**KDD Process**

**Knowledge Discovery Database in Data Mining**

**Data Mining** also known as Knowledge Discovery in Databases, refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases. The process of finding and interpreting patterns from data involves the repeated application of the following steps:

**Developing an understanding of:**

* The application domain
* Relevant prior knowledge
* The goals of the end-user

**Creating a target dataset:**

Selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.

**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.

**Data reduction and projection:**

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

**Choosing the data mining task:**

* Deciding whether the goal of the KDD process is [**classification**](https://data-flair.training/blogs/classification-algorithms/), regression, [**clustering**](https://data-flair.training/blogs/cluster-analysis-data-mining/), etc.

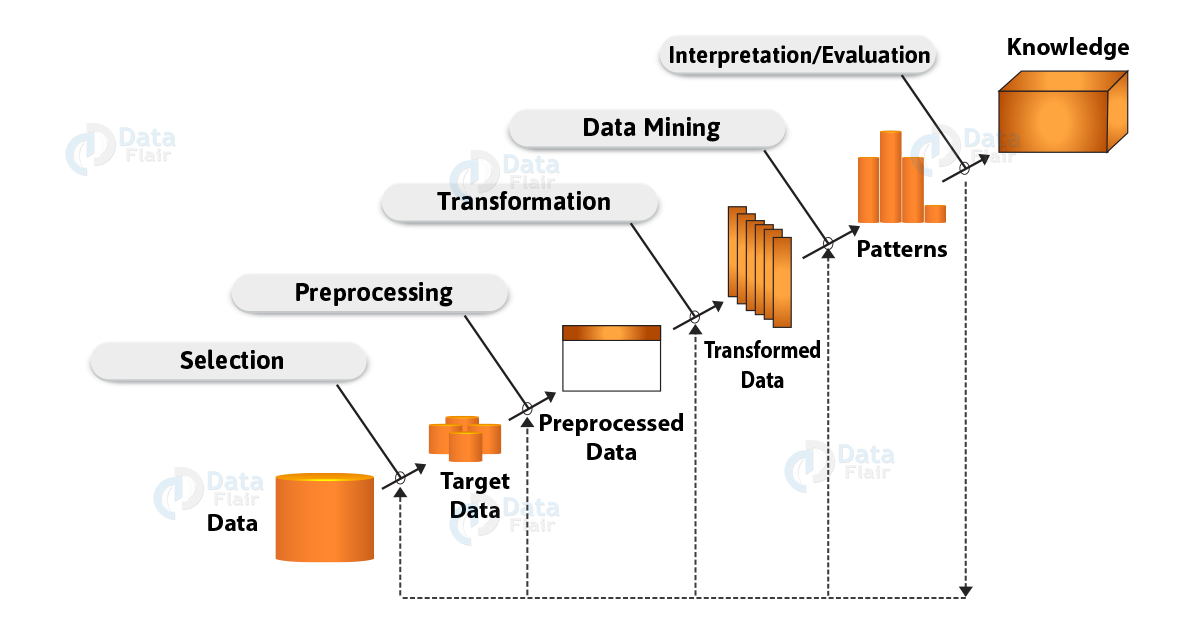
**Choosing the data mining algorithm(s):**

* Selecting 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 criteria of the KDD process.

**Data Mining:**

* Searching for patterns of interest in a particular representational form. Such representations as classification rules or trees, regression, clustering, and so forth.
* Interpreting mined patterns.
* Consolidating discovered knowledge.

**Steps Involved in KDD Process:**



1. **Data Cleaning**: Generally, the data we have collected is not clean and may contain errors, missing values, noisy or inconsistent data. Therefore we need to apply different techniques to get rid of such anomalies. Data cleaning is defined as removal of noisy and irrelevant data from collection.
   * Cleaning in case of **Missing values**.
   * Cleaning **noisy** data, where noise is a random or variance error.
   * Cleaning with **Data discrepancy detection** and **Data transformation tools**.
2. **Data Integration**: First of all the data is collected and integrated from all the different sources. Data integration is defined as heterogeneous data from multiple sources combined in a common source (Data Warehouse).
   * Data integration using **Data Migration tools**.
   * Data integration using **Data Synchronization tools**.
   * Data integration using **ETL** (Extract-Load-Transformation) process.
3. **Data Selection**: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
   * Data selection using **Neural network**.
   * Data selection using **Decision Trees**.
   * Data selection using **Naive bayes**.
   * Data selection using **Clustering**, **Regression**, etc.
4. **Data Transformation**: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure. Basically, the data even after cleaning is not ready for mining. Also, we need to transform them into forms appropriate for mining. Thus, the techniques used to do this are smoothing, aggregation, normalization etc.
5. **Data Transformation is a two step process:**
   * **Data Mapping**: Assigning elements from source base to destination to capture transformations.
   * **Code generation**: Creation of the actual transformation program.
6. **Data Mining**: Data mining is defined as clever techniques that are applied to extract patterns potentially useful. As now in this step, we are ready to apply [**data mining techniques**](https://data-flair.training/blogs/data-mining-techniques/) on the data. Basically, it is to discover the interesting patterns. Hence, clustering and association analysis are among the many different techniques present. Also, as we used for data mining.
   * Transforms task relevant data into **patterns**.
   * Decides purpose of model using **classification** or **characterization**.
7. **Pattern Evaluation**: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures. Generally, this step includes visualization, transformation, removing redundant patterns from the patterns we generated.
   * Find **interestingness score** of each pattern.
   * Uses **summarization** and **Visualization** to make data understandable by user.
8. **Knowledge representation**: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results. As this step is beneficial to us. Also, it helps to use the knowledge acquired to take better decisions.
   * Generate **reports**.
   * Generate **tables**.
   * Generate **discriminant rules**, **classification rules**, **characterization rules**, etc.