1.1 Introduction

Data Mining (DM)

- Non-trivial process of identifying **implicit**, **valid**, **novel** (measured by comparing to expected values), **potentially useful** and **understandable** patterns in date
- Step in the KDD process

CRISP-DM

Cross-Industry Standard Process for Data Mining

Objectives

- cheaper, faster and more reliable DM
- widespread adoption
- reduce skills required for data mining
- capture experience for reuse

Characteristics

- Non-proprietary
- application/industry neutral
- tool neutral
- focus on business issues
- framework for guidance
- experience based

Phases and Tasks

1. **Business Understanding**
    - Determine Business Objectives, Assess Situation, Determine Data Mining Goals, Produce Project Plan
2. **Data Understanding**
    - Collect Initial Data, Describe Data, Explore Data, Verify Data Quality
3. **Data Preparation**
    - Select Data, Clean Data, Construct Data, Integrate Data, Format Data
4. **Modelling**
    - Select Modeling Technique, Generate Test Design, Build Model, Assess Model
5. **Evaluation**
    - Evaluate Results, Review Process, Determine Next Steps

6. **Deployment**
    - Plan Deployment, Plan Monitoring & Maintenance, Produce Final Report, Review Project

SCRUM-DM

Scrum work management + CRISP-DM Data Mining

- 3 phases: Business Understanding, Sprint, Deployment
- 6 concepts: PO, SM, Dev. Team, DM Story, PBL, SBL

Last Remarks

- A DM project should always start with an analysis of the data with traditional query tools
    - 80% can be extracted with SQL
    - 20% (hidden information) requires more advanced techniques