Project on data Analytics

A project on data analytics does not imply only the use of one or more specific methods. It implies:

- understanding the problem to be solved
- defining the objectives of the project
- looking for the necessary data
- preparing these data so that they can be used
- identifying suitable methods and choosing between them
- tuning the hyper-parameters of each method (see below)
- analyzing and evaluating the results
- redoing the pre-processing tasks and repeating the experiments
- and so on.

Little History on Methodologies for Data Analytics

Machine learning, knowledge discovery from data and related areas experienced strong development in the 1990s. Both in academia and industry.

In the mid-1990s, both in academia and industry, different methodologies were presented.

The most successful methodology from academia came from the USA. This was the KDD process

The most successful tool from industry, was and still is the CRoss-Industry Standard Process for Data Mining (CRISP-DM)

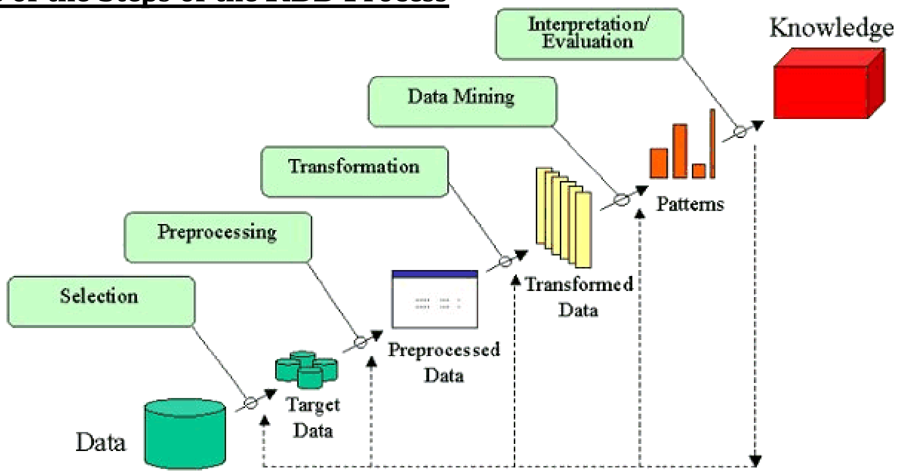
Other methodologies exist. Some of them are domain-specific: they assumethe use of a given tool for data analytics

This is not the case for SEMMA, which, despite has been created by SAS, is tool independent. Each letter of its name, SEMMA, refers to one of its five steps: Sample, Explore, Modify, Model and Assess.

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.

An Outline of the Steps of the KDD Process



The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:

1. Developing an understanding of
    - the application domain
    - the relevant prior knowledge
    - the goals of the end-user

2. 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.

3. 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.

4. Data reduction and projection.
    - Finding useful features to represent the data depending on the goal of the task.
    - Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

5. Choosing the data mining task.
    - Deciding whether the goal of the KDD process is classification, regression, clustering, etc.

6. 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 overall criteria of the KDD process.

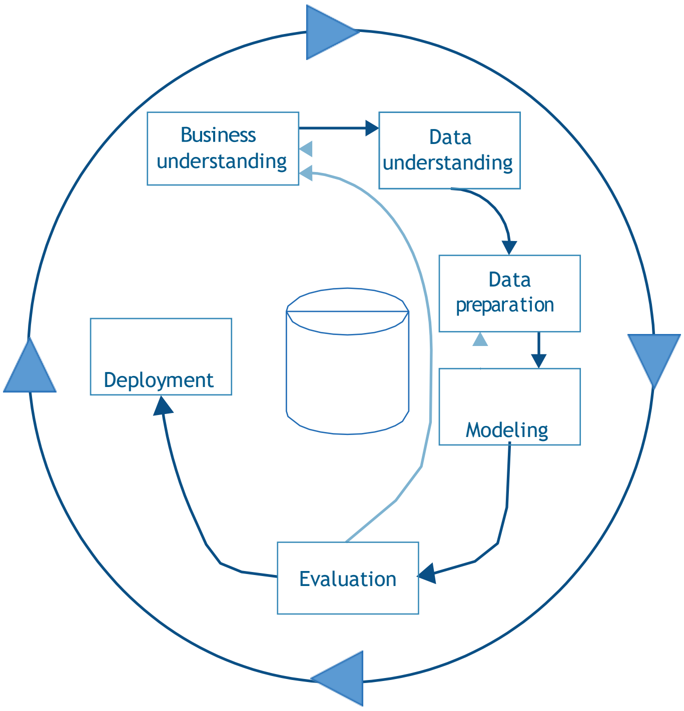
7. Data mining.
    - Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.

8. Interpreting mined patterns.

9. Consolidating discovered knowledge.

The CRISP-DM Methodology

CRoss-Industry Standard Process for Data Mining (CRISP-DM) is a six-step method, which, like the KDD process, uses a non-rigid sequential framework. Despite the six phases, CRISP-DM is seen as a perpetual process, used through- out the life of a company in successive iterations (Figure 1.3).



The CRISP-DM methodology

The six phases are:

1) *Business understanding:* This involves understanding the business domain, being able to define the problem from the business domain perspective, and finally being able to translate such business problems into a data analytics problem.
2) *Data understanding:* This involves collection of the necessary data and their initial visualization/summarization in order to obtain the first insights, par- ticularly but not exclusively, about data quality problems such as missing data or outliers.
3) *Data preparation:* This involves preparing the data set for the modeling tool, and includes data transformation, feature construction, outlier removal, missing data fulfillment and incomplete instances removal.
4) *Modeling:* Typically there are several methods that can be used to solve the same problem in analytics, often with specific data requirements. This implies that there may be a need for additional data preparation tasks that are method specific. In such case it is necessary to go back to the previous step. The modeling phase also includes tuning the hyper-parameters for each of the chosen method(s).
5) *Evaluation:* Solving the problem from the data analytics point of view is not the end of the process. It is now necessary to understand how its use ismeaningful from the business perspective; in other words, that the obtained solution answers to the business requirements.
6) *Deployment:* The integration of the data analytics solution in the business process is the main purpose of this phase. Typically, it implies the integration of the obtained solution into a decision-support tool, website maintenance process, reporting process or elsewhere.