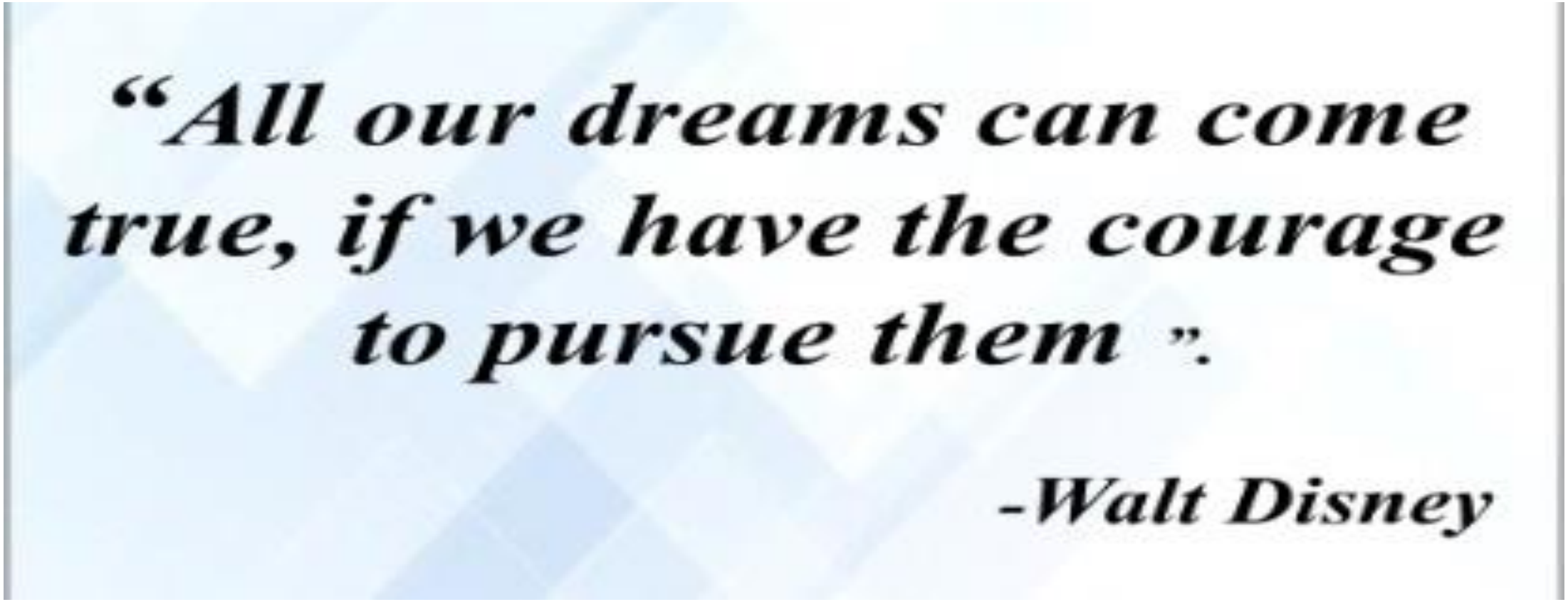
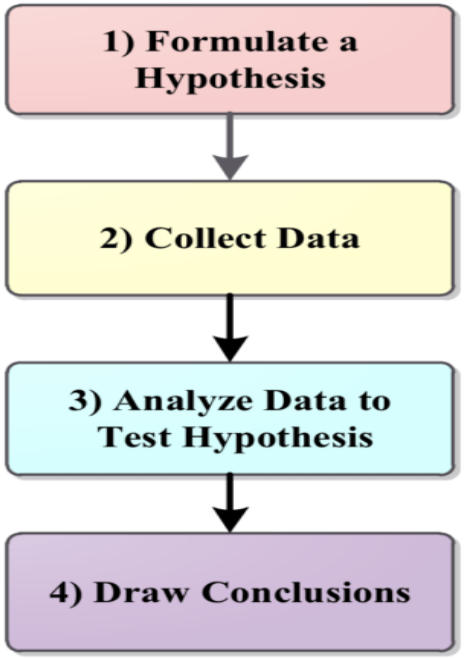
L 2-Data Analytics its types and classification of data



Difference between Data Analytics and Data Mining:

A key difference between data analytics and data mining is that data mining does not require any preconceived hypothesis or notions before tackling the data. It simply compiles it into useful formats. However, data analysis does need a hypothesis to test, as it is looking for answers to particular questions.



Data mining can be undertaken by a single specialist with excellent technological skills. With the right software, they are able to collect the data ready for further analysis. At this stage, a larger team simply isn't required. From here, a data mining specialist will usually report their findings to the client, leaving the next steps in someone else's hands.

However, when it comes to data analytics, a team of specialists may be needed. They need to assess the data, figure outpatterns, and draw conclusions. They may use machine learning to help with the processing, but this still has a human element involved. Data analytics teams need to know the right questions to ask –

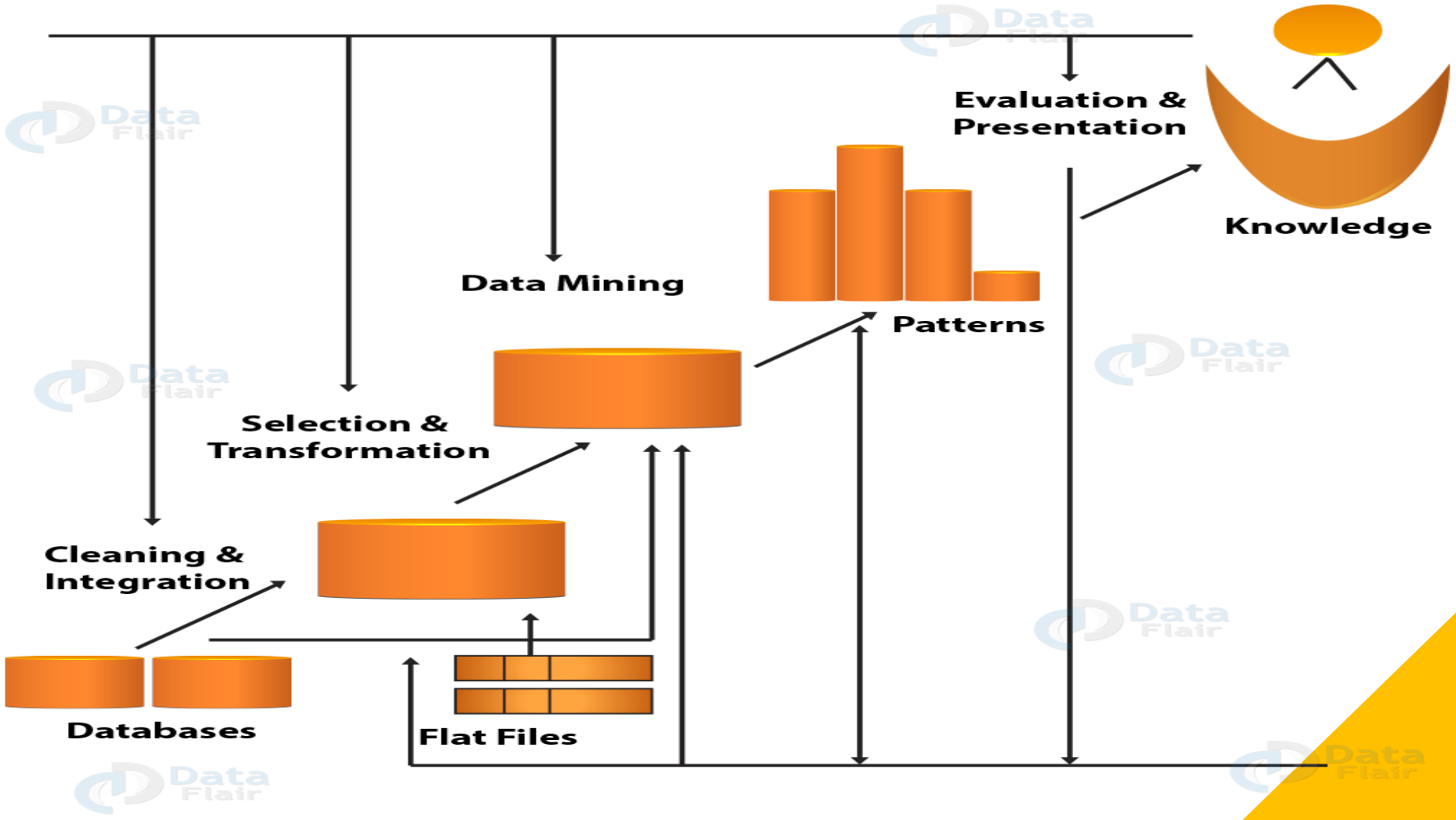
e.g. relation between reader gender and English paper

e.g. relation between trained and trained students versus committing errors

e.g. inoculated with vaccine, not inoculated vaccine versus died of disease , survived

Data mining usually does not need any visualizations, bar charts, graphs etc., whereas these visualizations /are the bread and butter of data analysis. Without a good representation of the data in question, all the efforts which are put into the analysis of the data would not come to fruition(the time when a plan, etc. starts to be successful).

Knowledge Discovery Database(KDD)



- **Data Integration**

First of all the data is collected and integrated from all the different sources.

- **Data Selection**

Generally, we may not all the data we have collected in the first step. Also, in this step, we select only those data which we think useful for data mining.

- **Data Cleaning**

Generally, the data we have collected is not clean and may contain errors, missing values, noisy or inconsistent data. Therefore we need to apply different techniques to get rid of such anomalies.

- **Data Transformation**

Basically, the data even after cleaning is not ready for mining. Also, we need to transform them into forms appropriate for mining. Thus, the techniques used to do this are smoothing, aggregation, normalization etc.

- **Data Mining**

As now in this step, we are ready to apply **data mining techniques** on the data. Basically, it is to discover the interesting patterns. Hence, clustering and association analysis are among the many different techniques present. Also, as we used for data mining.

- **Pattern Evaluation and Knowledge Presentation**

Generally, this step includes visualization.

- **Decisions / Use of Discovered Knowledge**

As this step is beneficial to us. Also, it helps to use the knowledge acquired to take better decisions.

patterns, and draw conclusions. They may use machine learning to help with the processing, but this still has a human element involved. Data analytics teams need to know the right questions to ask –

e.g. relation between reader gender and English paper

e.g. relation between trained and trained students versus committing errors

e.g. inoculated with vaccine, not inoculated vaccine versus died of disease , survived

Data mining usually does not need any visualizations, bar charts, graphs etc., whereas these visualizations /are the bread and butter of data analysis. Without a good representation of the data in question, all the efforts which are put into the analysis of the data would not come to fruition.

Quiz

1. Which of the following is usually the last step in the data mining process?

A. Visualization

B. Preprocessing

C. Modeling

D. Deployment

2. Name of a movie, can be considered as an attribute of type?

A. Nominal

B. Ordinal

C. Interval

D. Ratio

3. User rating given to a movie in a scale 1-10, can be considered as an attribute of type?

A. Nominal

B. Ordinal

C. Interval

D. Ratio

4. Which of the following operations cannot be performed on interval attributes?

A. Distinctness

B. Order

C. Addition

D. Multiplication

5. Which of the following operations can be performed on ratio attributes?

A. Addition

B. Multiplication

C. Both of the above

D. None of the above

6. Sales database of items in a supermarket can be considered as an example of:

A. Record data

B. Tree data

C. Graph data

D. None of the above

7. Rows of a data matrix storing record data usually represents?

A. Metadata

B. Objects

C. Attributes

D. Aggregates

8. Which of the following is an example of continuous attribute?

A. Weight of a person

B. Shoe size of a person

C. Gender of a person

D. None of the above

Types of Data Analytics:

Data analytics can be broken into four key types(DDPP):

Descriptive, which answers the question, “What happened?”
Diagnostic, which answers the question, “Why did this happen?”
Predictive, which answers the question, “What might happen in the future?”
Prescriptive, which answers the question, “What should we do next?”

Q.1 Amongst which of the following is / are the examples of descriptive analytics,

1.Traffic and Engagement Reports

2.Financial Statement Analysis

3.Demand Trends and Aggregated Survey Results

4.All of the mentioned above

Q.2 A manager at Amco Inc. wishes to know the company's revenue and profit in its previous quarter. Which of the following business analytics will help the manager?

A) prescriptive analytics

B) normative analytics

C) descriptive analytics

D) predictive analytics

Q.3 Amongst which of the following is /are the techniques that are used for predictive analytics,

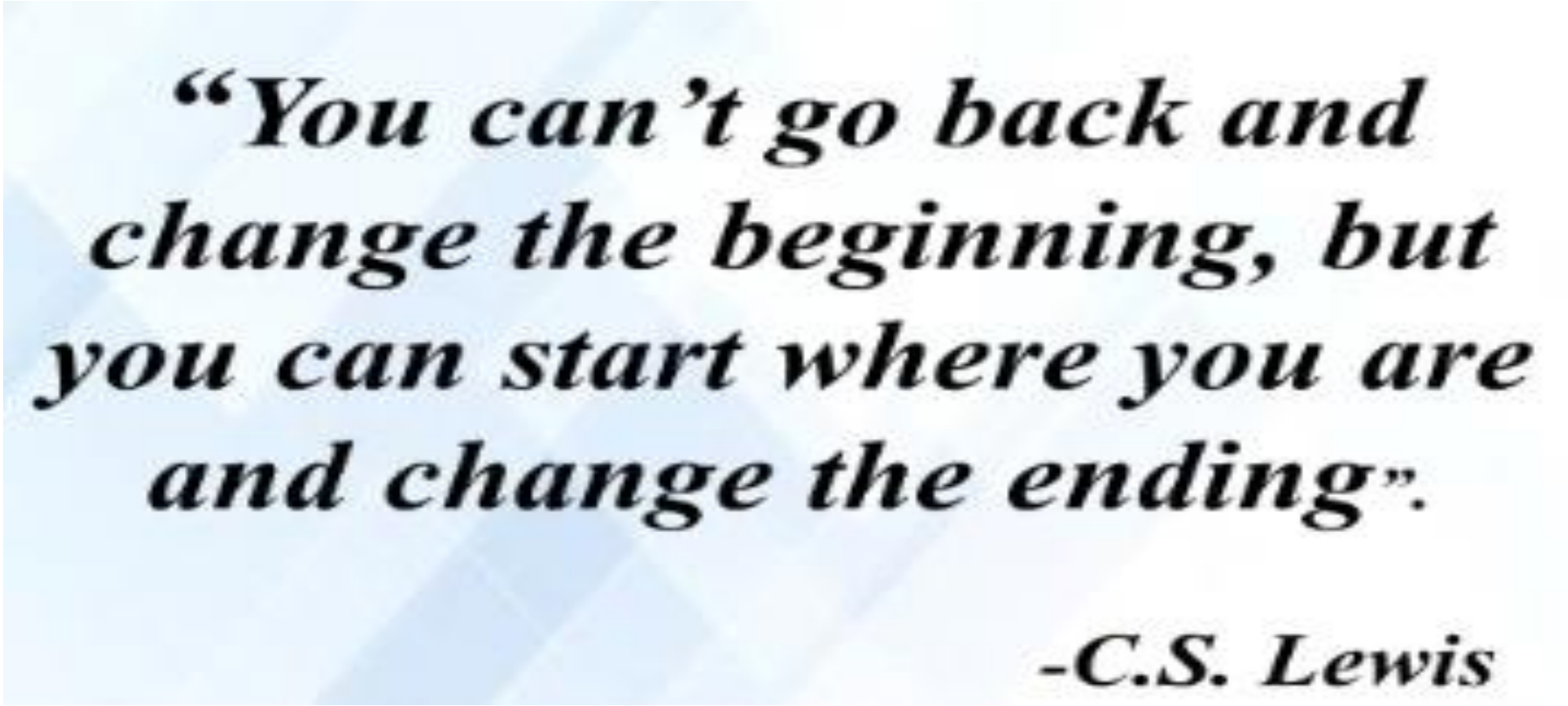
A. Linear Regression

B. Time series analysis and forecasting

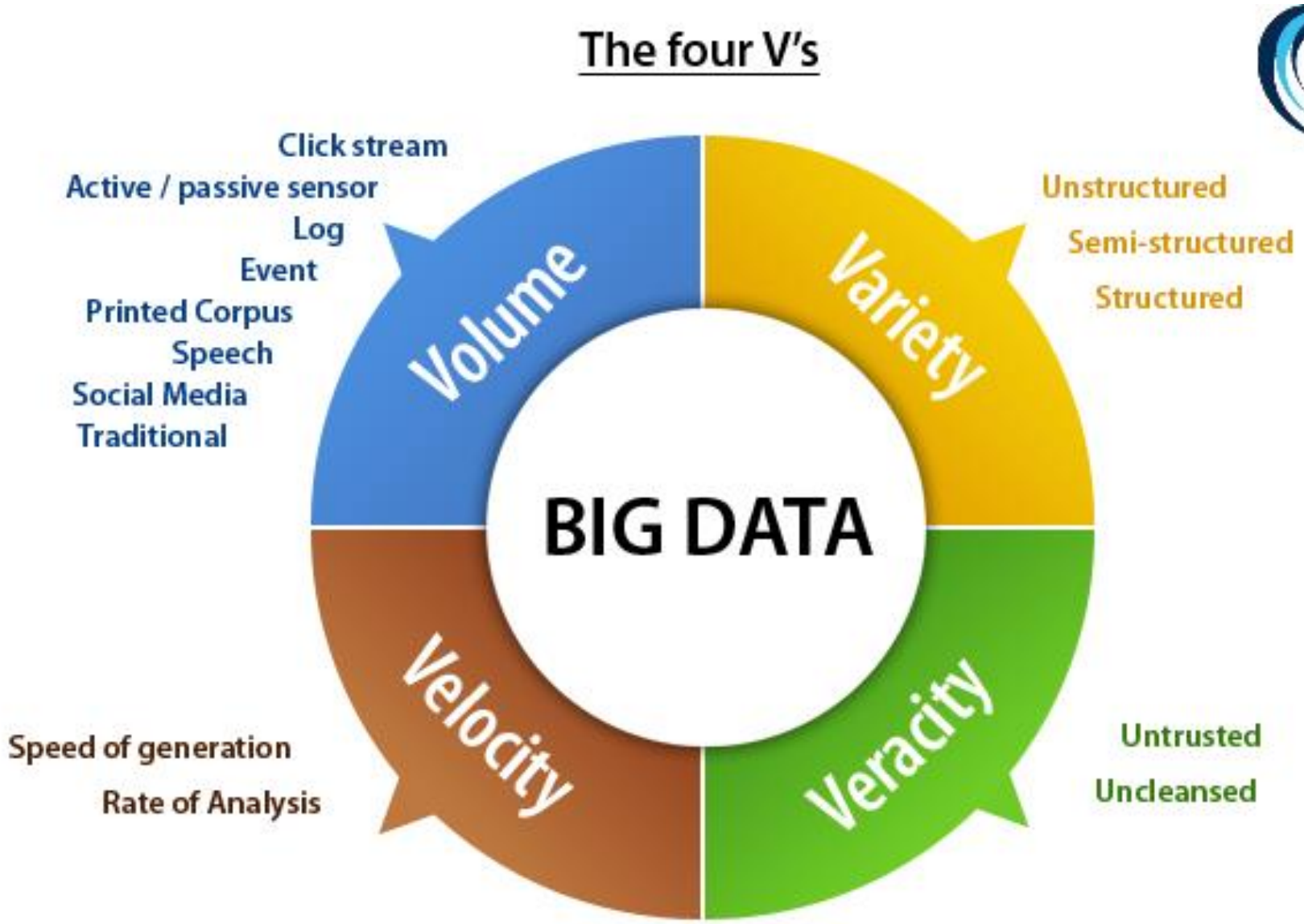
C. Data Mining

D. All of the mentioned above

L 3-Characteristics of Data/ Big Data, Classification of data, Data Analytics life cycle



Characteristics of Data/ Big Data



1. Variety

Variety of Big Data refers to structured, unstructured, and semi structured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more. Variety is one of the important characteristics of big data.

2. Velocity

Velocity essentially refers to the speed at which data is being created in real-time. In Big Data velocity data flows in from sources like networks, social media, mobile phones etc. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.

Example: There are more than 3.5 billion searches per day are made on Google. Also, Facebook users are increasing by 22%(Approx.) year by year. (Bits to stream)

3. Volume

Volume is one of the characteristics of big data. Big Data indicates huge 'volumes' of data that is being generated on a daily basis from various sources like social media platforms, business processes, networks, human interactions, etc. Such a large amount of data are stored in data warehouses.

Example: In the year 2016, the estimated global mobile traffic was 6.2 Exabytes (6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 Exabytes of data. (Mega Bytes to Exabytes)

4. Veracity

Veracity is concerned with the uncertainty or inaccuracy of the data. In many cases the data will be inaccurate hence filtering and selecting the data which is actually needed is a complicated task. A lot of statistical and analytical process has to go for data cleaning for choosing intrinsic data for decision-making.

Quiz

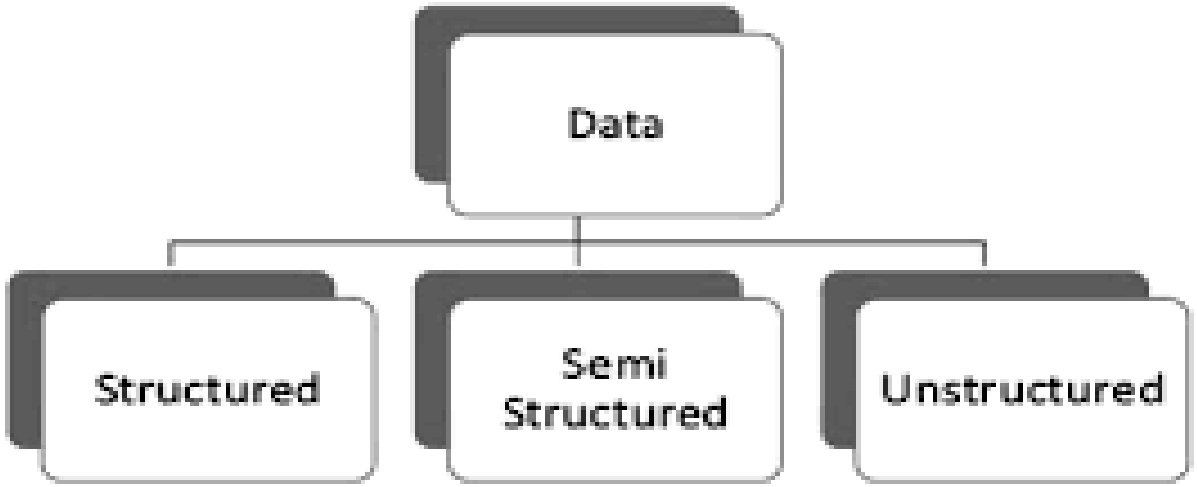
Q.1 Choose the primary characteristics of big data among the following
A. Velocity
B. Variety
C. Volume
D. All of the above

Q.2 Data in _____bytes size is called big data.
A. MB
B. Giga
C. Tera
D. Peta

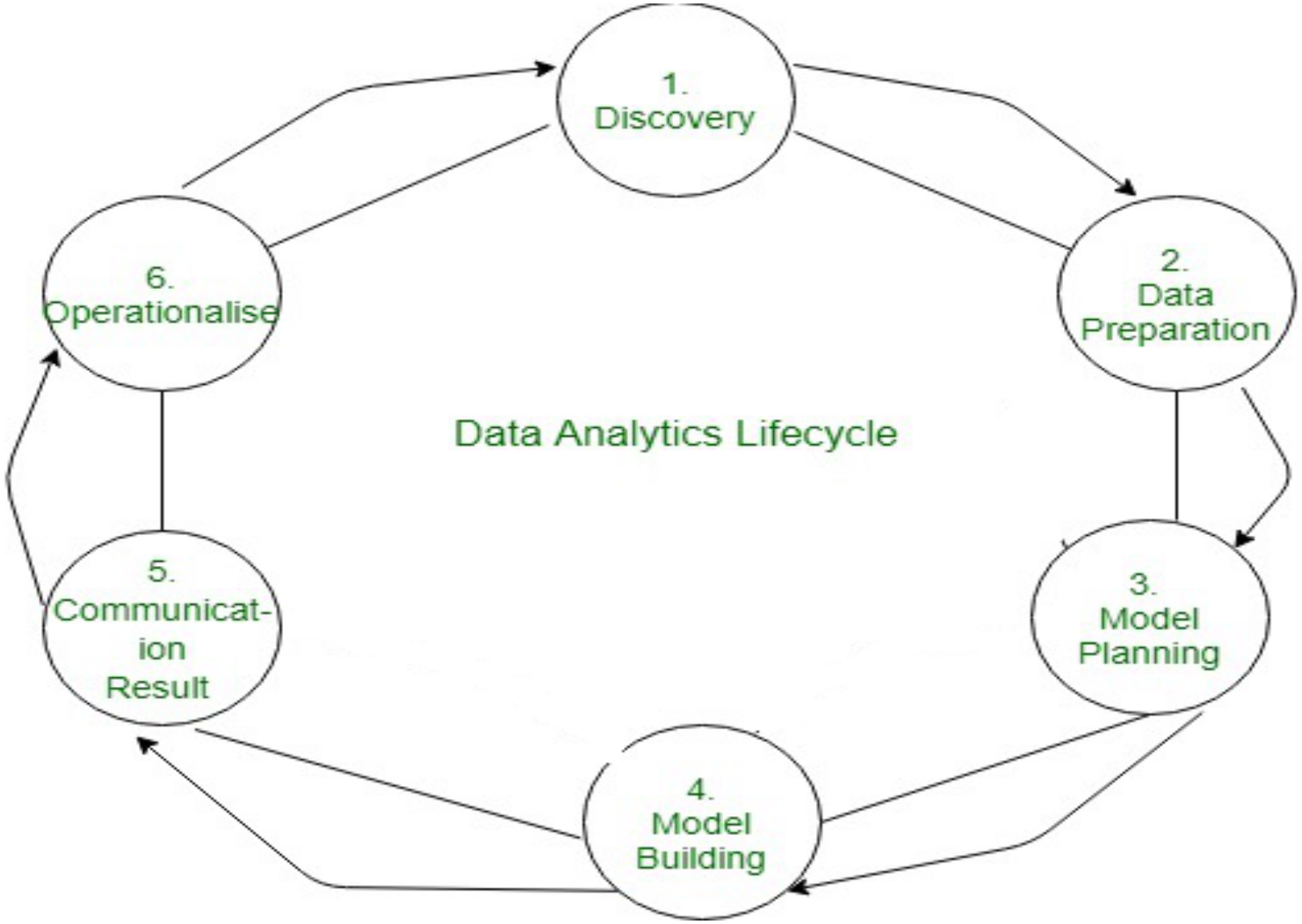
Q.3 Transaction of data of the bank is a type of
A. Unstructured B. Structured C. Both A and B D. None of these

Classification of Data

Data classification is the process of organizing structured and unstructured data into defined categories that represent different types of data.



Data Analytics Life Cycle:



1. Discovery:

- Team investigate the problem
- Develop understanding
- Know about data sources
- Formulate initial hypothesis which can be tested later.

2. Data Preparation:

It follows ETL process.

- Extract data from the different sources
- Transformation to achieve a common format
  date format, time unit, money unit, gender format etc.

3. Model planning:

Selection of the model, training data, test data and tool like Weka, MATLAB etc.

4. Model building:

Deploy model on the existing data and gets the patterns
Like Bread, Butter-> Milk

5. Result Communication:

- Analyze the patterns
- Communicate outcomes to various stakeholders

6. Operationalize:

- Deploy or modify the required process to improve the production or business gain
- Team delivers final reports or briefings etc.

Q.1 Which is the correct sequence of phases of Data Analytics life cycle

A) Discovery-->Model planning and building->data preparation- ->communicating result->operationalize

B) Model planning & building-->Data Preparation--> communicating result-->operationalize-->Discovery

C) Discovery-->Data Preparation-->model Planning & building-->communicating result->operationalize

D) Data Preparation-->model Planning&building-->operationalize --> communicating result->Discovery

Q. 2 Point out the correct statement.

A) Raw data is original source of data
B) Preprocessed data is original source of data
C) Raw data is the data obtained after processing steps
D) None of the mentioned