Notes 01-1: Introduction to Knowledge Discovery

This course is about discovering patterns in data using traditional data analysis techniques (i.e., commonly used and well-known statistical approaches) blended with sophisticated non-traditional techniques for searching and analyzing large volumes of data (i.e., data mining algorithms). *Data mining* is the process of automatically discovering useful information from, typically, large data repositories (e.g., databases, "flat" files, and streams). The process of data mining is actually one step in a much larger process called *knowledge discovery in databases* (KDD), but since data mining is such an important step, the terms are often used interchangeably.

KDD in a Nutshell

KDD, also known as *data mining*, has been almost universally accepted to be the non-trivial process of identifying previously unknown, valid, novel, potentially useful, and understandable patterns in data.

- *Non-trivial*: Implies that the process is not a simple one (i.e., it likely consumes significant time and computer resources and involves some complex process).

- *Previously unknown*: Implies that the discovered patterns represent new knowledge and/or confirm some hypothesis.

- *Valid*: Implies that the discovered patterns are reproducible and would likely be consistent with patterns discovered in similar data from other sources.

- *Novel*: Implies that the knowledge being sought is not immediately obvious and/or intuitive.

- *Potentially useful*: Implies that the knowledge provide some insight, have some applicability to solving a real-world problem, and/or affect decision-making processes.

- *Understandable*: Implies that results be suitable for human consumption.

What is the KDD Process?

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,

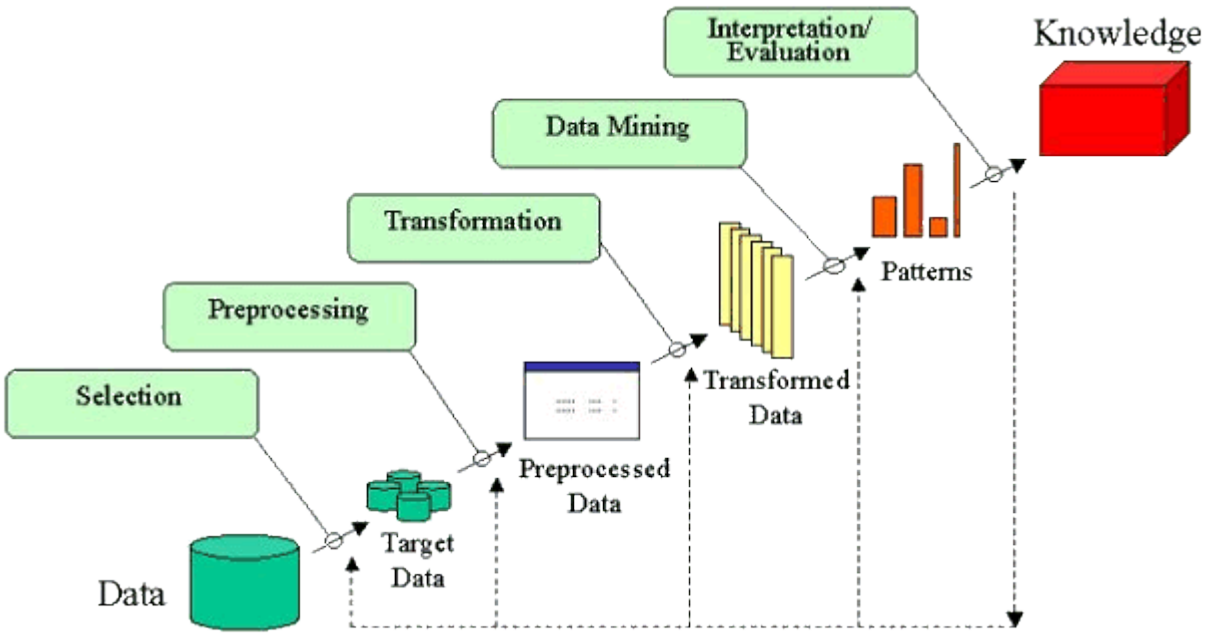
pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.

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.

More formally, perhaps, KDD is a systematic, automated process of autonomous generation and verification of new hypotheses that seeks to discover any articulated and justified truth about a domain represented by a formal language.

An Outline of the Steps of the KDD Process



The process of KDD typically includes many steps, but at a minimum, the following steps are usually required:

1. *Application*: Plan to use newly discovered knowledge as part of some problem-solving or decision-making process.

2. *Cleaning*: Attempts to eliminate errors, omissions, and inconsistencies through data hygiene operations.

3. *Pre-processing*: Re-formatting of the data to make it "easier" to process in subsequent steps.

4. *Integration*: Combines data from multiple sources.

5. *Selection*: Relies on tools such as SQL or a query-by-example template to describe the characteristics of the data to be retrieved from the data repository.

6. *Transformation*: Attempts to reduce the volume of data that needs to be considered in subsequent steps by aggregating and summarizing the data into a form suitable for the chosen data mining technique.

7. *Choosing the data mining task*: Deciding whether the goal of the KDD process is classification, regression, clustering, etc.

8. *Choosing the data mining algorithm(s)*: Selecting the method(s) to be used for searching for patterns in the data. Deciding which models and parameters may be appropriate. Matching a particular data mining method with the overall criteria of the KDD process.

9. *Data mining*: Applies a specific algorithm to search for, identify, and extract patterns. 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.

10. *Interpretation and evaluation*: Attempts to understand the results of the mining step by identifying "interesting" results (e.g., deviation from expected results, deviation from established norms, and quantifying the relative degree to which a result is surprising).

11. *Presentation*: Utilizes visualization and knowledge representation techniques to present the results in a manner suitable for human understanding.

12. *Refinement and repetition*: Attempts to use the newly discovered knowledge to improve the quality or focus of the mining step to make it more relevant to the problem-solving or decision-making process.

The Difference Between Knowledge Discovery and Data Mining

The terms *knowledge discovery* and *data mining* are distinct.

KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.

Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

Definitions Related to the KDD Process

Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

Data	A set of facts, *F*.
Pattern	An expression *E* in a language *L* describing facts in a subset F_E of *F*.
Process	KDD is a *multi-step process* involving data preparation, pattern searching, knowledge evaluation, and refinement with iteration after modification.
Valid	Discovered patterns should be true on new data with some degree of certainty. Generalize to the future (other data).
Novel	Patterns must be novel (should not be previously known).
Useful	Actionable; patterns should potentially lead to some useful actions.
Understandable	The process should lead to human insight. Patterns must be made understandable in order to facilitate a better understanding of the underlying data.

Interestingness is an overall measure of pattern value, combining validity, novelty, usefulness, and simplicity.