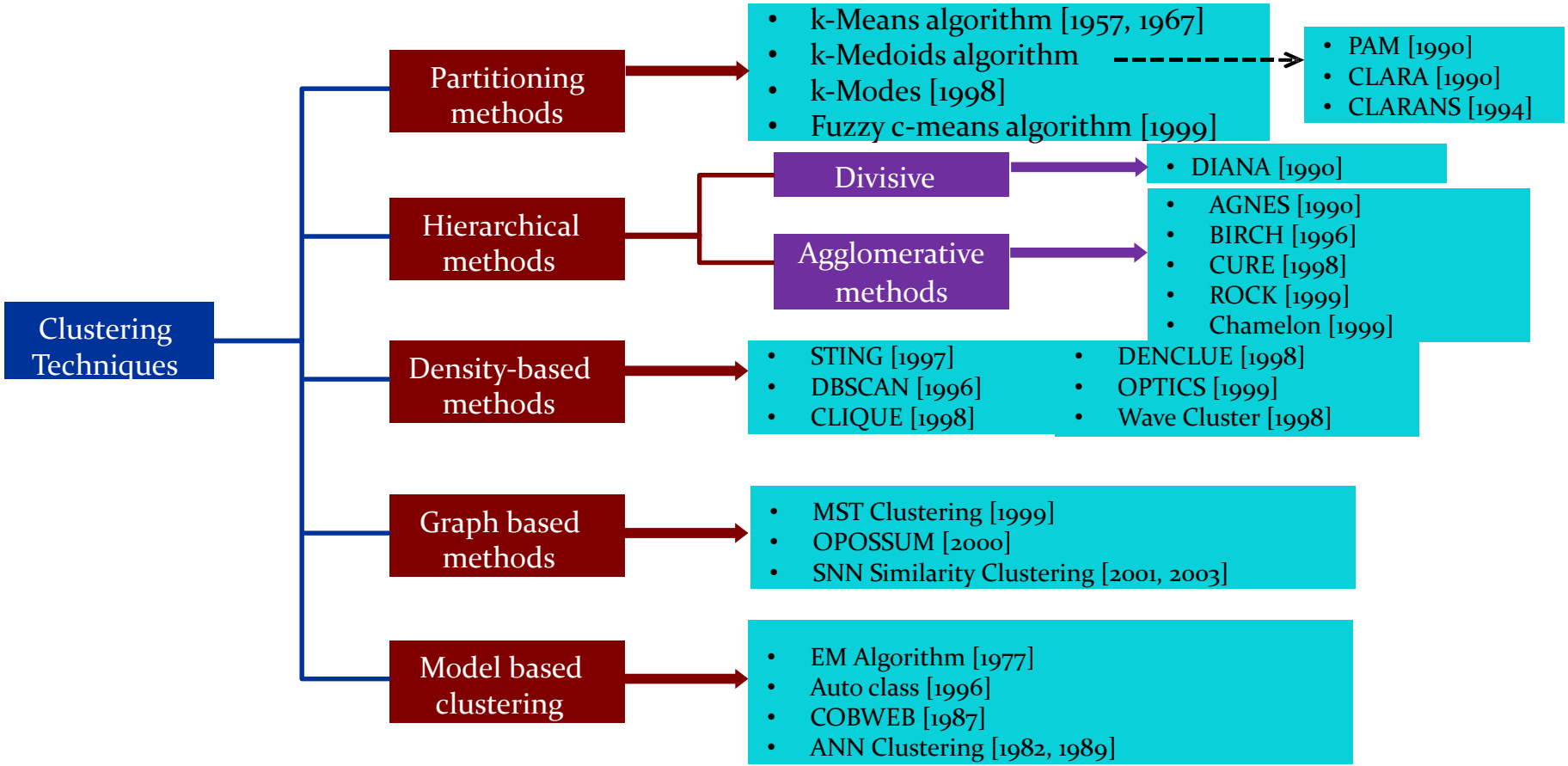
How many clusters?

Possible approaches

- Fix the number of cluster to k
- Find the best clustering according to the criterion function

Clustering techniques

- Clustering has been studied extensively for more than 40 years and across many disciplines due to its broad applications.
- As a result, many clustering techniques have been reported in the literature.
- Let us categorize the clustering methods. In fact, it is difficult to provide a crisp categorization because many techniques overlap to each other in terms of clustering paradigms or features.
- A broad taxonomy of existing clustering methods is shown in the next slide.
- It is not possible to cover all the techniques in this lecture series. We emphasize on major techniques belong to partitioning and hierarchical algorithms.


Clustering Techniques
Partitioning methods
k-Means algorithm [1957, 1967]
k-Medoids algorithm
k-Modes [1998]
Fuzzy c-means algorithm [1999]
PAM [1990]
CLARA [1990]
CLARANS [1994]
Hierarchical methods
Divisive
DIANA [1990]
Agglomerative methods
AGNES [1990]
BIRCH [1996]
CURE [1998]
ROCK [1999]
Chamelon [1999]
Density-based methods
STING [1997]
DBSCAN [1996]
CLIQUE [1998]
DENCLUE [1998]
OPTICS [1999]
Wave Cluster [1998]
Graph based methods
MST Clustering [1999]
OPOSSUM [2000]
SNN Similarity Clustering [2001, 2003]
Model based clustering
EM Algorithm [1977]
Auto class [1996]
COBWEB [1987]
ANN Clustering [1982, 1989]

k-Means Algorithm

- k-Means clustering algorithm proposed by J. Hartigan and M. A. Wong [1979].
- Given a set of *n* distinct objects, the k-Means clustering algorithm partitions the objects into *k* number of clusters such that intracluster similarity is high but the intercluster similarity is low.
- In this algorithm, user has to specify *k*, the number of clusters and consider the objects are defined with numeric attributes and thus using any one of the distance metric to demarcate the clusters.

k-Means Algorithm

The algorithm can be stated as follows.

- First it selects *k* number of objects at random from the set of n objects. These *k* objects are treated as the centroids or center of gravities of *k* clusters.
- For each of the remaining objects, it is assigned to one of the closest centroid. Thus, it forms a collection of objects assigned to each centroid and is called a cluster.
- Next, the centroid of each cluster is then updated (by calculating the mean values of attributes of each object).
- The assignment and update procedure is until it reaches some stopping criteria (such as, number of iteration, centroids remain unchanged or no assignment, etc.)

k-Means Algorithm

Algorithm 1: k-Means clustering

Input: D is a dataset containing *n* objects, *k* is the number of cluster

Output: A set of *k* clusters

Steps:

1. Randomly choose *k* objects from D as the initial cluster centroids.

2. **For** each of the objects in D **do**
   - Compute distance between the current objects and *k* cluster centroids
   - Assign the current object to that cluster to which it is closest.

3. Compute the "cluster centers" of each cluster. These become the new cluster centroids.

4. Repeat step 2-3 until the convergence criterion is satisfied

5. Stop

k-Means Algorithm

Note:

1) Objects are defined in terms of set of attributes. $A = \{A_1, A_2, \ldots\ldots, A_m\}$ where each A_i is continuous data type.

2) Distance computation: Any distance such as L_1, L_2, L_3 or cosine similarity.

3) Minimum distance is the measure of closeness between an object and centroid.

4) Mean Calculation: It is the mean value of each attribute values of all objects.

5) Convergence criteria: Any one of the following are termination condition of the algorithm.
   - Number of maximum iteration permissible.
   - No change of centroid values in any cluster.
   - Zero (or no significant) movement(s) of object from one cluster to another.
   - Cluster quality reaches to a certain level of acceptance.

Illustration of k-Means clustering algorithms

Table 3: 16 objects with two attributes A_1 and A_2.

A_1	A_2
6.8	12.6
0.8	9.8
1.2	11.6
2.8	9.6
3.8	9.9
4.4	6.5
4.8	1.1
6.0	19.9
6.2	18.5
7.6	17.4
7.8	12.2
6.6	7.7
8.2	4.5
8.4	6.9
9.0	3.4
9.6	11.1

Fig 3: Plotting data of Table 3

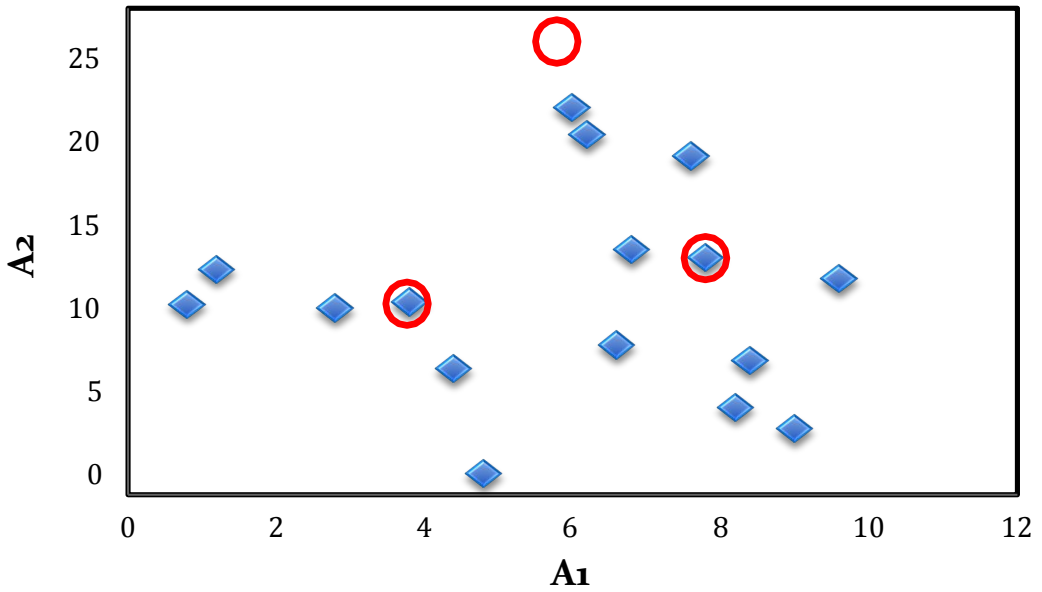


Illustration of k-Means clustering algorithms

- Suppose, k=3. Three objects are chosen at random shown as circled (see Fig 3). These three centroids are shown below.

Initial Centroids chosen randomly

Centroid	Objects	
	A1	A2
c_1	3.8	9.9
c_2	7.8	12.2
c_3	6.2	18.5

- Let us consider the Euclidean distance measure (L_2 Norm) as the distance measurement in our illustration.
- Let d_1, d_2 and d_3 denote the distance from an object to c_1, c_2 and c_3 respectively. The distance calculations are shown in Table 4.
- Assignment of each object to the respective centroid is shown in the right-most column and the clustering so obtained is shown in Fig 4.

Illustration of k-Means clustering algorithms

Table 4: Distance calculation

A_1	A_2	d_1	d_2	d_3	cluster
6.8	12.6	4.0	1.1	5.9	2
0.8	9.8	3.0	7.4	10.2	1
1.2	11.6	3.1	6.6	8.5	1
2.8	9.6	1.0	5.6	9.5	1
3.8	9.9	0.0	4.6	8.9	1
4.4	6.5	3.5	6.6	12.1	1
4.8	1.1	8.9	11.5	17.5	1
6.0	19.9	10.2	7.9	1.4	3
6.2	18.5	8.9	6.5	0.0	3
7.6	17.4	8.4	5.2	1.8	3
7.8	12.2	4.6	0.0	6.5	2
6.6	7.7	3.6	4.7	10.8	1
8.2	4.5	7.0	7.7	14.1	1
8.4	6.9	5.5	5.3	11.8	2
9.0	3.4	8.3	8.9	15.4	1
9.6	11.1	5.9	2.1	8.1	2

Fig 4: Initial cluster with respect to Table 4

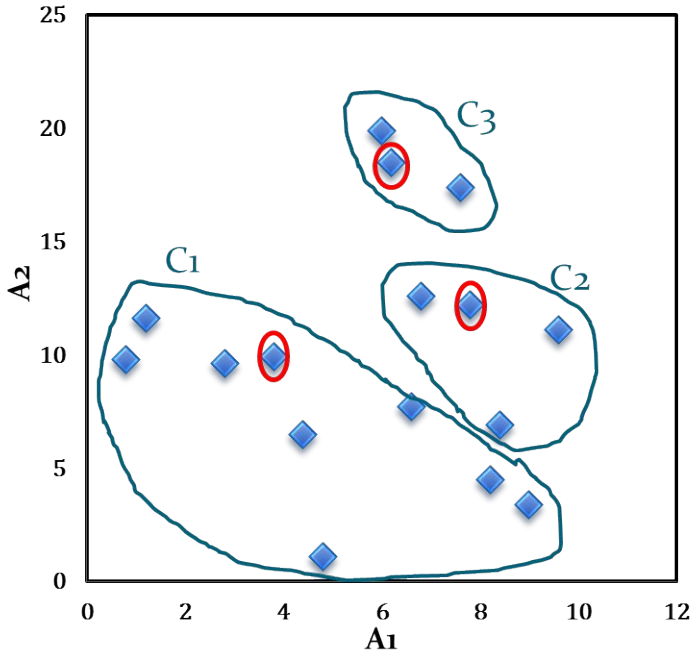


Illustration of k-Means clustering algorithms

The calculation new centroids of the three cluster using the mean of attribute values of A_1 and A_2 is shown in the Table below. The cluster with new centroids are shown in Fig 5.

Calculation of new centroids

New Centroid	Objects	
	A1	A2
c_1	4.6	7.1
c_2	8.2	10.7
c_3	6.6	18.6

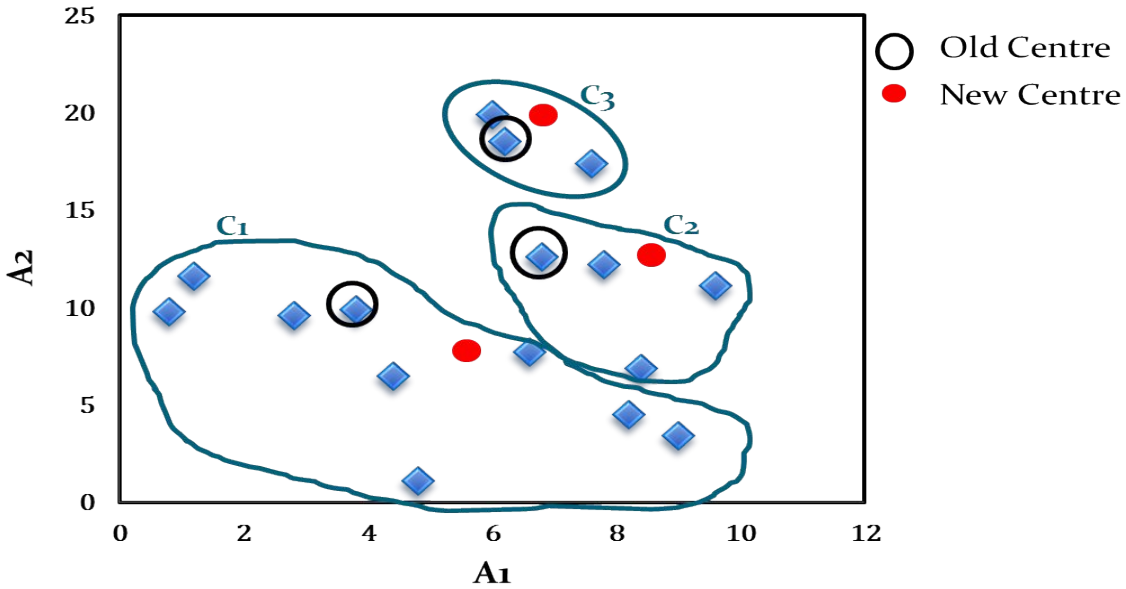


Fig 5: Initial cluster with new centroids

Illustration of k-Means clustering algorithms

We next reassign the 16 objects to three clusters by determining which centroid is closest to each one. This gives the revised set of clusters shown in Fig 6.

Note that point p moves from cluster C_2 to cluster C_1.

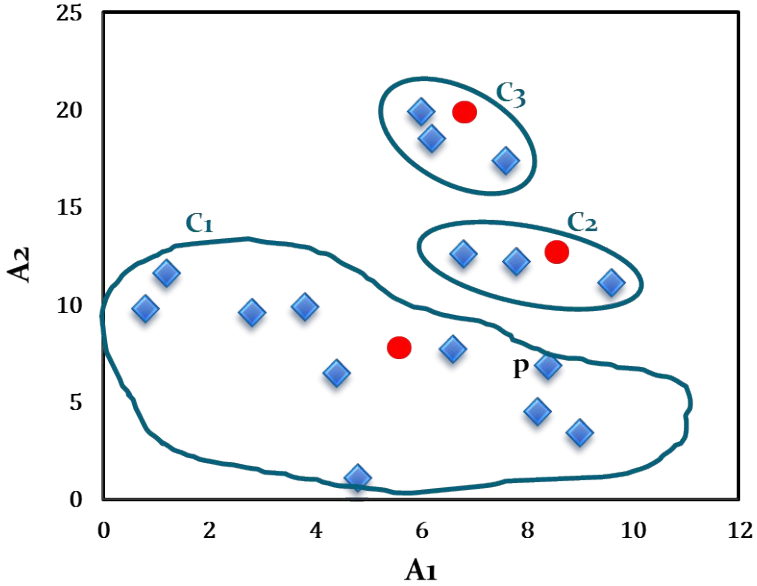


Fig 6: Cluster after first iteration

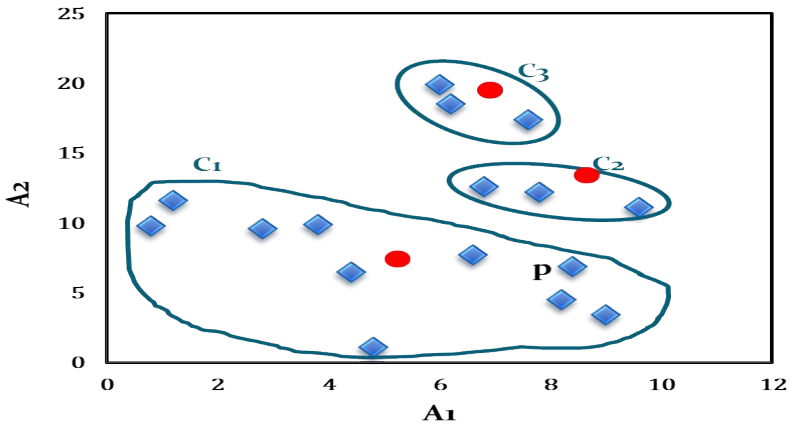
Illustration of k-Means clustering algorithms

- The newly obtained centroids after second iteration are given in the table below. Note that the centroid c_3 remains unchanged, where c_2 and c_1 changed a little.
- With respect to newly obtained cluster centres, 16 points are reassigned again. These are the same clusters as before. Hence, their centroids also remain unchanged.
- Considering this as the termination criteria, the k-means algorithm stops here. Hence, the final cluster in Fig 7 is same as Fig 6.

Cluster centres after second iteration

Centroid	Revised Centroids	
	A1	A2
c_1	5.0	7.1
c_2	8.1	12.0
c_3	6.6	18.6

Fig 7: Cluster after Second iteration



Comments on k-Means algorithm

Value of k:

- The k-means algorithm produces only one set of clusters, for which, user must specify the desired number, k of clusters.
- In fact, k should be the best guess on the number of clusters present in the given data. Choosing the best value of k for a given dataset is, therefore, an issue.
- We may not have an idea about the possible number of clusters for high dimensional data, and for data that are not scatter-plotted.
- Further, possible number of clusters is hidden or ambiguous in image, audio, video and multimedia clustering applications etc.
- There is no principled way to know what the value of k ought to be. We may try with successive value of k starting with 2.
- The process is stopped when two consecutive k values produce more-or-less identical results (with respect to some cluster quality estimation).
- Normally $k \ll n$ and there is heuristic to follow $k \approx \sqrt{n}$.

R Lab Session...

This tutorial serves as an introduction to the k-means clustering method.

1. Data Preparation: Preparing our data for cluster analysis
2. Clustering Distance Measures: Understanding how to measure differences in observations
3. K-Means Clustering: Calculations and methods for creating K subgroups of the data
4. Determining Optimal Clusters: Identifying the right number of clusters to group your data

Time Series Forecasting

Introduction

- Time series is a set of observations, each one being recorded at a specific time. (e.g., Annual GDP of a country, Sales figure, etc.)
- Discrete time series is one in which the set of time points at which observations are made is a discrete set. (e.g., All above including irregularly spaced data)
- Continuous time series are obtained when observations are made continuously over some time intervals. (e.g., ECG graph)
- Forecasting is estimating how the sequence of observations will continue in to the future. (e.g., Forecasting of major economic variables like GDP, Unemployment, Inflation, Exchange rates, Production and Consumption)
- Forecasting is very difficult, since it's about the future! (e.g., forecasts of daily cases of COVID-19)

Time Series Data

- A time series is a sequence of observations over time. What makes it distinguishable from other statistical analyses is the explicit recognition of the importance of the order in which the observations are made. Also, unlike many other problems where observations are independent, in time series observations are most often dependent.

- Why do we need special models for time series data?
  - Prediction of the future based on knowledge of the past (most important).
  - To control the process producing the series.
  - To have a description of the salient features of the series.

- Applications of time series forecasting
  - Economic planning
  - Sales forecasting
  - Inventory (stock) control
  - Exchange rate forecasting
  - Etc…

Use of Time Series Data

- To develop forecast model
  - What will the rate of inflation be next year?

- To estimate dynamic causal effects
  - If the rate of interest increases the interest rate now, what will be the effect on the rates of inflation and unemployment in 3 months? in 12 months?

  - What is the effect over time on electronics good consumption of a hike in the excise duty?

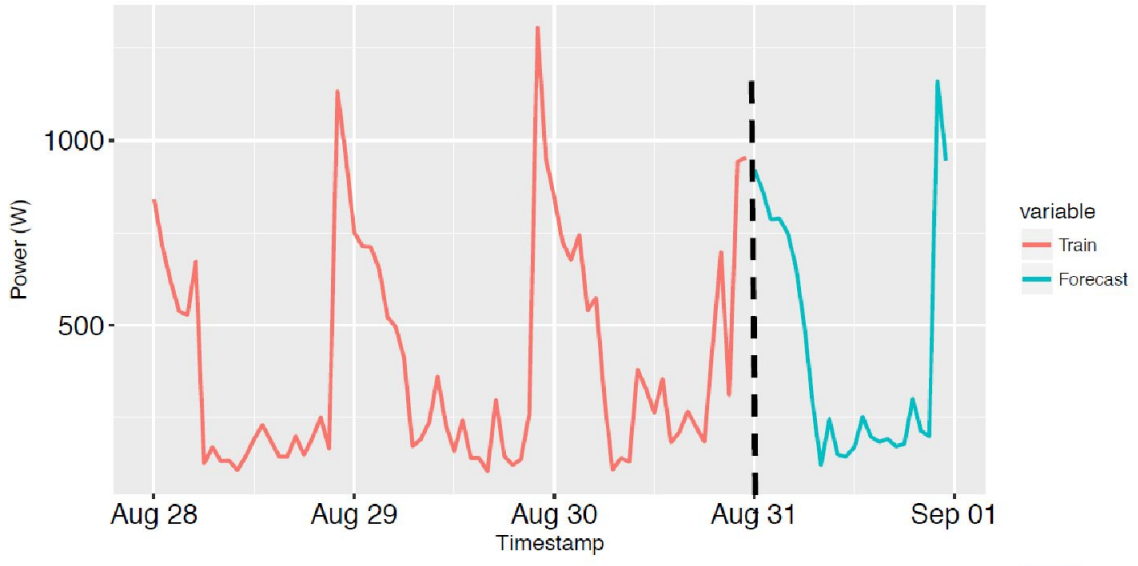
- Time dependent analysis
  - Rates of inflation and unemployment in the country can be observed only over time!

A Forecasting Problem: India / U.S. Foreign Exchange Rate (EXINUS)

- Source: FRED ECONOMICS DATA (Shaded areas indicate US recessions)
- Units: Indian Rupees to One U.S. Dollar, Not Seasonally Adjusted
- Frequency: Monthly (Averages of daily figures)

Forecasting: Assumptions

- **Time series Forecasting:** Data collected at regular intervals of time (e.g., Weather and Electricity Forecasting).
- **Assumptions:** (a) Historical information is available;
  (b) Past patterns will continue in the future.



Time Series Components

- Trend (T_t) : pattern exists when there is a long-term increase or decrease in the data.
- Seasonal (S_t) : pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week).
- Cyclic (C_t) : pattern exists when data exhibit rises and falls that are not of fixed period (duration usually of at least 2 years).
- Decomposition : $Y_t = f(T_t; S_t; C_t; I_t)$, where Y_t is data at period t and I_t is irregular component at period t.
- Additive decomposition: : $Y_t = T_t + S_t + C_t + I_t$
- Multiplicative decomposition: $Y_t = T_t * S_t * C_t * I_t$
- A stationary series is : roughly horizontal, constant variance and no patterns predictable in the long-term.

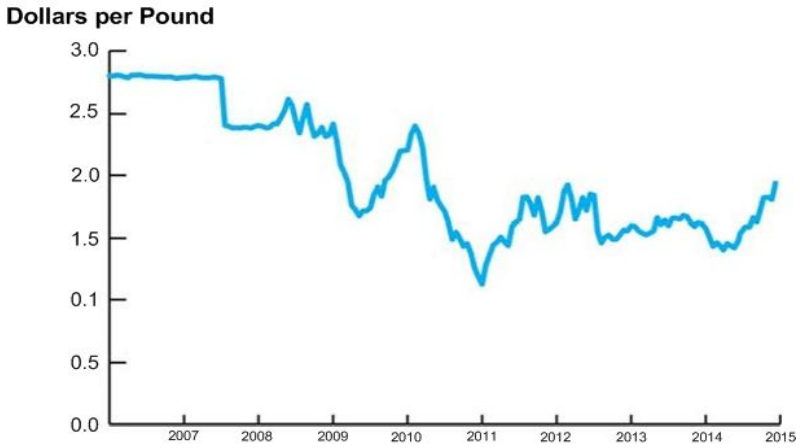
Auto Regression Analysis

- Regression analysis for time-ordered data is known as Auto-Regression Analysis
- Time series data are data collected on the same observational unit at multiple time periods

Example: Indian rate of price inflation

Modeling with Time Series Data

- Correlation over time
  - Serial correlation, also called autocorrelation
  - Calculating standard error
- To estimate dynamic causal effects
  - Under which dynamic effects can be estimated?
  - How to estimate?
- Forecasting model
  - Forecasting model build on regression model



Can we predict the tend at a time say 2017?

Some Notations and Concepts

- Y_t = Value of Y in a period t
- Data set [Y_1, Y_2, … Y_{T-1}, Y_T]: T observations on the time series random variable Y
- **Assumptions**
  - We consider only consecutive, evenly spaced observations
    - For example, monthly, 2000-2015, no missing months
  - A time series Y_t is **stationary** if its probability distribution does not change over time, that is, if the joint distribution of (Y_{i+1}, Y_{i+2}, …, Y_{i+T}) does not depend on i.
    - Stationary property implies that history is relevant. In other words, Stationary requires the future to be like the past (in a probabilistic sense).
    - Auto Regression analysis assumes that Y_t is stationary.

Some Notations and Concepts

- There are four ways to have the time series data for AutoRegression analysis
  - **Lag:** The first lag of Y_t is Y_{t-1}, its j-th lag is Y_{t-j}
  - **Difference:** The fist difference of a series, Y_t is its change between period t and $t-1$, that is, $y_t = Y_t - Y_{t-1}$
  - **Log difference:** $y_t = \log(Y_t) - \log(Y_{t-1})$
  - **Percentage:** $y_t = \frac{Y_{t-1}}{Y_t} \times 100$

Some Notations and Concepts

- **Autocorrelation**
  - The correlation of a series with its own lagged values is called autocorrelation (also called serial correlation)

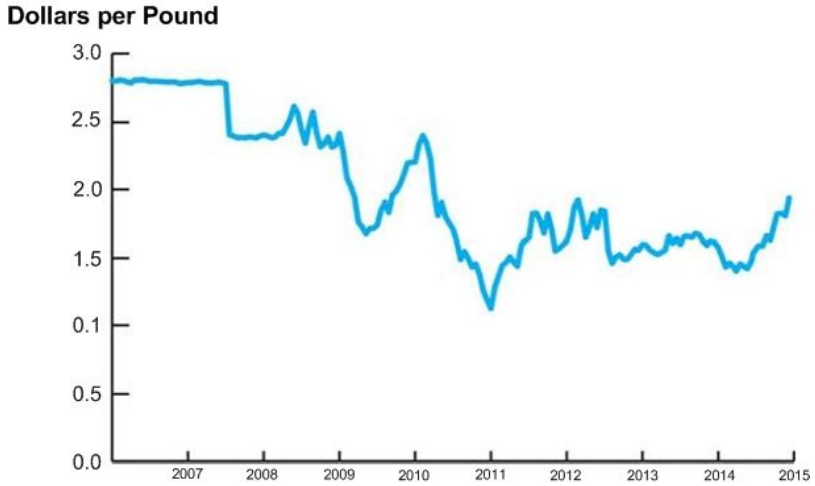
Definition : *j*-th Autocorrelation

The j-th autocorrelation, denoted by ρ_j is defined as

$$\rho_j = \frac{Cov(Y_t, Y_{t-j})}{\sigma_{Y_t}\sigma_{Y_{t-j}}}$$

Where, $Cov(Y_t, Y_{t-j})$ is the **j-th autocovariance**.

Some Notations and Concepts



- For the given data, say $\rho_1 = 0.84$
  - This implies that the Dollars per Pound is highly serially correlated

- Similarly, we can determine ρ_2, ρ_3 .... etc., and hence different regression analyses

Auto-Regression Model for Forecasting

- A natural starting point for forecasting model is to use past values of Y, that is, Y_{t-1}, Y_{t-2}, … to predict Y_t
- An autoregression is a regression model in which Y_t is regressed against its own lagged values.
- The number of lags used as regressors is called the **order of autoregression**
  - In first order autoregression (denoted as AR(1)), Y_t is regressed against Y_{t-1}
  - In p-th order autoregression (denoted as AR(p)), Y_t is regressed against, Y_{t-1}, Y_{t-2}, …,Y_{t-p} .

p-th Order AutoRegression Model

Definition : *p*-th AutoRegression Model

In general, the p-th order autregression model is defined as

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \varepsilon_t$$

where, $\beta_0, \beta_1, \ldots, \beta_p$ is called autoregression coefficients and ε_t is the noise term or residue and in practice it is assumed to Gausian white noise.

- For example, AR(1) is $Y_t = \beta_0 + \beta_1 Y_t + \varepsilon_t$
- The task in AR analysis is to derive the 'best' possible values for β_i given a time series Y_t.

Computing AR Coefficients

- A number of techniques known for computing the AR coefficients
- The most common method is called **Least Squares Method** (LSM)
- The LSM is based upon the Yule-Walker equations

$$\begin{bmatrix} 1 & r_1 & r_2 & r_3 & r_4 & \cdots & r_{p-2} & r_{p-1} \\ r_1 & 1 & r_1 & r_2 & r_3 & \cdots & r_{p-3} & r_{p-2} \\ r_2 & r_1 & 1 & r_1 & r_2 & \cdots & r_{p-4} & r_{p-3} \\ r_3 & r_2 & r_1 & 1 & r_2 & \cdots & r_{p-5} & r_{p-4} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ r_{p-1} & r_{p-2} & r_{p-3} & r_{p-4} & r_{p-5} & \cdots & r_1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \beta_{p-1} \\ \beta_p \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ r_{p-1} \\ r_p \end{bmatrix}$$

- Here, r_i (i = 1, 2 , 3, …, p-1) denotes the i^{th} auto correlation coefficient.
- β_0 can be chosen empirically, usually taken as zero.

AutoRegressive Integrated Moving Average (ARIMA) Model

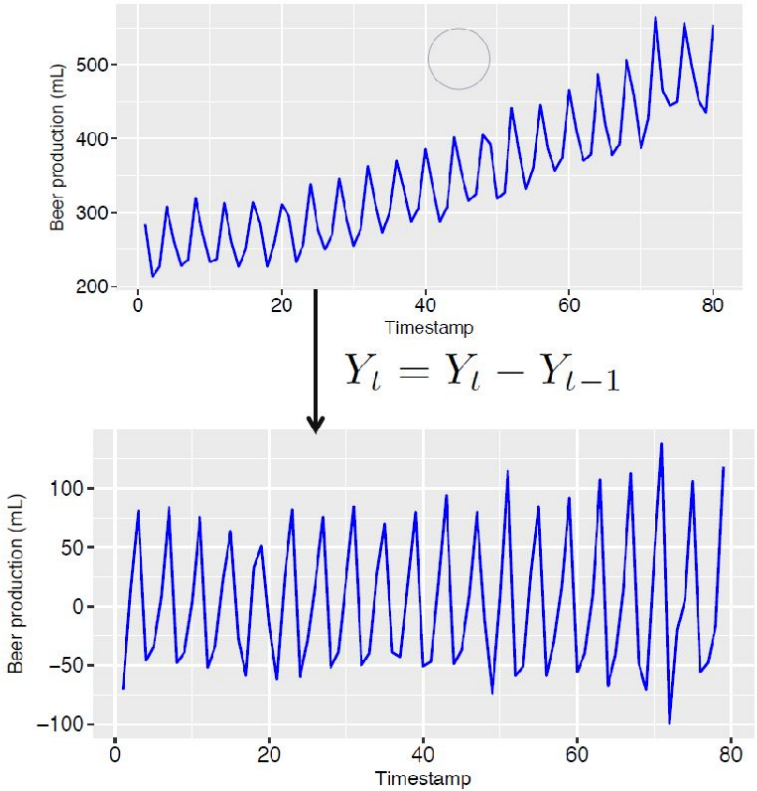
- The ARIMA model, introduced by Box and Jenkins (1976), is a linear regression model indulged in tracking linear tendencies in stationary time series data.
- AR: autoregressive (lagged observations as inputs) I: integrated (differencing to make series stationary) MA: moving average (lagged errors as inputs).
- The model is expressed as ARIMA (p, d, q) where p, d *and* q are integer parameter values that decide the structure of the model.
- More precisely, p *and* q are the order of the AR model and the MA model respectively, and parameter d is the level of differencing applied to the data.
- The mathematical expression of the ARIMA model is as follows:

$$y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

where y_t is the actual value, ε_t is the random error at time t, ϕ_i and θ_j are the coefficients of the model.

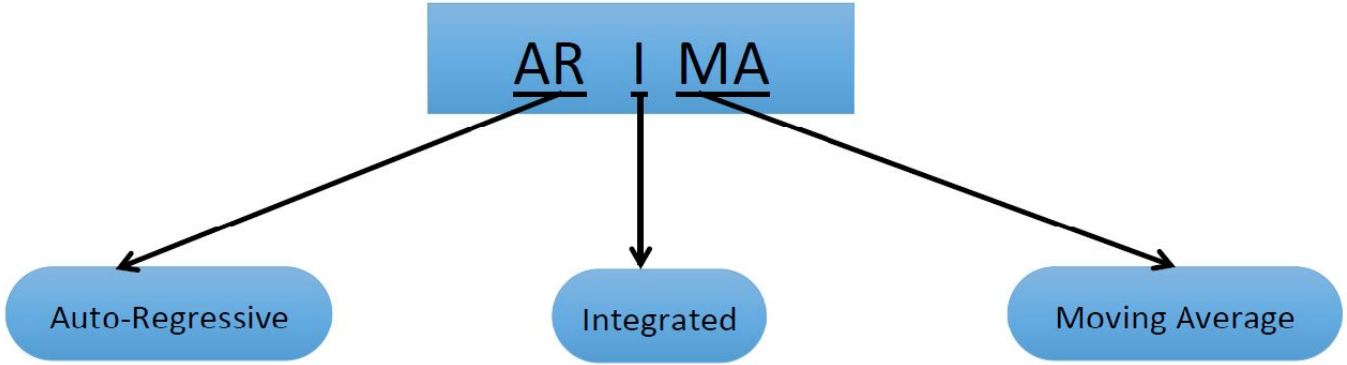
- It is assumed that $\varepsilon_{t-1} (\varepsilon_{t-1} = y_{t-1} - \hat{y}_{t-1})$ has zero mean with constant variance, and satisfies the i.i.d. condition.
- Three basic Steps: Model identification, Parameter Estimation, and Diagnostic Checking

Differencing in ARIMA Model



Differencing order (*d*): Number of times differencing is done

ARIMA model



$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \epsilon_t \quad \text{[Order p]}$$

$$Y_t = Y_t - Y_{t-1} \quad \text{[Order d]}$$

$$Y_t = \beta_0 + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \ldots + \phi_q \epsilon_{t-q} \quad \text{[Order q]}$$

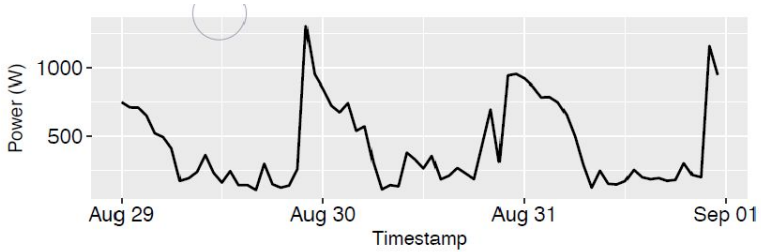
ARIMA is defined by a tuple (*p, d, q*)

ACF / PACF Plots

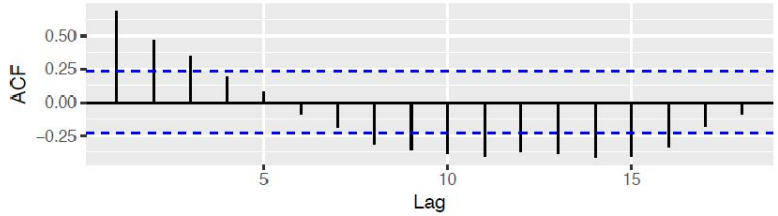
1. Auto-Correlation Function (ACF) Plot:
   - Correlation coefficients of time-series at different lags
   - Defines *q* order of MA model

2. Partial Auto-correlation Function (PACF) Plot:
   - Partial correlation coefficients of time series at different lags
   - Defines *p* order of AR model

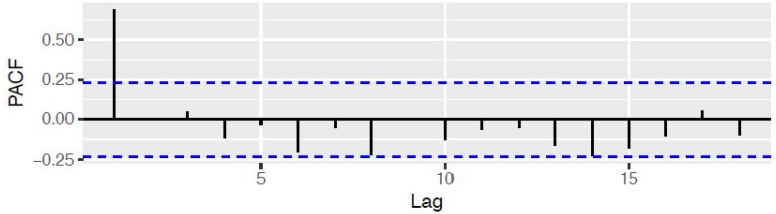
ACF / PACF Plots : Example



Data



ACF plot



PACF plot

Forecast Evaluation

Performance metrics such as mean absolute error (MAE), root mean square error (RMSE), and mean absolute percent error (MAPE) are used to evaluate the performances of different forecasting models for the unemployment rate data sets:

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|;$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|,$$

Where y_i is the actual output, $\hat{y}_i$ is the predicted output, and n denotes the number of data points.

By definition, the lower the value of these performance metrics, the better is the performance of the concerned forecasting model.

Time Series Analysis using R

Time Series Plot:

The graphical representation of time series data by taking time on x axis & data on y axis.

A plot of data over time

Example

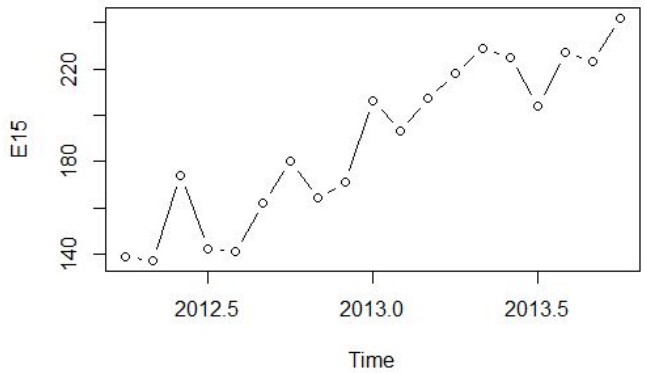
The demand for a commodity E15 for last 20 months from April 2012 to October 2013 is given in *E15demand.csv* file. Draw the time series plot

Month	Demand	Month	Demand
1	139	11	193
2	137	12	207
3	174	13	218
4	142	14	229
5	141	15	225
6	162	16	204
7	180	17	227
8	164	18	223
9	171	19	242
10	206	20	239

Reading data to R

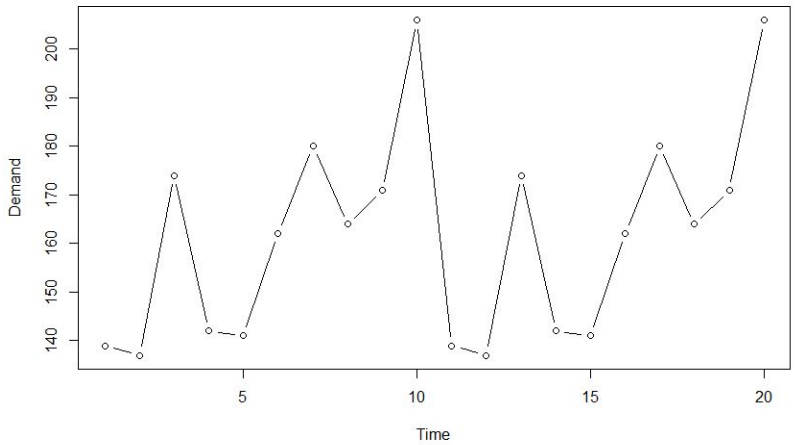
```
mydata <- read.csv("E15demand.csv")
E15 = ts(mydata$Demand, start = c(2012,4), end = c(2013,10), frequency = 12)
E15
plot(E15, type = "b")
```

For quarterly data, frequency = 4
For monthly data, frequency = 12



Reading data to R

```
E15 = ts(mydata$Demand)
E15
plot(E15, type = "b")
```

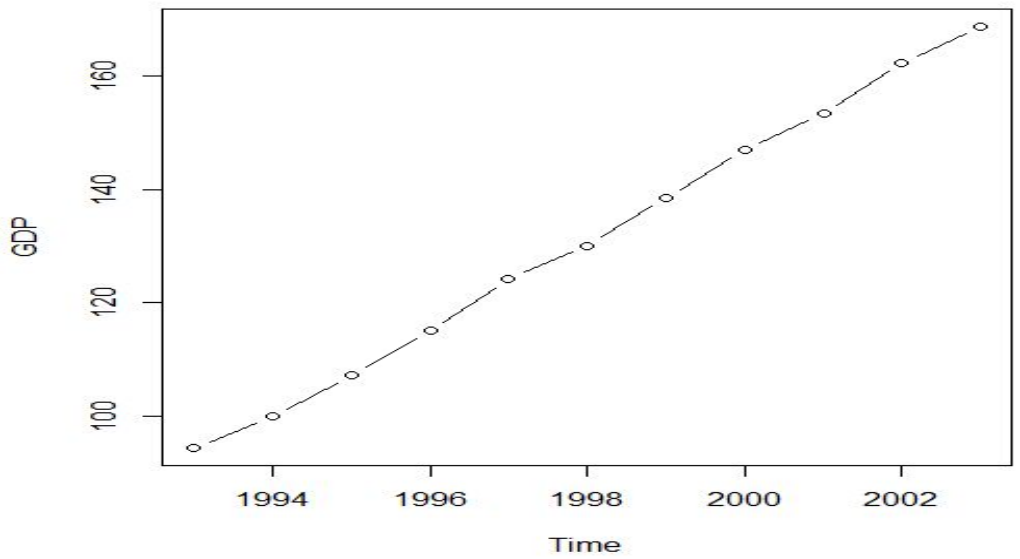


Trend:

A long term increase or decrease in the data

Example: The data on Yearly average of Indian GDP during 1993 to 2005.

Year	GDP
1993	94.43
1994	100.00
1995	107.25
1996	115.13
1997	124.16
1998	130.11
1999	138.57
2000	146.97
2001	153.40
2002	162.28
2003	168.73



Seasonal Pattern:

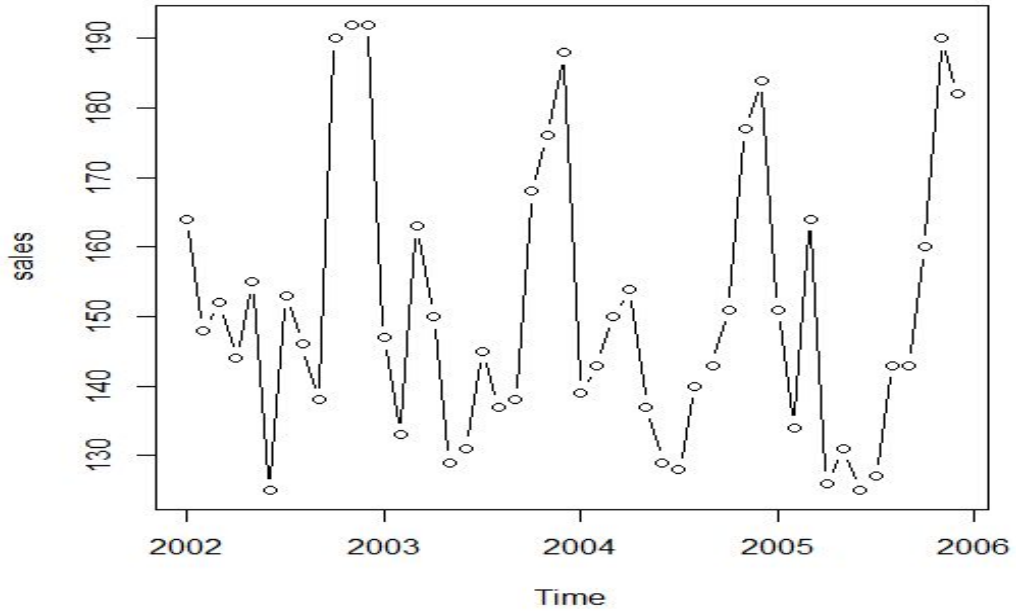
The time series data exhibiting rises and falls influenced by seasonal factors

Example: The data on monthly sales of a branded jackets

Month	Sales	Month	Sales	Month	Sales	Month	Sales
Jan-02	164	Jan-03	147	Jan-04	139	Jan-05	151
Feb-02	148	Feb-03	133	Feb-04	143	Feb-05	134
Mar-02	152	Mar-03	163	Mar-04	150	Mar-05	164
Apr-02	144	Apr-03	150	Apr-04	154	Apr-05	126
May-02	155	May-03	129	May-04	137	May-05	131
Jun-02	125	Jun-03	131	Jun-04	129	Jun-05	125
Jul-02	153	Jul-03	145	Jul-04	128	Jul-05	127
Aug-02	146	Aug-03	137	Aug-04	140	Aug-05	143
Sep-02	138	Sep-03	138	Sep-04	143	Sep-05	143
Oct-02	190	Oct-03	168	Oct-04	151	Oct-05	160
Nov-02	192	Nov-03	176	Nov-04	177	Nov-05	190
Dec-02	192	Dec-03	188	Dec-04	184	Dec-05	182

Seasonal Pattern:

The time series data exhibiting rises and falls influenced by seasonal factors

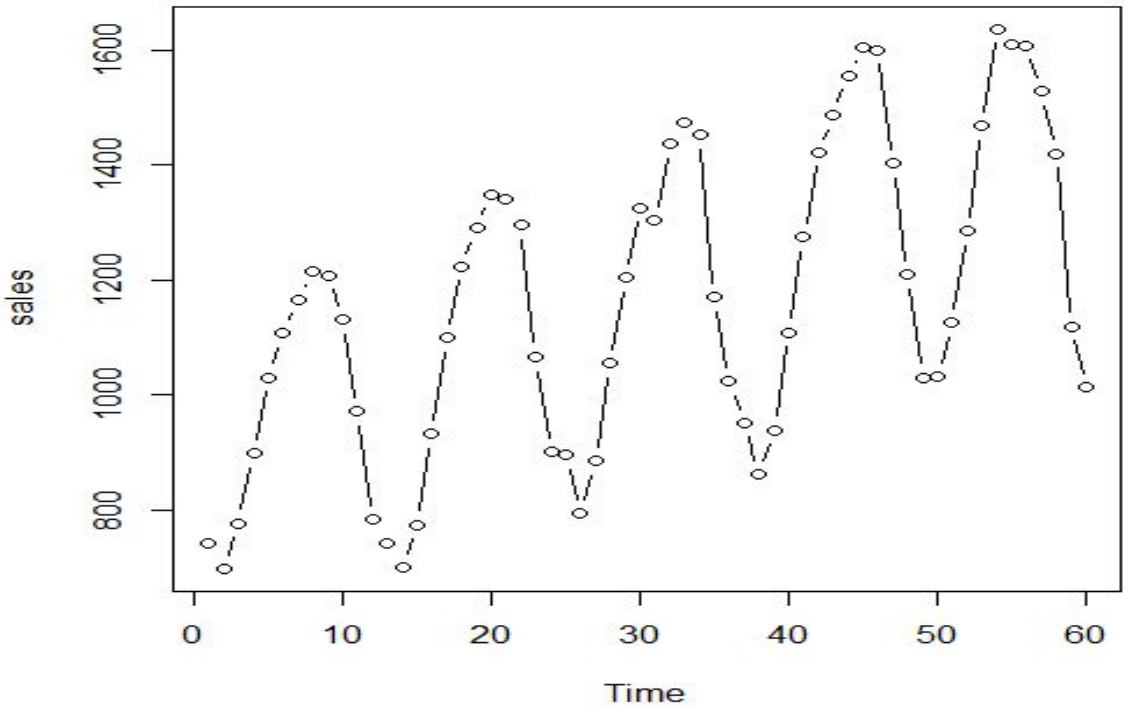


Trend and Seasonal Patterns Combined

The time series data may include a combination of trend and seasonal patterns

Example: The data on monthly sales of an aircraft component is given below:

Month	Sales	Month	Sales	Month	Sales
1	742	21	1341	41	1274
2	697	22	1296	42	1422
3	776	23	1066	43	1486
4	898	24	901	44	1555
5	1030	25	896	45	1604
6	1107	26	793	46	1600
7	1165	27	885	47	1403
8	1216	28	1055	48	1209
9	1208	29	1204	49	1030
10	1131	30	1326	50	1032
11	971	31	1303	51	1126
12	783	32	1436	52	1285
13	741	33	1473	53	1468
14	700	34	1453	54	1637
15	774	35	1170	55	1611
16	932	36	1023	56	1608
17	1099	37	951	57	1528
18	1223	38	861	58	1420
19	1290	39	938	59	1119
20	1349	40	1109	60	1013


sales
800
1000
1200
1400
1600
0
10
20
30
40
50
60
Time

Stationary Series

- A series free from trend and seasonal patterns
- A series exhibits only random fluctuations around mean

Differencing

Differencing: A method for making series stationary

A differenced series is the series of difference between each observation Y_t and the previous observation Y_{t-1}

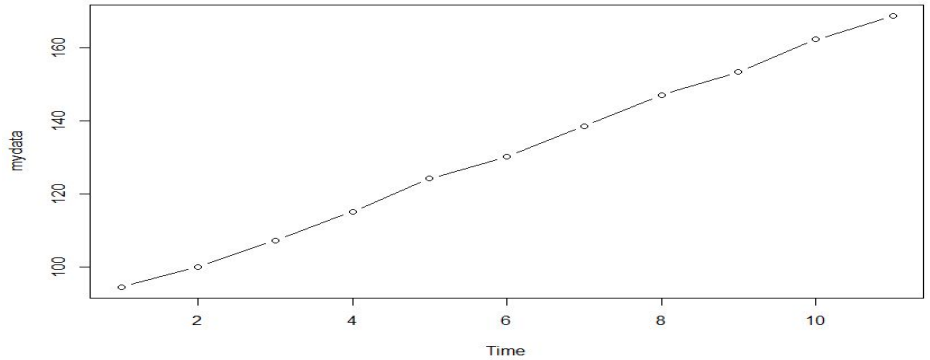
$$Y_t' = Y_t - Y_{t-1}$$

Example: Is it possible to make the GDP data given in GDP.csv stationary?

Differencing: A method for making series stationary

Example: Is it possible to make the GDP data given in GDP.csv stationary?

R Code
>mydata = ts(GDP$GDP)
> plot(mydata, type = "b")



Conclusion
Series has a linear trend hence the data is not stationary

Differencing: A method for making data stationary

Example: Is it possible to make the GDP data given in GDP.csv stationary?

Identify the number of differencing required

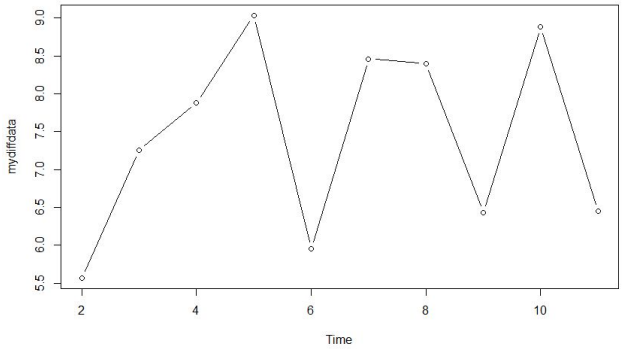
R Code

```
install.packages("forecast")
library(forecast)
ndiffs(GDP)
```

Differencing required is 1

Yt' = Yt – Yt-1

```
mydiffdata = diff(GDP, difference = 1)
plot(mydiffdata, type = "b")
```



Time Series Modeling

General form of linear model

y is modeled in terms of x's

$$Y = a + b_1x_1 + b_2x_2 + - - - + b_kx_k$$

Step 1: Check Correlation between y and x's

y should be correlated with some of the x's

Time series model

Generally there will not be any x's

Hence patterns in y series is explored

y will be modeled in terms of previous values of y

$$y_t = a + b_1y_{t-1} + b_2y_{t-2} + - -$$

Step 1: Check correlation between y_t and y_{t-1}, etc

correlation between y and previous values of y are called autocorrelation

Time Series Modeling

Example: Check the auto-correlation up to 3 lags in GDP data

Year	GDP(y_t)	y_{t-1}	y_{t-2}	y_{t-3}
1993	94.43			
1994	100	94.43		
1995	107.3	100	94.43	
1996	115.1	107.3	100	94.43
1997	124.2	115.1	107.3	100
1998	130.1	124.2	115.1	107.3
1999	138.6	130.1	124.2	115.1
2000	147	138.6	130.1	124.2
2001	153.4	147	138.6	130.1
2002	162.3	153.4	147	138.6
2003	168.7	162.3	153.4	147

Lag	variables	Auto Correlation
1	y_t vs y_{t-1}	0.9985
2	y_t vs y_{t-2}	0.9984
3	y_t vs y_{t-3}	0.9982

Time Series Modeling

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

- Widely used and very effective modeling approach
- Proposed by George Box and Gwilym Jenkins
- Also known as Box – Jenkins model or ARIMA(p,d,q)

where,

p: number of auto regressive (AR) terms

q: number of moving average (MA) terms

d: level of differencing

Forecast Method

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

General Form

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + - - - + \theta_1 e_{t-1} + + \theta_2 e_{t-2} - - - -$$

Where

c: constant

$\varphi_1, \varphi_2, \theta_1, \theta_2$, - - - are model parameters

$e_{t-1} = y_{t-1} - s_{t-1}$, e_t are called errors or residuals

s_{t-1} : predicted value for the t-1th observation (y_{t-1})

Forecast Method

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 1:

Draw time series plot and check for trend, seasonality, etc

Step 2:

Draw Auto Correlation Function (ACF) and Partially Auto Correlation Function (PACF) graphs to identify auto correlation structure of the series

Step 3:

Check whether the series is stationary.

If series is non stationary do differencing or transform the series

Forecast Method

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 4:

Identify the model using ACF and PACF or automatically

The best model is one which minimizes AIC or BIC or both

Step 5:

Estimate the model parameters using maximum likelihood method (MLE)

Forecast Method

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Example: The number of visitors to a web page is given in Visits.csv. Develop a model to predict the daily number of visitors?

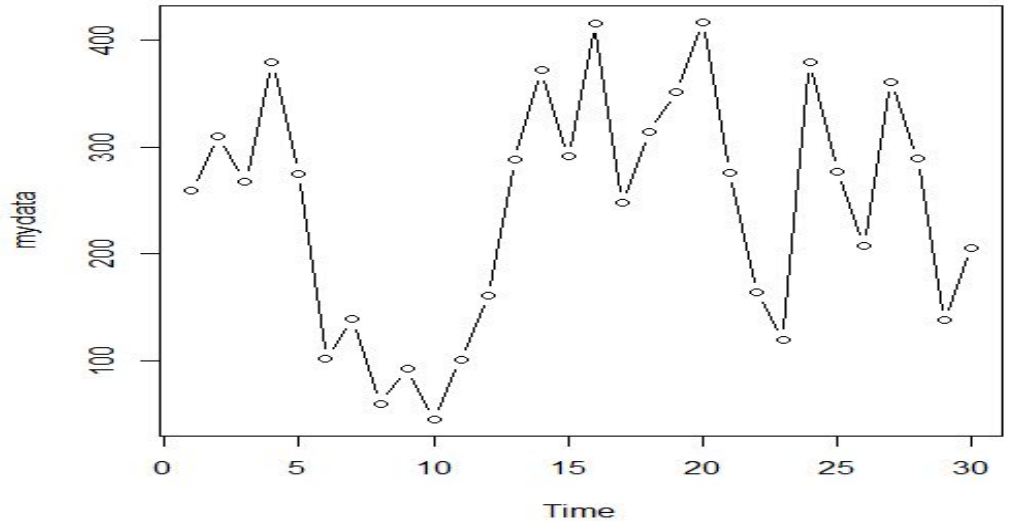
SL No.	Data	SL No.	Data
1	259	16	416
2	310	17	248
3	268	18	314
4	379	19	351
5	275	20	417
6	102	21	276
7	139	22	164
8	60	23	120
9	93	24	379
10	45	25	277
11	101	26	208
12	161	27	361
13	288	28	289
14	372	29	138
15	291	30	206

Forecast Method

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 1: Read and plot the series

```
mydata <- read.csv("Visits.csv")
mydata <- ts(mydata$Data)
plot(mydata, type = "b")
```



Forecast Method

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 2: Descriptive Statistics

summary(mydata)

Statistic	Value
Minimum	45
Quartile 1	144.5
Median	271.5
Mean	243.6
Quartile 3	313
Maximum	417

Forecast Method

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 3: Check whether the series is stationary

```
ndiffs(mydata)
```

Number of differences required = 0
Hence the series is stationary

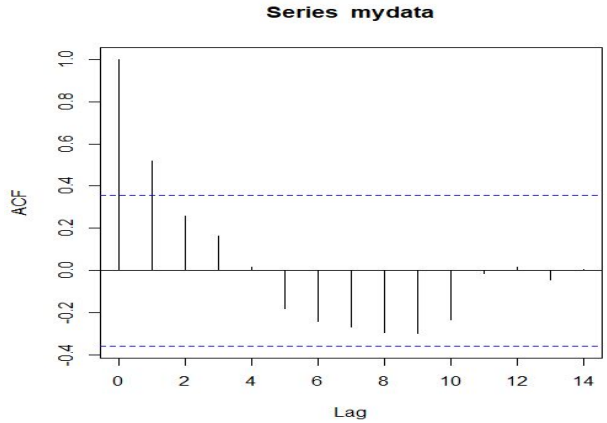
Forecast Method

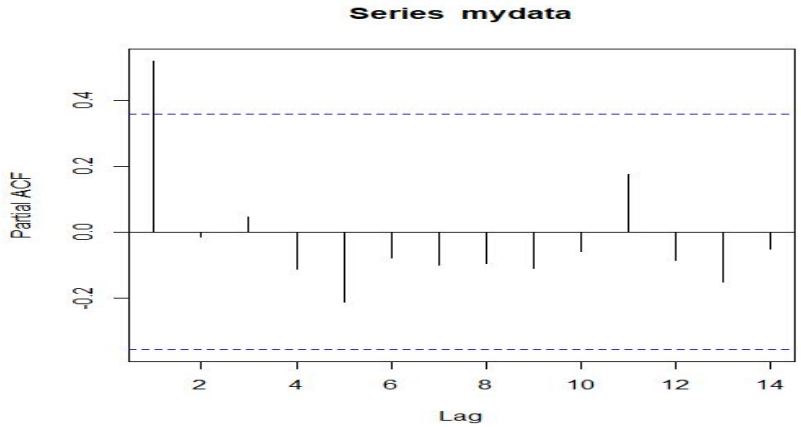
Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 4: Draw ACF & PACF Graphs

acf(mydata)

pacf(mydata)





Potential Models

ARMA(1,0) since acf at lag 1 is crossing 95% confidence interval

ARMA(0,1) since pacf at lag 1 is crossing 95% confidence interval

ARMA(1,1) since both acf and pacf at lag 1 is crossing 95% confidence interval

Forecast Method

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 5: Identification of model automatically

```
library(forecast)
mymodel = auto.arima(mydata)
mymodel
```

Model	Log likelihood	AIC	BIC
ARIMA(1,0,0)	-178.31	362.62	366.82

Model Parameters	Value
Intercept	242.8594
AR1	0.5064

Forecast Method

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 6: Identification of model manually

arima(mydata, c(0,0,1))
arima(mydata, c(1,0,0))
arima(mydata, c(1,0,1))

Model		Log likelihood	AIC
p=0,q=1	ARIMA(0,0,1)	-179.07	364.15
p=1,q=0	ARIMA(1,0,0)	-178.31	362.62
p=1,q=1	ARIMA(1,0,1)	-178.31	364.62

Conclusion:

The best model which minimizes AIC & BIC is p=1, q=0 or ARIMA(1,0,0)

Identified automatically

Forecast Method

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 7: Estimation of parameters

ARIMA(1,0,0) Parameters	Value	Std Error
Intercept	242.8594	32.8552
AR1	0.5064	0.1520

The model is

$$Y_t = 242.8594 + 0.5064\, Y_{t-1}$$

Forecast Method

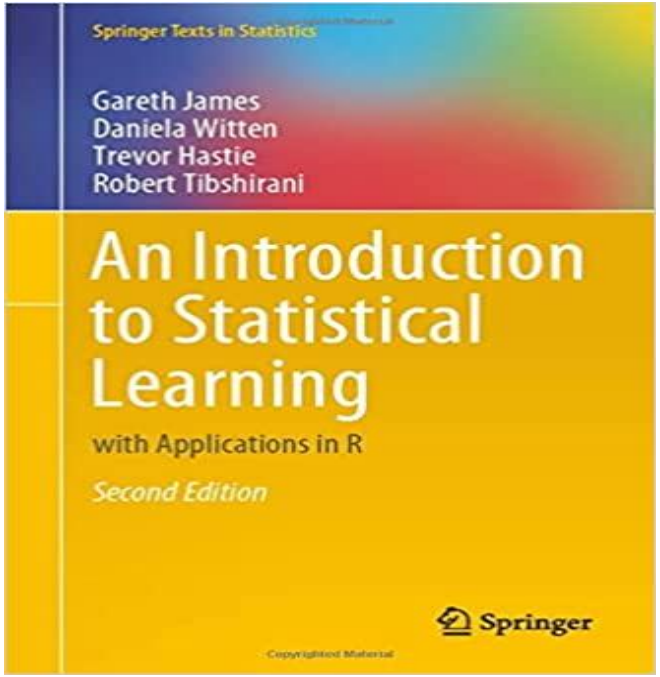
Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 8: Model Diagnostics

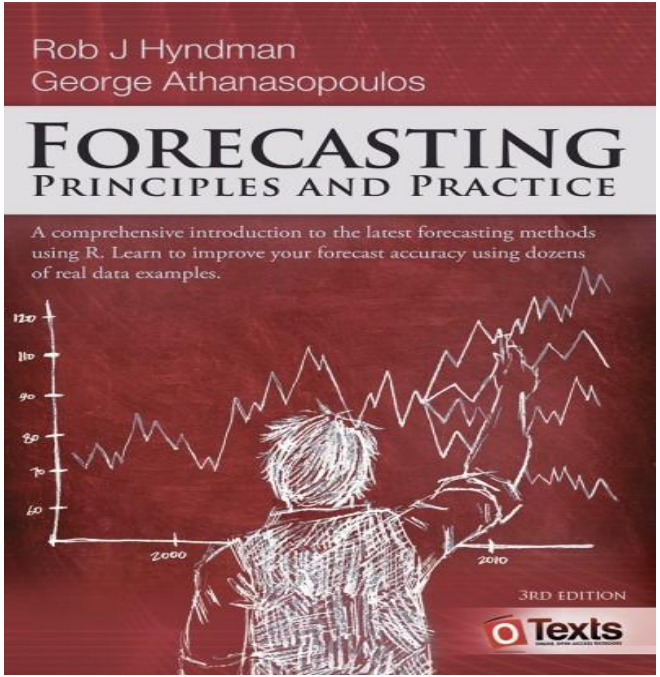
summary(mymodel)

Statistic	Description	Value
ME	Residual average	-0.3470709
MAE	Average of absolute residuals	76.90398
RMSE	Root mean square of residuals	91.81328
MAPE	Mean absolute percent error	47.78088

References



Read ISLR online



Read FPP Online

For any query drop an email at

madhurima.panja@gmail.com