MALLA REDDY INSTITUTE OF TECHNOLOGY & SCIENCE
(SPONSORED BY MALLA REDDY EDUCATIONAL SOCIETY)
Affiliated to JNTUH & Approved by AICTE, New Delhi
NAAC with 'A' Grade, NBA Accredited, ISO 9001:2015 Certified, Approved by UK Accrediation Centre
Granted Status of 2(f) & 12(b) under UGC Act, 1956, Govt. of India.

DATA ANALYTICS

COURSE FILE

B.Tech (CSE,IT & CSE(NW)) III Year – I

SemesterR18 Regulation

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING(2022-2023)

Compiled by

Dr. M. Jaganathan
Professor
Computer Science and Engineering

CS513PE: DATA ANALYTICS (Professional Elective - I)

III Year B.Tech. CSE I-Sem

L	T	P	C
3	0	0	3

Prerequisites

1. A course on "Database Management Systems".
2. Knowledge of probability and statistics.

Course Objectives:

- To explore the fundamental concepts of data analytics.
- To learn the principles and methods of statistical analysis
- Discover interesting patterns, analyze supervised and unsupervised models and estimate theaccuracy of the algorithms.
- To understand the various search methods and visualization techniques.

Course Outcomes: After completion of this course students will be able to

- Understand the impact of data analytics for business decisions and strategy
- Carry out data analysis/statistical analysis
- To carry out standard data visualization and formal inference procedures
- Design Data Architecture
- Understand various Data Sources

UNIT - I

Data Management: Design Data Architecture and manage the data for analysis, understand various sources of Data like Sensors/Signals/GPS etc. Data Management, Data Quality(noise, outliers, missing values, duplicate data) and Data Processing & Processing.

UNIT - II

Data Analytics: Introduction to Analytics, Introduction to Tools and Environment, Application of Modelingin Business, Databases & Types of Data and variables, Data Modeling Techniques, Missing Imputationsetc. Need for Business Modeling.

UNIT - III

Regression – Concepts, Blue property assumptions, Least Square Estimation, Variable Rationalization, and Model Building etc.

Logistic Regression: Model Theory, Model fit Statistics, Model Construction, Analytics applications to various Business Domains etc.

UNIT - IV

Object Segmentation: Regression Vs Segmentation – Supervised and Unsupervised Learning, Tree Building – Regression, Classification, Overfitting, Pruning and Complexity, Multiple Decision Trees etc. Time Series Methods: Arima, Measures of Forecast Accuracy, STL approach, Extract features from generated model as Height, Average Energy etc and Analyze for prediction

UNIT - V

Data Visualization: Pixel-Oriented Visualization Techniques, Geometric Projection Visualization Techniques, Icon-Based Visualization Techniques, Hierarchical Visualization Techniques,

Visualizing Complex Data and Relations.

TEXT BOOKS:

1. Student's Handbook for Associate Analytics – II, III.
2. Data Mining Concepts and Techniques, Han, Kamber, 3rd Edition, Morgan KaufmannPublishers.

REFERENCE BOOKS:

1. Introduction to Data Mining, Tan, Steinbach and Kumar, Addision Wisley, 2006.
2. Data Mining Analysis and Concepts, M. Zaki and W. Meira
3. Mining of Massive Datasets, Jure Leskovec Stanford Univ. Anand Rajaraman Milliway LabsJeffrey D Ullman Stanford Univ.

CS513PE: DATA ANALYTICS (Professional Elective - I)

BASIC TERMINOLOGIES

BIG DATA

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

4V PROPERTIES OF BIG DATA

- Volume
- Variety
- Velocity
- Veracity

Volume of Big Data

The volume of data refers to the size of the data sets that need to be analyzed and processed, which are now frequently larger than terabytes and petabytes. The sheer volume of the data requires distinct and different processing technologies than traditional storage and processing capabilities. In other words, this means that the data sets in Big Data are too large to process with a regular laptop or desktop processor. An example of a high-volume data set would be all credit card transactions on a day within Europe.

Velocity of Big Data

Velocity refers to the speed with which data is generated. High velocity data is generated with such a pace that it requires distinct (distributed) processing techniques. An example of a data that is generated with high velocity would be Twitter messages or Facebook posts.

Variety of Big Data

Variety makes Big Data really big. Big Data comes from a great variety of sources and generally is one out of three types: structured, semi structured and unstructured data. The variety in data types frequently requires distinct processing capabilities and specialist algorithms. An example of high variety data sets would be the CCTV audio and video files that are generated at various locations in a city.

Veracity of Big Data

Veracity refers to the quality of the data that is being analyzed. High veracity data has many records that are valuable to analyze and that contribute in a meaningful way to the overall results. Low veracity data, on the other hand, contains a high percentage of meaningless data.

The non-valuable in these data sets is referred to as noise. An example of a high veracity data set would be data from a medical experiment or trial.

Data that is high volume, high velocity and high variety must be processed with advanced tools (analytics and algorithms) to reveal meaningful information. Because of these characteristics of the data, the knowledge domain that deals with the storage, processing, and analysis of these data sets has been labeled Big Data.

FORMS OF DATA

- Collection of information stored in a particular file is represented as forms of data.
  - **STRUCTURED FORM**
    - Any form of relational database structure where relation between attributes is possible. That there exists a relation between rows and columns in the database with a table structure. Eg: using database programming languages **(sql, oracle, mysql etc**).
  - **UNSTRUCTURED FORM.**
    - Any form of data that does not have predefined structure is represented as unstructured form of data. Eg: **video, images, comments, posts, few websites such as blogs and wikipedia**
  - **SEMI STRUCTURED DATA**
    - Does not have form tabular data similar to rdbms.
    - Predefined organized formats available.
    - Eg: csv, xml, json, txt file with tab separator etc..

SOURCES OF DATA

– There are two types of sources of data available.

- **PRIMARY SOURCE OF DATA**
  - Eg: data created by individual or a business concern on their own.
- **SECONDARY SOURCE OF DATA**
  - Eg: data can be extracted from cloud servers, website sources (kaggle, uci, aws, google cloud, twitter, facebook, youtube, github etc..)

DATA ANALYSIS

Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making.

DATA ANALYTICS

- Data analytics is the science of analyzing raw data in order to make conclusions about that information..... This information can then be used to optimize processes to increase the overall efficiency of a business or system.

Types:

– **Descriptive analytics** Eg: (observation, case-study, surveys)

In descriptive statistics the result is always going lead with probability among 'n' number of options where each option has an equal chance of probability.

– **Predictive analytics** Eg: healthcare, sports, weather, insurance, social media analysis.

This type of analytics deals with predicting past data to make decisions based on certain algorithms. In case of a doctor the doctor questions the patient about the past to correct his illness through already existing procedures.

– **Prescriptive analytics** Eg: healthcare, banking.

Prescriptive analytics works with predictive analytics, which uses data to determine near-term outcomes. Prescriptive analytics makes use of machine learning to help businesses decide a course of action based on a computer program's predictions.

Fig 0.1Relation between Social Media, Data Analysis and Big Data

Social media data are used in number of domains such as health and political trending and forecasting, hobbies, ebusiness,cyber-crime, counter terrorism, time-evolving opinion mining, social net-work analysis, and human machineinteractions.

Finally, summarizing all the above concepts processing for social media data can be categorized into 3 parts as shown infigure 0.1. The first part consists of social media websites, the second part consists of data analysis part and the thirdpart consists of big data management layer and schedules the jobs across the cluster.

DIFFERENCE BETWEEN DATA ANALYTICS AND DATA ANALYSIS

Characteristics	Data Analytics	Data Analysis
Form	Used in business to make decision from data – Data Driven	Form of data analytics in business to identify useful information in data.
Structure	It is a process of data collection with various strategies	Cleaning, transforming the data
Tools	Excel, python, R etc..	KNIME, NodeXL, Rapid Miner etc..
Prediction	Analytics means we are trying to find conclusions about future.	Analysis means we analyze always what has happened in the past

MACHINE LEARNING

- Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
- Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

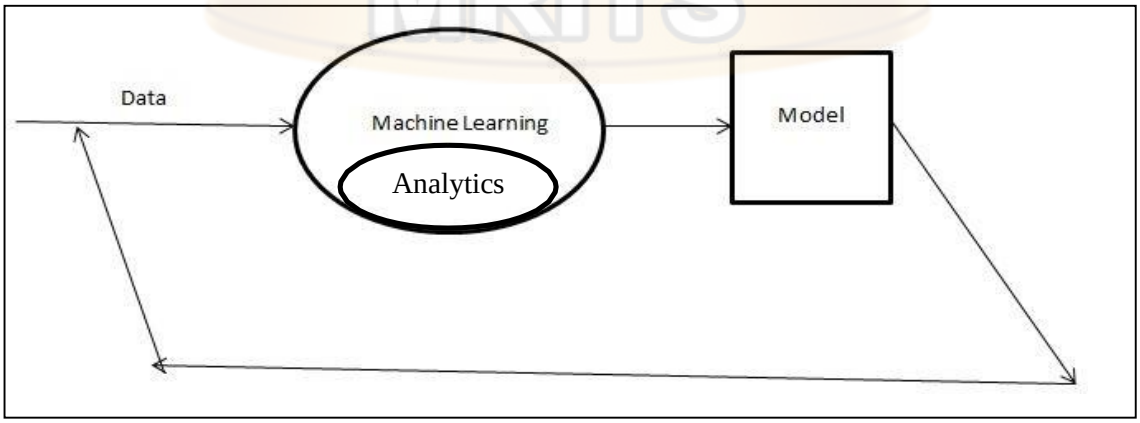


Fig 0.2Relation betweenmachine learning and data analytics

In general data is passed to a machine learning tool to perform descriptive data analytics through set of algorithms built in it. Here both data analytics and data analysis is done by the tool automatically. Hence we can say that Data analysis is a sub component of data analytics. And data analytics is a sub component of machine learning tool. All these are described in figure 0.2. The output of this machine learning tool generates a model. And from this model predictive analytics and prescriptive analytics can be performed because the model gives output as data to machine learning tool. This cycle continues till we get an efficient output.

UNIT - I

1.1 DESIGN DATA ARCHITECTURE AND MANAGE THE DATA FOR ANALYSIS

Data architecture is composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organizations. Data is usually one of several architecture domains that form the pillars of an enterprise architecture or solution architecture.

Various constraints and influences will have an effect on data architecture design. These include enterprise requirements, technology drivers, economics, business policies and data processing needs.

- **Enterpriserequirements**

These will generally include such elements as economical and effective system expansion, acceptable performance levels (especially system access speed), transaction reliability, and transparent data management. In addition, the conversion of raw data such as transaction records and image files into more useful information forms through such features as data warehouses is also a common organizational requirement, since this enables managerial decision making and other organizational processes. One of the architecture techniques is the split between managing transaction data and (master) reference data. Another one is splitting data capture systems from data retrieval systems (as done in a datawarehouse).

- **Technologydrivers**

These are usually suggested by the completed data architecture and database architecture designs. In addition, some technology drivers will derive from existing organizational integration frameworks and standards, organizational economics, and existing site resources (e.g. previously purchased software licensing).

- **Economics**

These are also important factors that must be considered during the data architecture phase. It is possible that some solutions, while optimal in principle, may not be potential candidates due to their cost. External factors such as the business cycle, interest rates, market conditions, and legal considerations could all have an effect on decisions relevant to data architecture.

- **Businesspolicies**

Business policies that also drive data architecture design include internal organizational policies, rules of regulatory bodies, professional standards, and applicable governmental laws that can vary by applicable agency. These policies and rules will help describe the manner in which enterprise wishes to process their data.

- **Data processingneeds**

These include accurate and reproducible transactions performed in high volumes, data warehousing for the support of management information systems (and potential data mining), repetitive periodic reporting, ad hoc reporting, and support of various organizational initiatives as required (i.e. annual budgets, new productdevelopment).

The General Approach is based on designing the Architecture at three Levels of Specification as shown below in figure 1.1

- The LogicalLevel
- The PhysicalLevel
- The ImplementationLevel

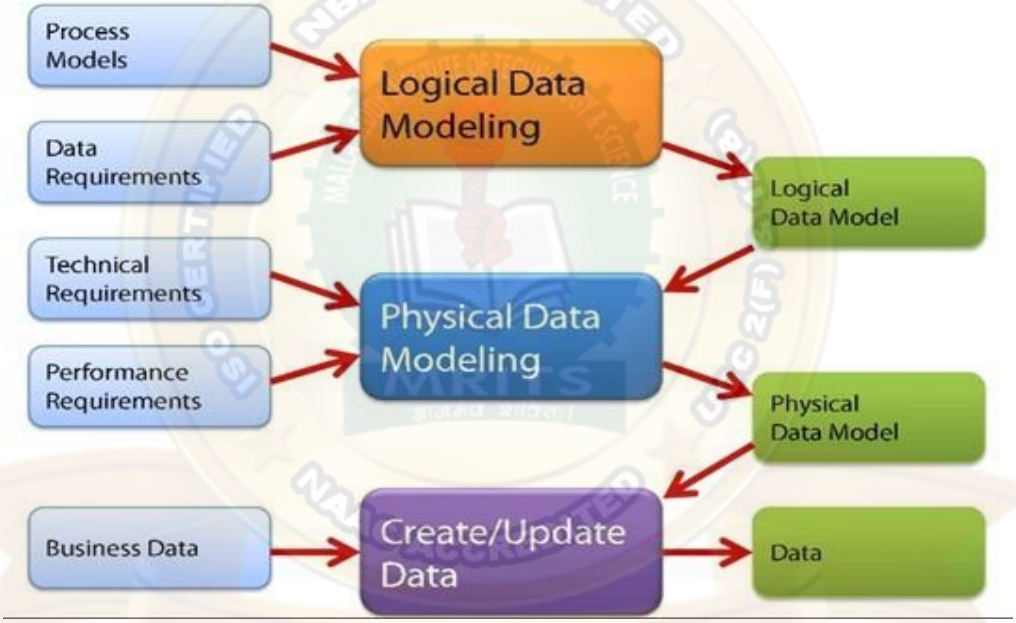


Fig 1.1: Three levels architecture in data analytics.

The **logical view/user's view**, of a data analytics represents data in a format that is meaningful to a user and to the programs that process those data. That is, the logical view tells the user, in user terms, what is in the database. Logical level consists of data requirements and process models which are processed using any data modelling techniques to result in logical data model.

Physical level is created when we translate the top level design in physical tables in the database. This model is created by the **database architect, software architects, software developers or database administrator**. The input to this level from logical level and various data modeling techniques are used here with input from software developers or database administrator. These data modelling techniques are various formats of representation of data

such as relational data model, network model, hierarchical model, object oriented model, Entity relationship model.

Implementation level contains details about modification and presentation of data through the use of **various data mining tools** such as (R-studio, WEKA, Orange etc). Here each tool has a specific feature how it works and different representation of viewing the same data. These tools are very helpful to the user since it is user friendly and it does not require much programming knowledge from the user.

1.2 Various Sources of Data

Understand various primary sources of the Data

Data can be generated from two types of sources namely Primary and Secondary Sources of Primary Data

The sources of generating primary data are -

- ObservationMethod
- SurveyMethod
- ExperimentalMethod

Observation Method:

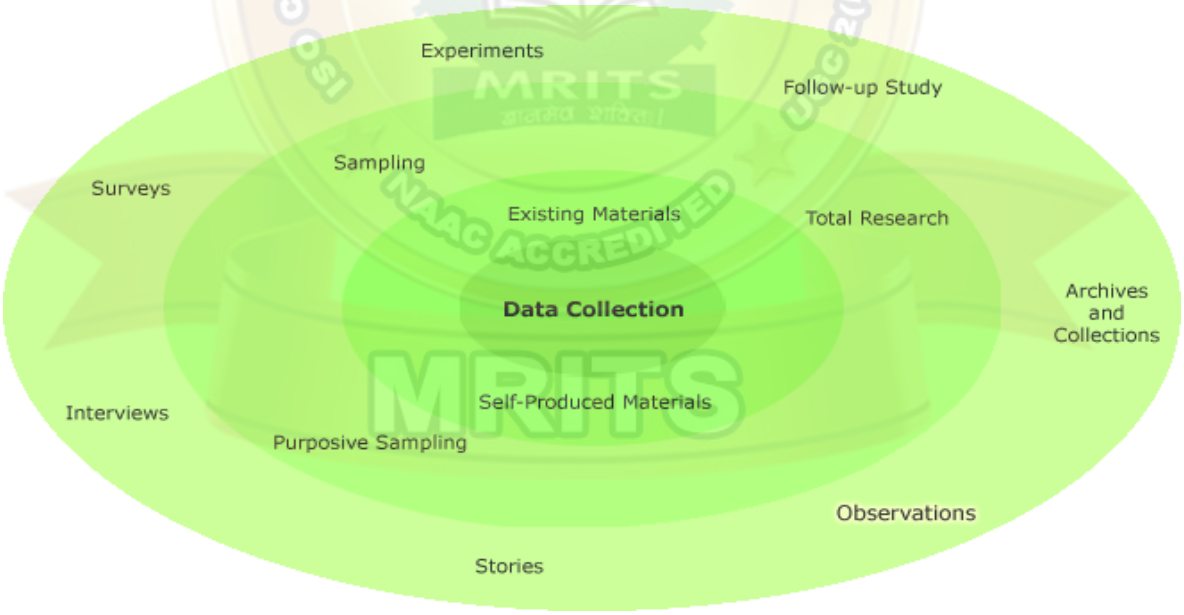


Fig 1.2: Data collections

An **observation is a data collection method**, by which you gather knowledge of the researched phenomenon through making observations of the phenomena, as and when it occurs. The main aim is to focus on observations of human behavior, the use of the phenomenon and human interactions related to the phenomenon. We can also make observations on verbal and nonverbal expressions. In making and documenting observations,

we need to clearly differentiate our own observations from the observations provided to us by other people. The range of data storage genre found in Archives and Collections, is suitable for documenting observations e.g. **audio, visual, textual and digital including sub-genres of note taking, audio recording and video recording**.

There exist various observation practices, and our role as an observer may vary according to the research approach. We make observations from either the outsider or insider point of view in relation to the researched phenomenon and the observation technique can be structured or unstructured. The degree of the outsider or insider points of view can be seen as a movable point in a continuum between the extremes of outsider and insider. If you decide to take the insider point of view, you will be a participant observer *in situ* and actively participate in the observed situation or community. The activity of a Participant observer *in situ* is called field work. This observation technique has traditionally belonged to the data collection methods of ethnology and anthropology. If you decide to take the outsider point of view, you try to try to distance yourself from your own cultural ties and observe the researched community as an outsider observer. These details are seen in figure 1.2.

Experimental Designs

There are number of experimental designs that are used in carrying out and experiment. However, Market researchers have used 4 experimental designs most frequently. These are –

CRD - Completely Randomized Design

A **completely randomized design (CRD)** is one where the treatments are assigned completely at random so that each experimental unit has the same chance of receiving any one treatment. For the CRD, any difference among experimental units receiving the same treatment is considered as experimental error. Hence, CRD is appropriate only for experiments with homogeneous experimental units, such as laboratory experiments, where environmental effects are relatively easy to control. For field experiments, where there is generally large variation among experimental plots in such environmental factors as soil, the CRD is rarely used. CRD is mainly used in agricultural field.

Step 1. Determine the total number of experimental plots (n) as the product of the number of treatments (t) and the number of replications (r); that is, $n = rt$. For our example, $n = 5 \times 4 = 20$. Here, one pot with a single plant in it may be called a plot. In case the number of replications is not the same for all the treatments, the total number of experimental pots is to be obtained as the sum of the replications for each treatment. *i.e.*,

$$n = \sum_{i=1}^{t} r_i$$

where r_i, is the number of times the ith treatment replicated

Step 2. Assign a plot number to each experimental plot in any convenient manner; for example, consecutively from 1 to *n*.

Step 3. Assign the treatments to the experimental plots randomly using a table of random numbers.

Example 1: Assume that a farmer wishes to perform the experiment to determine which of his 3 fertilizers to use on 2800 tress. Assuming that farmer has a farm divided in to 3 terraces, where those 2800 trees can be divided in the below format

Lower Terrace	1200
Middle Terrace	1000
Upper Terrace	600

Design a CRD for this experiment

Solution
Scenario 1

First we divide the 2800 trees in to random assignment of almost 3 equal parts
Random Assignment1: 933 trees
Random Assignment2: 933 trees
Random Assignment3: 934 trees
So for example random assignment1 we can assign fertilizer1, random assignment2 we can assign fertilizer2, random assignment3 we can assign fertilizer3.

Scenario 2

2800 trees is divided into terrace as shown below

Total no of trees	**Terrace**	**Random assignment**	**Fertilizer usage**
2800 Trees	Upper Terrace (600 trees)	200	fertilizer1
		200	fertilizer2
		200	fertilizer3
	Middle Terrace (1200 trees)	400	fertilizer1
		400	fertilizer2
		400	fertilizer3
	Lower Terrace (1000 trees)	333	fertilizer1
		333	fertilizer2
		334	fertilizer3

Thus the farmer will be able analyze and compare various fertilizer performance on different terrace.

Example 2:

A company wishes to test 4 different types of tyre. The tyres lifetime as determined from their threads are given. Where each tyre has been tried on 6 similar automobiles assigned at random to their tyres. Determine whether there is a significant difference between tyres at .05 level.

Tyres	Automobile 1	Automobile 2	Automobile 3	Automobile 4	Automobile 5	Automobile 6
A	33	38	36	40	31	35
B	32	40	42	38	30	34
C	31	37	35	33	34	30
D	29	34	32	30	33	31

Solution:

Null Hypothesis: There is no difference between the tyres in their life time.

We choose a **random value closest to the average of all values** in the table and subtract that for each tyre in the automobile, for example by choosing 35

Tyres	Automobile 1	Automobile 2	Automobile 3	Automobile 4	Automobile 5	Automobile 6	Total
A	-2	3	1	5	-4	0	3
B	-3	5	7	3	-5	-1	6
C	-4	2	0	-2	-1	-5	-10
D	6	-1	-3	-5	-2	-4	-21
	T = Sum(X) =						-22

N = no of samples = 24 (4 rows * 6 columns)

Correction factor = $\frac{T*T}{N}$ = 20.16

Square the values to find

Tyres	Automobile 1	Automobile 2	Automobile 3	Automobile 4	Automobile 5	Automobile 6	Total
A	4	9	1	25	16	0	55
B	9	25	9	49	25	1	118
C	16	4	0	4	1	25	50
D	36	1	9	25	4	16	91
	T = Sum(X^2) =						314

Total sum of squares (SST) = sum(X^2) – Correction factor
= 314 – 20.16 = 293.84
Sum of Squares between Treatments (SST_r) =
= $((3)^2/6 + (6)^2/6 + (10)^2/6 + (21)^2/6)$ – Correction factor = 77.50
Sum of Squares Error (SSE) SST – SST_r = 293.84 – 77.50 = 216.34

Now by using ANOVA (one way classification) Table, We calculate the F- Ratio.

F-Ratio:

The F ratio is the ratio of two mean square values. If the null hypothesis is true, you expect F to have a value close to 1.0 most of the time. A large F ratio means that the variation among group mean is more than you'd expect to see by chance

If the value of F-Ratio is closer to 1 it means that null hypothesis is true. If F-ratio is greater than then we assume that the null hypothesis is false.

Source of variation	**Sum of squares**	**Degrees of freedom**	**Mean of sum of squares**	**F - Ratio**
Between treatments	SST_r = 77.50	No of treatment – 1 = 4-1 =3	MST_r = SST_r/ Degrees of Freedom = 77.50/3 = 25.83	
				F-ratio = MST_r/MSE = 25.83/10.87 = 2.376
Within treatments	SSE = 216.34	No of values – no of treatment = 24 – 4 =20	MSE = SSE/Degrees of Freedom =216.34/20 =10.87	

In this scenario the value of F-ratio is greater than 1. This indicates there will be variation between samples. So assumed null hypothesis will be false
Level of significance = 0.05 (given in question)
Degrees of Freedom = (3, 20)
Critical value = 3.10 (calculated from 5 percentage table)
F-Ratio >critical value (i.e) 2.376> 3.10

Hence assumed null hypothesis is false. This indicates there is life time difference between tyres.

A **randomized block design**, the experimenter divides subjects into subgroups called **blocks**, such that the **variability within blocks is less than the variability between blocks**. Then, subjects within each block are randomly assigned to treatment conditions. Compared to a completely randomized design, this design reduces variability within treatment conditions and potential confounding, producing a better estimate of treatment effects.

The table below shows a randomized block design for a hypothetical medical experiment.

Gender	**Treatment**	
	Placebo	**Vaccine**
Male	250	250
Female	250	250

Subjects are assigned to blocks, based on gender. Then, within each block, subjects are randomly assigned to treatments (either a placebo or a cold vaccine). For this design, 250 men get the placebo, 250 men get the vaccine, 250 women get the placebo, and 250 women get the vaccine.

It is known that men and women are physiologically different and react differently to medication. This design ensures that each treatment condition has an equal proportion of men and women. As a result, differences between treatment conditions cannot be attributed to gender. This randomized block design removes gender as a potential source of variability and as a potential confounding variable.

LSD - Latin Square Design - A Latin square is one of the experimental designs which has a balanced two-way classification scheme say for example - 4 X 4 arrangement. In this scheme each letter from A to D occurs only once in each row and also only once in each column. The balance arrangement, it may be noted that, will not get disturbed if any row gets changed with the other.

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

The balance arrangement achieved in a Latin Square is its main strength. In this design, the comparisons among treatments, will be free from both differences between rows and columns. Thus the magnitude of error will be smaller than any other design.

FD - Factorial Designs - This design allows the experimenter to test two or more variables simultaneously. It also measures interaction effects of the variables and analyzes the impacts of each of the variables.

In a true experiment, randomization is essential so that the experimenter can infer cause and effect without any bias.

Sources of Secondary Data

While primary data can be collected through questionnaires, depth interview, focus group interviews, case studies, experimentation and observation; The secondary data can be obtained through

- Internal Sources - These are within theorganization
- External Sources - These are outside theorganization
- Internal Sources ofData

Internal sources

If available, internal secondary data may be obtained with less time, effort and money than the external secondary data. In addition, they may also be more pertinent to the situation at hand since they are from within the organization. The internal sources include

Accounting resources- This gives so much information which can be used by the marketing researcher. They give information about internal factors.

Sales Force Report- It gives information about the sale of a product. The information provided is of outside theorganization.

Internal Experts- These are people who are heading the various departments. They can give an idea of how a particular thing isworking

Miscellaneous Reports- These are what information you are getting from operational reports.If the data available within the organization are unsuitable or inadequate, the marketer should extend the search to external secondary data sources.

External Sources of Data

External Sources are sources which are outside the company in a larger environment. Collection of external data is more difficult because the data have much greater variety and the sources are much more numerous.

External data can be divided into following classes.

Government Publications- Government sources provide an extremely rich pool of data for the researchers. In addition, many of these data are available free of cost on internet websites. There are number of government agencies generating data. These are:

Registrar General of India- It is an office which generates demographic data. It includes details of gender, age, occupation etc.

Central Statistical Organization- This organization publishes the national accounts statistics. It contains estimates of national income for several years, growth rate, and rate of major economic activities. Annual survey of Industries is also published by the CSO. It gives information about the total number of workers employed, production units, material used and value added by themanufacturer.

Director General of Commercial Intelligence- This office operates from Kolkata. It gives information about foreign trade i.e. import and export. These figures are provided region-wise and country-wise.

Ministry of Commerce and Industries- This ministry through the office of economic advisor provides information on wholesale price index. These indices may be related to a number of sectors like food, fuel, power, food grains etc. It also generates All India Consumer Price Index numbers for industrial workers, urban, non-manual employees and cultural labourers.

Planning Commission- It provides the basic statistics of Indian Economy.

Reserve Bank of India- This provides information on Banking Savings and investment. RBI also prepares currency and finance reports.

Labour Bureau- It provides information on skilled, unskilled, white collared jobs etc.
National Sample Survey- This is done by the Ministry of Planning and it provides social, economic, demographic, industrial and agricultural statistics.

Department of Economic Affairs- It conducts economic survey and it also generates information on income, consumption, expenditure, investment, savings and foreign trade.

State Statistical Abstract- This gives information on various types of activities related to the state like - commercial activities, education, occupation etc.

Non-Government Publications- These includes publications of various industrial and trade associations, such as

The Indian Cotton Mill Association Various chambers of commerce

The Bombay Stock Exchange (it publishes a directory containing financial accounts, key profitability and other relevant matter)

Various Associations of Press Media. Export Promotion Council.

Confederation of Indian Industries (CII)
Small Industries Development Board of India

Different Mills like - Woolen mills, Textile mills etc

The only disadvantage of the above sources is that the data may be biased. They are likely to colour their negative points.

Syndicate Services- These services are provided by certain organizations which collect and tabulate the marketing information on a regular basis for a number of clients who are the subscribers to these services. So the services are designed in such a way that the information suits the subscriber. These services are useful in television viewing, movement of consumer goods etc. These syndicate services provide information data from both household as well as institution.

In collecting data from household they use three approaches Survey- They conduct surveys regarding - lifestyle, sociographic, general topics. Mail Diary Panel- It may be related to 2 fields - Purchase and Media.

Electronic Scanner Services- These are used to generate data on volume.
They collect data for Institutions from Whole sellers, Retailers, and Industrial Firms

Various syndicate services are **Operations Research Group** (ORG) and The **Indian Marketing Research Bureau (IMRB).**

Importance of Syndicate Services

Syndicate services are becoming popular since the constraints of decision making are changing and we need more of specific decision-making in the light of changing environment. Also Syndicate services are able to provide information to the industries at a low unit cost.

Disadvantages of Syndicate Services

The information provided is not exclusive. A number of research agencies provide customized services which suits the requirement of each individual organization.

International Organization- These includes

The International Labour Organization (ILO)- It publishes data on the total and active population, employment, unemployment, wages and consumer prices

The Organization for Economic Co-operation and development (OECD) - It publishes data on foreign trade, industry, food, transport, and science andtechnology.

The International Monetary Fund (IMA) - It publishes reports on national and international foreign exchange regulations.

1.2.1 Comparison of sources of data

Based on various features (cost, data, process, source time etc.) various sources of data can be compared as per table 1.

Table 1: Difference between primary data and secondary data.

Comparison Feature	**Primary data**	**Secondary data**
Meaning	Data that is collected by a researcher.	Data that is collected by other people.
Data	Real time data	Past data.
Process	Very involved	Quick and easy
Source	Surveys, interviews, or experiments, questionnaire, interview etc..	Books, journals, publications etc..
Cost effectiveness	Expensive	Economical
Collection time	Long	Short
Specific	Specific to researcher need	May not be specific to researcher need
Available	Crude form	Refined form
Accuracy and reliability	More	Less

1.3 Understanding Sources of Data from Sensor

Sensor data is the output of a device that detects and responds to some type of input from the physical environment. The output may be used to provide information or input to another system or to guide a process. Examples are as follows

- A **photosensor** detects the presence of visible light, infrared transmission (IR) and/or ultraviolet (UV) energy.
- **Lidar, a laser-based method of detection**, range finding and mapping, typically uses a low-power, eye-safe pulsing laser working in conjunction with a camera.
- A **charge-coupled device** (CCD) stores and displays the data for an image in such a way that each pixel is converted into an electrical charge, the intensity of which is related to a color in the color spectrum.
- **Smart grid sensors** can provide real-time data about grid conditions, detecting outages, faults and load and triggering alarms.
- **Wireless sensor networks** combine specialized transducers with a communications infrastructure for monitoring and recording conditions at diverse locations. Commonly monitored parameters include temperature, humidity, pressure, wind direction and speed, illumination intensity, vibration intensity, sound intensity, powerline voltage, chemical concentrations, pollutant levels and vital body functions.

1.4 Understanding Sources of Data from Signal

The simplest form of **signal** is a **direct current (DC)** that is switched on and off; this is the principle by which the early telegraph worked. More complex signals consist of an **alternating-current (AC)** or electromagnetic carrier that contains one or more data streams.

Data must be transformed into electromagnetic signals prior to transmission across a network. **Data and signals can be either analog or digital**. A signal is periodic if it consists of a continuously repeating pattern.

1.5 Understanding Sources of Data from GPS

The Global Positioning System (GPS) is a **space based navigation system** that provides location and time information in all weather conditions, anywhere on or near the Earth where there is an unobstructed line of sight to four or more GPS satellites. The system provides critical capabilities to military, civil, and commercial users around the world. The United States government created the system, maintains it, and makes it freely accessible to anyone with a **GPS receiver**.

1.6 Data Management

Data management is the development and execution of architectures, policies, practices and procedures in order to manage the information lifecycle needs of an enterprise in an effective manner.

1.7 Data Quality

Data quality refers to the quality of data. ***Data quality*** refers to the state of qualitative or quantitative pieces of information. There are many definitions of ***data quality*** but ***data*** is generally considered high ***quality*** if it is "fit for [its] intended uses in operations, decision making and planning

The seven characteristics that define data quality are:

1. Accuracy and Precision
2. Legitimacy and Validity
3. Reliability and Consistency
4. Timeliness and Relevance
5. Completeness and Comprehensiveness
6. Availability and Accessibility
7. Granularity and Uniqueness

Accuracy and Precision: This characteristic refers to the exactness of the data. It cannot have any erroneous elements and must convey the correct message without being misleading. This accuracy and precision have a component that relates to its intended use. Without understanding how the data will be consumed, ensuring accuracy and precision could be off-

target or more costly than necessary. For example, **accuracy in healthcare** might be more important than in another industry (which is to say, inaccurate data in healthcare could have more serious consequences) and, therefore, justifiably worth higher levels of investment.

Legitimacy and Validity: Requirements governing data set the boundaries of this characteristic. For example**, on surveys, items such as gender, ethnicity, and nationality** are typically limited to a set of options and **open answers are not permitted**. Any answers other than these would not be considered valid or legitimate based on the survey's requirement. This is the case for most data and must be carefully considered when determining its quality. The people in each department in an organization understand what data is valid or not to them, so the requirements must be leveraged when evaluating data quality.

Reliability and Consistency: Many systems in today's environments use and/or collect the same source data. Regardless of what source collected the data or where it resides, it cannot contradict a value residing in a different source or collected by a different system. There must be a stable and steady mechanism that collects and stores the data without contradiction or unwarranted variance.

Timeliness and Relevance: There must be a valid reason to collect the data to justify the effort required, which also means it has to be collected at the right moment in time. **Data collected too soon or too late could misrepresent a situation and drive inaccurate decisions**.

Completeness and Comprehensiveness: Incomplete data is as dangerous as inaccurate data. Gaps in data collection lead to a partial view of the overall picture to be displayed. Without a complete picture of how operations are running, uninformed actions will occur. It's important to understand the complete set of requirements that constitute a comprehensive set of data to determine whether or not the requirements are being fulfilled.

Availability and Accessibility: This characteristic can be tricky at times due to legal and regulatory constraints. Regardless of the challenge, though, individuals need the right level of access to the data in order to perform their jobs. This **presumes that the data exists and is available for access to be granted**.

Granularity and Uniqueness: The level of detail at which data is collected is important, because confusion and inaccurate decisions can otherwise occur**. Aggregated, summarized and manipulated collections of data could offer a different meaning than the data implied at a lower level**. An appropriate level of granularity must be defined to provide sufficient uniqueness and distinctive properties to become visible. This is a requirement for operations to function effectively.

Noisy data is meaningless data. The term has often been used as a synonym for **corrupt data**. However, its meaning has expanded to include any data that cannot be understood and interpreted correctly by machines, such as unstructured text.

An **outlier** is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.

In statistics, **missing data, or missing values**, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Missing values can be replaced by following techniques:

- Ignore the record with missing values.
- Replace the missing term with constant.
- Fill the missing value manually based on domain knowledge.
- Replace them with mean (if data is numeric) or frequent value (if data is categorical)
- Use of modelling techniques such decision trees, baye`s algorithm, nearest neighbor algorithm Etc.

In computing, **data deduplication** is a specialized data compression technique for eliminating duplicate copies of repeating data. Related and somewhat synonymous terms are intelligent (data) compression and single instance (data) storage.

Noisy data

For objects, noise is considered an extraneous object.

For attributes, noise refers to modification of original values.

- Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen
- We can talk about **signal to noise ratio.**
  Left image of 2 sine waves has low or zero SNR; the right image are the two waves combined with noise and has high SNR

Origins of noise

- **outliers** -- values seemingly out of the normal range of data
- **duplicate records** -- good database design should minimize this (use DISTINCT on SQL retrievals)
- **incorrect attribute values** -- again good db design and integrity constraints should minimize this
- **numeric only**, deal with rogue strings or characters where numbers should be.
- **null handling** for attributes (nulls=missing values)

Missing Data Handling

Many causes: malfunctioning equipment, changes in experimental design, collation of different data sources, measurement not possible. People may wish to not supply information. Information is not applicable (childen don't have annual income)

- **Discard** records with missing values
- **Ordinal-continuous** data, could **replace with attribute means**
- **Substitute** with a value from a similar instance
- **Ignore** missing values, i.e., just proceed and let the tools deals with them
- **Treat** missing values **as equals** (all share the same missing value code)
- **Treat** missing values **as unequal values**

BUT...Missing (null) values may have significance in themselves (e.g. missing test in a medical examination, deathdate missing means still alive!)

Missing completely at random (MCAR)

- Missingness of a value is independent of attributes
- Fill in values based on the attribute as suggested above (e.g. attribute mean)
- Analysis may be unbiased overall

Missing at Random (MAR)

- Missingness is related to other variables
- Fill in values based other values (e.g., from similar instances)
- Almost always produces a bias in the analysis

Missing Not at Random (MNAR)

- Missingness is related to unobserved measurements
- Informative or non-ignorable missingness

Duplicate Data

Data set may include data objects that are duplicates, or almost duplicates of one another

A major issue when merging data from multiple, heterogeneous sources

- Examples: Same person with multiple email addresses

1.8 Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

Data goes through a series of steps during preprocessing:

- **Data Cleaning:** Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
- **Data Integration:** Data with different representations are put together and conflicts within the data are resolved.
- **Data Transformation:** Data is normalized, aggregated and generalized.
- **Data Reduction:** This step aims to present a reduced representation of the data in a data warehouse.

Data Discretization: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

Steps Involved in Data Preprocessing:

1. Data Cleaning:
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

(a). Missing Data:
This situation arises when some data is missing in the data. It can be handled in various ways.
Some of them are:

1. **Ignore the tuples:**
   This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
2. **Fill the Missing values:**
   There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

(b). Noisy Data:
Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

1. **Binning Method:**
   This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
2. **Regression:**
   Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
3. **Clustering:**
   This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation:
This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
2. **Attribute Selection:**
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
3. **Discretization:**
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
4. **Concept Hierarchy Generation:**
   Here attributes are converted from level to higher level in hierarchy. For Example- The attribute "city" can be converted to "country".

3. Data Reduction:
Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. **Data Cube Aggregation:**
   Aggregation operation is applied to data for the construction of the data cube.
2. **Attribute Subset Selection:**
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.
3. **Numerosity Reduction:**
   This enable to store the model of data instead of whole data, for example: Regression Models.
4. **Dimensionality Reduction:**
   This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Componenet Analysis).

UNIT – II
INTRODUCTION TO ANALYTICS

2.1 Introduction to Analytics

As an enormous amount of data gets generated, the need to extract useful insights is a must for a business enterprise. Data Analytics has a key role in improving your business. Here are 4 main factors which signify the need for Data Analytics:

- **Gather Hidden Insights** – Hidden insights from data are gathered and then analyzed with respect to business requirements.
- **Generate Reports** – Reports are generated from the data and are passed on to the respective teams and individuals to deal with further actions for a high rise in business.
- **Perform Market Analysis** – Market Analysis can be performed to understand the strengths and the weaknesses of competitors.
- **Improve Business Requirement** – Analysis of Data allows improving Business to customer requirements and experience.

Data Analytics refers to the techniques to analyze data to enhance productivity and business gain. Data is extracted from various sources and is cleaned and categorized to analyze different behavioral patterns. The techniques and the tools used vary according to the organization or individual.

Data analysts translate numbers into plain English. A Data Analyst delivers value to their companies by **taking information** about specific topics and then **interpreting, analyzing**, and presenting findings in comprehensive **reports**. So, if you have the capability to collect data from various sources, analyze the data, gather hidden insights and generate reports, then you can become a Data Analyst. Refer to the image below:

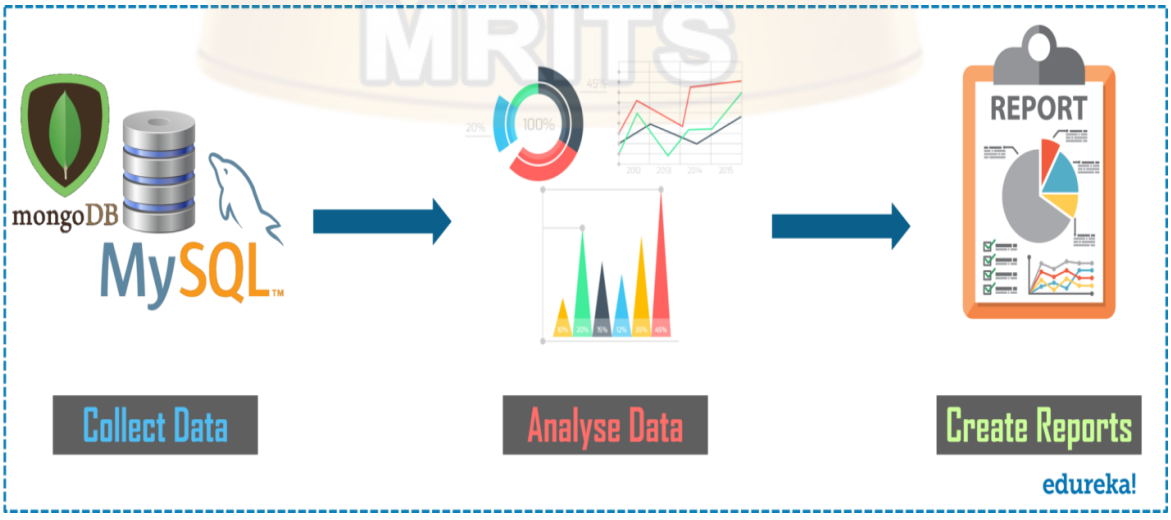


Fig 2.1 Data Analytics

In general data analytics also deals with bit of human knowledge as discussed below in figure 2.2 in this under each type of analytics there is a part of human knowledge required in prediction. Descriptive analytics requires the highest human input while predictive analytics requires less human input. In case of prescriptive analytics no human input is required since all the data is predicted.

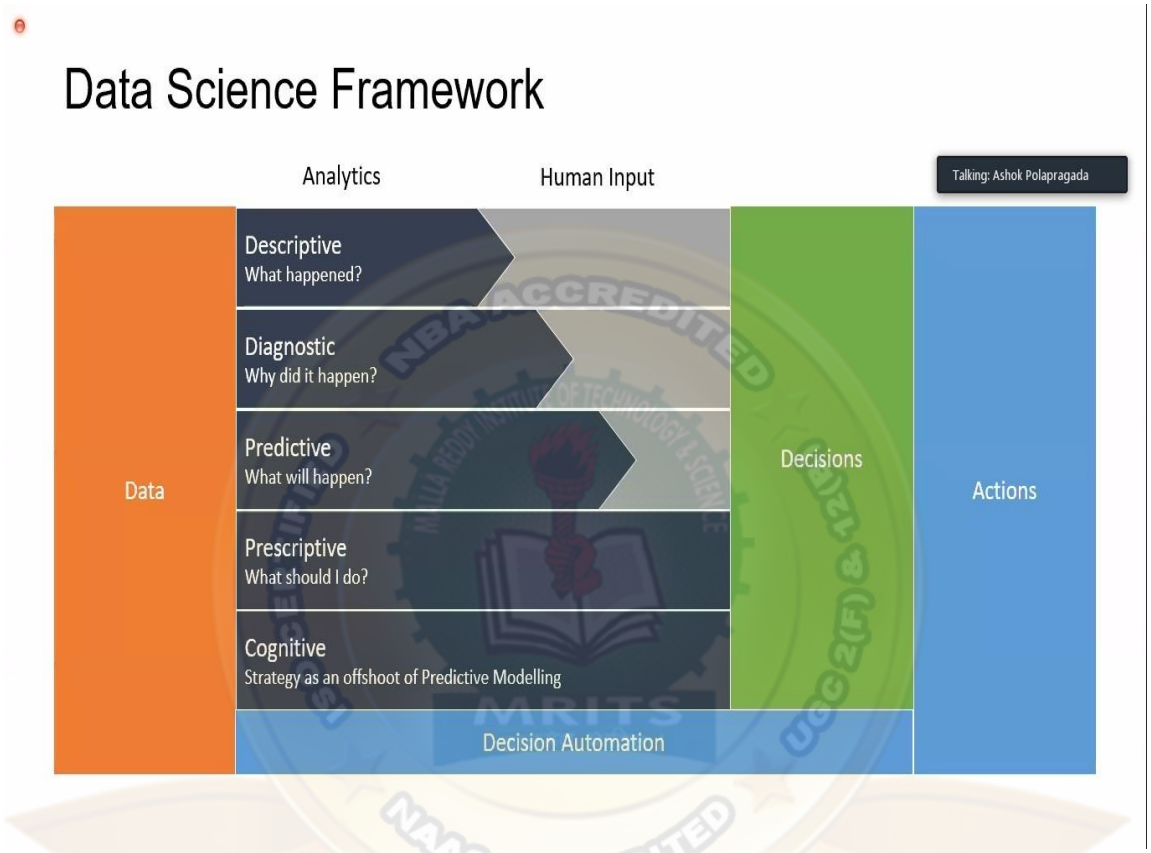


Fig 2.3 Data and Human work

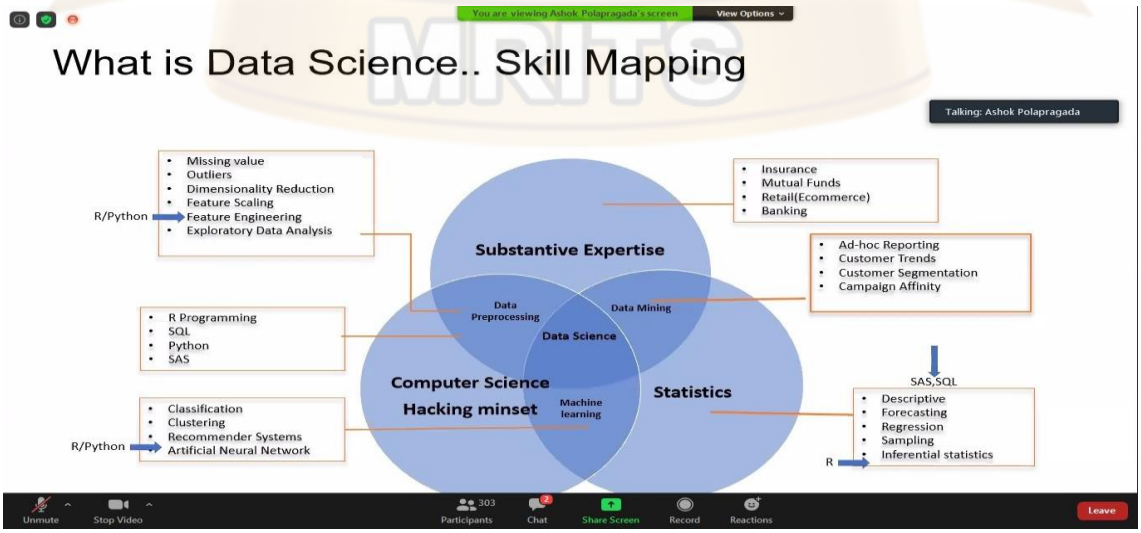


Fig 2.3 Data Analytics

2.2 Introduction to Tools and Environment

In general data analytics deals with three main parts, subject knowledge, statistics and person with computer knowledge to work on a tool to give insight in to the business. In the mainly used tool is Rand Phyton as shown in figure 2.3

With the increasing demand for Data Analytics in the market, many tools have emerged with various functionalities for this purpose. Either open-source or user-friendly, the top tools in the data analytics market are as follows.

- **R programming** – This tool is the leading analytics tool used for statistics and data modeling. R compiles and runs on various platforms such as UNIX, Windows, and Mac OS. It also provides tools to automatically install all packages as per user-requirement.
- **Python** – Python is an open-source, object-oriented programming language which is easy to read, write and maintain. It provides various machine learning and visualization libraries such as Scikit-learn, TensorFlow, Matplotlib, Pandas, Keras etc. It also can be assembled on any platform like SQL server, a MongoDB database or JSON
- **Tableau Public** – This is a free software that connects to any data source such as Excel, corporate Data Warehouse etc. It then creates visualizations, maps, dashboards etc with real-time updates on the web.
- **QlikView** – This tool offers in-memory data processing with the results delivered to the end-users quickly. It also offers data association and data visualization with data being compressed to almost 10% of its original size.
- **SAS** – A programming language and environment for data manipulation and analytics, this tool is easily accessible and can analyze data from different sources.
- **Microsoft Excel** – This tool is one of the most widely used tools for data analytics. Mostly used for clients' internal data, this tool analyzes the tasks that summarize the data with a preview of pivot tables.
- **RapidMiner** – A powerful, integrated platform that can integrate with any data source types such as Access, Excel, Microsoft SQL, Tera data, Oracle, Sybase etc. This tool is mostly used for predictive analytics, such as data mining, text analytics, machine learning.
- **KNIME** – Konstanz Information Miner (KNIME) is an open-source data analytics platform, which allows you to analyze and model data. With the benefit of visual programming, KNIME provides a platform for reporting and integration through its modular data pipeline concept.
- **OpenRefine** – Also known as GoogleRefine, this data cleaning software will help you clean up data for analysis. It is used for cleaning messy data, the transformation of data and parsing data from websites.
- **Apache Spark** – One of the largest large-scale data processing engine, this tool executes applications in Hadoop clusters 100 times faster in memory and 10 times faster on disk. This tool is also popular for data pipelines and machine learning model development.

Apart from the above-mentioned capabilities, a Data Analyst should also possess skills such as Statistics, Data Cleaning, Exploratory Data Analysis, and Data Visualization. Also, if you have knowledge of Machine Learning, then that would make you stand out from the crowd.

2.3 Application of modelling a business & Need for business Modelling

Data analytics is mainly involved in field of business in various concerns for the following purpose and it varies according to business needs and it is discussed below in detail. Nowadays majority of the business deals with prediction with large amount of data to work with.

Using big data as fundamental factor of making decision which need new capability, most firms are far away from accessing all data resources. Companies in various sectors have acquired crucial insight from the structured data collected from different enterprise systems and anatomize by commercial database management systems. Eg:

1.) Facebook and Twitter to standard the instantaneous influence on campaign and to examine consumer opinion about their products
2.) Some companies, like Amazon, eBay, and Google, considered as early commandants, examining factors that control performance to define what raise sales revenue and user interactivity.

2.3.1 Utilizing Hadoop in Big Data Analytics.

Hadoop is an open source software platform that enables processing of large data sets in a distributed computing environment", it discusses some concepts according to big data, the rules for building, organizing and analyzing huge data-sets in the business environment, they offered 3 architecture layers and also they indicate some graphical tools to explore and represent unstructured-data, the authors specified how the famous companies could improve their business. Eg: Google, Twitter and Facebook show their attention in processing big data within cloud-environment

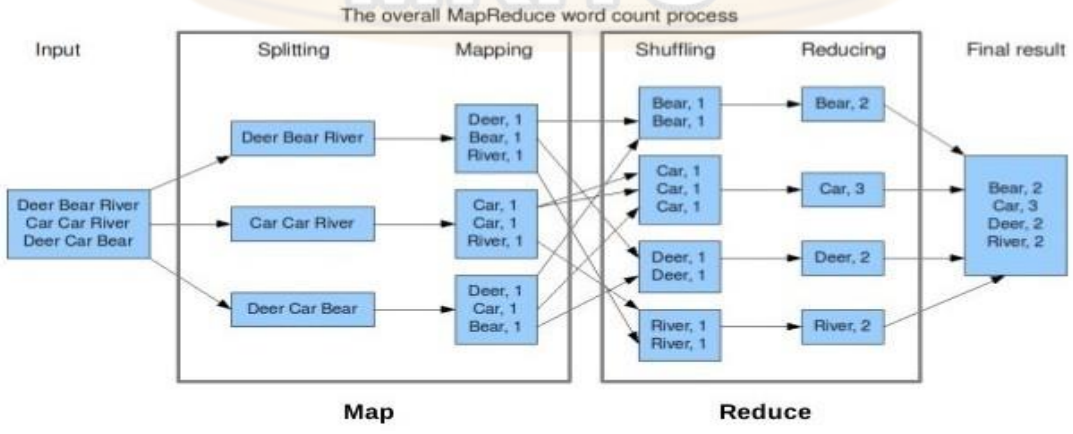


Fig 2.4: Working of Hadoop – With Map Reduce Concept

The **Map()** step: Each worker node applies the Map() function to the local data and writes the output to atemporary storage space. The Map() code is run exactly once for each **K1** key value, generating output that isorganized by key values **K2**. A master node arranges it so that for redundant copies of input data only one isprocessed.

The **Shuffle ()**step: The map output is sent to the reduce processors, which assign the **K2** key value that eachprocessor should work on, and provide that processor with all of the map-generated data associated with that keyvalue, such that all data belonging to one key are located on the same worker node.

The **Reduce()** step: Worker nodes process each group of output data(perkey) in parallel, executing the userprovidedReduce() code; each function is run exactly once for each **K2** key value pro-duced by the map step.

Produce the **final output**: The MapReduce system collects all of the reduce outputs and sorts them by **K2** to producethe final out-come.

Fig.2.4 shows the classical "word count problem" using the MapReduce paradigm. As shown in Fig.2.4, initially aprocess will split the data into a subset of chunks that will later be processed by the *mappers*. Once the key/values aregenerated by mappers, a shuffling process is used to mix (combine) these key values (combining the same keys in the sameworker node). Finally, the *reduce* functions are used to count the words that generate a common output as a result of thealgorithm. As a result of the execution or wrappers/reducers, the out-put will generate a sorted list of word counts from theoriginal text input.

2.3.2 The Employment of Big Data Analytics on IBM.

IBM and Microsoft are prominent representatives. IBM represented many big data options that enable users to storing, managing, and analyzing data through various resources; it has a good rendering on business-intelligence also healthcare areas. Compared with IBM, also Microsoft showed powerful work in the area of cloud computing activities and techniques another example is Face-book and Twitter, who are collecting various data from user's profiles and using it to increase their revenue

2.3.3 The Performance of Data Driven Companies.

Big data analytics and Business intelligence are united fields which became widely significant in the business and academic area, companies are permanently trying to make insight from the extending the three V's (variety, volume and velocity) to support decision making

2.4 Databases

Database *is* an organized collection of structured information, or data, typically stored electronically in a computer system. A database is usually controlled by

a database management system (DBMS)

The database can be divided into various categories such as text databases, desktop database programs, relational database management systems (RDMS), and NoSQL and object-oriented databases

A **text database** is a system that maintains a (usually large) text collection and provides fast and accurate access to it. **Eg: Text book, magazine, journals, manuals, etc..**

A **desktop database** is a database system that is made to run on a single computer or PC. These simpler solutions for data storage are much more limited and constrained than larger data center or data warehouse systems, where primitive database software is replaced by sophisticated hardware and networking setups. **Eg: Microsoft excel, open access, etc**.

A **relational database** (RDB) is a collective set of multiple data sets organized by tables, records and columns. RDBs establish a well-defined relationship between database tables. Tables communicate and share information, which facilitates data searchability, organization and reporting. **Eg: sql, oracle,Db2, DbaaS etc**

NoSQL databases are non-tabular, and store data differently than relational tables. ***NoSQL databases*** come in a variety of types based on their data model. The main types are document, key-value, wide-column, and graph. Eg: JSON,Mango DB,CouchDB etc

O**bject-oriented databases (OODB)** are databases that represent data in the form of objects and classes. In object-oriented terminology, an object is a real-world entity, and a class is a collection of objects. Object-oriented databases follow the fundamental principles of object-oriented programming (OOP). Eg: c++, java, c#, small talk, LISP etc..

2.5 Types of Data and variables

In any database we will be working with data to perform any kind of analysis and predication. In relational data base management system we normally use rows to represent data and columns to represent the attribute.

In terms of big data we represent the columns from RDMS as an attribute or a variable. This variable can be divided in to two types' **categorical data** or **qualitative data** and **continuous or discrete data** called as **quantitative data**. As shown below in figure 2.5.

Qualitative data or Categorical data is normally represented as variable that holds characters. And this is divided in to two types' **nominal data and ordinal data.**

In**Nominal Data** there is no natural ordering in values in the attribute of the dataset. Eg: color, Gender, nouns (name, place, animal, thing). These categories cannot be predefined with order for example there is no specific way to arrange gender of 50 students in a class. In this case the first student can be male or female similarly for all 50 students. So ordering

cannot be valid.

In **Ordinal Data** there isnatural ordering in values in the attribute of the dataset. Eg: size (S, M, L, XL, XXL), rating (excellent, good, better, worst). In the above example we can quantify the amount of data after performing ordering which gives valuable insights into the data.

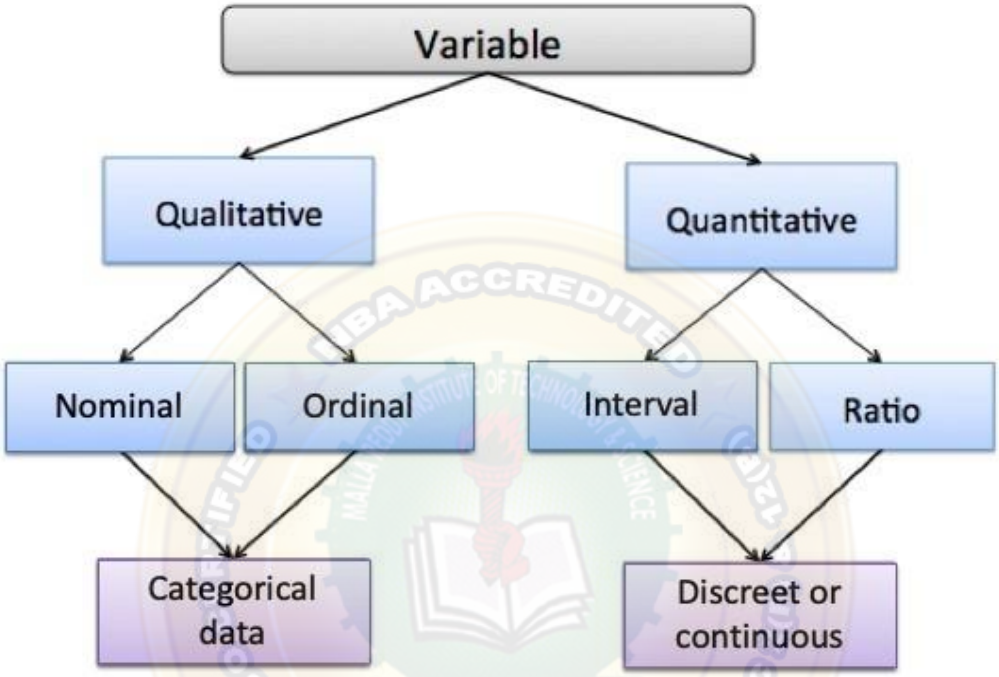


Fig 2.5: Types of Data Variables

Quantitative data or (discrete or continuous data) can be further divided in to two types' **discrete attribute and continuous attribute**.

Discrete Attribute which takes only finite number of numerical values (integers). Eg: number of buttons, no of days for product delivery etc.. These data can be represented at every specific interval in case of time series data mining or even in ratio based entries.

Continuous Attribute which takes finite number of fractional values. Eg: price, discount, height, weight, length, temperature, speed etc….. These data can be represented at every specific interval in case of time series data mining or even in ratio based entries.

2.5 Data Modelling Techniques

Data modelling is nothing but a process through which data is stored structurally in a format in a database. Data modelling is important because it enables organizations to make data-driven decisions and meet varied business goals.

The entire process of data modelling is not as easy as it seems, though. You are required to have a deeper understanding of the structure of an organization and then propose

a solution that aligns with its end-goals and suffices it in achieving the desired objectives.

Types of Data Models

Data modeling can be achieved in various ways. However, the basic concept of each of them remains the same. Let's have a look at the commonly used data modeling methods:

Hierarchical model

As the name indicates, this data model makes use of hierarchy to structure the data in a tree-like format as shown in figure 2.6. However, retrieving and accessing data is difficult in a hierarchical database. This is why it is rarely used now.

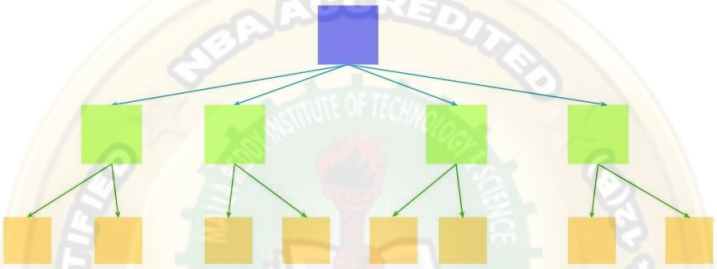


Fig 2.6: Hierarchical Model Structure

Relational model

Proposed as an alternative to hierarchical model by an IBM researcher, here data is represented in the form of tables. It reduces the complexity and provides a clear overview of the data as shown below in figure 2.7.

ID	First Name	Last Name
581-8463	Yan	Smith
962-6743	Marie	Johnston
826-272	Geoff	Lutter

Plan ID	Plan Provider
98374578	Provider A
82638367	Provider B
19274021	Provider C

ID	Plan ID	Type	Date
581-8463	98374578	R-5	12/04/2019
962-6743	82638367	M-9	09/08/2019
826-272	19274021	L-4	11/10/2019

Fig 2.7: Relational Model Structure

Network model

The network model is inspired by the hierarchical model. However, unlike the hierarchical model, this model makes it easier to convey complex relationships as each record can be linked with multiple parent records as shown in figure 2.8. In this model data can be shared easily and the computation becomes easier.

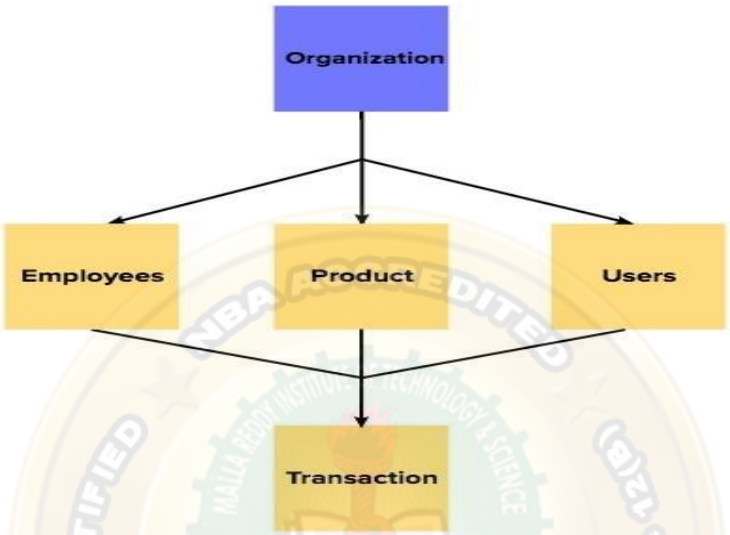


Fig 2.8: Network Model Structure

Object-oriented model

This database model consists of a collection of objects, each with its own features and methods. This type of database model is also called the post-relational database model as shown in figure 2.8.

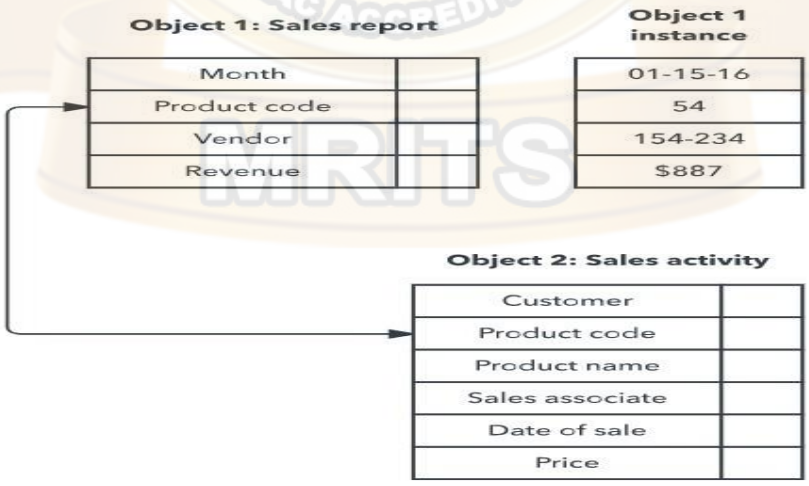


Fig 2.9: Object-Oriented Model Structure

Entity-relationship model

Entity-relationship model, also known as ER model, represents entities and their relationships in a graphical format. An entity could be anything – a concept, a piece of data, or an object.

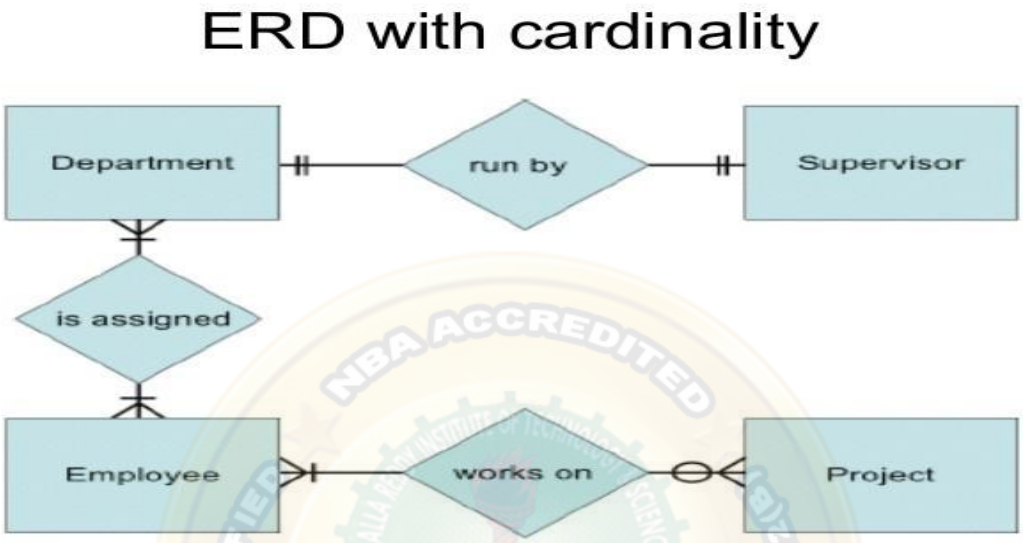


Fig 2.10: Entity Relationship Diagram

The entity relationship diagram explains relation between variables and with their primary key and foreign key as shown in figure 2.10. along with this it also explains the multiple instances of relation between tables.

Now that we have a basic understanding of data modeling, let's see why it is important.

Importance of Data Modeling

- A clear representation of data makes it easier to analyze the data properly. It provides a quick overview of the data which can then be used by the developers in varied applications.
- Data modeling represents the data properly in a model. It rules out any chances of data redundancy and omission. This helps in clear analysis and processing.
- Data modeling improves data quality and enables the concerned stakeholders to make data-driven decisions.

Since a lot of business processes depend on successful data modeling, it is necessary to adopt the right data modeling techniques for the best results.

Best Data Modeling Practices to Drive Your Key Business Decisions

Have a clear understanding of your end-goals and results

You will agree with us that the main goal behind data modeling is to equip your business and contribute to its functioning. As a data modeler, you can achieve this objective only when you know the needs of your enterprise correctly.
It is essential to make yourself familiar with the varied needs of your business so that you can prioritize and discard the data depending on the situation.

Key takeaway: Have a clear understanding of your organization's requirements and organize your data properly.

Keep it sweet and simple and scale as you grow

Things will be sweet initially, but they can become complex in no time. This is why it is highly recommended to keep your data models small and simple, to begin with.

Once you are sure of your initial models in terms of accuracy, you can gradually introduce more datasets. This helps you in two ways. First, you are able to spot any inconsistencies in the initial stages. Second, you can eliminate them on the go.

Key takeaway: Keep your data models simple. The best data modeling practice here is to use a tool which can start small and scale up as needed.

Organize your data based on facts, dimensions, filters, and order

You can find answers to most business questions by organizing your data in terms of four elements – facts, dimensions, filters, and order.

Let's understand this better with the help of an example. Let's assume that you run four e-commerce stores in four different locations of the world. It is the year-end, and you want to analyze which e-commerce store made the most sales.

In such a scenario, you can organize your data over the last year. Facts will be the overall sales data of last 1 year, the dimensions will be store location, the filter will be last 12 months, and the order will be the top stores in decreasing order.

This way, you can organize all your data properly and position yourself to answer an array of business intelligence questions without breaking a sweat.

Key takeaway: It is highly recommended to organize your data properly using individual tables for facts and dimensions to enable quick analysis.

Keep as much as is needed

While you might be tempted to keep all the data with you, do not ever fall for this trap! Although storage is not a problem in this digital age, you might end up taking a toll over your

machines' performance.

More often than not, just a small yet useful amount of data is enough to answer all the business-related questions. Spending huge on hosting enormous data of data only leads to performance issues, sooner or later.

Key takeaway: Have a clear opinion on how much datasets you want to keep. Maintaining more than what is actually required wastes your data modeling, and leads to performance issues.

Keep crosschecking before continuing

Data modeling is a big project, especially when you are dealing with huge amounts of data. Thus, you need to be cautious enough. Keep checking your data model before continuing to the next step.

For example, if you need to choose a primary key to identify each record in the dataset properly, make sure that you are picking the right attribute. Product ID could be one such attribute. Thus, even if two counts match, their product ID can help you in distinguishing each record. Keep checking if you are on the right track. Are product IDs same too? In those aces, you will need to look for another dataset to establish the relationship.
Key takeaway: It is the best practice to maintain one-to-one or one-to-many relationships. The many-to-many relationship only introduces complexity in the system.

Let them evolve
Data models are never written in stone. As your business evolves, it is essential to customize your data modeling accordingly. Thus, it is essential that you keep them updating over time. The best practice here is to store your data models in as easy-to-manage repository such that you can make easy adjustments on the go.

Key takeaway: Data models become outdated quicker than you expect. It is necessary that you keep them updated from time to time.

The Wrap Up

Data modeling plays a crucial role in the growth of businesses, especially when you organizations to base your decisions on facts and figures. To achieve the varied business intelligence insights and goals, it is recommended to model your data correctly and use appropriate tools to ensure the simplicity of the system.

2.6 Missing Imputations

In statistics, **imputation** is the process of replacing **missing data** with substituted **values**. ... Because **missing data** can create problems for analyzing **data**, **imputation** is seen as a way

to avoid pitfalls involved with list-wise deletion of cases that have **missing values**.

I. Do nothing to missing data
II. Fill the missing values in the dataset using mean, median.

Eg: for sample dataset given below

SNo	Column 1	Column2	Column 3
1	3	6	NAN
2	5	10	12
3	6	11	15
4	NAN	12	14
5	6	NAN	NAN
6	10	13	16

Can be replaced as using column mean as follows

SNo	Column 1	Column2	Column 3
1	3	6	9.5
2	5	10	12
3	6	11	15
4	5	12	14
5	6	8.66	9.5
6	10	13	16

Advantages:

- Works well with numerical dataset.
- Very fast and reliable.

Disadvantage:

- Does not work with categorical attributes
- Does not correlate relation between columns
- Not very accurate.
- Does not account for any uncertainty in data

III. Imputations using (most frequent) or (zero / constant) values

This can be used for categorical attributes.

Disadvantage:

- Does not correlate relation between columns
- Creates bias in data.

IV. Imputation using KNN

It creates a basic mean impute then uses the resulting complete list to construct a KDTree. Then, it uses the resulting KDTree to compute nearest neighbours (NN). After it finds the k-NNs, it takes the weighted average of them.

The *k* nearest neighbours is an algorithm that is used for simple classification. The algorithm uses '**feature similarity**' to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. This can be very useful in making predictions about the missing values by finding the *k's* closest neighbours to the observation with missing data and then imputing them based on the non-missing values in the neighbourhood.

Advantage:

- This method is very accurate than mean, median and mode

Disadvantage:

- Sensitive to outliers

UNIT-3
BLUE Property assumptions

- The Gauss Markov theorem tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the Best Linear Unbiased Estimate (BLUE) possible.

- There are five Gauss Markov assumptions (also called conditions):

- **Linearity**:
  - The parameters we are estimating using the OLS method must be themselves linear.
- **Random**:
  - Our data must have been randomly sampled from the population.
- **Non-Collinearity:**
  - The regressors being calculated aren't perfectly correlated with each other.
- **Exogeneity**:
  - The regressors aren't correlated with the error term.
- **Homoscedasticity:**
  - No matter what the values of our regressors might be, the error of the variance is constant.

Purpose of the Assumptions

- The Gauss Markov assumptions guarantee the validity of ordinary least squares for estimating regression coefficients.

- Checking how well our data matches these assumptions is an important part of estimating regression coefficients.

- When you know where these conditions are violated, you may be able to plan ways to change your experiment setup to help your situation fit the ideal Gauss Markov situation more closely.

- In practice, the Gauss Markov assumptions are rarely all met perfectly, but they are still useful as a benchmark, and because they show us what 'ideal' conditions would be.

- They also allow us to pinpoint problem areas that might cause our estimated regression coefficients to be inaccurate or even unusable.

The Gauss-Markov Assumptions in Algebra

- We can summarize the Gauss-Markov Assumptions succinctly in algebra, by saying that a linear regression model represented by

$$y_i = x_i' \beta + \varepsilon_i$$

- and generated by the ordinary least squares estimate is the best linear unbiased estimate (BLUE) possible if
  - $E\{\varepsilon_i\} = 0, i = 1, \ldots, N$
  - $\{\varepsilon_1 \ldots\ldots \varepsilon_n\}$ and $\{x_1 \ldots\ldots, x_N\}$ are independent
  - $cov\{\varepsilon_i, \varepsilon_j\} = 0, i, j = 1, \ldots., N \; I \neq j.$
  - $V\{\varepsilon_1 = \sigma^2, i = 1, \ldots.N$

- The first of these assumptions can be read as "The expected value of the error term is zero.". The second assumption is collinearity, the third is exogeneity, and the fourth is homoscedasticity.

Regression Concepts

Regression

- It is a Predictive modeling technique where the target variable to be estimated is continuous.

Examples of applications of regression

- predicting a stock market index using other economic indicators
- forecasting the amount of precipitation in a region based on characteristics of the jet stream
- projecting the total sales of a company based on the amount spent for advertising
- estimating the age of a fossil according to the amount of carbon-14 left in the organic material.

- Let D denote a data set that contains N observations,

$$D = \{(\mathbf{x}_i, y_i) |\ i = 1, 2, \ldots, N\}.$$

- Each x_i corresponds to the set of attributes of the i_{th} observation (known as explanatory variables) and y_i corresponds to the target (or response) variable.
- The explanatory attributes of a regression task can be either discrete or continuous.

Regression (Definition)

- Regression is the task of learning a target function *f* that maps each attribute set *x* into a continuous-valued output *y*.

The goal of regression

- To find a target function that can fit the input data with minimum error.
- The error function for a regression task can be expressed in terms of the sum of absolute or squared error:

$$\text{Absolute Error} = \sum_i |y_i - f(\mathbf{x}_i)| \qquad \text{(D.1)}$$

$$\text{Squared Error} = \sum_i (y_i - f(\mathbf{x}_i))^2 \qquad \text{(D.2)}$$

Simple Linear Regression

- Consider the physiological data shown in Figure D.1.
- The data corresponds to measurements of heat flux and skin temperature of a person during sleep.
- Suppose we are interested in predicting the skin temperature of a person based on the heat flux measurements generated by a heat sensor.
- The two-dimensional scatter plot shows that there is a strong linear relationship between the two variables.

Heat Flux	Skin Temperature
10.858	31.002
10.617	31.021
10.183	31.058
9.7003	31.095
9.652	31.133
10.086	31.188
9.459	31.226
8.3972	31.263
7.6251	31.319
7.1907	31.356
7.046	31.412
6.9494	31.468
6.7081	31.524

Heat Flux	Skin Temperature
6.3221	31.581
6.0325	31.618
5.7429	31.674
5.5016	31.712
5.2603	31.768
5.1638	31.825
5.0673	31.862
4.9708	31.919
4.8743	31.975
4.7777	32.013
4.7295	32.07
4.633	32.126
4.4882	32.164

Heat Flux	Skin Temperature
4.3917	32.221
4.2951	32.259
4.2469	32.296
4.0056	32.334
3.716	32.391
3.523	32.448
3.4265	32.505
3.3782	32.543
3.4265	32.6
3.3782	32.657
3.3299	32.696
3.3299	32.753
3.4265	32.791

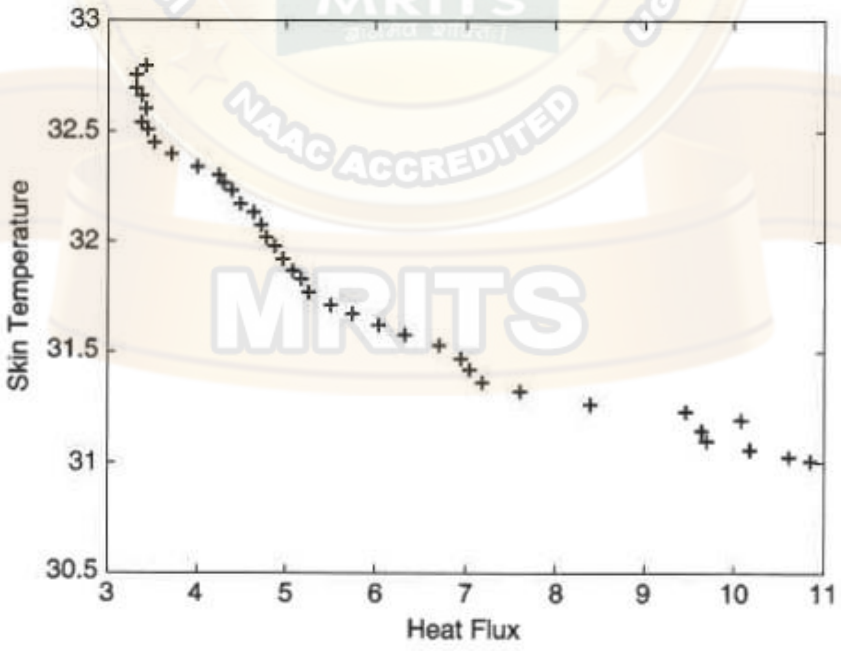


Figure D.1. Measurements of heat flux and skin temperature of a person.

Least Square Estimation or Least Square Method

- Suppose we wish to fit the following linear model to the observed data:

$$f(x) = \omega_1 x + \omega_0, \qquad (D.3)$$

- where w0 and w1 are parameters of the model and are called the regression coefficients.
- A standard approach for doing this is to apply the method of least squares, which attempts to find the parameters (w0,w1) that minimize the sum of the squared error

$$SSE = \sum_{i=1}^{N} [y_i - f(x_i)]^2 = \sum_{i=1}^{N} [y_i - \omega_1 x - \omega_0]^2, \qquad (D.4)$$

- which is also known as the residual sum of squares.
- This optimization problem can be solved by taking the partial derivative of E with respect to w0 and w1, setting them to zero, and solving the corresponding system of linear equations.

$$\begin{aligned} \frac{\partial E}{\partial \omega_0} &= -2 \sum_{i=1}^{N} [y_i - \omega_1 x_i - \omega_0] = 0 \\ \frac{\partial E}{\partial \omega_1} &= -2 \sum_{i=1}^{N} [y_i - \omega_1 x_i - \omega_0] x_i = 0 \end{aligned} \qquad (D.5)$$

- These equations can be summarized by the following matrix equation' which is also known as the normal equation:

$$\begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \begin{pmatrix} \omega_0 \\ \omega_1 \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}. \qquad (D.6)$$

- Since

$$\sum_i x_i = 229.9, \sum_i x_i^2 = 1569.2, \sum_i y_i = 1242.9, \text{ and } \sum_i x_i y_i = 7279.7,$$

- the normal equations can be solved to obtain the following estimates for the parameters.

$$\begin{pmatrix} \hat{\omega}_0 \\ \hat{\omega}_1 \end{pmatrix} = \begin{pmatrix} 39 & 229.9 \\ 229.9 & 1569.2 \end{pmatrix}^{-1} \begin{pmatrix} 1242.9 \\ 7279.7 \end{pmatrix}$$
$$= \begin{pmatrix} 0.1881 & -0.0276 \\ -0.0276 & 0.0047 \end{pmatrix} \begin{pmatrix} 1242.9 \\ 7279.7 \end{pmatrix}$$
$$= \begin{pmatrix} 33.1699 \\ -0.2208 \end{pmatrix}$$

- Thus, the linear model that best fits the data in terms of minimizing the SSE is

$$f(x) = 33.17 - 0.22x.$$

- Figure D.2 shows the line corresponding to this model.

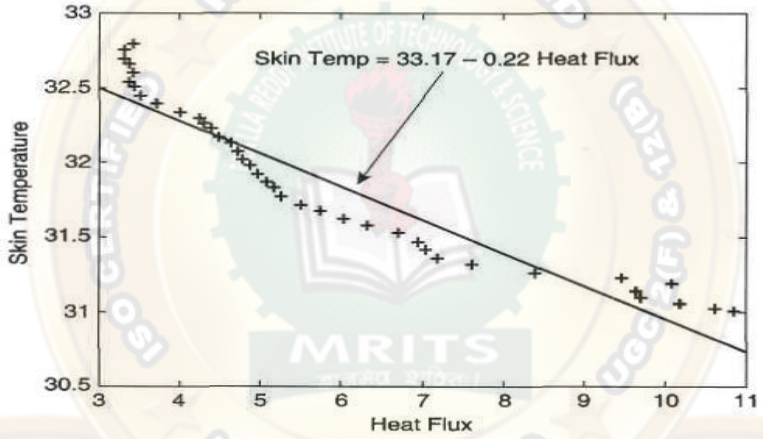


Figure D.2. A linear model that fits the data given in Figure D.1.

- We can show that the general solution to the normal equations given in D.6 can be expressed as follow:

$$\hat{\omega}_0 = \overline{y} - \hat{\omega}_1 \overline{x}$$
$$\hat{\omega}_1 = \frac{\sigma_{xy}}{\sigma_{xx}} \quad \text{(D.7)}$$

where $\overline{x} = \sum_i x_i/N$, $\overline{y} = \sum_i y_i/N$, and

$$\sigma_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y}) \quad \text{(D.8)}$$
$$\sigma_{xx} = \sum_i (x_i - \overline{x})^2 \quad \text{(D.9)}$$
$$\sigma_{yy} = \sum_i (y_i - \overline{y})^2 \quad \text{(D.10)}$$

- Thus, linear model that results in the minimum squared error is given by

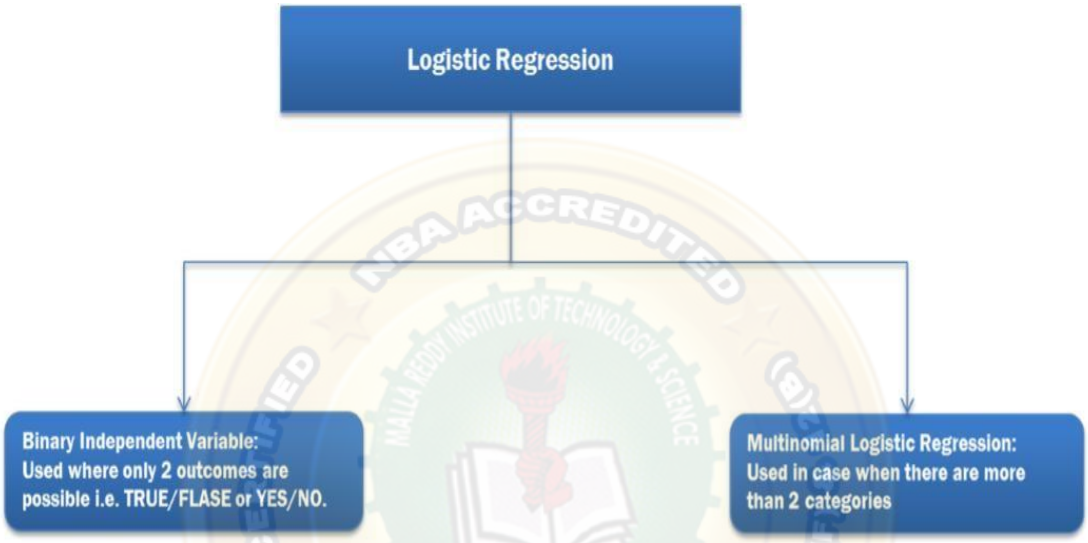
$$f(x) = \overline{y} + \frac{\sigma_{xy}}{\sigma_{xx}}[x - \overline{x}]. \qquad \text{(D.11)}$$

- In summary, the least squares method is a systematic approach to fit a linear model to the response variable g by minimizing the squared error between the true and estimated value of g.
- Although the model is relatively simple, it seems to provide a reasonably accurate approximation because a linear model is the first-order Taylor series approximation for any function with continuous derivatives.

Logistic Regression

Logistic regression, or Logit regression, or Logit model

- is a regression model where the dependent variable (DV) is categorical.
- was developed by statistician David Cox in 1958.



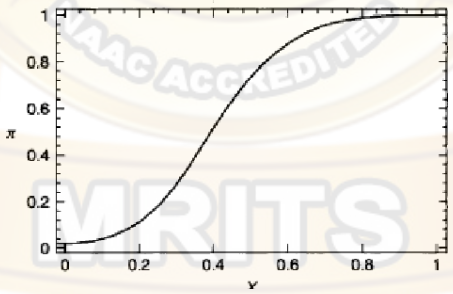
- The response variable Y has been regarded as a continuous quantitative variable.
- There are situations, however, where the response variable is qualitative.
- The predictor variables, however, have been both quantitative, as well as qualitative.
- Indicator variables fall into the second category.

- Consider a procedure in which individuals are selected on the basis of their scores in a battery of tests.
- After five years the candidates are classified as "good" or "poor."
- We are interested in examining the ability of the tests to predict the job performance of the candidates.
- Here the response variable, performance, is dichotomous.
- We can code "good" as 1 and "poor" as 0, for example.
- The predictor variables are the scores in the tests.

- In a study to determine the risk factors for cancer, health records of several people were studied.
- Data were collected on several variables, such as age, gender, smoking, diet, and the family's medical history.
- The response variable was the person had cancer (Y = 1) or did not have cancer (Y = 0).

- The relationship between the probability π and X can often be represented by a logistic response function.
- It resembles an S-shaped curve.
- The probability π initially increases slowly with increase in X, and then the increase accelerates, finally stabilizes, but does not increase beyond 1.
- Intuitively this makes sense.
- Consider the probability of a questionnaire being returned as a function of cash reward, or the probability of passing a test as a function of the time put in studying for it.
- The shape of the S-curve can be reproduced if we model the probabilities as follows:

$$\pi = \Pr(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}},$$



- A sigmoid function is a bounded differentiable real function that is defined for all real input values and has a positive derivative at each point.

- It has an “S” shape. It is defined by below function:

$$S(t) = \frac{1}{1 + e^{-t}}.$$

- The process of linearization of logistic regression function is called Logit Transformation.

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

- Modeling the response probabilities by the logistic distribution and estimating the parameters of the model given below constitutes fitting a logistic regression.
- In logistic regression the fitting is carried out by working with the logits.
- The Logit transformation produces a model that is linear in the parameters.
- The method of estimation used is the maximum likelihood method.
- The maximum likelihood estimates are obtained numerically, using an iterative procedure.

$$\begin{aligned}\pi &= \Pr(Y = 1|X_1 = x_1, \cdots, X_p = x_p) \\ &= \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}.\end{aligned}$$

OLS:

- The ordinary least squares, or OLS, can also be called the linear least squares.
- This is a method for approximately determining the unknown parameters located in a linear regression model.
- According to books of statistics and other online sources, the ordinary least squares is obtained by minimizing the total of squared vertical distances between the observed responses within the dataset and the responses predicted by the linear approximation.
- Through a simple formula, you can express the resulting estimator, especially the single regressor, located on the right-hand side of the linear regression model.
- For example, you have a set of equations which consists of several equations that have unknown parameters.
- You may use the ordinary least squares method because this is the most standard approach in finding the approximate solution to your overly determined systems.
- In other words, it is your overall solution in minimizing the sum of the squares of errors in your equation.
- Data fitting can be your most suited application. Online sources have stated that the data that best fits the ordinary least squares minimizes the sum of squared residuals.
- "Residual" is "the difference between an observed value and the fitted value provided by a model."

Maximum likelihood estimation, or MLE,

- is a method used in estimating the parameters of a statistical model, and for fitting a statistical model to data.
- If you want to find the height measurement of every basketball player in a specific location, you can use the maximum likelihood estimation.
- Normally, you would encounter problems such as cost and time constraints.
- If you could not afford to measure all of the basketball players' heights, the maximum likelihood estimation would be very handy.
- Using the maximum likelihood estimation, you can estimate the mean and variance of the height of your subjects.
- The MLE would set the mean and variance as parameters in determining the specific parametric values in a given model.

Multinomial Logistic Regression

- We have n independent observations with p explanatory variables.
- The qualitative response variable has k categories.
- To construct the logits in the multinomial case one of the categories is considered the base level and all the logits are constructed relative to it. Any category can be taken as the base level.
- We will take category k as the base level in our description of the method.
- Since there is no ordering, it is apparent that any category may be labeled k. Let 7rj denote the multinomial probability of an observation falling in the jth category.
- We want to find the relationship between this probability and the p explanatory variables, Xl, X 2 , ... ,Xp. The multiple logistic regression model then is

$$\ln\left(\frac{\pi_j(x_i)}{\pi_k(x_i)}\right) = \beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \cdots + \beta_{pj}x_{pi}; \quad \begin{matrix} j = 1, 2, \cdots, (k-1), \\ i = 1, 2, \cdots, n. \end{matrix}$$

- Since all the 7r'S add to unity, this reduces to

$$\ln\left(\pi_j(x_i)\right) = \frac{\exp\left(\beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \cdots + \beta_{pj}x_{pi}\right)}{1 + \sum_{j=1}^{k-1}\exp\left(\beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \cdots + \beta_{pj}x_{pi}\right)},$$

- For j = 1,2,⋯, (k - 1). The model parameters are estimated by the method of maximum likelihood. Statistical software is available to do this fitting.

UNIT-4

Regression vs. Segmentation

- **Regression analysis** focuses on finding a relationship between a dependent variable and one or more independent variables.
- Predicts the value of a dependent variable based on the value of at least one independent variable.
- Explains the impact of changes in an independent variable on the dependent variable.

- We use linear or logistic regression technique for developing accurate models for predicting an outcome of interest.
- Often, we create separate models for separate segments.
- **Segmentation methods** such as CHAID or CRT is used to judge their effectiveness
- Creating separate model for separate segments may be time consuming and not worth the effort.
- But, creating separate model for separate segments may provide higher predictive power.

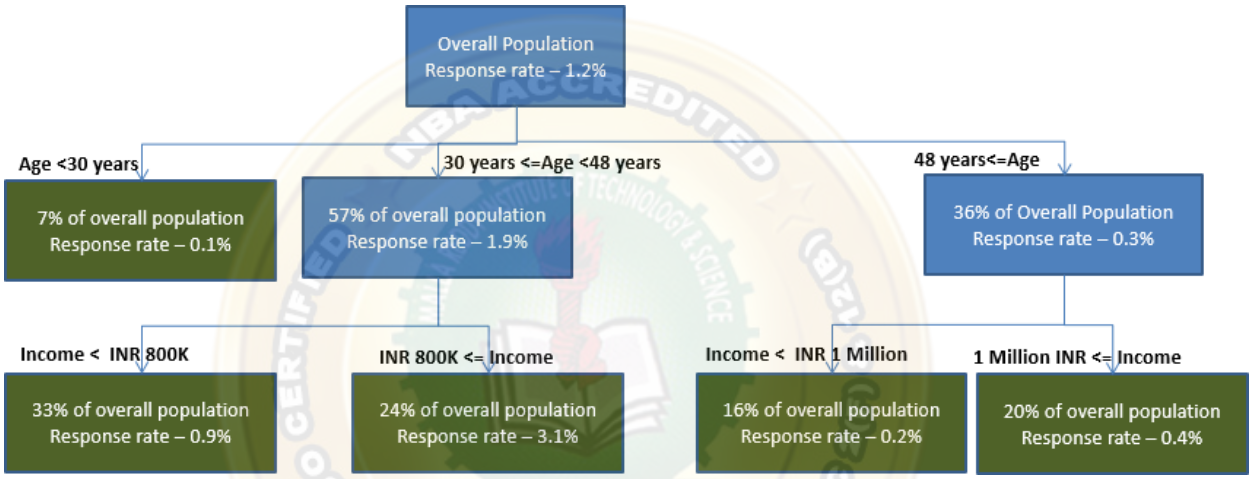
- **Market Segmentation**
  - Dividing the target market or customers on the basis of some significant features which could help a company sell more products in less marketing expenses.

  - Companies have limited marketing budgets. Yet, the marketing team is expected to makes large number of sales to ensure rising revenue & profits.
  - A product is created in two ways:
    - Create a product after analyzing (research) the needs and wants of target market – For example: Computer. Companies like Dell, IBM, Microsoft entered this market after analyzing the enormous market which this product upholds.
    - Create a product which evokes the needs & wants in target market – For example: iPhone.
  - Once the product is created, the ball shifts to the marketing team's court.
  - As mentioned above, they make use of market segmentation techniques.
  - This ensures the product is positioned to the right segment of customers with high propensity to buy.

- **How to create segments for model development?**
  - Commonly adopted methodology
  - Let us consider an example.
  - Here we'll build a logistic regression model for predicting likelihood of a customer to respond to an offer.

— A very similar approach can also be used for developing a linear regression model.

— **Logistic regression** uses 1 or 0 indicator in the historical campaign data, which indicates whether the customer has responded to the offer or not.
— Usually, one uses the target (or 'Y' known as dependent variable) that has been identified for model development to undertake an objective segmentation.
— Remember, a separate model will be built for each segment.
— A segmentation scheme which provides the maximum difference between the segments with regards to the objective is usually selected.
— Below is a simple example of this approach.



— Fig: Sample segmentation for building a logistic regression – commonly adopted methodology
— The above segmentation scheme is the best possible objective segmentation developed, because the segments demonstrate the maximum separation with regards to the objectives (i.e. response rate).

Supervised and Unsupervised Learning

There are two broad set of methodologies for segmentation:

- Objective (supervised) segmentation
- Non-Objective (unsupervised) segmentation

Objective Segmentation

- Segmentation to identify the type of customers who would respond to a particular offer.
- Segmentation to identify high spenders among customers who will use the e-commerce channel for festive shopping.
- Segmentation to identify customers who will default on their credit obligation for a loan or credit card.

Non-Objective Segmentation

- Segmentation of the customer base to understand the specific profiles which exist within the customer base so that multiple marketing actions can be personalized for each segment
- Segmentation of geographies on the basis of affluence and lifestyle of people living in each geography so that sales and distribution strategies can be formulated accordingly.
- Segmentation of web site visitors on the basis of browsing behavior to understand the level of engagement and affinity towards the brand.
- Hence, it is critical that the segments created on the basis of an objective segmentation methodology must be different with respect to the stated objective (e.g. response to an offer).
- However, in case of a non-objective methodology, the segments are different with respect to the "generic profile" of observations belonging to each segment, but not with regards to any specific outcome of interest.
- The most common techniques for building non-objective segmentation are cluster analysis, K nearest neighbor techniques etc.
- Each of these techniques uses a distance measure (e.g. Euclidian distance, Manhattan distance, Mahalanobis distance etc.)
- This is done to maximize the distance between the two segments.
- This implies maximum difference between the segments with regards to a combination of all the variables (or factors).

Tree Building

- Decision tree learning
  - is a method commonly used in data mining.
  - is the construction of a decision tree from class-labeled training tuples.

- goal
  - to create a model that predicts the value of a target variable based on several input variables.

- Decision trees used in data mining are of two main types.
  - *Classification tree analysis*
  - *Regression tree analysis*

    - Classification tree analysis is when the predicted outcome is the class to which the data belongs.
    - Regression tree analysis is when the predicted outcome can be considered a real number. (e.g. the price of a house, or a patient's length of stay in a hospital).

- A decision tree
  - is a flow-chart-like structure
  - each internal (non-leaf) node denotes a test on an attribute
  - each branch represents the outcome of a test,
  - each leaf (or terminal) node holds a class label.
  - The topmost node in a tree is the root node.

- Decision-tree algorithms:
  - ID3 (Iterative Dichotomiser 3)
  - C4.5 (successor of ID3)
  - CART (**Classification and Regression Tree**)
  - CHAID (CHI-squared Automatic Interaction Detector). Performs multi-level splits when computing **classification trees**.
  - MARS: extends decision trees to handle numerical data better. Conditional Inference Trees.

- Statistics-based approach that uses non-parametric tests as splitting criteria, corrected for multiple testing to avoid over fitting.
- This approach results in unbiased predictor selection and does not require pruning.
- ID3 and CART follow a similar approach for learning decision tree from training tuples.

CHAID (CHI-squared Automatic Interaction Detector)

- A simple method for fitting trees to predict a quantitative variable proposed by Morgan and Sonquist (1963).
- They called the method AID, for Automatic Interaction Detection.
- The algorithm performs stepwise splitting.
- It begins with a single cluster of cases and searches a candidate set of predictor variables for a way to split this cluster into two clusters.
- Each predictor is tested for splitting as follows:
    - Sort all the n cases on the predictor and examine all n-1 ways to split the cluster in two.
    - For each possible split, compute the within-cluster sum of squares about the mean of the cluster on the dependent variable.
    - Choose the best of the n-1 splits to represent the predictor's contribution. Now do this for every other predictor.
    - For the actual split, choose the predictor and its cut point which yields the smallest overall within-cluster sum of squares.
    - Categorical predictors require a different approach. Since categories are unordered, all possible splits between categories must be considered.
    - For deciding on one split of k categories into two groups, this means that 2k-1 possible splits must be considered.
    - Once a split is found, its suitability is measured on the same within-cluster sum of squares as for a quantitative predictor.
- Morgan and Sonquist called their algorithm AID because it naturally incorporates interaction among predictors. Interaction is not correlation.
- It has to do instead with conditional discrepancies.
- In the analysis of variance, interaction means that a trend within one level of a variable is not parallel to a trend within another level of the same variable.

- In the ANOVA model, interaction is represented by cross-products between predictors.
- In the tree model, it is represented by branches from the same nodes which have different splitting predictors further down the tree.

No interaction (left) and interaction (right) trees.

- Regression trees parallel regression/ANOVA modeling in which the dependent variable is quantitative.
- Classification trees parallel discriminant analysis and algebraic classification methods.
- Kass (1980) proposed a modification to AID called CHAID for categorized dependent and independent variables.
- His algorithm incorporated a sequential merge and split procedure based on a chi-square test statistic.
- Kass was concerned about computation time, so he decided to settle for a sub-optimal split on each predictor instead of searching for all possible combinations of the categories.
- Kass's algorithm is like sequential cross-tabulation.
    - For each predictor:
        1. cross tabulate the m categories of the predictor with the k categories of the dependent variable,
        2. find the pair of categories of the predictor whose 2xk sub-table is least significantly different on a chi-square test and merge these two categories;
        3. if the chi-square test statistic is not "significant" according to a preset critical value, repeat this merging process for the selected predictor until no non-significant chi-square is found for a sub-table, and pick the predictor variable whose chi-square is largest and split the sample into subsets, where l is the number of categories resulting from the merging process on that predictor;
        4. Continue splitting, as with AID, until no "significant" chi-squares result. The CHAID algorithm saves some computer time, but it is not guaranteed to find the splits which predict best at a given step. Only by searching all possible category subsets can we do that. CHAID is also limited to categorical predictors, so it cannot be used for quantitative or mixed categorical quantitative models.

CART (Classification And Regression Tree)

- CART algorithm was introduced in Breiman et al. (1986).
- A CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample.
- The CART growing method attempts to maximize within-node homogeneity.
- The extent to which a node does not represent a homogenous subset of cases is an indication of impurity.
- For example, a terminal node in which all cases have the same value for the dependent variable is a homogenous node that requires no further splitting because it is "pure."
- For categorical (nominal, ordinal) dependent variables the common measure of impurity is Gini, which is based on squared probabilities of membership for each category.

- Splits are found that maximize the homogeneity of child nodes with respect to the value of the dependent variable.

- Impurity Measure:

- GINI Index Used by the CART (classification and regression tree) algorithm, Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset.
- Gini impurity can be computed by summing the probability fi of each item being chosen times the probability 1-fi of a mistake in categorizing that item.
- It reaches its minimum (zero) when all cases in the node fall into a single target category.
- To compute Gini impurity for a set of items, suppose i ε {1, 2... m}, and let fi be the fraction of items labeled with value i in the set.

$$I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i) = \sum_{i=1}^{m} (f_i - f_i^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i^2 = 1 - \sum_{i=1}^{m} f_i^2 = \sum_{i \neq k} f_i f_k$$

Advantages of Decision Tree:

- *Simple to understand and interpret*. People are able to understand decision tree models after a brief explanation.
- *Requires little data preparation*. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed.
- *Able to handle both numerical and categorical data*. Other techniques are usually specialized in analysing datasets that have only one type of variable.
- *Uses a white box model*. If a given situation is observable in a model the explanation for the condition is easily explained by Boolean logic.
- *Possible to validate a model using statistical tests*. That makes it possible to account for the reliability of the model.
- *Robust.* Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.
- Performs well with large datasets. Large amounts of data can be analyzed using standard computing resources in reasonable time.

Tools used to make Decision Tree:

- Many data mining software packages provide implementations of one or more decision tree algorithms.
- Several examples include:
  - Salford Systems CART
  - IBM SPSS Modeler
  - Rapid Miner
  - SAS Enterprise Miner
  - Matlab
  - R (an open source software environment for statistical computing which includes several CART implementations such as rpart, party and random Forest packages)
  - Weka (a free and open-source data mining suite, contains many decision tree algorithms)
  - Orange (a free data mining software suite, which includes the tree module orngTree)
  - KNIME
  - Microsoft SQL Server
  - Scikit-learn (a free and open-source machine learning library for the Python programming language).

- Pruning
- After building the decision tree, a tree-pruning step can be performed to reduce the size of the decision tree.
- Pruning helps by trimming the branches of the initial tree in a way that improves the generalization capability of the decision tree.

- The errors committed by a classification model are generally divided into two types:
  - training errors
  - generalization errors.

- Training error
  - also known as resubstitution error or apparent error.
  - it is the number of misclassification errors committed on training records.
- generalization error
  - is the expected error of the model on previously unseen records.
  - A good classification model must not only fit the training data well, it must also accurately classify records it has never seen before.
- A good model must have low training error as well as low generalization error.

- Model overfitting
  - Decision trees that are too large are susceptible to a phenomenon known as overfitting.
  - A model that fits the training data too well can have a poorer generalization error than a model with a higher training error.
  - Such a situation is known as model overfitting.

- Model underfitting
  - The training and test error rates of the model are large when the size of the tree is very small.
  - This situation is known as model underfitting.
  - Underfitting occurs because the model has yet to learn the true structure of the data.

- Model complexity
  - To understand the overfitting phenomenon, the training error of a model can be reduced by increasing the model complexity.
  - Overfitting and underfitting are two pathologies that are related to the model complexity.

ARIMA (Autoregressive Integrated Moving Average)

- ARIMA model is a generalization of an autoregressive moving average (ARMA) model, in time series analysis,
- These models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting).
- They are applied in some cases where data show evidence of non-stationary, wherein initial differencing step (corresponding to the "integrated" part of the model) can be applied to reduce the non-stationary.

- *Non-seasonal ARIMA models*
    - These are generally denoted ARIMA(p, d, q) where parameters p, d, and q are non-negative integers, p is the order of the Autoregressive model, d is the degree of differencing, and q is the order of the Moving-average model.

- *Seasonal ARIMA models*
    - These are usually denoted ARIMA(p, d, q)(P, D, Q)_m, where m refers to the number of periods in each season, and the uppercase P, D, Q refer to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model.

- ARIMA models form an important part of the Box-Jenkins approach to time-series modeling.

- *Applications*
    - ARIMA models are important for generating forecasts and providing understanding in all kinds of time series problems from economics to health care applications.
    - In quality and reliability, they are important in process monitoring if observations are correlated.
    - designing schemes for process adjustment
    - monitoring a reliability system over time
    - forecasting time series
    - estimating missing values
    - finding outliers and atypical events
    - understanding the effects of changes in a system

Measure of Forecast Accuracy

- Forecast Accuracy can be defined as the deviation of Forecast or Prediction from the actual results.

Error = Actual demand – Forecast

OR

ei = At – Ft

- We measure Forecast Accuracy by 2 methods :
  - *Mean Forecast Error (MFE)*
    - For n time periods where we have actual demand and forecast values:

$$MFE = \frac{\sum_{i=1}^{n}(e_i)}{n}$$

    - Ideal value = 0;
    - MFE > 0, model tends to under-forecast
    - MFE < 0, model tends to over-forecast

  - *Mean Absolute Deviation (MAD)*
    - For n time periods where we have actual demand and forecast values:

$$MAD = \frac{\sum_{i=1}^{n}|e_i|}{n}$$

- While MFE is a measure of forecast model bias, MAD indicates the absolute size of the errors

Uses of Forecast error:

- Forecast model bias
- Absolute size of the forecast errors
- Compare alternative forecasting models
- Identify forecast models that need adjustment

ETL Approach

- Extract, Transform and Load (ETL) refers to a process in database usage and especially in data warehousing that:
  - Extracts data from homogeneous or heterogeneous data sources
  - Transforms the data for storing it in proper format or structure for querying and analysis purpose
  - Loads it into the final target (database, more specifically, operational data store, data mart, or data warehouse)
- Usually all the three phases execute in parallel since the data extraction takes time, so while the data is being pulled another transformation process executes, processing the already received data and prepares the data for loading and as soon as there is some data ready to be loaded into the target, the data loading kicks off without waiting for the completion of the previous phases.
- ETL systems commonly integrate data from multiple applications (systems), typically developed and supported by different vendors or hosted on separate computer hardware.
- The disparate systems containing the original data are frequently managed and operated by different employees.
- For example, a cost accounting system may combine data from payroll, sales, and purchasing.

- Commercially available ETL tools include:
  - Anatella
  - Alteryx
  - CampaignRunner
  - ESF Database Migration Toolkit
  - InformaticaPowerCenter
  - Talend
  - IBM InfoSphereDataStage
  - Ab Initio
  - Oracle Data Integrator (ODI)
  - Oracle Warehouse Builder (OWB)
  - Microsoft SQL Server Integration Services (SSIS)
  - Tomahawk Business Integrator by Novasoft Technologies.
  - Pentaho Data Integration (or Kettle) opensource data integration framework Stambia
  - Diyotta DI-SUITE for Modern Data Integration
  - FlyData
  - Rhino ETL
  - SAP Business Objects Data Services

- SAS Data Integration Studio
- SnapLogic
- Clover ETL opensource engine supporting only basic partial functionality and not server
- SQ-ALL - ETL with SQL queries from internet sources such as APIs
- North Concepts Data Pipeline

- Various steps involved in ETL.
  - Extract
  - Transform
  - Load

  - Extract
    - The Extract step covers the data extraction from the source system and makes it accessible for further processing.
    - The main objective of the extract step is to retrieve all the required data from the source system with as little resources as possible.
    - The extract step should be designed in a way that it does not negatively affect the source system in terms or performance, response time or any kind of locking.

    - There are several ways to perform the extract:
      - Update notification - if the source system is able to provide a notification that a record has been changed and describe the change, this is the easiest way to get the data.
      - Incremental extract - some systems may not be able to provide notification that an update has occurred, but they are able to identify which records have been modified and provide an extract of such records. During further ETL steps, the system needs to identify changes and propagate it down. Note, that by using daily extract, we may not be able to handle deleted records properly.
      - Full extract - some systems are not able to identify which data has been changed at all, so a full extract is the only way one can get the data out of the system. The full extract requires keeping a copy of the last extract in the same format in order to be able to identify changes. Full extract handles deletions as well.
      - When using Incremental or Full extracts, the extract frequency is extremely important. Particularly for full extracts; the data volumes can be in tens of gigabytes.

- Clean - The cleaning step is one of the most important as it ensures the quality of the data in the data warehouse. Cleaning should perform basic data unification rules, such as:
- Making identifiers unique (sex categories Male/Female/Unknown, M/F/null, Man/Woman/Not Available are translated to standard Male/Female/Unknown)
- Convert null values into standardized Not Available/Not Provided value
- Convert phone numbers, ZIP codes to a standardized form
- Validate address fields, convert them into proper naming, e.g. Street/St/St./Str./Str
- Validate address fields against each other (State/Country, City/State, City/ZIP code, City/Street).

- Transform
  - The transform step applies a set of rules to transform the data from the source to the target.
  - This includes converting any measured data to the same dimension (i.e. conformed dimension) using the same units so that they can later be joined.
  - The transformation step also requires joining data from several sources, generating aggregates, generating surrogate keys, sorting, deriving new calculated values, and applying advanced validation rules.

- Load
  - During the load step, it is necessary to ensure that the load is performed correctly and with as little resources as possible.
  - The target of the Load process is often a database.
  - In order to make the load process efficient, it is helpful to disable any constraints and indexes before the load and enable them back only after the load completes.
  - The referential integrity needs to be maintained by ETL tool to ensure consistency.

- Managing ETL Process

  - The ETL process seems quite straight forward.
  - As with every application, there is a possibility that the ETL process fails.
  - This can be caused by missing extracts from one of the systems, missing values in one of the reference tables, or simply a connection or power outage.

- Therefore, it is necessary to design the ETL process keeping fail-recovery in mind.

- Staging
  - It should be possible to restart, at least, some of the phases independently from theothers.
  - For example, if the transformation step fails, it should not be necessary to restart the Extract step.
  - We can ensure this by implementing proper staging. Staging means that the data is simply dumped to the location (called the Staging Area) so that it can then be read by the next processing phase.
  - The staging area is also used during ETL process to store intermediate results of processing.
  - This is ok for the ETL process which uses for this purpose.
  - However, the staging area should be accessed by the load ETL process only.
  - It should never be available to anyone else; particularly not to end users as it isnot intended for data presentation to the end-user.
  - May contain incomplete or in-the-middle-of-the-processing data.

UNIT-5
Data Visualization

Data Visualization

- Data visualization is the art and practice of gathering, analyzing, and graphically representing empirical information.
- They are sometimes called information graphics, or even just charts and graphs.
- The goal of visualizing data is to tell the story in the data.
- Telling the story is predicated on understanding the data at a very deep level, and gathering insight from comparisons of data points in the numbers

Why data visualization?

- Gain insight into an information space by mapping data onto graphical primitives Provide qualitative overview of large data sets
- Search for patterns, trends, structure, irregularities, and relationships among data.
- Help find interesting regions and suitable parameters for further quantitative analysis.
- Provide a visual proof of computer representations derived.

Categorization of visualization methods

- Pixel-oriented visualization techniques
- Geometric projection visualization techniques
- Icon-based visualization techniques
- Hierarchical visualization techniques
- Visualizing complex data and relations

Pixel-Oriented Visualization Techniques

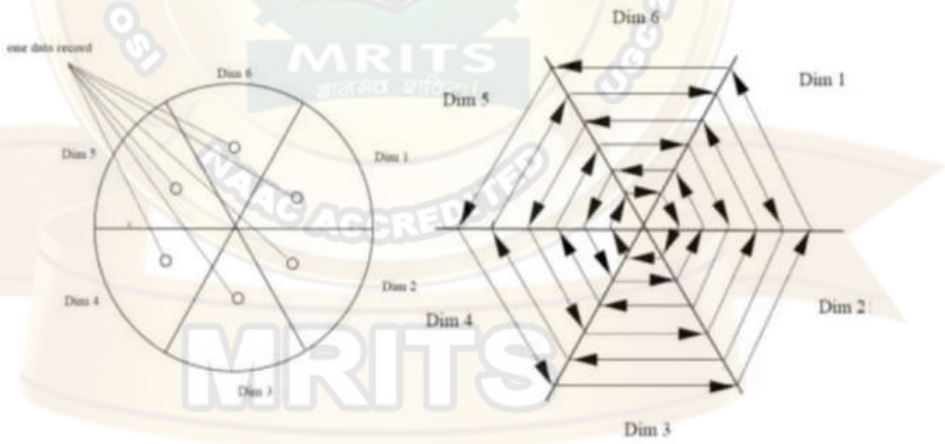
- For a data set of m dimensions, create m windows on the screen, one for each dimension**.**
- The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows**.**
- The colors of the pixels reflect the corresponding values.

(a) Income (b) Credit Limit (c) transaction volume (d) age

Pixel-Oriented Visualization Techniques

Laying Out Pixels in Circle Segments

- To save space and show the connections among multiple dimensions, space filling is often done in a circle segment.



(a) Representing a data record in circle segment (b) Laying out pixels in circle segment

Laying Out Pixels in Circle Segments

Geometric Projection Visualization Techniques

Visualization of geometric transformations and projections of the data**.**

Methods

- Direct visualization
- Scatterplot and scatterplot matrices
- Landscapes Projection pursuit technique: Help users find meaningful projections of multidimensional data
- Prosection views
- Hyperslice
- Parallel coordinates

Scatter Plots

- A scatter plot displays 2-D data points using Cartesian coordinates.
- A third dimension can be added using different colors or shapes to represent different data points
- Through this visualization, in the adjacent figure, we can see that points of types "+" and "×" tend to be colocated

Scatterplot Matrices

- The scatter-plot matrix is an extension to the scatter plot.
- For k-dimensional data a minimum of (k2-k)/2 scatterplots of 2D will be required.
- There can be maximum of k2 plots of 2D
- In the adjoining figure , there are k2 plots.
- Out of these, k are X-X plots, and all X-Y plots (where X, Y are distinct dimensions) are given in 2 orientations (X vs Y and Y vs, X)

Parallel Coordinates

- The scatter-plot matrix becomes less effective as the dimensionality increases.
- Another technique, called parallel coordinates, can handle higher dimensionality
- n equidistant axes which are parallel to one of the screen axes and correspond to the attributes (i.e. n dimensions)
- The axes are scaled to the [minimum, maximum]: range of the corresponding attribute
- Every data item corresponds to a polygonal line which intersects each of the axes at the point which corresponds to the value for the attribute

Icon-Based Visualization Techniques

- Visualization of the data values as features of icons
- Typical visualization methods
  - Chernoff Faces
  - Stick Figures
- General techniques
  - Shape coding: Use shape to represent certain information encoding
  - Color icons: Use color icons to encode more information
  - Tile bars: Use small icons to represent the relevant feature vectors in document retrieval

Chernoff Faces

- A way to display variables on a two-dimensional surface, e.g., let x be eyebrow slant, y be eye size, z be nose length, etc.
- The figure shows faces produced using 10 characteristics–head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening): Each assigned one of 10 possible values.

Chernoff Faces

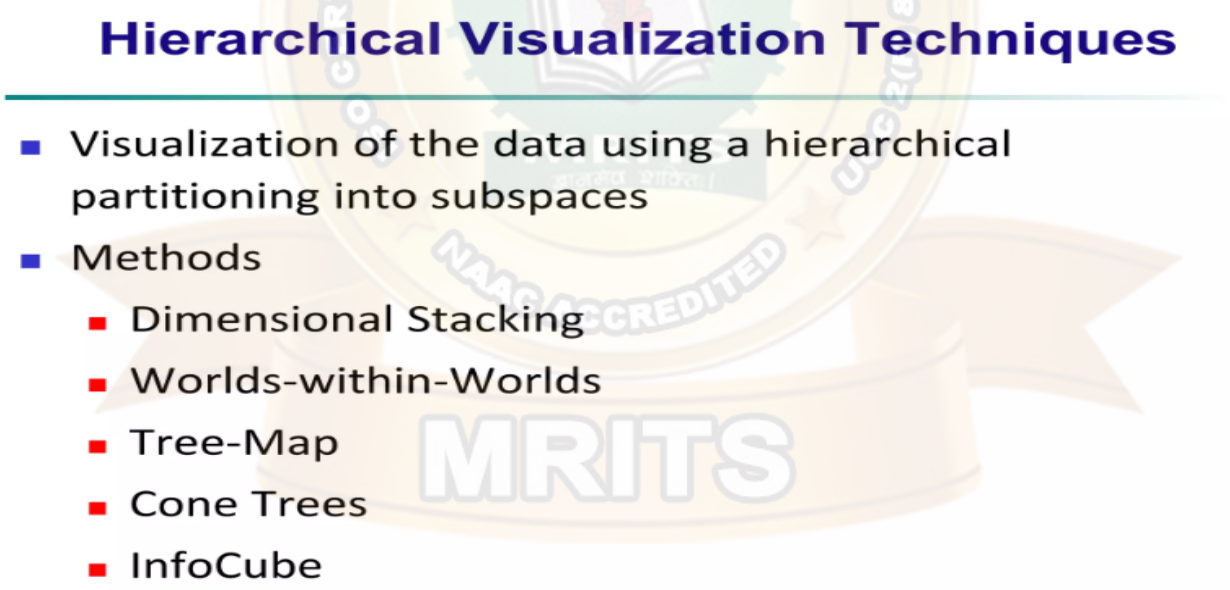
Stick Figure

Stick Figure

- A census data figure showing age, income, gender, education
- A 5-piece stick figure (1 body and 4 limbs w. different angle/length)
- Age, income are indicated by position of the figure.
- Gender, education are indicated by angle/length.
- Visualization can show a texture pattern

Hierarchical Visualization

- For a large data set of high dimensionality, it would be difficult to visualize all dimensions at the same time.
- Hierarchical visualization techniques partition all dimensions into subsets (i.e., subspaces).
- The subspaces are visualized in a hierarchical manner
- "Worlds-within-Worlds," also known as n-Vision, is a representative hierarchical visualization method.
- To visualize a 6-D data set, where the dimensions are F,X1,X2,X3,X4,X5.
- We want to observe how F changes w.r.t. other dimensions. We can fix X3,X4,X5 dimensions to selected values and visualize changes to F w.r.t. X1, X2

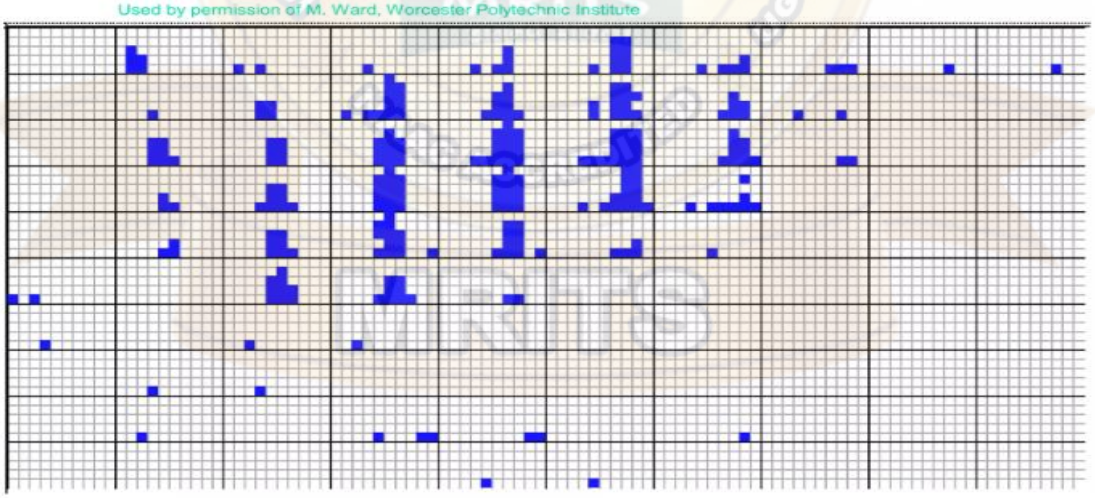


Dimensional Stacking

- Partitioning of the n-dimensional attribute space in 2-D subspaces, which are 'stacked' into each other
- Partitioning of the attribute value ranges into classes. The important attributes should be used on the outer levels.
- Adequate for data with ordinal attributes of low cardinality
- But, difficult to display more than nine dimensions
- Important to map dimensions appropriately

Dimensional Stacking

Used by permission of M. Ward, Worcester Polytechnic Institute

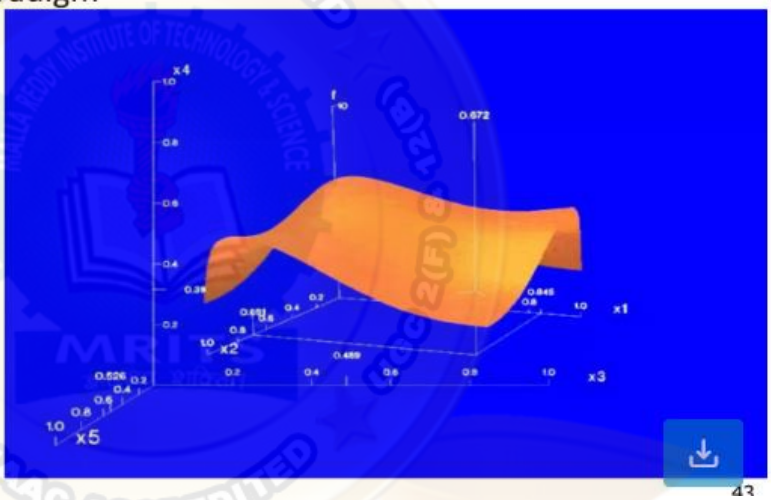


Visualization of oil mining data with longitude and latitude mapped to the outer x-, y-axes and ore grade and depth mapped to the inner x-, y-axes

Worlds-within-Worlds

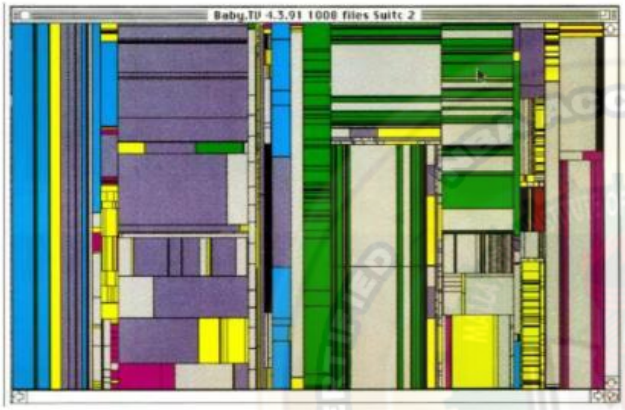
- Assign the function and two most important parameters to innermost world
- Fix all other parameters at constant values - draw other (1 or 2 or 3 dimensional worlds choosing these as the axes)
- Software that uses this paradigm

- N–vision: Dynamic interaction through data glove and stereo displays, including rotation, scaling (inner) and translation (inner/outer)
- Auto Visual: Static interaction by means of queries

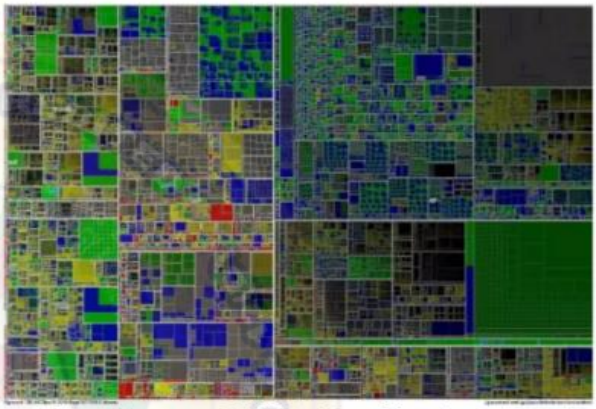


Tree-Map

- Screen-filling method which uses a hierarchical partitioning of the screen into regions depending on the attribute values
- The x- and y-dimension of the screen are partitioned alternately according to the attribute values (classes)



Schneiderman@UMD: Tree-Map of a File System



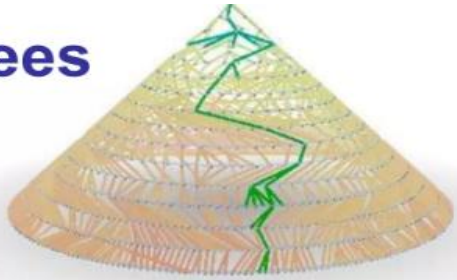
Schneiderman@UMD: Tree-Map to support large data sets of a million items

44

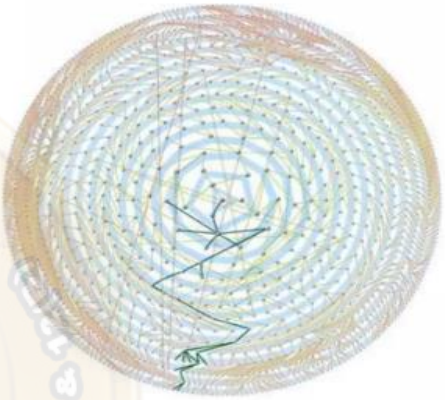
InfoCube

- A 3-D visualization technique where hierarchical information is displayed as nested semi-transparent cubes
- The outermost cubes correspond to the top level data, while the subnodes or the lower level data are represented as smaller cubes inside the outermost cubes, and so on

Three-D Cone Trees



- *3D cone tree* visualization technique works well for up to a thousand nodes or so
- First build a *2D circle tree* that arranges its nodes in concentric circles centered on the root node
- Cannot avoid overlaps when projected to 2D
- G. Robertson, J. Mackinlay, S. Card. "Cone Trees: Animated 3D Visualizations of Hierarchical Information", *ACM SIGCHI'91*
- Graph from Nadeau Software Consulting website: Visualize a social network data set that models the way an infection spreads from one person to the next



4

Visualizing Complex Data and Relations

- Most visualization techniques were mainly for numeric data.
- Recently, more and more non-numeric data, such as text and social networks, havebecome available.
- Many people on the Web tag various objects such as pictures, blog entries, and productreviews.
- A tag cloud is a visualization of statistics of user-generated tags.
- Often, in a tag cloud, tags are listed alphabetically or in a user-preferred order.
- The importance of a tag is indicated by font size or color.

Visualizing Complex Data and Relations

Visualizing Complex Data and Relations

- Visualizing non-numerical data: text and social networks
- Tag cloud: visualizing user-generated tags
  - The importance of tag is represented by font size/color
- Besides text data, there are also methods to visualize relationships, such as visualizing social networks

Newsmap: Google News Stories in 2005

Reference:

https://www.slideserve.com/eben/introduction-to-information-visualization