QUESTION 1 (5pts)

1. Describe shortly all stages of the Data Mining Process (you can also draw a picture).

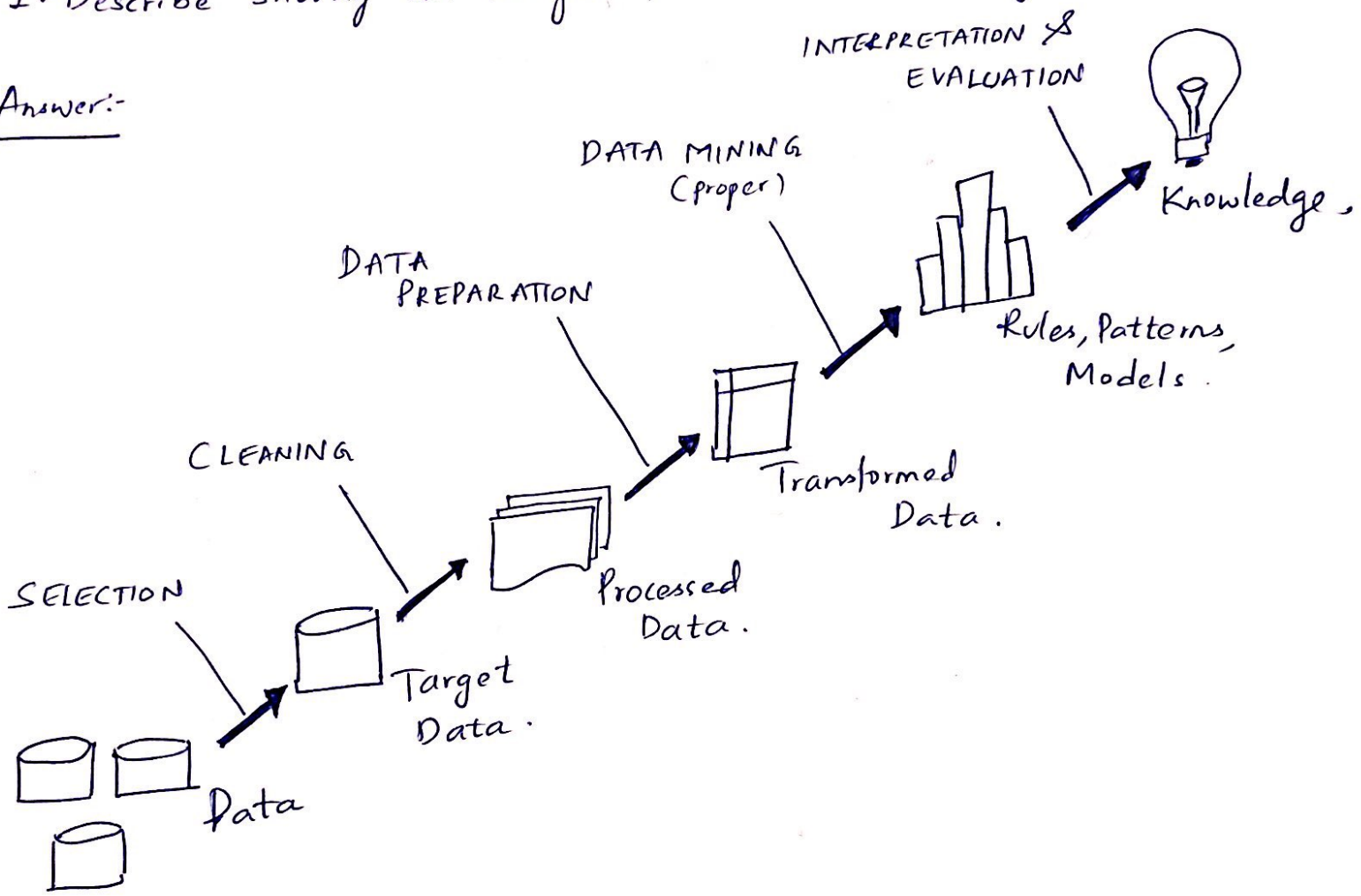
2. When you re-iterate the DM Process?

PART-1

Question1:-

1. Describe shortly all stages of the Data Mining Process?

Answer:-



The Data Mining Process is divided into three stages i.e (i) Preprocessing (ii) Data Mining (proper) (iii) Interpretation

⇒ PRE-PROCESSING: It includes all the operations that needs to be performed before a data mining algorithm is applied. The preprocessing involves data selection,

data cleaning, data integration, data reduction, data transformation and data discretization.

* DATA SELECTION :- In this step, data relevant to the analysis task are retrived from the database.

* DATA CLEANING: This step involves the removal of noisy or incomplete data from the collection. This step carries out the routine cleaning work by Filling in missing values, removing the noisy data [i.e Smooth noisy data by binning, clustering, regression techniques & Identify or Remove Outliers] and also resolve inconsistencies.

* DATA INTEGRATION :- In this step, multiple heterogeneous data sources such as databases, data cubes or files are combined for analysis. This step is helpful in improving the accuracy & speed of DM process.

* DATA REDUCTION :- This technique is applied to obtain relevant data for analysis from the collection of data. The size of representation is much smaller in volume but yet produce the same/almost same analytical results. Some strategies are data compression, dimensionality reduction etc.

* DATA TRASFORMATION: In this step, data are transformed or consolidated into forms appropriate

for mining. Data is consolidated so that the DM process is more efficient and patterns are easier to understand, Techniques such as normilization & aggregation are used here.

* DATA DISCRETIZATION :- This step is mainly important for numerical data. This involves reducing the number of values of attributes.

⇒ DATA MINING PROPER : This is an essential process where inteligent patterns are applied to extract the data patterns. The data is represented in the form of patterns & models are structured using classification & clustering techniques.

⇒ INTERPRETATION & EVALUATION : Visualization & knowledge representation techniques are used to represent the mined data to the user. Pattern evaluation is a process, where we identify the truly interesting patterns representing knowledge based on some interesting measures. The user decides if it is necessary to re-iterate the algorithms.

2. When you re-iterate the DM process?

Answer

→ Data Mining Propeo is a step in Data Mining process, in which algorithms / intelligent patterns are applied to extract the data patterns. The data is represented in the form of patterns and models are structured using classification & clustering techniques.

→ This process is re-iterated inorder to get more efficient results. We can also add new data to re-iterate

→ The results of any one step may indicate that a previous step needs to be refined. For example; Error rate of the pattern recognition step could be high, indicating that the features extracted are not representative enough of the patterns being considered or that the quality of training data could be improved. So, in this case inorder to improve efficiency of results, we re-iterate the DM process.