1.KDD Process:
With the advancement of technology, information is being collected and used in a very drastic way. There is an urgent need for a new generation of computational theo- ries and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD). At an abstract level, the KDD field is con- cerned with the development of methods and techniques for making sense of data.

A Brief Overview of KDD Process:
The KDD process is interactive and iterative, involving numerous steps with many decisions made by the user. Brachman and Anand (1996) give a practical view of the KDD process, emphasizing the interactive nature of the process. Here, we broadly outline some of its basic steps:

First is developing an understanding:

Developing the understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the customer's viewpoint.

Second is 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.

Third is data cleaning and preprocessing:

Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes.

Fourth is data reduction and projection: finding useful features to represent the data depending on the goal of the task. With dimensionality reduction or transformation methods, the effective number of variables under consideration can be reduced, or in- variant representations for the data can be found.

Fifth is matching the goals: Matching the goals of KDD process (step 1) to a particular data-mining method.

Sixth is exploratory analysis and model and hypothesis selection: choosing the data- mining algorithm(s) and selecting method(s) to be used for searching for data patterns. This process includes deciding which models and parameters might be appropriate (for ex- ample, models of categorical data are different than models of vectors over the reals) and matching a particular data-mining method with the overall criteria of the KDD process (for example, the end user might be more interested in understanding the model than its predictive capabilities).

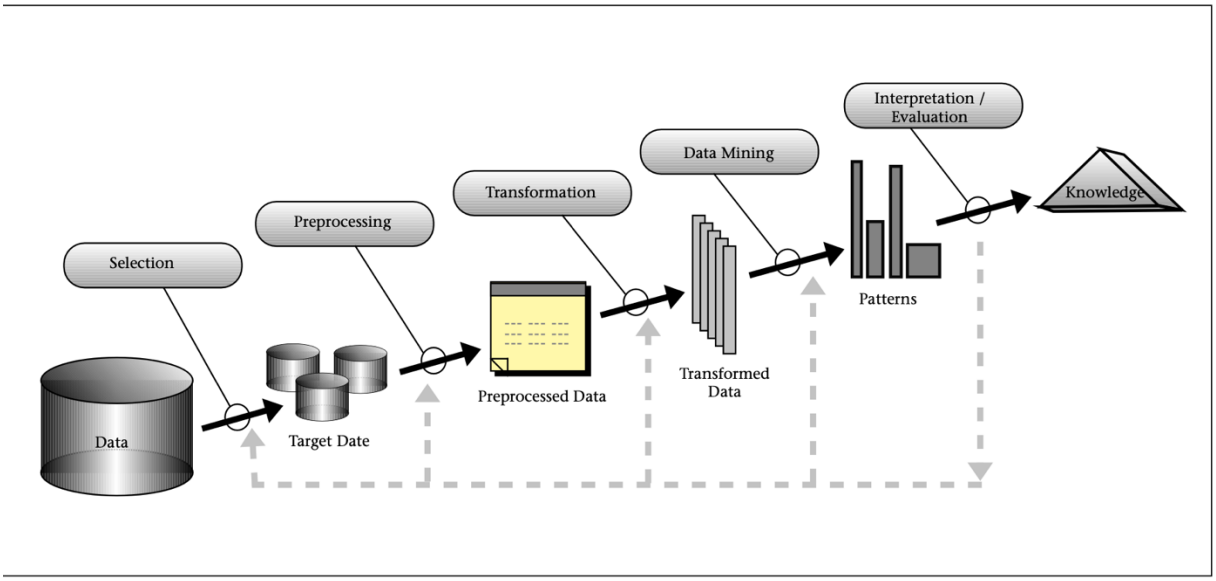
Seventh is data mining: searching for pat- terns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data-mining method by correctly performing the preceding steps.

Eighth is interpretation: Interpreting mined patterns, possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models.

Ninth is acting on the discovered knowledge: using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

The KDD process can involve significant iteration and can contain loops between any two steps. The basic flow of steps (although not the potential multitude of iterations and loops) is illustrated in figure 1. Most previous work on

KDD has focused on step 7, the data mining. However, the other steps are as important (and probably more so) for the successful application of KDD in practice. Having defined the basic notions and introduced the KDD process, we now focus on the data-mining component, which has, by far, received the most attention in the literature.



The KDD process involves using the database along with any required selection, preprocessing, subsampling, and transformations of it; applying data-mining methods (algorithms) to enumerate patterns from it; and evaluating the products of data mining to identify the subset of the enumerated patterns deemed knowledge. The data-mining component of the KDD process is concerned with the algorithmic means by which patterns are extracted and enumerated from data. The overall KDD process includes the evaluation and possible interpretation of the mined patterns to determine which patterns can be considered new knowledge. The KDD process also includes all the additional steps described in the next section.

2. We can outline some of the current primary re- search and application challenges for KDD. This list is by no means exhaustive and is in- tended to give the reader a feel for the types of problems.

- Larger Databases
- High Dimensionality
- Overfitting
- Assessing of Statistical Significance
- Changing Data and Knowledge
- Missing and Noisy Data
- Complex relationships between fields
- Understandability of Patterns
- User-interaction and Prior Knowledge
- Integration with Other Systems

In my opinion, the biggest challenge for KDD is Missing and Noisy data. This problem is especially acute in business databases. U.S. census data reportedly have error rates as great as 20 percent in some fields. Important attributes can be missing if the database was not designed with discovery in mind. Possible solutions include more sophisticated statistical strategies to identify hidden variables and dependencies (Heckerman 1996; Smyth et al. 1996). This a big challenge for KDD because when data gets lost it is extremely difficult to retrieve it again. And also when noise gets mixed into data, extracting the original version from the faulty is difficult. It can distort the data. Some most of the important part such as attributes, variables can get missing. Also removing this challenge is not fully efficient. So, there is always a chance that it can be a threat for data transmission.

3. If I am a machine learning researcher at Amazon.com, where I have to predict what a user would like to buy I would like to choose the Probabilistic Graphic Dependency Models. Suppose the problem is if a user buys body wash, then the user is likely to buy facewash, shampoo or other bathing items. So, the prediction has to be other bathing items based on the user's previous behavior.

First is developing an understanding:

Developing the understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the customer's viewpoint. Here if the user buys body wash, then the user is likely to buy facewash, shampoo or other bathing items.

Second is 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. Here the data sample will be different bathing items.

Third is data cleaning and preprocessing:

Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes. Now, the user may not buy all the bathing items. So, need to clean up the unnecessary data and keep which are more similar to the user's previous behavior.

Fourth is data reduction and projection: finding useful features to represent the data depending on the goal of the task. With dimensionality reduction or transformation methods, the effective number of variables under consideration can be reduced, or in- variant representations for the data can be found. Now there will be still multiple data which are similar to the user's previous behavior. Let's say for this problem the user need to see the items which are close range of price of his previous purchasement.

Fifth is matching the goals: Matching the goals of KDD process (step 1) to a particular data-mining method. Here I would like to choose Probabilistic Graphic Dependency Models.

Sixth is exploratory analysis and model and hypothesis selection: choosing the data- mining algorithm(s) and selecting method(s) to be used for searching for data patterns. This process includes deciding which models and parameters might be appropriate (for ex- ample, models of categorical data are different than models of vectors over the reals) and matching a particular data-mining method with the overall criteria of the KDD process (for example, the end user might be more interested in understanding the model than its predictive capabilities).

Probabilistic Graphic Dependency Models:

Graphic models specify probabilistic dependencies using a graph structure (Whittaker 1990; Pearl 1988). In its simplest form, the model specifies which variables are directly de- pendent on each other. Typically, these models are used with categorical or discrete-valued variables, but extensions to special cases, such as Gaussian densities, for real-valued variables are also possible. Within the AI and statistical communities, these models were initially developed within the framework of probabilistic expert systems; the structure of the model and the parameters (the conditional probabilities attached to the links of the graph) were elicited from experts. Recently, there has been significant work in both the AI and statistical communities on methods whereby both the structure and the parameters of graphic models can be learned directly from databases (Buntine 1996; Heckerman 1996). Model-evaluation criteria are typically Bayesian in form, and parameter estimation can be a mixture of closed-form estimates and iterative methods depending on whether a variable is directly observed or hidden. Model search can consist of greedy hill-climbing methods over various graph structures. Prior knowledge, such as a partial ordering of the variables based on causal relations, can be useful in terms of reducing

the model search space. Although still primarily in the research phase, graphic model induction methods are of particular interest to KDD because the graphic form of the model lends itself easily to human interpretation

Seventh is data mining: searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data-mining method by correctly performing the preceding steps. Here we will be using trees for the Probabilistic Graphic Dependency Models.

Eighth is interpretation: Interpreting mined patterns, possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models, or visualization of the data given the extracted models. Here, we can visualize what the user would like to buy. We will use the methods such as: (1) model representation, (2) model evaluation, and (3) search.

Ninth is acting on the discovered knowledge: using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge. Here, we can use the discovered knowledge for a more accurate prediction in future.

The KDD process can involve significant iteration and can contain loops between any two steps. The basic flow of steps (although not the potential multitude of iterations and loops) is illustrated in figure 1. Most previous work on KDD has focused on step 7, the data mining. However, the other steps are as important (and probably more so) for the successful application of KDD in practice. Having defined the basic notions and introduced the KDD process, we now focus on the data-mining component, which has, by far, received the most attention in the literature.