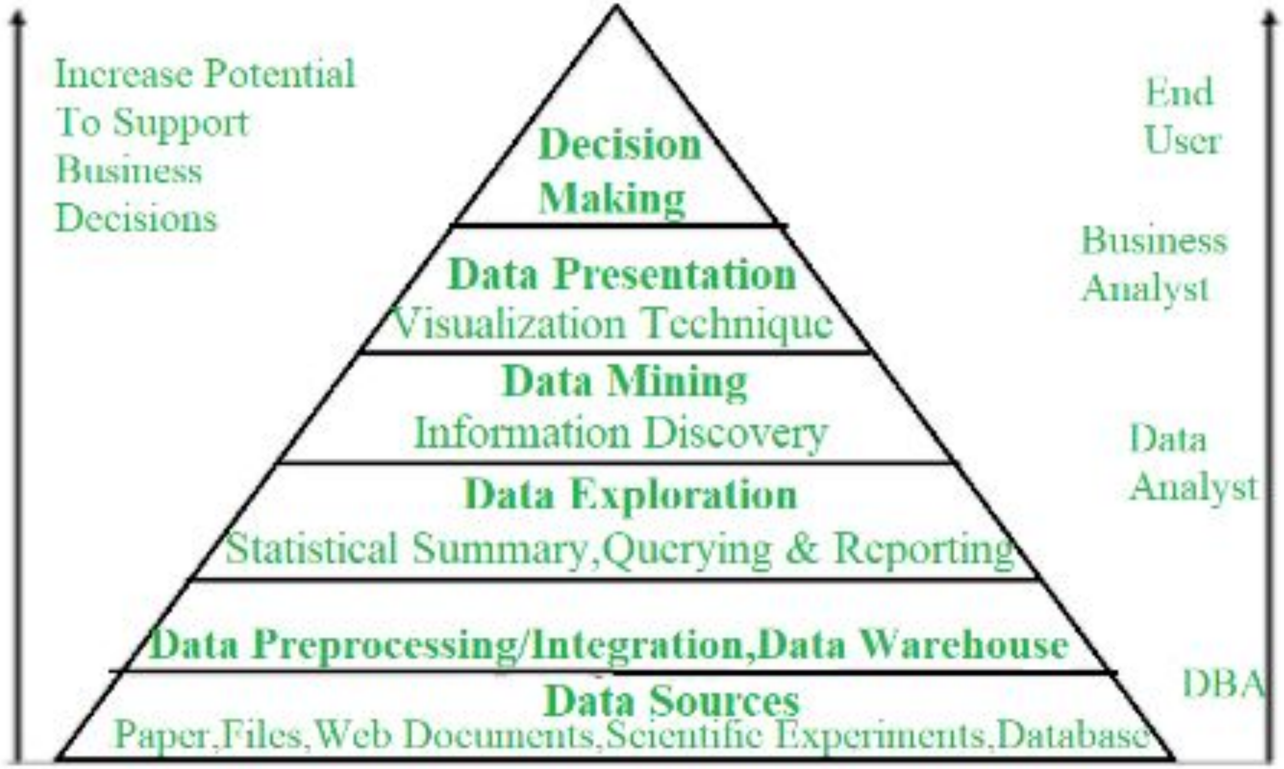
Stages of the Data Mining Process

- Data Mining refers to extracting or mining knowledge from large amounts of data. The term is actually a misnomer. Thus, data mining should have been more appropriately named as knowledge mining which emphasis on mining from large amounts of data. It is computational process of discovering patterns in large data sets involving methods at intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of data mining process is to extract information from a data set and transform it into an understandable structure for further use. It is also defined as extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from a huge amount of data. Data mining is a rapidly growing field that is concerned with developing techniques to assist managers and decision-makers to make intelligent use of a huge amount of repositories.

Alternative names for Data Mining :

- 1. Knowledge discovery (mining) in databases (KDD)
- 2. Knowledge extraction
- 3. Data/pattern analysis
- 4. Data archaeology
- 5. Data dredging
- 6. Information harvesting
- 7. Business intelligence

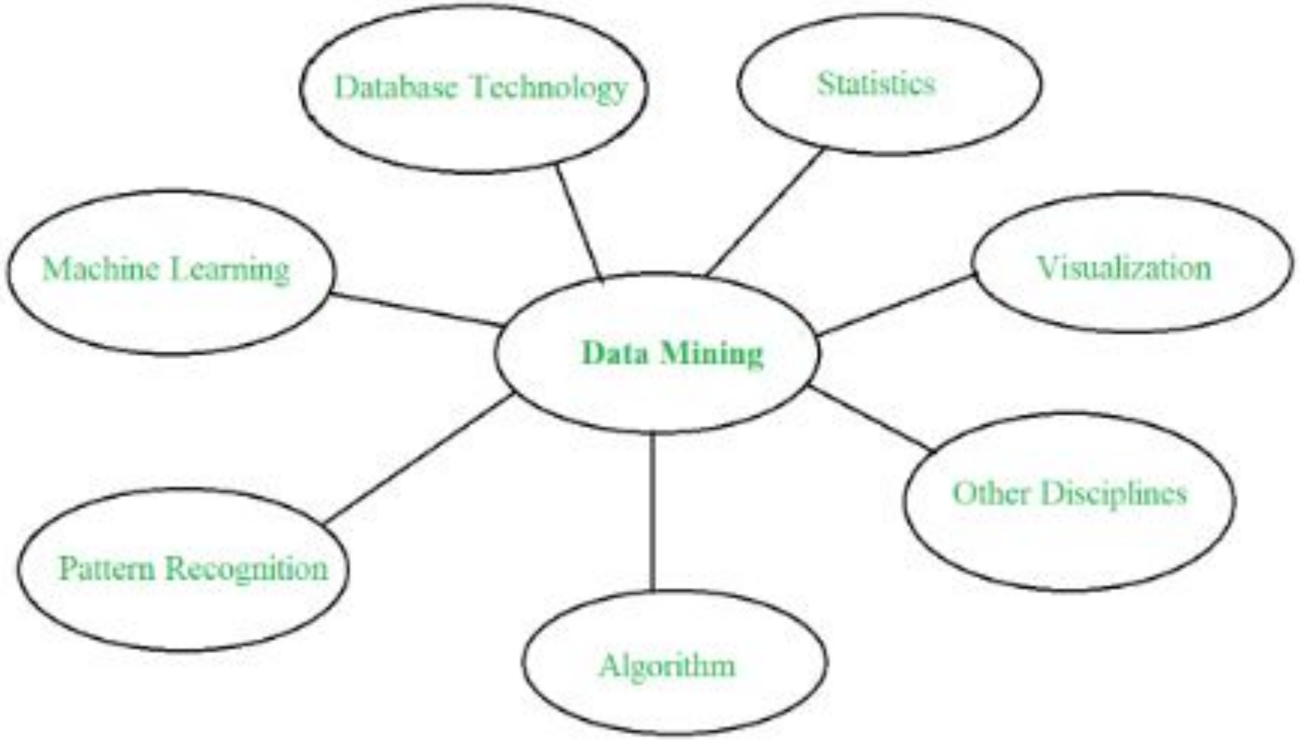
Data Mining and Business Intelligence :



Key properties of Data Mining

- 1. Automatic discovery of patterns
- 2. Prediction of likely outcomes
- 3. Creation of actionable information
- 4. Focus on large datasets and databases

Data Mining : Confluence of Multiple Disciplines –



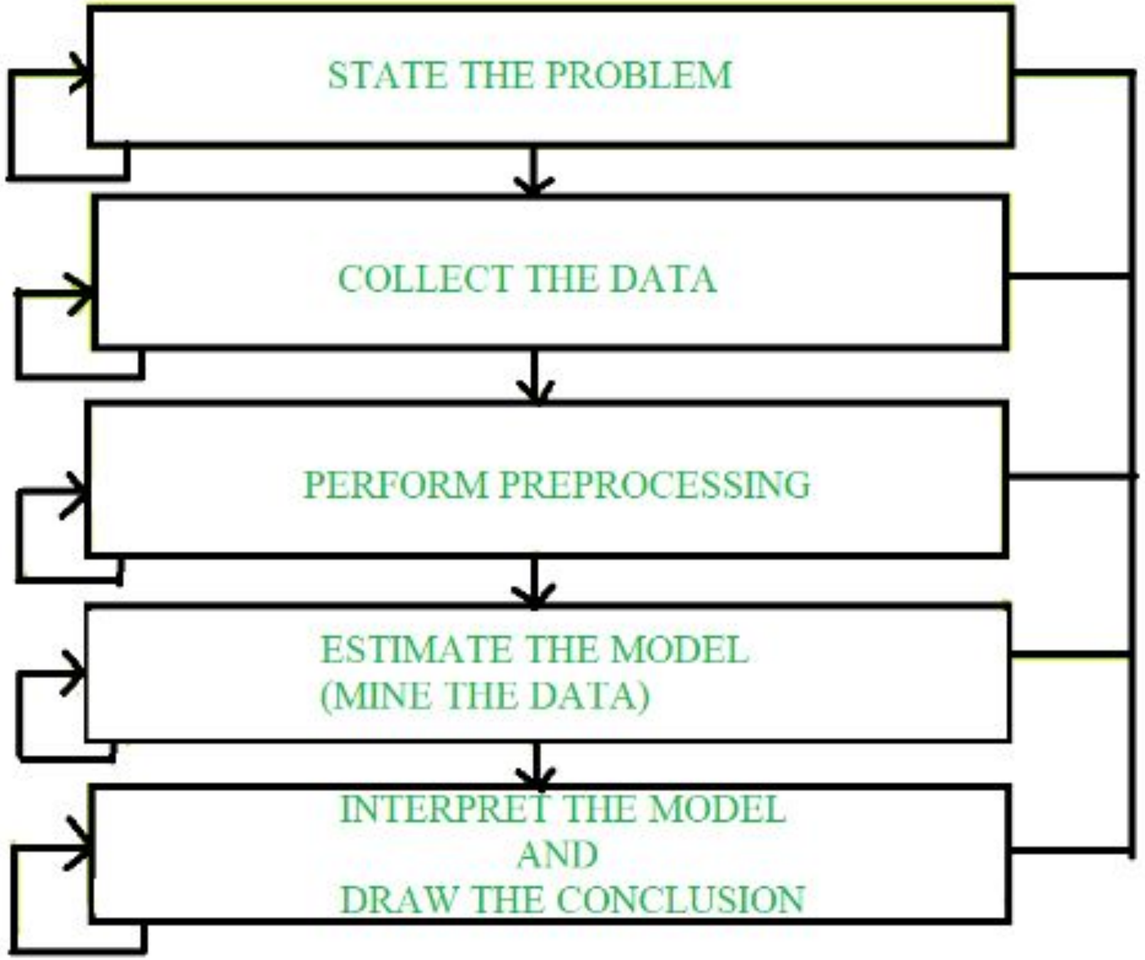
- Mining is a process of discovering various models, summaries, and derived values from a given collection of data. The general experimental procedure adapted to data-mining problem involves following steps :

1. **State problem and formulate hypothesis –** In this step, a modeler usually specifies a group of variables for unknown dependency and, if possible, a general sort of this dependency as an initial hypothesis. There could also be several hypotheses formulated for one problem at this stage. The primary step requires combined expertise of an application domain and a data-mining model. In practice, it always means an in-depth interaction between data-mining expert and application expert. In successful data-mining applications, this cooperation does not stop within initial phase. It continues during whole data-mining process.
2. **Collect data –** This step cares about how information is generated and picked up. Generally, there are two distinct possibilities. The primary is when data-generation process is under control of an expert (modeler). This approach is understood as a designed experiment. The second possibility is when expert cannot influence data generation process. This is often referred to as observational approach. An observational setting, namely, random data generation, is assumed in most data-mining applications. Typically, sampling distribution is totally unknown after data are collected, or it is partially and implicitly given within data-collection procedure. It is vital, however, to know how data collection affects its theoretical distribution since such a piece of prior knowledge is often useful for modeling and, later, for ultimate interpretation of results. Also, it is important to form sure that information used for estimating a model and therefore data used later for testing and applying a model come from an equivalent, unknown, sampling distribution. If this is often not case, estimated model cannot be successfully utilized in a final application of results.

1. **Data Preprocessing –** In the observational setting, data is usually “collected” from prevailing databases, data warehouses, and data marts. Data preprocessing usually includes a minimum of two common tasks :
    1. **(i) Outlier Detection (and removal) :** Outliers are unusual data values that are not according to most observations. Commonly, outliers result from measurement errors, coding, and recording errors, and, sometimes, are natural, abnormal values. Such non-representative samples can seriously affect model produced later. There are two strategies for handling outliers : Detect and eventually remove outliers as a neighborhood of preprocessing phase. And Develop robust modeling methods that are insensitive to outliers.
    2. **(ii) Scaling, encoding, and selecting features :** Data preprocessing includes several steps like variable scaling and differing types of encoding. For instance, one feature with range [0, 1] and other with range [100, 1000] will not have an equivalent weight within applied technique. They are going to also influence ultimate data-mining results differently. Therefore, it is recommended to scale them and convey both features to an equivalent weight for further analysis. Also, application-specific encoding methods usually achieve dimensionality reduction by providing a smaller number of informative features for subsequent data modeling.

3. Estimate model – The selection and implementation of acceptable data-mining technique is that main task during this phase. This process is not straightforward. Usually, in practice, implementation is predicated on several models, and selecting simplest one is a further task.

4. **Interpret model and draw conclusions –** In most cases, data-mining models should help in deciding. Hence, such models got to be interpretable so as to be useful because humans are not likely to base their decisions on complex "black-box" models. Note that goals of accuracy of model and accuracy of its interpretation are somewhat contradictory. Usually, simple models are more interpretable, but they are also less accurate. Modern data-mining methods are expected to yield highly accurate results using high dimensional models. The matter of interpreting these models, also vital, is taken into account a separate task, with specific techniques to validate results.



STATE THE PROBLEM

COLLECT THE DATA

PERFORM PREPROCESSING

ESTIMATE THE MODEL
(MINE THE DATA)

INTERPRET THE MODEL
AND
DRAW THE CONCLUSION

Classification of Data Mining Systems :

- 1. Database Technology
- 2. Statistics
- 3. Machine Learning
- 4. Information Science
- 5. Visualization

Major issues in Data Mining

1. **Mining different kinds of knowledge in databases –** The need for different users is not same. Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery tasks.
2. **Interactive mining of knowledge at multiple levels of abstraction –** The data mining process needs to be interactive because it allows users to focus on search for patterns, providing and refining data mining requests based on returned results.
3. **Incorporation of background knowledge –** To guide discovery process and to express discovered patterns, background knowledge can be used to express discovered patterns not only in concise terms but at multiple levels of abstraction.
4. **Data mining query languages and ad-hoc data mining –** Data Mining Query language that allows user to describe ad-hoc mining tasks should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
5. **Presentation and visualization of data mining results –** Once patterns are discovered it needs to be expressed in high-level languages, visual representations. These representations should be easily understandable by users.

Major issues in Data Mining

6. **Handling noisy or incomplete data –** The data cleaning methods are required that can handle noise, incomplete objects while mining data regularities. If data cleaning methods are not there then accuracy of discovered patterns will be poor.
7. **Pattern evaluation –** It refers to interestingness of problem. The patterns discovered should be interesting because either they represent common knowledge or lack of novelty.
8. **Efficiency and scalability of data mining algorithms –** In order to effectively extract information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
9. **Parallel, distributed, and incremental mining algorithms –** The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate development of parallel and distributed data mining algorithms. These algorithms divide data into partitions that are further processed parallel. Then results from partitions are merged. The incremental algorithms update databases without having mined data again from scratch.