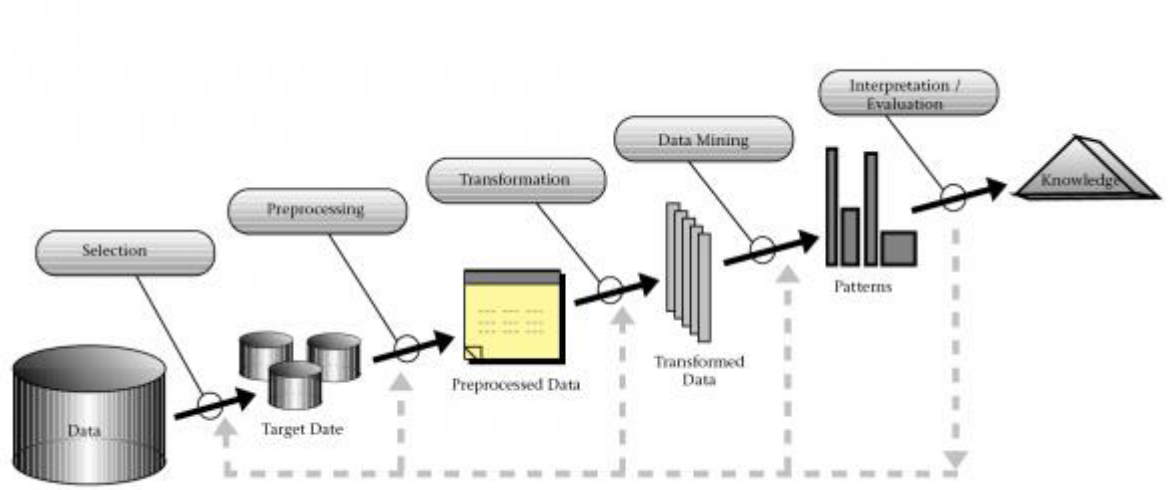
Data Mining

Process model

The KDD process model

1. Selection
2. Preprocessing
3. Transformation
4. *Data Mining*
5. Interpretation / Evaluation



Fayyad, U. M. et al. 1996. From data mining to knowledge discovery: an overview. In Fayyad, U. M. et al (Eds.), Advances in knowledge discovery and data mining. AAAI Press / The MIT Press.

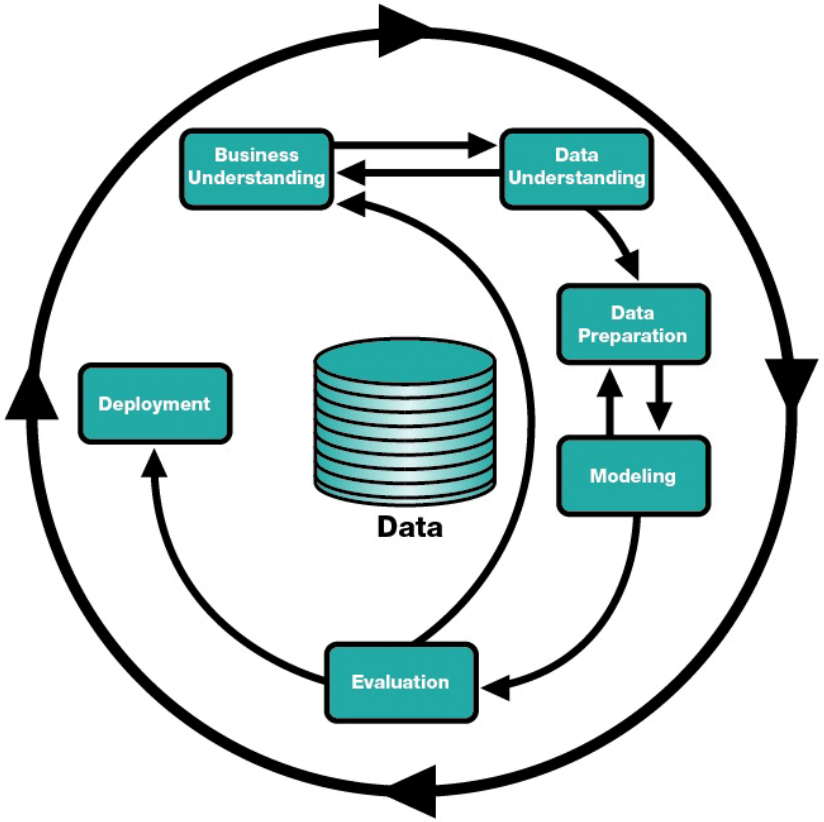
The KDD process model

- **Selection**: identification and selection of all external and internal sources of information and selection of the subset of data or variables needed for the KDD process.
- **Preprocessing**: includes the removal of data with extreme values (outliers), filling in missing values, etc.
- **Transformation**: converting data into a format suitable for Data Mining algorithms.
- **Data Mining**: in this step the specialized tools seek, through specialized algorithms, existing patterns in the data. This search can be performed automatically or interactively systems through the aid of the analyst responsible for the generation of the hypotheses. At the end of the process, the DM system should generate a report of the analysis carried out in order to enable analysts to verify the results obtained.
- **Interpretation/evaluation**: this step should be performed in conjunction with business analysts. If the knowledge generated is not satisfactory, analysts can form a new set of experiments giving rise to a new iteration of the process.

The CRISP-DM process model

CRISP-DM (CRoss Industry Standard Process for Data Mining)

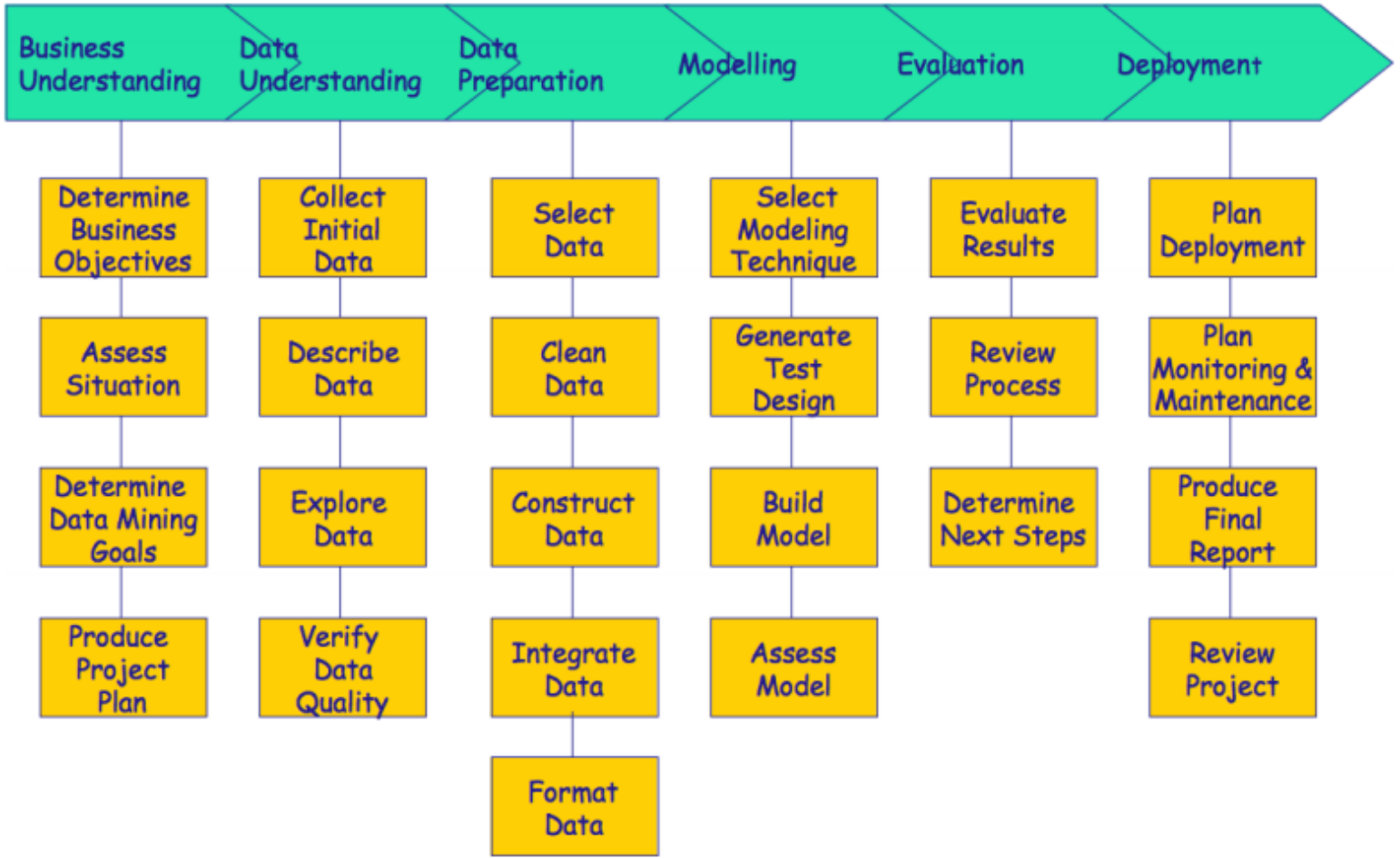
1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment



The CRISP-DM process model

1. **Understand the Business**: focuses on understanding the objective of the project from a business perspective, defining a preliminary plan to achieve the goals.
2. **Understand the data**: data gathering and early activities to better understand the data, identifying problems or interesting sets.
3. **Data preparation**: construction of the final data set from the initial one. Normally occurs several times in the process.
4. **Modeling**: several modeling techniques are applied, and its parameters calibrated for optimization. Thus, it is common to return to Data Preparation during this phase.
5. **Evaluation**: a model that seems to have great quality in a data analysis perspective was constructed. However, it is necessary to verify if the model reaches the goals of the business.
6. **Deployment**: the knowledge acquired by the model is organized and presented in a way that the customer can use.

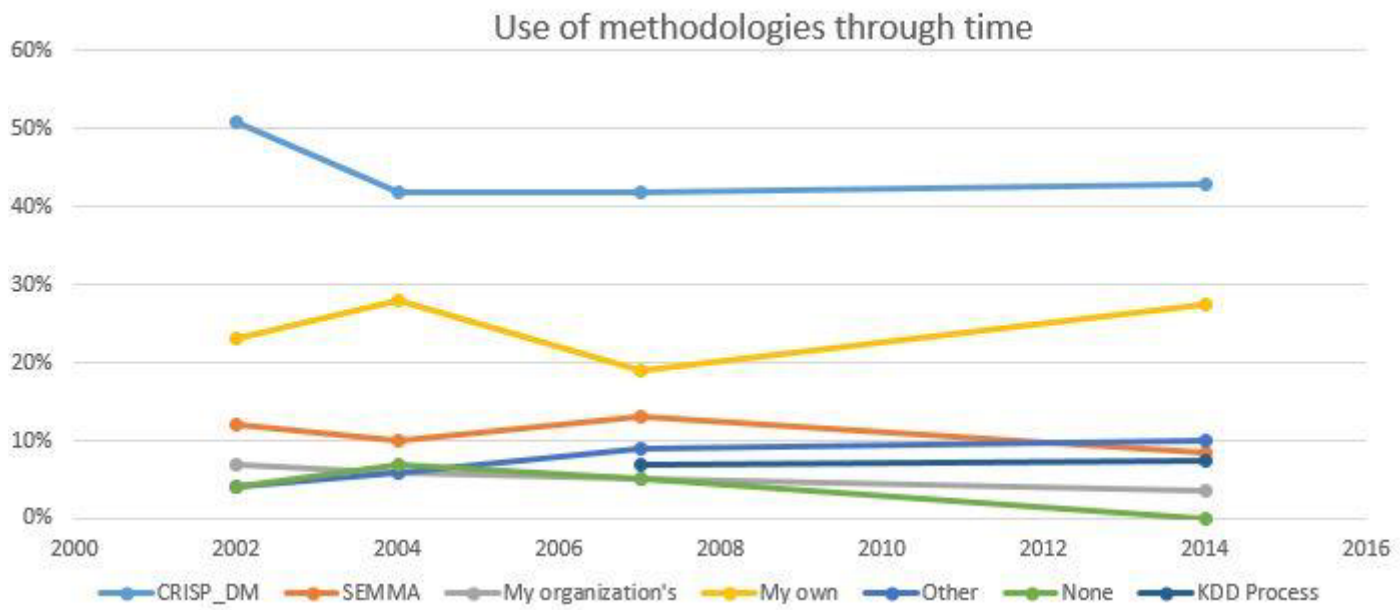
The CRISP-DM process model



Process models

KDD	CRISP-DM
Pre-process KDD	Business understanding
Selection	Data understanding
Preprocessing	
Transformation	Data preparation
Data Mining	Modeling
Interpretation/Evaluation	Evaluation
Post-process KDD	Deployment

Process models



References

Books

- João Moreira, André de Carvalho, Tomas Horvath, A general introduction to data analytics, Wiley, 2018: o ebook está disponível na biblioteca da FEUP
- Matthew North; Data mining for the masses, 2012. ISBN: 0615684378: está no moodle de IACEC
- Aggarwal Charu C.; Data mining. ISBN: 978-3-319-14142-8