# Data Mining

**Why data mining?**

Today, huge amount of data is collected every day. Because of the many advances in data collection, there has been an explosive growth of data. The amount of data collection, the better data storage abilities we have today and because data is more available have made it so that traditional data analysis may not always work. Because of this, data mining is needed which is an automated analysis of massive data sets. Data mining blends traditional data analysis methods with new algorithms that can process large volumes of data.

**What is data mining?**

Data mining is the process of automatically discover useful and interesting information from large data sets. Data mining is “knowledge mining from data” and is a technique for extracting interesting (non-trivial, previously unknown) patterns or knowledge from a lot of data. Data mining is important today since we have a lot of data and it can be difficult to find useful information in that amount of data with traditional techniques.

Data mining tasks could be divided into two major categories:

* Descriptive tasks – the objective is to derive patterns that summarize the underlying relationships in the data.
* Predictive tasks – the objective is to predict the value of the target based on the values of the attribute.

**What is the purpose of data mining?**

The purpose of data mining is to find interesting information that was previously unknown from large databases. Data mining is used when traditional data analysis may not work, for example when the amount of data is huge, of high dimensionality and/or when data is of high complexity. The purpose of data mining is to find interesting patterns and knowledge (previously unknown) from large amount of data.

**Typical process for the knowledge discovery process (KDD)**

Data mining is the core of the KDD process, which is an iterative sequence of the following steps:

1. Data cleaning - first data is cleaned to remove noise and inconsistent data. Missing data is handled
2. Data integration – multiple data sources are combined
3. Data selection – data that is relevant to the analysis task are retrieved from the database, task-relevant data is selected
4. Data transformation – data is transformed in a way so it is appropriate for the specific data mining analysis
5. Data mining – this is where data mining is used to find the interesting patterns
6. Pattern evaluation – the extracted patterns from the data mining step is evaluated to identify the interesting patterns based on interestingness measures
7. Knowledge presentation – visualization and other knowledge presentation techniques are used to present the found knowledge to users

Step 1-4 are different forms of preprocessing the data. This is where data is prepared for the mining and is actually the hardest part of the process.

**Dirty data and preprocessing**

Data preprocessing is important because data in the real world is dirty.

One problem where data is dirty is when it is incomplete. Missing values is very common and is a huge problem that need to be handled before analysis. Incomplete data could be when an attribute is missing some values for certain observations. This may have been caused in the process of collecting data, when for example information was not collected for all observation. Another type of incomplete data is when some of the attributes of interest is missing. A simple example of this could be when the difference of salaries between women and men are to be analyzed and there’s no gender attribute is in the data. Another problem with incomplete data is when only aggregated data is saved. Incomplete data may come from when data can not be collected for some attribute because it’s not applicable for all observations. It could also be due to human error or software problems.

Another type of dirty data is noisy data that contains a lot of extreme values (outliers) or errors. Noise is a random error or variance in a measured attribute. The reason for noisy data could be that there are errors when data is collected or that a few observations differ greatly from the others. Noisy data may come from human or computer error at data entry or broken collection instruments.

Data could also be inconsistent, like containing disagreements between different attributes for an observation. An example of this is when age=”42”, but birth year is “2018”. Inconsistent data could also be due to a change in collecting data, like when a ranking system has changed the way it ranks, ex was rating “1,2,3”, but changed to “A,B,C”. Inconsistent data may be due to different data sources of because linked data was modified.

Dirty data is a problem in all kinds of analysis and is the reason why preprocessing is needed in data mining. If the data is of bad quality, the data mining results will also be of bad quality.

Forms of data preprocessing – cleaning, integration, transformation and reduction

**What kinds of data can be mined?**

Data mining can be used on any kind of data as long as the data is meaningful for the task at hand. The most common forms of data to be mined are database data, data warehouse data and transactional data.

Data warehouse – a repository of information collected from multiple sources stored under a unified representation.

Transactional database – contains data from transactions

**What kind of patterns can be mined?**

Concept/class – Data entries can be associated with classes or concepts. We can describe individual classes and concepts in summarized terms. This is called class/concept descriptions.

* Characterization – summarizing the data of the class under study in general terms. A summary of general characteristics or features of the target class of data. Ex characteristics of customers spending more than 10 000 kr per year
* Discrimination – comparing target class with other (contrasting) classes. Ex Compare the characteristics of products that had a sales increase to products that had a sales decrease last year

Frequent patterns, association, correlations – frequent patterns are patters that occur frequently in the data

Classification and prediction – construct models (functions) that describe and distinguish classes or concepts for future prediction. The derived model is based on analyzing training data. Predict some unknown or missing numerical values.

Cluster analysis – class label is unknown: group data to form new classes. Maximizing intra-class similarity and minimizing interclass similarity

Outlier analysis – detect outliers in the data

**When are patterns interesting?**

Interesting patterns represent knowledge. When data mining is used, a lot of patterns can be found, but not all of them are interesting. Only a small number of patterns are really that interesting for the task.

A pattern is interesting if we can understand it, it’s valid on new data (test data), if it’s useful for the task and if it confirms a hypothesis that the used wants to confirm. But the concept of “interesting” can be hard to define. That’s why there’s measures in data mining called “interestingness measures” in data mining. There’s two types of this, objective and subjective interestingness measures.

Subjective measures are based on what the user believes is important in the data, like if there’s unexpected patterns or if the patterns give the user information on how to act (actionability).

Objective measures on the other hand, are not based on the user’s belief and is only based on the structures and statistics of the patterns. Some example of objective interestingness measures is support and confidence in association analysis.

Objective measures help to find interesting patterns but are not enough. We also need subjective measures that reflect the user’s interests. Also, many patterns that are objectively interesting may just represent common sense and therefore actually be uninteresting.