CRISP-DM

UNIVERSITÀ DI PISA

What is CRISP-DM?

- It stays for **CRoss-Industry Standard Process for Data Mining**
- A methodology that provides a structured approach to planning a **data mining** project
- An open standard process model that describes common approaches used by data mining experts
- Introduced in 1996 and widely adopted
- IBM incorporated the CRISP-Dm model in its SPSS Modeler product

Why Should There be a Standard Process?

- The data mining process must be **reliable** and **repeatable** by people with little data mining background.

Why Should There be a Standard Process?

- Framework for **recording experience**
  - Allows projects to be replicated
  - Aid to project planning and management
  - “Comfort factor” for new adopters
  - Demonstrates maturity of Data Mining
  - Encourage **best practices** and help to obtain better results

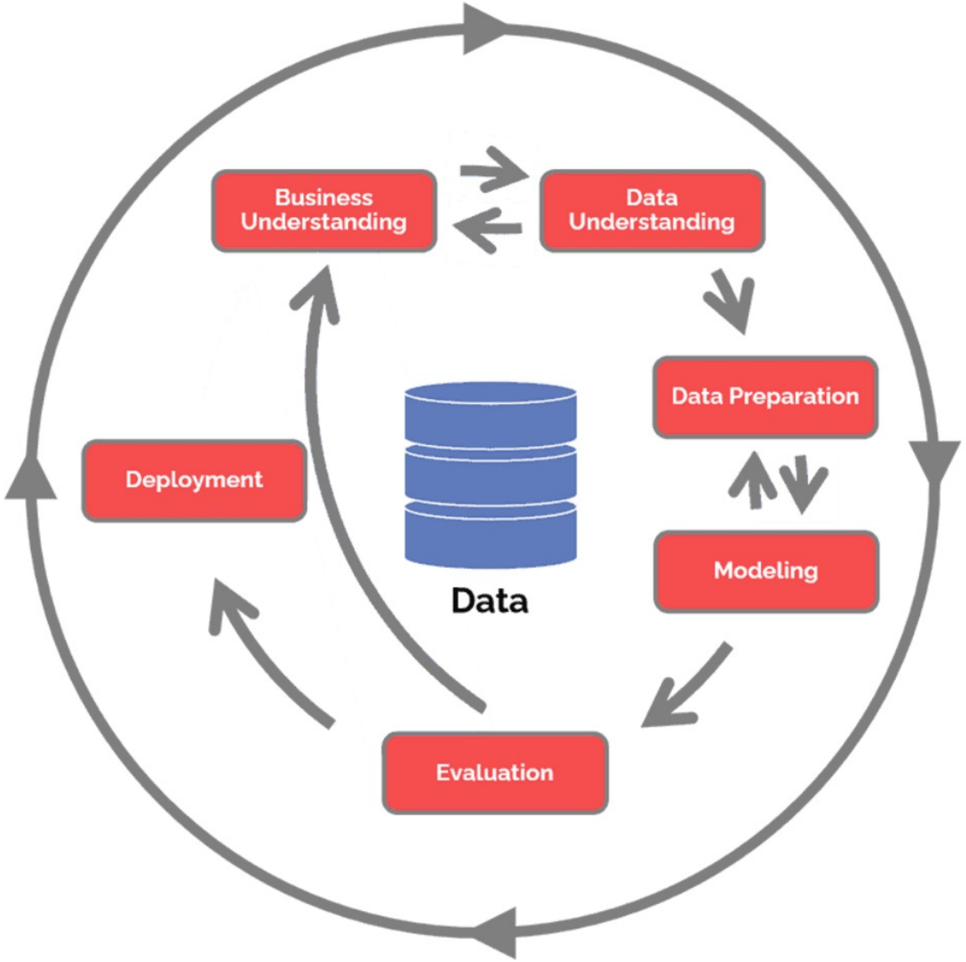
Properties of the methodology

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
    - As well as technical analysis
- Framework for guidance
- Experience base
    - Templates for Analysis

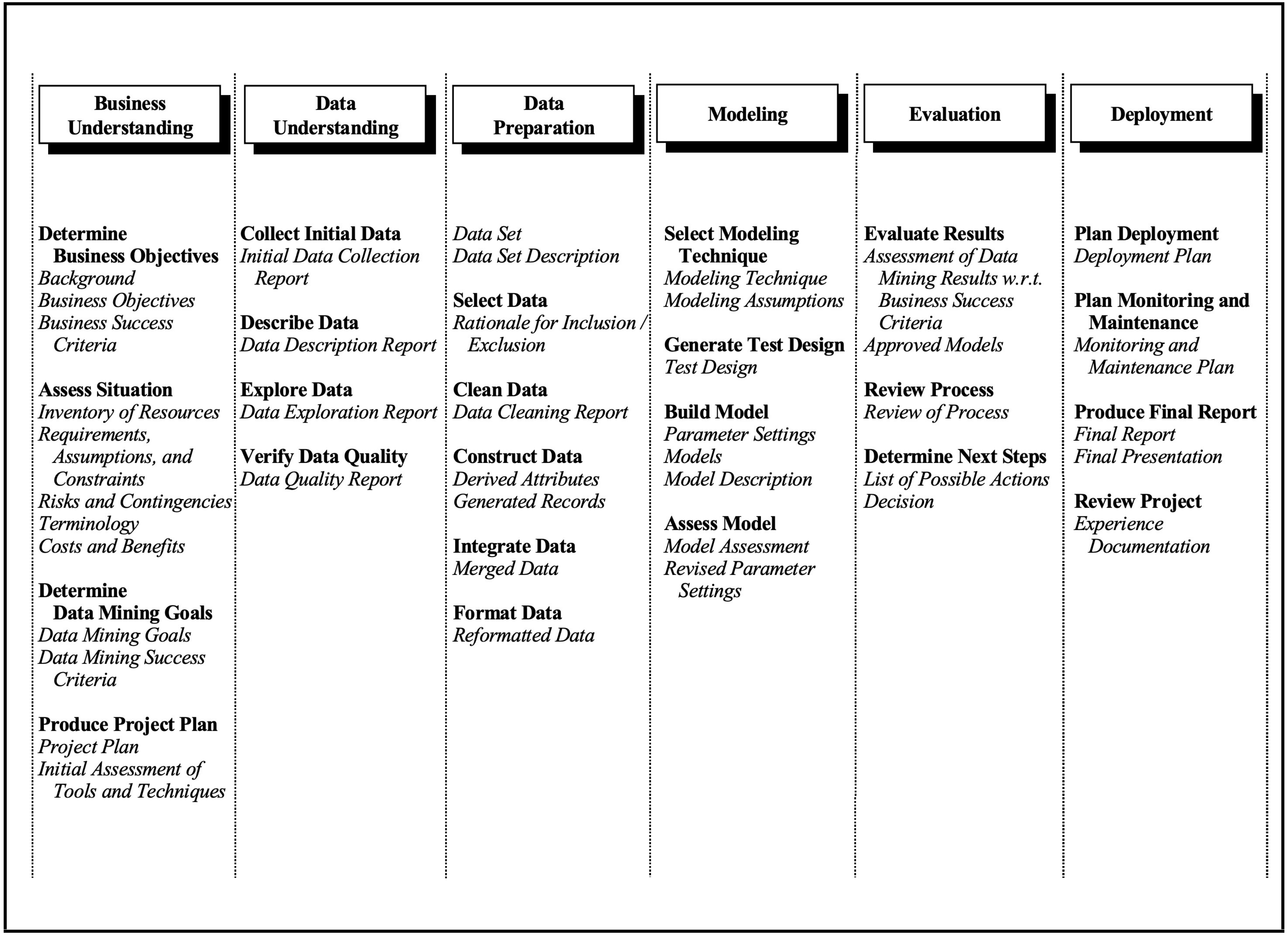
Six Phases

A set of guardrails to help you plan, organize, and implement your data mining project

CRISP-DM breaks down the life cycle of a data mining project into six phases.



Phases & Tasks



Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion / Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Phase 1 - Business Understanding

- Statement of Business Objective
  - States goal in business terminology
- Statement of Data Mining objective
  - States objectives in technical terms
- Statement of Success Criteria

GOAL: Focuses on **understanding the project objectives** and **requirements from a business perspective**, then **converting** this knowledge **into a data mining problem** definition and a **preliminary plan** designed to achieve the objectives

Questions:

- What the client really wants to accomplish?
- Which are important factors (constraints, competing objectives, ...)?
  - constraints, competing objectives to be balances

Phase 1 - Business Understanding - Determine business objectives

- Thoroughly **understand**, from **a business perspective**, what the client really wants to accomplish
- Describe the **primary objective** from a business perspective
- **Uncover important factors**, at the beginning, that can **influence the outcome of the project**

Example

Primary goal

- keep current customers by predicting when they are prone to move to a competitor

Related Business Questions

- Does the channel used affect whether customers stay or go?
- Will lower ATM fees significantly reduce the number of high-value customers who leave?

Phase 1 - Business Understanding - Determine business objectives

- **Determine business objectives**
  - Key persons and their roles?
  - Is there a steering committee. Internal sponsor (financial, domain expert)
  - Business units impacted by the project (sales, finance,...) ?
  - Business success criteria and who assesses it?
  - Users’ needs and expectations
  - Describe problem in general terms. Business questions, Expected benefits.

Phase 1 - Business Understanding - Assess situation

- **Inventory of resources** – List the resources available to the project including:
  - Personnel (business experts, data experts, technical support, data mining experts)
  - Data (fixed extracts, access to live, warehoused, or operational data)
  - Computing resources (hardware platforms)
  - Software (data mining tools, other relevant software)
- **Requirements, assumptions and constraints**
  - Schedule of completion
  - Required comprehensibility and quality of results
  - Any data security concerns as well as any legal issues
  - Make sure that you are allowed to use the data
  - List the assumptions made by the project (non-verifiable and verifiable by DM)
- **Risks**
  - List the risks or events that might delay the project or cause it to fail. List the corresponding contingency plans – what action will you take if these risks or events take place?
- **Costs and benefits**
  - Construct a cost-benefit analysis for the project which compares the costs of the project with the potential benefits to the business if it is successful.

Phase 1 - Business Understanding – DM goals

- A business goal states **objectives in business terminology**
- A data mining goal states **objectives in technical terms**

A business goal: “Increase catalog sales to existing customers.”

A data mining goal: “Predict how many widgets a customer will buy, given their purchases over the past three years, demographic information (age, salary, city) and the price of the item.”

- Specify data mining problem type (e.g., classification, prediction and clustering)
- Specify criteria for model assessment – criteria of assessment e.g., accuracy of predictive task

Phase 1 - Business Understanding – Project Plan

- **Describe the intended plan** for achieving the data mining goals and thereby achieving the business goals.
- The plan should **specify the anticipated set of steps to be performed** during the rest of the project including an initial selection of tools and techniques
- **Project Plan:**
  - Stages of the project with duration, resources, input, output and dependencies
  - Analysis of dependencies between time schedule and risks
  - Identify actions and recommendations if the risks are manifested
  - Decide the valuation strategy will be used in the evaluation phase
- This document is dynamic because at the end of phase there is a review
- of the progress and achievements and the plan should be updated

Phase 1 - Business Understanding – Project Plan

- **Initial assessment of tools and techniques**
  - select a data mining tool that supports various methods for different stages of the process.
  - It is important to assess tools and techniques early in the process since the selection of tools and techniques may influence the entire project.

Phase 2. Data Understanding

- Acquire the data
  - Document locations, methods for colletion, problems enountered and solutions achieved
- Describe data
  - Document the description of their structure, attributes, properties accessibility
- Explore data & Verify data quality
  - All the analysis described in the data understanding

Phase 3 - Data Preparation

Covers all activities to construct the final dataset from the initial raw data

- Select data
- Clean data
- Construct data
- Integrate data
- Format Data

Phase 4 – Modeling

- **Select modeling techniques:** Determine which algorithms to try (e.g. regression, neural net).
  - Select technique
  - Identify any built-in assumptions made by the technique about the data (e.g. quality, format, distribution).
  - Compare these assumptions with those in the Data Description Report and make sure that these assumptions hold.
  - Preparation Phase if necessary.

Phase 4 – Modeling

- **Generate test design:**
  - Before actually building a model generate a procedure or mechanism to test the model’s quality and validity
  - **Example**: In classification, it is common to use error rates as quality measures for data mining models. Therefore, typically separate the dataset into train and test set, build the model on the train set and estimate its quality on the separate test set
  - **Describe the intended plan** for train, test and evaluate the models
    - How to divide the dataset into training, test and validation sets
    - Decide on necessary steps (number of iterations, number of folds etc.)

Phase 4 – Modeling

- **Build model:**
  - Set initial parameters and document reasons for choosing those values
  - Run the selected technique on the input dataset
  - Post-process data mining results (eg. editing rules, display trees)
  - Record parameter settings used to produce the model
  - Describe the model, its special features, behaviour and interpretation

- **Assess model:**
  - Evaluate result with respect to evaluation criteria.
  - Rank results with respect to success and evaluation criteria and select best models Interpret results in business terms. Get comments by domain experts. Check plausibility of model
  - Check model against given knowledge base (*discovered info. novel and useful?)*
  - Check result reliability. Analyze potentials for deployment of each result

Phase 5 – Evaluation

- **Evaluate results**
- **Review process**
- **Determine next steps**

- Thoroughly **evaluate the model** and **review the steps executed to construct the model** to be certain it properly achieves the business objectives.
- A key objective is to **determine** if there is some **important business issue that has not been sufficiently considered**.
- At the end of this phase, a **decision on the use of the data mining results** should be reached

Phase 5 – Evaluation – Evaluate results

- Assesses the degree to which the model **meets the business objectives**
- Rank results according to **business success criteria.**
- Seeks to determine if there is some business reason **why this model is deficient**
- Test the model(s) on **test applications** in the real application if time and budget constraints permit
- Assesses other data mining results generated
- Unveil additional challenges, information or hints for **future directions**

Phase 5 – Evaluation – Review process

- **Summarize the process review**
  - Some activities are missed?
  - Some acvities should be repeated?
- **Overview data mining process**
  - Is there any overlooked factor or task?
  - **Example**: *did we correctly build the model*? *Did we only use attributes that we are allowed to use and that are available for future analyses*?
- Identify **failures**, misleading steps, possible alternative actions, unexpected paths
- Review data mining results with respect to business success

Phase 5 – Evaluation – Next Steps & Decision

- **Determine next steps**
  - Analyze potential for deployment of each result.
  - Estimate potential for improvement of current process.
  - Check remaining resources to determine if they allow additional process iterations (or whether additional resources can be made available)
  - Recommend alternative continuations. Refine process plan.
- **Decision**
  - According to the results and process review, it is decided how to proceed to the next stage (remaining resources and budget)
  - Rank the possible actions. Select one of the possible actions.
  - Document reasons for the choice.

Phase 6 – Deployment

- Determine **how** the results need to be utilized
- **Who** needs to use them?
- **How often** do they need to be used
- Deploy Data Mining results

The knowledge gained will need to **be organized and presented in a way that the customer can use it**. However, **depending on the requirements**, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

Phase 6 – Deployment

- **Plan deployment**
  - in order to deploy the data mining result(s) into the business, **takes the evaluation results and concludes a strategy for deployment**
  - **document the procedure** for later deployment
  - **Identify possible problems** when deploying the data mining results
- **Plan monitoring and maintenance**
  - **helps to avoid unnecessarily long periods of incorrect usage of data mining results**
  - needs a detailed on monitoring process for performance of the models
  - takes into account the specific type of deployment
  - Consider the **change over the time**

Phase 6 – Deployment

- **Produce final report**
    - the project leader and his team **write up a final report**
    - Identify reports needed (*slide presentation*, *management summary*, *detailed findings, explanation of models*, etc.)
    - How well initial data mining goals have been met.
    - **Identify target groups for reports.** Outline structure and contents of reports.
    - **Select findings** to be included in the reports. Write a report
- **Review project**
    - **Interview people involved in project. Interview end users**.
        - *What could have been done better?*
        - *Do they need additional support?*
        - **Summarize feedback** and write the experience documentation
    - **Analyze the process**
        - what went right or wrong, what was done well and what needs to be improved
    - **Document the specific data mining process**
        - How can results and experience of applying the model be fed back into the process?.
        - Abstract from details to make the experience useful for future projects.

Summary

- The data mining process must be reliable and repeatable by people with little data mining skills

- CRISP-DM provides a uniform framework for
  - guidelines
  - experience documentation

- CRISP-DM is flexible to account for differences
  - Different business/agency problems
  - Different data

References

- CRISP-DM 1.0 - Step-by-step data mining guide
- Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler)
- http://www.crisp-dm.org/CRISPWP-0800.pdf
- The CRISP-DM Model: The New Blueprint for Data Mining, Colin Shearer, JOURNAL of Data Warehousing, Volume 5, Number 4, pag. 13-22, 2000
- Introduction to Data Mining, Prof. Chris Clifton, http://www.cs.purdue.edu/homes/clifton/cs490d/Process.ppt
- CRISP – DM, Yi-Li, http://www.cs.ualberta.ca/~yli/CRISPDM.ppt