III B.Tech I- Semester

Data Analytics (R18) 2020-21

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.

Data Management:

Data Management is an administrative process that includes acquiring, validating, storing, protecting and processing required data to ensure the accessibility, reliability and timelines of the data for its users.

Design Data Architecture and manage the Data for analysis Data

architecture

- Data architecture is composed of
  - models,
  - policies,
  - rules or standards

  that govern which data

  - is collected,
  - how it is stored,
  - arranged,
  - integrated,
  - 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.

Constraints and influences

Various constraints and influences that will have an effect on data architecture design are

- enterprise requirements
- technology drivers
- economics
- business policies
- Data processing needs.

Enterprise requirements

- These will generally include such elements as

  - economical and effective system expansion,
  - acceptable performance levels (especially system access speed),
  - transaction reliability,
  - 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,
- This enables managerial decision making and other organizational processes.
- Architecture techniques
  - Split between managing transaction data and (master) reference data.
  - Splitting data capture systems from data retrieval systems.

Technology drivers

- These are usually suggested by
  - completed data architecture
  - database architecture designs
- In addition, some technology drivers will derive from
  - existing organizational integration frameworks and standards,
  - organizational economics,
  - 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
  - business cycle,
  - interest rates,
  - market conditions,
  - legal considerations

Could all have an effect on decisions relevant to data architecture?

Business policies

- Business policies that also drive data architecture design include
  - internal organizational policies,
  - rules of regulatory bodies,
  - professional standards,
  - 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.

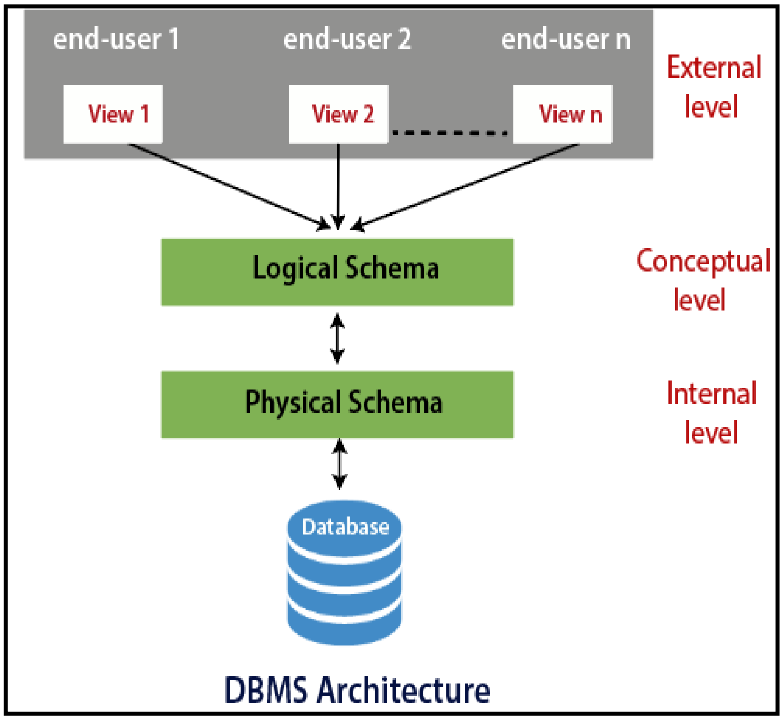
Data processing needs

- 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,
  - support of various organizational initiatives as required (i.e. annual budgets, new product development).

General Approach

The General Approach is based on designing the Architecture at three Levels of Specification:

- The Logical Level
- The Physical Level
- The Implementation Level



UNDERSTAND VARIOUS SOURCES OF THE DATA

Data can be generated from different sources.

- ✓ **Sensor**
- ✓ **Signals**
- ✓ **Global Positioning System(GPS)**
- ✓ **Social networking sites**
- ✓ **E-commerce site**
- ✓ **Weather Station**
- ✓ **Telecom company**
- ✓ **Share Market**

Sources of Big Data

These data come from many sources like

- **Sensor Data:** 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.
- The **Global Positioning System** (**GPS**) has been developed in order to allow accurate determination of geographical locations by military and civil users. It is based on the **use** of satellites in Earth orbit that transmit information which allow to measure the distance between the satellites and the user.
- **Social networking sites:** Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.
- **E-commerce site:** Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.
- **Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
- **Telecom company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.
- **Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.

3V's of Big Data

1. **Velocity:** The data is increasing at a very fast rate. It is estimated that the volume of data will double in every 2 years.
2. **Variety:** Now a days data are not stored in rows and column. Data is structured as well as unstructured. Log file, CCTV footage is unstructured data. Data which can be saved in tables are structured data like the transaction data of the bank.
3. **Volume:** The amount of data which we deal with is of very large size of Peta bytes.

Two types of sources

Data can be generated from two types of sources

- Primary
- Secondary

SOURCES OF PRIMARY DATA

- The sources of generating primary data are
  - Observation Method
  - Survey Method
  - Experimental Method

Observation method

- The observation method involves human or mechanical observation of what people actually do or what events take place during a buying or consumption situation.
- "Information is collected by observing process at work**."**
- The following are a few situations:
  - Service Stations-
    - Pose as a customer,
    - go to a service station and observe.
  - To evaluate the effectiveness of display of Dunlop Pillow Cushions-
  - In a departmental store, observer notes:-
    - How many pass by;
    - How many stopped to look at the display;
    - How many decide to buy.
  - Super Market-
    - Which is the best location in the shelf? Hidden cameras are used.
    - To determine typical sales arrangement and find out sales enthusiasm shown by various salesmen-
  - Normally this is done by an investigator using a concealed tape-recorder.
- Advantages of Observation Method
  - If the researcher observes and record events, it is not necessary to rely on the willingness and ability of respondents to report accurately.
  - The biasing effect of interviewers is either eliminated or reduced. Data collected by observation are, thus, more objective and generally more accurate.
- Disadvantages of Observation Method
  - The most limiting factor in the use of observation method is the inability to observe such things such as attitudes, motivations, customers/consumers state of mind, their buying motives and their images.
  - It also takes time for the investigator to wait for a particular action to take place.

- Personal and intimate activities, such as watching television late at night, are more easily discussed with questionnaires than they are observed.
- Cost is the final disadvantage of observation method.
- Under most circumstances, observational data are more expensive to obtain than other survey data.
- The observer has to wait doing nothing, between events to be observed.
- The unproductive time is an increased cost.

Survey Method

There are mainly 4 methods by which we can collect data through the Survey Method

- Telephonic Interview
- Personal Interview
- Mail Interview
- Electronic Interview

Telephonic Interview

- Best method for gathering quickly needed information.
- Responses are collected from the respondents by the researcher on telephone.
- Advantages of Telephonic Interview
  - It is very fast method of data collection.
  - It has the advantage over “Mail Questionnaire” of permitting the interviewer to talk to one or more persons and to clarifying his questions if they are not understood.
  - Response rate of telephone interviewing seems to be a little better than mail questionnaires
  - The quality of information is better
  - It is less costly method and there are less administration problems
- Disadvantages of Telephonic Interview
  - They can’t handle interview which need props
  - It can’t handle unstructured interview
  - It can’t be used for those questions which requires long descriptive answers
  - Respondents cannot be observed
  - People are reluctant to disclose personal information on telephone
  - People who don’t have telephone facility cannot be approached

Personal Interviewing

- It is the most versatile of the all methods. They are used when props are required along with the verbal response non-verbal responses can also be observed.

- Advantages of Personal Interview
  - The person interviewed can ask more questions and can supplement the interview with personal observation.
  - They are more flexible. Order of questions can be changed

  - Knowledge of past and future is possible.
  - In-depth research is possible.
  - Verification of data from other sources is possible.
  - The information obtained is very reliable and dependable and helps in establishing cause and effect relationship very early.
- Disadvantages of Personal Interview
  - It requires much more technical and administrative planning and supervision
  - It is more expensive
  - It is time consuming
  - The accuracy of data is influenced by the interviewer
  - A number of call banks may be required
  - Some people are not approachable

Mail Survey

- Questionnaires are sent to the respondents; they fill it up and send it back.
- Advantages of Mail Survey
  - It can reach all types of people.
  - Response rate can be improved by offering certain incentives.
- Disadvantages of Mail Survey
  - It cannot be used for unstructured study.
  - It is costly.
  - It requires established mailing list.
  - It is time consuming.
  - There is problem in case of complex questions.

Electronic Interview

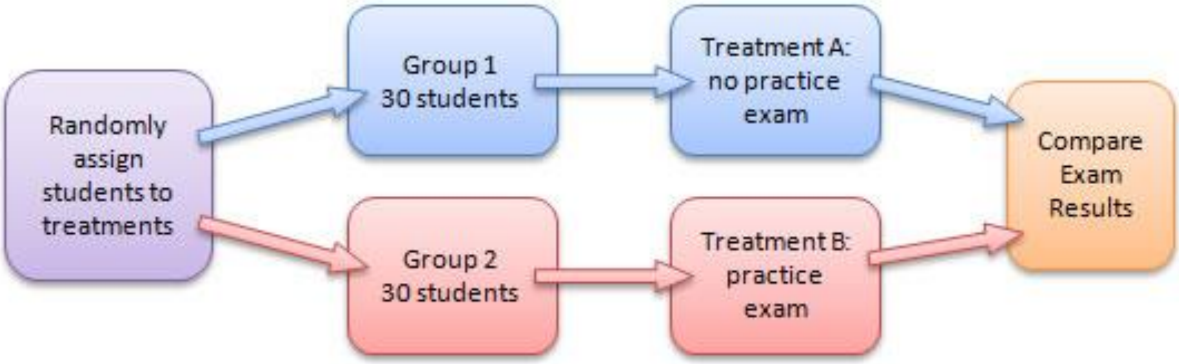
- Electronic interviewing is a process of recognizing and noting people, objects, and occurrences rather than asking for information.
- For example-When you go to store, you notice which product people like to use.
- The Universal Product Code (UPC) is also a method of observing what people are buying.
- Advantages of Electronic Interview
  - There is no relying on willingness or ability of respondent.
  - The data is more accurate and objective.
- Disadvantages of Electronic Interview
  - Attitudes cannot be observed.
  - Those events which are of long duration cannot be observed.
  - There is observer bias. It is not purely objective.
  - If the respondents know that they are being observed, their response can be biased.
  - It is a costly method.

Experimental Method

- 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
  - RBD - Randomized Block Design
  - LSD - Latin Square Design
  - FD - Factorial Designs

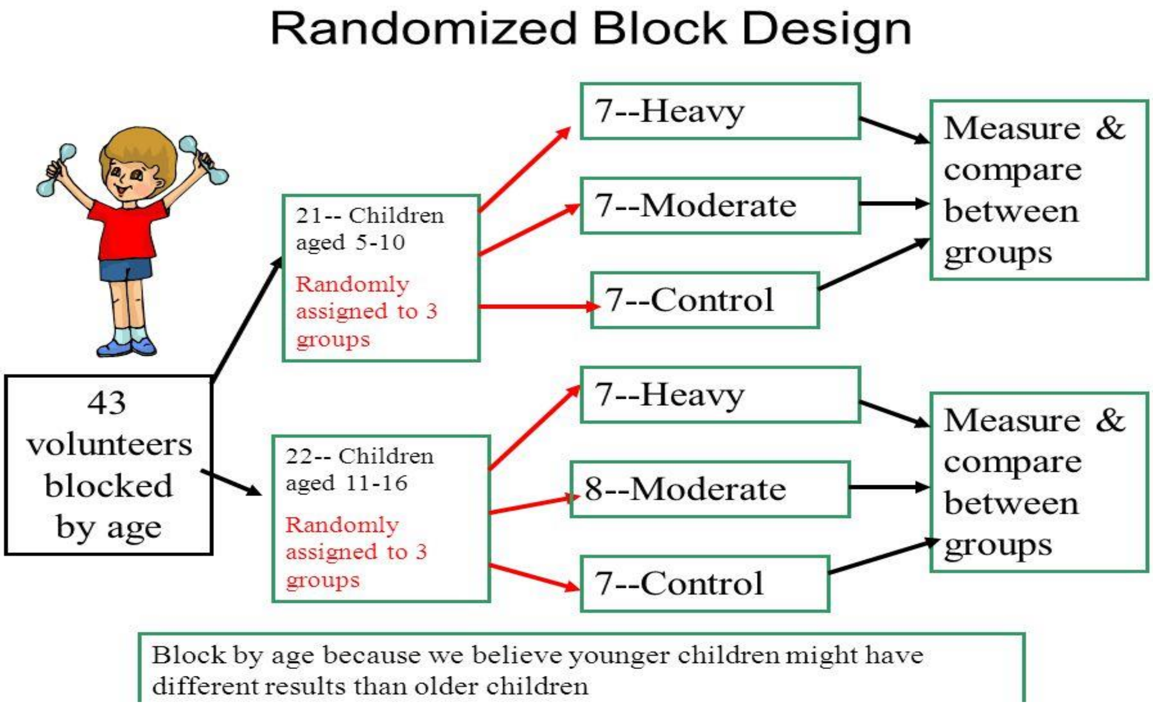
CRD - Completely Randomized Design

- 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.


Randomly assign students to treatments
Group 1 30 students
Treatment A: no practice exam
Group 2 30 students
Treatment B: practice exam
Compare Exam Results

RBD - Randomized Block Design

- The term Randomized Block Design has originated from agricultural research.
- In this design several treatments of variables are applied to different blocks of land to ascertain their effect on the yield of the crop.
- Blocks are formed in such a manner that each block contains as many plots as a number of treatments so that one plot from each is selected at random for each treatment.
- The production of each plot is measured after the treatment is given.
- These data are then interpreted and inferences are drawn by using the analysis of Variance Technique so as to know the effect of various treatments like different dozes of fertilizers, different types of irrigation etc.



Randomized Block Design

43 volunteers blocked by age

21-- Children aged 5-10
Randomly assigned to 3 groups

7--Heavy

7--Moderate

7--Control

Measure & compare between groups

22-- Children aged 11-16
Randomly assigned to 3 groups

7--Heavy

8--Moderate

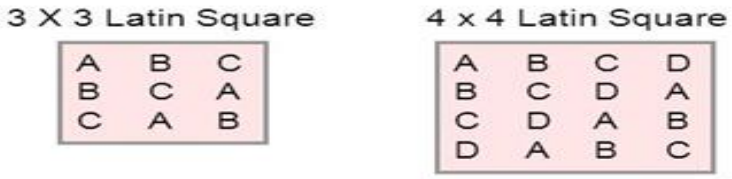
7--Control

Measure & compare between groups

Block by age because we believe younger children might have different results than older children

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.



3 X 3 Latin Square

A	B	C
B	C	A
C	A	B

4 x 4 Latin Square

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

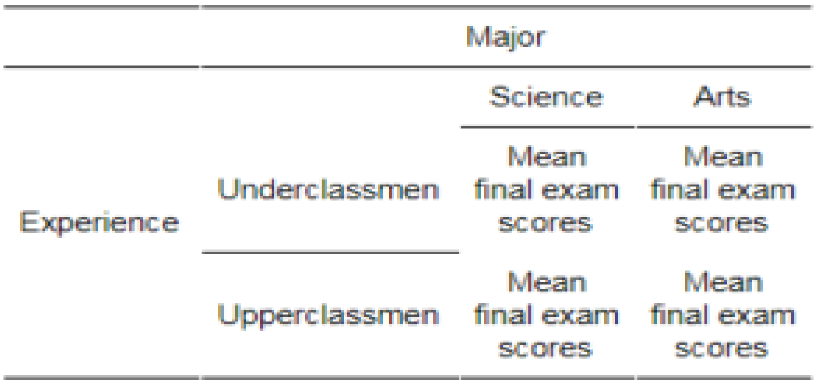
A B C D B
C D A C D
A B D A B
C

- 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.



		Major	
		Science	Arts
Experience	Underclassmen	Mean final exam scores	Mean final exam scores
	Upperclassmen	Mean final exam scores	Mean final exam scores

SOURCES OF SECONDARY DATA

The secondary data can be obtained through

1. Internal Sources - These are within the organization
2. External Sources - These are outside the organization

Internal Sources of Data

- Obtained with less time, effort and money than the external secondary data.
- 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 Re**port- It gives information about the sale of a product. The information provided is of outside the organization.
  - **Internal Experts**- These are people who are heading the various departments. They

can give an idea of how a particular thing is working

- **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. 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
  - 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.
  - It gives information about the total number of workers employed, production units, material used and value added by the manufacturer.
- Director General of Commercial Intelligence
  - 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
  - It provides information on wholesale price index.
  - Indices related to a number of sectors like food, fuel, power, food grains etc.
  - Also generates All India Consumer Price Index numbers for industrial workers, urban, non manual employees and cultural laborers.
- Planning Commission
  - It provides the basic statistics of Indian Economy..
- Reserve Bank of India
  - Provides information on Banking Savings and investment.
  - Also prepares currency and finance reports.
- National Sample Survey

  - Provides social, economic, demographic, industrial and agricultural statistics.
- Labour Bureau
  - provides information on skilled, unskilled, white collared jobs etc
- Department of Economic Affairs
  - Conducts economic survey and it also generates information on income, consumption, expenditure, investment, savings and foreign trade.
- State Statistical Abstract
  - Gives information on various types of activities related to the state like - commercial activities, education, occupation etc.
- Non Government Publications
  - includes publications of various industrial and trade associations, such as The Indian Cotton Mill Association Various chambers of commerce
- The Bombay Stock Exchange
  - 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

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.
- 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 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 includes
    - The International Labour Organization (ILO)
        - publishes data on the total and active population, employment, unemployment, wages and consumer prices
    - The Organization for Economic Co-operation and development (OECD)
        - Publishes data on foreign trade, industry, food, transport, and science and technology.
    - The International Monetary Fund (IMA)
        - Publishes reports on national and international foreign exchange regulations.

DATA QUALITY

Data Quality is a Perception or an assessment of data's fitness to serve its purpose in a given context.

What are the Typical Measures of Data Quality?
Accuracy
Completeness
Consistency
Uniqueness
Timeliness
Validity

✓ Improved data quality leads to better decision-making across an organization. The more high-quality data you have, the more confidence you can have in your decisions. Good data decreases risk and can result in consistent improvements in results.

It is described by several dimensions like

Correctness/Accuracy : Accuracy of data is the degree to which the captured data correctly describes the real world entity.

Consistency: This is about the single version of truth. Consistency means data throughout the enterprise should be sync with each other.

Completeness: It is the extent to which the expected attributes of data are provided.

Timeliness: Right data to the right person at the right time is important business.

Metadata: Data about data.

- Data mining applications are often applied to data that was collected for another purpose, or for future, but unspecified applications.
- For that reason data mining cannot usually take advantage of the significant benefits of "addressing quality issues at the source."
- In contrast, much of statistics deals with the design of experiments or surveys that achieve a pre specified level of data quality.
- Because preventing data quality problems is typically not an option, data mining focuses on
    1. Detection and correction (called data cleaning) of data quality problems
    2. Use of algorithms that can tolerate poor data quality.

MEASUREMENT AND DATA COLLECTION ISSUES

- It is unrealistic to expect that data will be perfect.
- There may be problems due to human error, limitations of measuring devices, or flaws in the data collection process.
- Values or even entire data objects may be missing.
- In other cases, there may be spurious or duplicate objects; i.e., multiple data objects that all correspond to a single "real" object.
- For example, there might be two different records for a person who has recently lived at two different addresses.
- Even if all the data is present and "looks fine," there may be inconsistencies-a person has a height of 2 meters, but weighs only 2 kilograms.

Measurement and Data Collection Errors

- **Measurement error**
    - Refers to any problem resulting from the measurement process.
    - A common problem is that the value recorded differs from the true value to some extent.
    - For continuous attributes, the numerical difference of the measured and true value is called the error.
- **Data collection error**
    - It refers to errors such as omitting data objects or attribute values, or

inappropriately including a data object.

  - For example, a study of animals of a certain species might include animals of a related species that are similar in appearance to the species of interest.

- Both measurement errors and data collection errors can be either systematic or random.
- Within particular domains, there are certain types of data errors that are commonplace, and there often exist well-developed techniques for detecting and/or correcting these errors.
- For example, keyboard errors are common when data is entered manually
- As a result, many data entry programs have techniques for detecting and, with human intervention, correcting such errors.

Noise and Artifacts

- Noise is the random component of a measurement error.
- It may involve the distortion of a value or the addition of false objects.
- If a bit more noise were added to the time series, its shape would be lost.
- The term noise is often used in connection with data that has a spatial or temporal component.
- In such cases, techniques from signal or image processing can frequently be used to reduce noise and thus, help to discover patterns (signals) that might be "lost in the noise."
- However, the elimination of noise is frequently difficult, and much work in data mining focuses on devising robust algorithms that produce acceptable results even when noise is present.
- Data errors may be the result of a more deterministic phenomenon, such as a streak in the same place on a set of photographs.
- Such deterministic distortions of the data are often referred to as artifacts.

Precision, Bias, and Accuracy

- In statistics and experimental science, the quality of the measurement process and the resulting data are measured by precision and bias.
- **Precision**
  - The closeness of repeated measurements (of the same quantity) to one another.
  - Precision is often measured by the standard deviation of a set of values.
- **Bias**
  - A systematic quantity being measured.
  - Bias is measured by taking the difference between the mean of the set of values and the known value of the quantity being measured.
  - Bias can only be determined for objects whose measured quantity is known by means external to the current situation.
- **Accuracy**
  - It is common to use the more general term, accuracy, to refer to the degree of

measurement error in data.

- The closeness of measurements to the true value of the quantity being measured.
- Accuracy depends on precision and bias.
- But there is no specific formula for accuracy in terms of these two quantities.
- One important aspect of accuracy is the use of significant digits.
- The goal is to use only as many digits to represent the result of a measurement or calculation as are justified by the precision of the data.

Outliers

- Outliers are either
  1. data objects that have characteristics that are different from most of the other data objects in the data set, or
  2. Values of an attribute that are unusual with respect to the typical values for that attribute.
- Outliers can be legitimate data objects or values.
- Unlike noise, outliers may sometimes be of interest.

Missing Values

- It is not unusual for an object to be missing one or more attribute values.
- In some cases, the information was not collected; e.g., some people decline to give their age or weight.
- In other cases, some attributes are not applicable to all objects; e.g., often, forms have conditional parts that are filled out only when a person answers a previous question in a certain way, but for simplicity, all fields are stored.
- Missing values should be taken into account during the data analysis.

- Strategies for dealing with missing data, each of which may be appropriate in certain circumstances:

  Eliminate Data Objects or Attributes

  - A simple and effective strategy is to eliminate objects with missing values.
  - However, even a partially specified data object contains some information, and if many objects have missing values, then a reliable analysis can be difficult or impossible.
  - However, if a data set has only a few objects that have missing values, then it may be convenient to omit them.
  - A related strategy is to eliminate attributes that have missing values.
  - This should be done with caution, however, since the eliminated attributes may be the ones that are critical to the analysis.

Estimate Missing Values

- Sometimes missing data can be reliably estimated.
- For example, consider a time series that changes in a reasonably smooth fashion, but has a few, widely scattered missing values.
- In such cases, the missing values can be estimated using the remaining values.
- As another example, consider a data set that has many similar data points.
- In this situation, the attribute values of the points closest to the point with the missing value are often used to estimate the missing value.
- If the attribute is continuous, then the average attribute value of the nearest neighbors is used
- if the attribute is categorical, then the most commonly occurring attribute value can be taken.

- **Ignore the Missing Value during Analysis**
  - Many data mining approaches can be modified to ignore missing values.
  - For example, suppose that objects are being clustered and the similarity between pairs of data objects needs to be calculated.
  - If one or both objects of a pair have missing values for some attributes, then the similarity can be calculated by using only the attributes that do not have missing values.

Inconsistent Values

- Data can contain inconsistent values.
- Consider an address field, where both a zip code and city are listed, but the specified zip code area is not contained in that city.
- It may be that the individual entering this information transposed two digits, or perhaps a digit was misread when the information was scanned from a handwritten form.
- It is important to detect and, if possible, correct such problems.
- Some types of inconsistencies are easy to detect. For instance, a person's height should not be negative.
- In some cases, it can be necessary to consult an external source of information.
- For example, when an insurance company processes claims for reimbursement, it checks the names and addresses on the reimbursement forms against a database of its customers.
- Once an inconsistency has been detected, it is sometimes possible to correct the data.
- A product code may have "check" digits, or it may be possible to double-check a product code against a list of known product codes, and then correct the code if it is incorrect, but close to a known code.
- The correction of an inconsistency requires additional or redundant information.

Duplicate Data

- A data set may include data objects that are duplicates, or almost duplicates, of one another.
- Many people receive duplicate mailings because they appear in a database multiple times under slightly different names.
- To detect and eliminate such duplicates, two main issues must be addressed.
    - First, if there are two objects that actually represent a single object, then the values of corresponding attributes may differ, and these inconsistent values must be resolved.
    - Second, care needs to be taken to avoid accidentally combining data objects that are similar, but not duplicates, such as two distinct people with identical names.
- The term duplication is often used to refer to the process of dealing with these issues.
- In some cases, two or more objects are identical with respect to the attributes measured by the database, but they still represent different objects.
- Here, the duplicates are legitimate, but may still cause problems for some algorithms if the possibility of identical objects is not specifically accounted for in their design.

ISSUES RELATED TO APPLICATIONS

- Data quality issues can also be considered from an application viewpoint.
- There are many issues that are specific to particular applications and fields.
    - **Timeliness**
        - Some data starts to age as soon as it has been collected.
        - In particular, if the data provides a snapshot of some ongoing phenomenon or process, such as the purchasing behavior of customers or Web browsing patterns, then this snapshot represents reality for only a limited time.
        - If the data is out of date, then so are the models and patterns that are based on it.
    - **Relevance**
        - The available data must contain the information necessary for the application.
        - Consider the task of building a model that predicts the accident rate for drivers.
        - If information about the age and gender of the driver is omitted, then it is likely that the model will have limited accuracy unless this information is indirectly available through other attributes.
        - Making sure that the objects in a data set are relevant is also challenging.
    - **Sampling bias**
        - Which occurs when a sample does not contain different types of objects in proportion to their actual occurrence in the population?

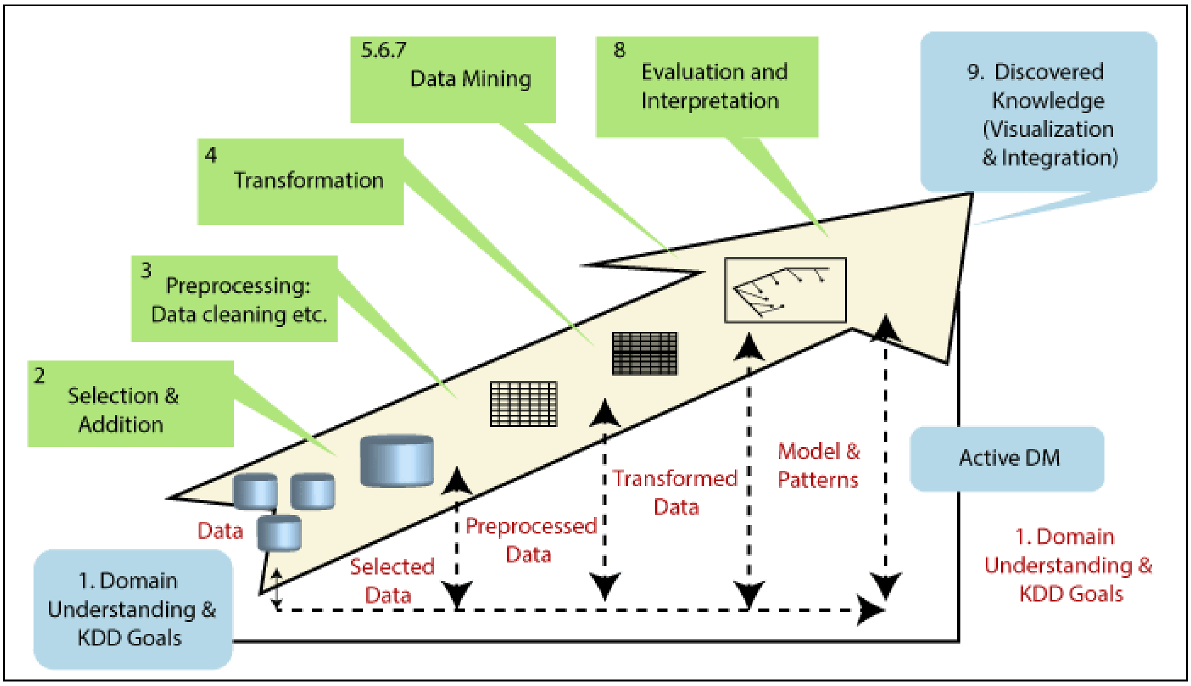
    - For example, survey data describes only those who respond to the survey.
    - Because the results of a data analysis can reflect only the data that is present, sampling bias will typically result in an erroneous analysis.

- **Knowledge about the Data**
    - Ideally, data sets are accompanied by documentation that describes different aspects of the data.
    - The quality of this documentation can either aid or hinder the subsequent analysis.
    - For example, if the documentation identifies several attributes as being strongly related, these attributes are likely to provide highly redundant information, and we may decide to keep just one.
    - If the documentation is poor, however, and fails to tell us, for example, that the missing values for a particular field are indicated with a -9999, then our analysis of the data may be faulty.
    - Other important characteristics are the precision of the data, the type of features (nominal, ordinal, interval, ratio), the scale of measurement (e.g., meters or feet for length), and the origin of the data.

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 Preprocessing is one of the most Data Mining steps which deals with data preparation and transformation of the data set and seeks at the same time to make knowledge discovery more efficient.


5.6.7 Data Mining
8 Evaluation and Interpretation
9. Discovered Knowledge (Visualization & Integration)
4 Transformation
3 Preprocessing: Data cleaning etc.
2 Selection & Addition
Active DM
Data
Selected Data
Preprocessed Data
Transformed Data
Model & Patterns
1. Domain Understanding & KDD Goals
1. Domain Understanding & KDD Goals

Why We Need Data Preprocessing?

The real-world data tend to be incomplete, noisy, and inconsistent. This can lead to a poor quality of collected data and further to a low quality of models built on such data. In order to address these issues, Data Preprocessing provides operations which can organize the data into a proper form for better understanding in data mining process.

What are the Techniques Provided in Data Preprocessing?

They are Data Cleaning/Cleansing, Data Integration, Data Transformation, and Data Reduction.

1. Data Cleaning/Cleansing

Real-world data tend to be incomplete, noisy, and inconsistent. Data Cleaning/Cleansing routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

Data can be noisy, having incorrect attribute values. Owing to the following, the data collection instruments used may be fault. Maybe human or computer errors occurred at data entry. Errors in data transmission can also occur.

Cleaning "dirty" data

"Dirty" data can cause confusion for the mining procedure. Although most mining routines have some procedures, they deal incomplete or noisy data, which are not always robust. Therefore, a useful Data Preprocessing step is to run the data through some Data Cleaning/Cleansing routines.

2. Data Integration

Data Integration is involved in data analysis task which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files. The issue to be considered in Data Integration is schema integration. It is tricky.

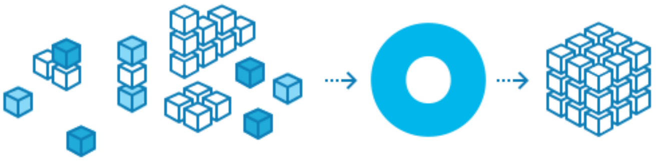
How can real-world entities from multiple data sources be 'matched up'? This is referred as entity identification problem. For example, how can a data analyst be sure that customer_id in one database and cust_number in another refer to the same entity? The answer is metadata. Databases and data warehouses typically have metadata. Simply, metadata is data about data.

Combining data from multiple sources

Metadata is used to help avoiding errors in schema integration. Another important issue is redundancy. An attribute may be redundant, if it is derived from another table. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

3. Data Transformation

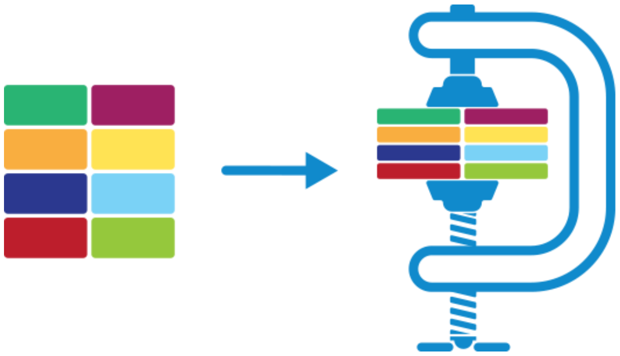
Data are transformed into appropriate forms of mining. Data Transformation involves the following:



Constructing data cube

1. In Normalisation, where the attribute data are scaled to fall within a small specified range, such as -1.0 to 1.0, or 0 to 1.0.
2. Smoothing works to remove the noise from the data. Such techniques include binning, clustering, and regression.
3. In Aggregation, summary or aggregation operations are applied to the data. For example, daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities.
4. In Generalisation of the Data, low level or primitive/raw data are replaced by higher level concepts through the use of concept hierarchies. For example, categorical attributes are generalised to higher level concepts street into city or country. Similarly, the values for numeric attributes may be mapped to higher level concepts like, age into young, middle-aged, or senior.

4. Data Reduction



Reducing representation of data set.

Complex data analysis and mining on huge amounts of data may take a very long time, making such analysis impractical or infeasible. Data Reduction techniques are helpful in analysing the reduced representation of the data set without compromising the integrity of the original data and yet producing the qualitative knowledge. Strategies for data reduction include the following:

1. In Data Cube Aggregation, aggregation operations are applied to the data in the construction of a data cube.
2. In Dimension Reduction, irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.

3. In Data Compression, encoding mechanisms are used to reduce data set size. The methods used for Data Compression are Wavelet Transform and Principle Component Analysis.
4. In Numerosity Reduction, data is replaced or estimated by alternative and smaller data representations such as parametric models (which store only the model parameters instead of the actual data, e.g. Regression and Log-Linear Models) or non-parametric methods (e.g. Clustering, Sampling, and the use of histograms).
5. In Discretisation and Concept Hierarchy Generation, raw data values for attributes are replaced by ranges or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of abstraction and are powerful tools for data mining.

Data Preprocessing

- Steps that should be applied to make the data more suitable for data mining.
- Consists of a number of different strategies and techniques that are interrelated in complex ways.

Strategies and techniques

(1) Aggregation
(2) Sampling
(3) Dimensionality reduction
(4) Feature subset selection
(5) Feature creation
(6) Discretization and binarization
(7) Variable transformation

Two categories of Strategies and techniques

- Selecting data objects and attributes for the analysis.
- Creating/changing the attributes.

***Goal*:**

- To improve the data mining analysis with respect to time, cost, and quality.

Aggregation

- Quantitative attributes are typically aggregated by taking a sum or an average.
- A qualitative attribute can either be omitted or summarized.

Motivations for aggregation

- Smaller data sets resulting from data reduction require less memory and processing time.
- Hence aggregation uses more expensive data mining algorithms.
- Aggregation can act as a change of scope or scale by providing a high-level view of the data instead of a low-level view.

Disadvantage of aggregation

- Potential loss of interesting details.

Sampling

- An approach for selecting a subset of the data objects to be analyzed.

Key principle for effective sampling

- a sample is representative if it has approximately the same property of interest as the original set of data.
- If the mean (average) of the data objects is the property of interest, then a sample is representative if it has a mean that is close to that of the original data.

Sampling Approaches

- Random sampling.
- Progressive or Adaptive Sampling

Random sampling

- Sampling without replacement: as each item is selected, it is removed from the set of all objects that together constitute the population.
- Sampling with replacement: objects are not removed from the population as they are selected for the sample. Same object can be picked more than once.

Progressive or Adaptive Sampling

- Difficult to determine proper sample size.
- So adaptive or progressive sampling schemes are used.
- These approaches start with a small sample, and then increase the sample size until a sample of sufficient size has been obtained.

Dimensionality reduction

- Data mining algorithms work better if the dimensionality - the number of attributes in the data - is lower.
- Eliminate irrelevant features and reduce noise.
- Lead to a more understandable model due to fewer attributes.
- Allow the data to be more easily visualized.
- Amount of time and memory required by the data mining algorithm is reduced.

Feature subset selection

- The reduction of dimensionality by selecting new attributes that are a subset of the old.

Discretization and Binarization

Discretization

- Transform a continuous attribute into a categorical attribute.

Binarization

- Transform both continuous and discrete attributes into one or more binary attributes.

Variable transformation

- A transformation that is applied to all the values of a variable.