cct | College Dublin
Computing • IT • Business

Machine Learning for Data Analysis
MSc in Data Analytics
CCT College Dublin

Project Management Methodologies
Week 5

Lecturer: Dr. Muhammad Iqbal*

Email: miqbal@cct.ie

Agenda

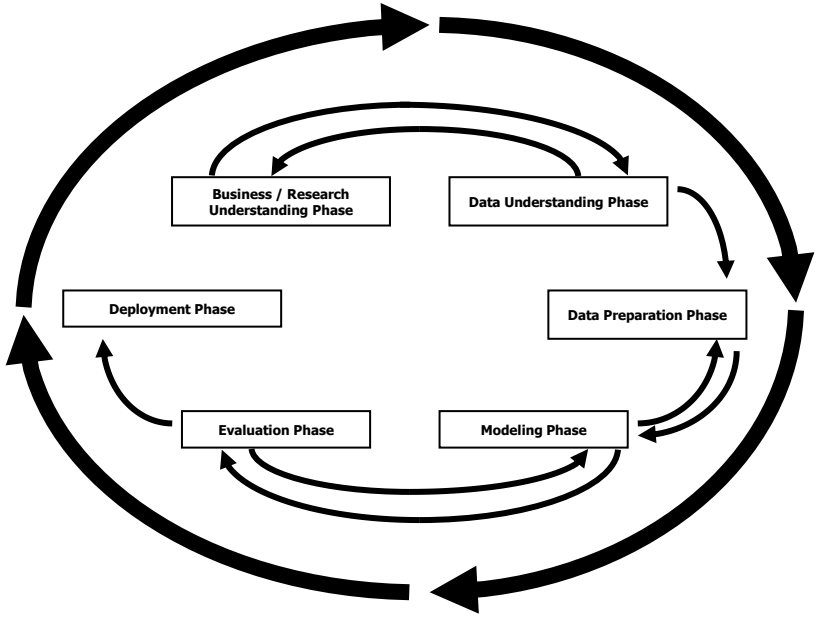
cct College Dublin
Computing • IT • Business

- CRISP-DM
- KDD
- SEMMA
- Project management methodologies in Machine learning

Cross Industry Standard Process
CRISP-DM

cct College Dublin
Computing • IT • Business

- **Cross-Industry Standard Process for Data Mining (CRISP-DM) developed in 1996**
  - Fits data exploration into the general problem-solving strategy of business/research unit
  - Industry, tool and application neutral
  - Data exploration projects follow iterative, adaptive life cycle consisting of 6 phases



CRISP-DM Lifecycle

- Iterative CRIP-DM process shown in outer circle
- Most significant dependencies between phases shown
- Next phase depends on results from preceding phase
- Returning to earlier phase possible before moving forward

Cross Industry Standard Process
CRISP-DM

cct | College Dublin
Computing • IT • Business

1. Business/ Research Understanding Phase
- Define project requirements and objectives
- Translate objectives into data exploration problem definition
- Prepare preliminary strategy to meet objectives

2. Data Understanding Phase
- Collect data
- Perform exploratory data analysis (EDA)
- Assess data quality
- Optionally, select interesting subsets

3. Data Preparation Phase
- Prepares for modeling in subsequent phases
- Select cases and variables appropriate for analysis
- Cleanse and prepare data so it is ready for modeling tools
- Perform transformation of certain variables, if needed

4. Modeling Phase
- Select and apply one or more modeling techniques
- Calibrate model settings to optimize results

5. Evaluation Phase
- Evaluate one or more models for effectiveness
- Determine whether defined objectives achieved
- Make decision regarding data exploration results before deploying to field

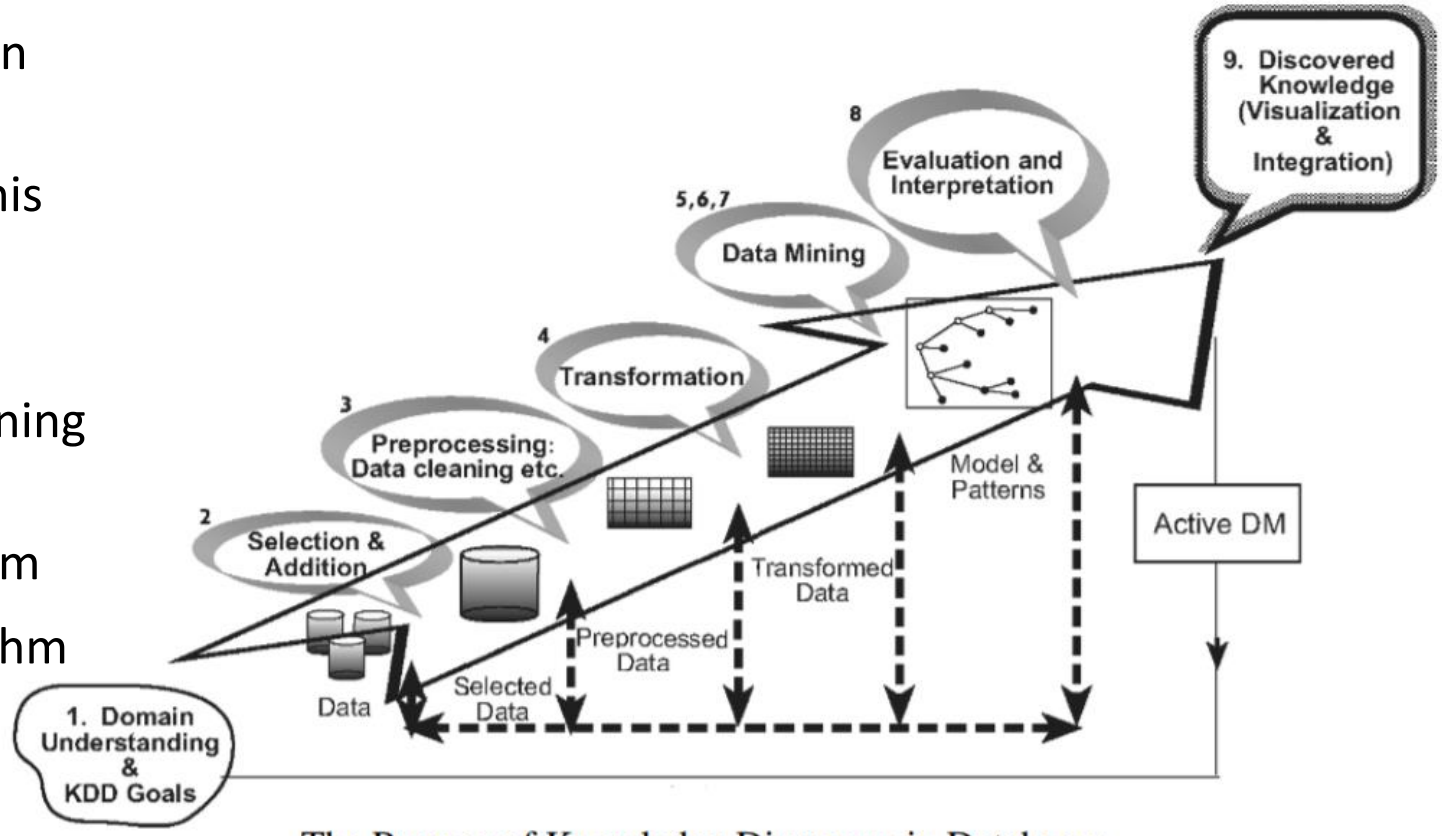
6. Deployment Phase
- Make use of models created
- Simple deployment example: generate report
- Complex deployment example: implement parallel data exploration effort in another department
- In businesses, customer often carries out deployment based on your model

Knowledge Discovery in Databases

KDD

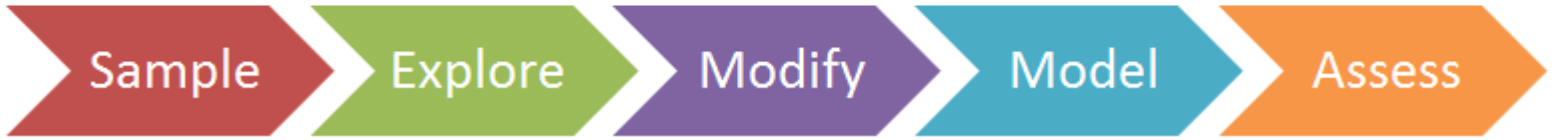
cct College Dublin
Computing • IT • Business

1. Developing an understanding of the application domain
2. Selecting and creating a data set on which discovery will be performed
3. Pre-processing and cleansing. In this stage, data reliability is enhanced
4. Data transformation
5. Choosing the appropriate Data Mining task
6. Choosing the Data Mining algorithm
7. Employing the Data Mining algorithm
8. Evaluation
9. Using the discovered knowledge

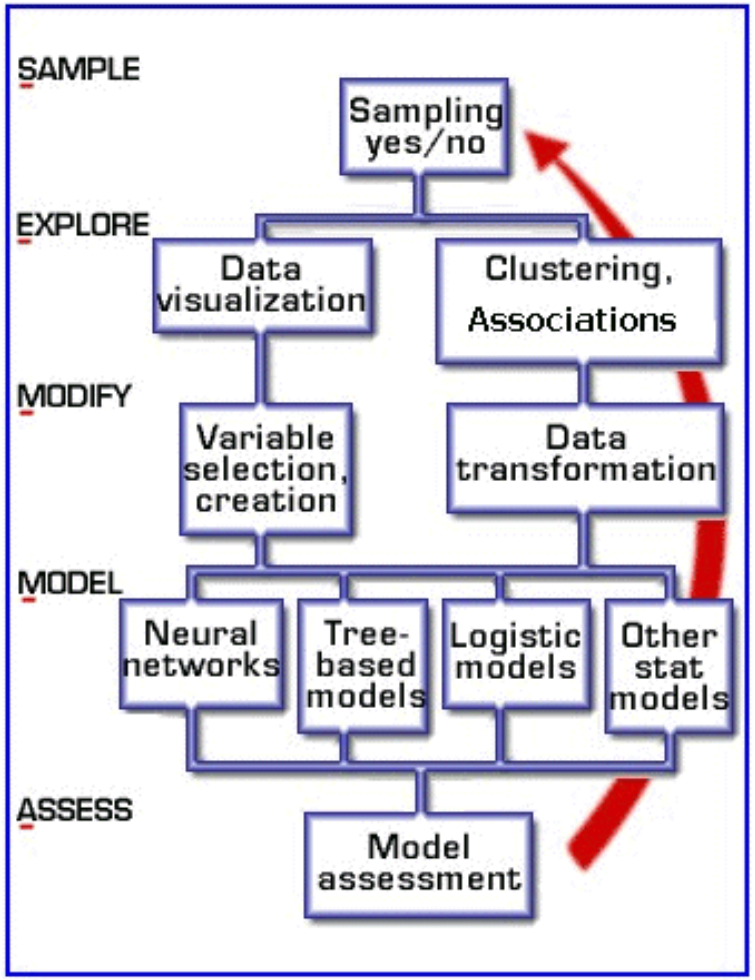


The Process of Knowledge Discovery in Databases.

SEMMA



- A graphical user interface (GUI) provides a user-friendly front end to the SEMMA data mining process:
- Sample the data by creating one or more data tables. The samples should be large enough to contain the significant information, yet small enough to process.
- Explore the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas.
- Modify the data by creating, selecting, and transforming the variables to focus the model selection process.
- Model the data by using the analytical tools to search for a combination of the data that reliably predicts a desired outcome.
- Assess the data by evaluating the usefulness and reliability of the findings from the data mining process.



Project Management Methodologies

cct | College Dublin
Computing • IT • Business

- There are several data mining processes and machine learning projects that can be applied to modern Data Science/ Machine Learning projects.
- We discussed the most common of them are CRISP-DM, KDD and SEMMA.

Resources/ References

- Introduction to Machine Learning with Python, Andreas C. Müller and Sarah Guido, O'Reilly Media, Inc. October 2016.
- Data Science project management methodologies | by Quantum | DataDrivenInvestor