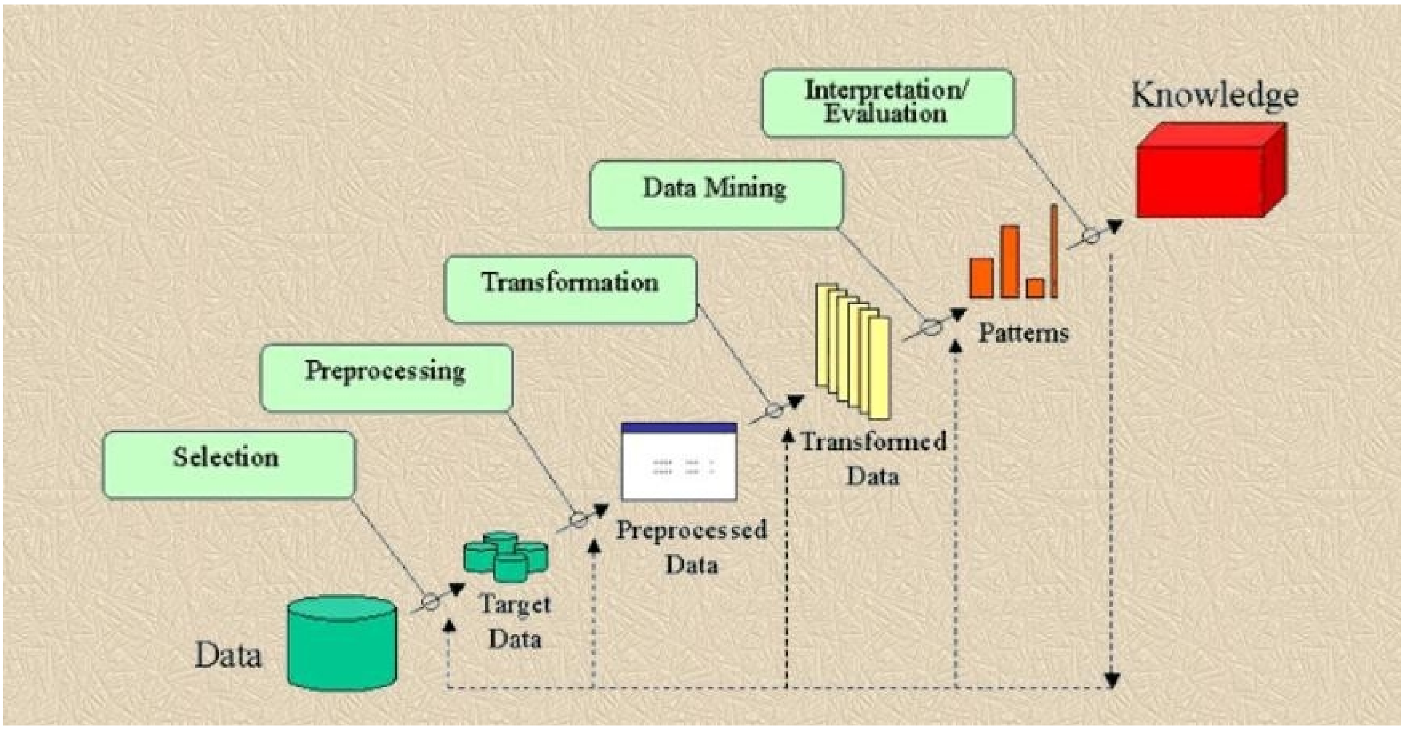
Unit 1: Introduction to BI part II

5. Knowledge discovery in databases(KDD).

The term *Knowledge Discovery in Databases*, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.

The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, subsampling, and transformations of that database.



The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:

Before starting the process, an unsaid step is Developing an understanding of

- the application domain
- the relevant prior knowledge
- the goals of the end-user

Step 1- Selection

Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.

Step 2- Preprocessing

Data cleaning and preprocessing.

- Removal of noise or outliers.
- Collecting necessary information to model or account for noise.
- Strategies for handling missing data fields.
- Accounting for time sequence information and known changes.

Step 3- Transformation

Data reduction and projection.

- Finding useful features to represent the data depending on the goal of the task.
- Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

Step 4- Data Mining

- Deciding whether the goal of the KDD process is classification, regression, clustering, etc.
- Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.

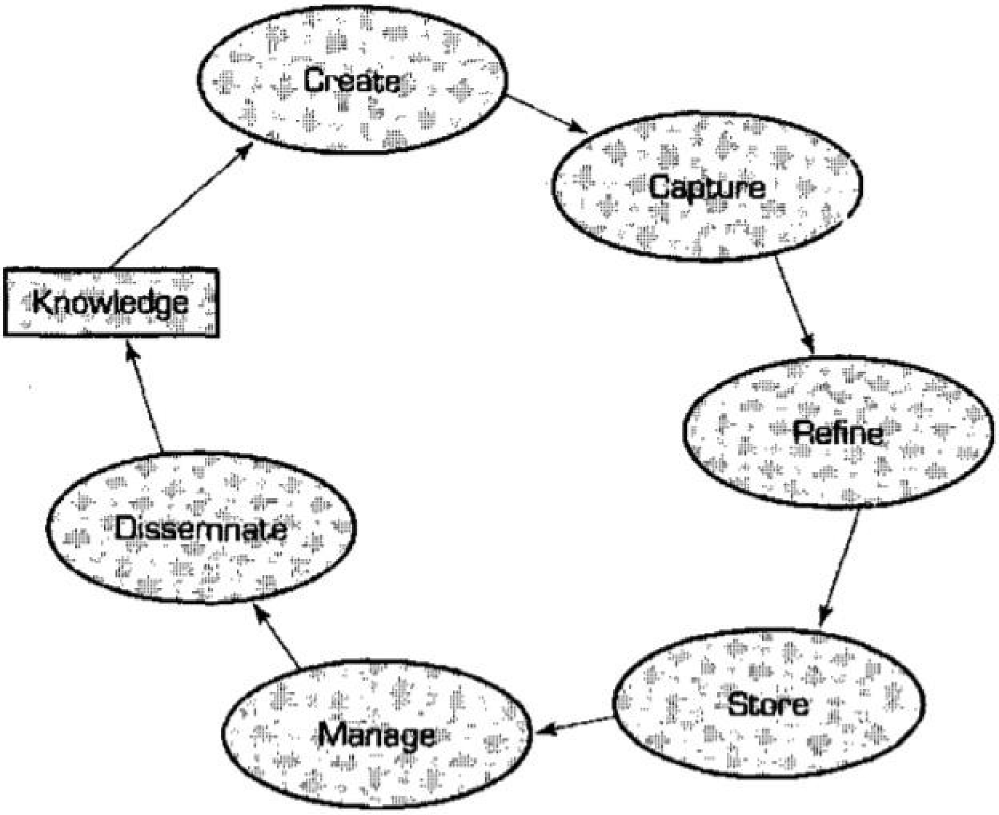
Step 5- Interpretation and Evaluation

Out of the patterns found from above step, various findings are noted down. And patterns are mined for final results that is Knowledge. All the knowledge is then consolidated.

6. **Knowledge Management Cycle.**

A functioning knowledge management system follows six steps in a cycle (Figure 9.2). The reason for the cycle is that knowledge is dynamically refined over time. The knowledge in a good KM system is never finished because the environment changes over time, and the knowledge must be updated to reflect the changes. The cycle works as follows:

1. **Create knowledge:** Knowledge is created as people determine new ways of doing things or develop know-how. Sometimes external knowledge is brought in. Some of these new ways may become best practices.



2. **Capture knowledge:** New knowledge must be identified as valuable and be represented in a reasonable way.

3. **Refine knowledge:** New knowledge must be placed in context so that it is actionable. This is where human insights (tacit qualities) must be captured along with explicit facts.

4. **Store knowledge:** Useful knowledge must then be stored in a reasonable format in a knowledge repository so that others in the organization can access it.

5. **Manage knowledge:** Like a library, the knowledge must be kept current. It must be reviewed to verify that it is relevant and accurate.

6. **Disseminate knowledge:** Knowledge must be made available in a useful format to anyone in the organization who needs it, anywhere and anytime.

As knowledge is disseminated, individuals develop, create, and identify new knowledge or update old knowledge which they replenish into the system. Knowledge is a resource that is ncft consumed when used, though it can age. (For example, driving a car in 1900 was different from driving one now, but many of the basic principles still apply.) Knowledge must be updated. Thus, the amount of knowledge grows over time.

Normalization with example.

Normalization is a database design technique which organizes tables in a manner that reduces redundancy and dependency of data.

It divides larger tables to smaller tables and links them using relationships.

1st NF:

- All values in the cell should be atomic that is single valued.
- The table must have a primary key.

2nd NF:

- It should be in 1st NF.
- All non key attributes should be dependent on Key attributes.

3rd NF:

- It should be in 2nd NF.
- It has no transitive functional dependencies.

Check the normalization example in excel sheet.