Welcome students to the unit 4 basics of data mining, module name, Data Mining and Knowledge Discovery Module number 12. I am Edwin D'Souza Associate Professor, St. Xavier's College, Mapusa Goa.

In this module we cover, Knowledge discovery in databases, namely KDD.

- The knowledge discovery process.
- Data mining as a part of knowledge discovery process.
- The iterative nature of the knowledge discovery process.

After completing this module you will be able to :

- Describe knowledge discovery in databases
- List the Steps of knowledge discovery process.
- Explain the knowledge discovery process
- Explain the iterative nature of the knowledge discovery process.
- Differentiate between data mining and knowledge discovery
- Understand the knowledge discovery process.

We are discovering knowledge from multiple databases. The multiple databases are also called data collections. Knowledge discovery in databases is also called KDD.

Data collection goes as input to the knowledge discovery process and the output is knowledge.

These are the steps of the knowledge discovery process.

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining
6. Pattern evaluation.
7. Knowledge representation

There are seven steps. The 5th step here is data mining.

Let's look at the steps off the knowledge discovery process.

The first step is, Data Cleaning :

Noise data and irrelevant data is removed from the individual data collection. Please note that there are multiple data collections from various database sources. Noise can be instances with outliers. Irrelevant data can be redundant and duplicate data.

The next step, step 2 , is data integration.
Data from multiple data sources is combined into a common source, also called a unified schema.
The multiple data collections are often heterogeneous.

For example, the individual data collections can be from various sources like Oracle, Microsoft SQL Server, Microsoft Excel, Google Forms.

Each of these have their own proprietary format to store their data, so they are heterogeneous in nature.
After completion of step one and two.

We get that is data cleaning and data integration. We are actually created the data warehouse.

The Datawarehouse is a common unified schema of data which will be the data source for the later steps of the knowledge discovery process.

The third step of the knowledge discovery process is data selection. Only the data which is relevant to the analysis is retrieved from the data warehouse in this step.

The fourth step of the knowledge discovery process is data transformation. The selected data is transformed into forms appropriate for the mining procedure.

Normalization, aggregation, generalization is done with the attributes of the data set to get the data set into appropriate format for the mining algorithms.

This is the most important step of the knowledge discovery process.

The 5th Step data mining. Here clever techniques are applied to extract patterns which are potentially useful.

Selected and transformed data from the previous step is given as input to the data mining step and potentially useful patterns outputted. An example of clever technique is Association rule mining.

The 6th step of the knowledge discovery process is pattern evaluation.

Interesting patterns representing knowledge identified based on given matches.

Potentially useful patterns which are outputted by the datamining step, are given as input to pattern evaluation and interesting patterns outputted.

In Association rule mining technique, conviction, lift are some of the metrics which are used to evaluate the patterns given by data mining.

These patterns are potentially useful patterns. The last step of the knowledge discovery process is knowledge representation. Discovered knowledge is visually represented to the user.

Visualization techniques helps users to understand and interpret the data mining results. So here in knowledge representation, interested patterns are given as input and they are visually represented as knowledge. This visual representation can be done via pie charts or bar graphs.

Knowledge discovery is a iterative process. If the user is not happy with the knowledge or the potential patterns the steps can be iteratively repeated, updating the inputs to the steps till the user is satisfied with the knowledge obtained.

Knowledge discovery as an iterative process. The pattern evaluation matches can be enhanced. Data mining can be further refined. New data can be selected or further transformed. New data sources can be integrated in order to get different and more appropriate results.

Data mining is actually just a part of the knowledge discovery process. The seven steps of the knowledge discovery process are listed here, and data mining is the fifth step. It is just a step in the knowledge discovery process.

Data mining and knowledge discovery.
We can compare the same. Data mining became the accepted customary term and very rapidly trend that even overshadowed more general terms such as knowledge discovery process, that is KDD, that describe a more complete process, even though data mining is just a step in the knowledge discovery process. The KDD process is more often known as data mining in data science.

Other terms referring to data mining are Data dredging, Knowledge extraction and pattern discovery. Here's a pictorial representation of the knowledge discovery process. We are taking the database data collection as input and the final output is knowledge. The steps cleaning and integration create the data warehouse, next we have the step selection and transformation. In data mining step Patterns output which are evaluated and represented as knowledge in the next step.

Let's summarize. KDD is a knowledge discovery in databases. The knowledge discovery process outputs knowledge from various data sources, which can be heterogeneous. But they can be homogeneous as well. Data mining is just a step of the KDD process. Data mining is more popular than the KDD process among data scientists and is used more often to describe knowledge discovery. KDD Process is an iterative process. The steps are repeated till the user is satisfied with the patterns. The knowledge output. These are my references. Thank you.