Data Mining

Part 1. Introduction

1.4 CRISP-DM

Fall 2009

Instructor: Dr. Masoud Yaghini

Outline

- Introduction
- Phase 1. Business Understanding
- Phase 2. Data Understanding
- Phase 3. Data Preparation
- Phase 4. Modeling
- Phase 5. Evaluation
- Phase 6. Deployment
- References

Introduction

Introduction

- CRISP-DM:
  - CRoss-Industry Standard Process for Data Mining
- The data mining process must be reliable and repeatable by people with little data mining skills
- Why should there be a standard process?
  - Framework for recording experience and allows projects to be replicated
  - Aid to project planning and management
  - Comfort factor for new adopters with little data mining background and reduces dependency on “stars”

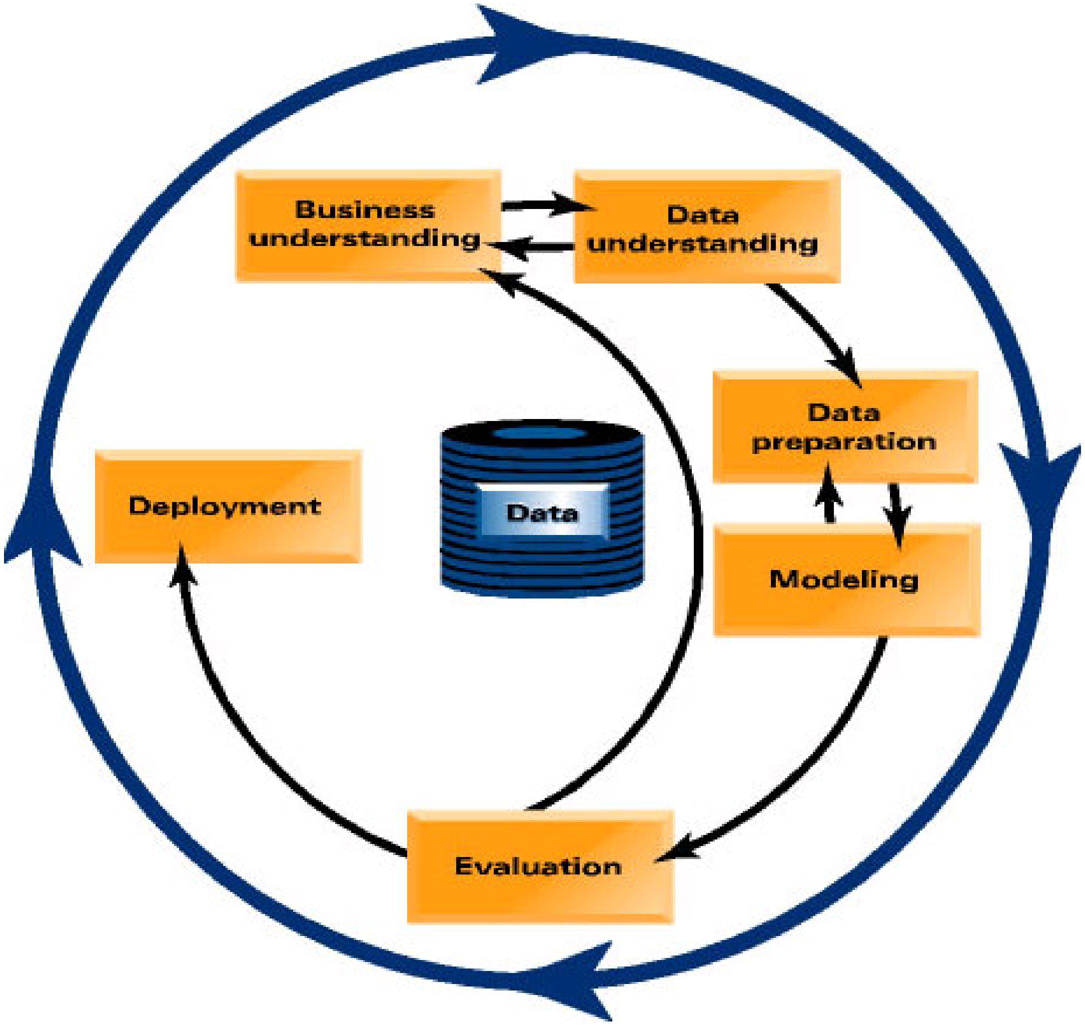
Process Standardization

- Initiative launched in late 1996 by three experienced organizations in data mining market:
  - Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) , NCR
- Developed and refined through series of workshops (from 1997-1999)
- Over 300 organization contributed to the process model.
- Published **CRISP-DM 1.0** (1999)
- Over 200 members of the CRISP-DM special interest group (SIG) worldwide:
  - **DM Vendors**: SPSS, NCR, IBM, SAS, SGI, Data Distilleries, Syllogic, etc.
  - **System Suppliers / consultants :** Cap Gemini, ICL Retail, Deloitte & Touche, etc.
  - **End Users**: BT, ABB, Lloyds Bank, AirTouch, Experian, etc.

CRISP-DM: Overview

- Data Mining methodology
- Process Model
- For anyone
- Provides a complete blueprint
- Life cycle: 6 phases

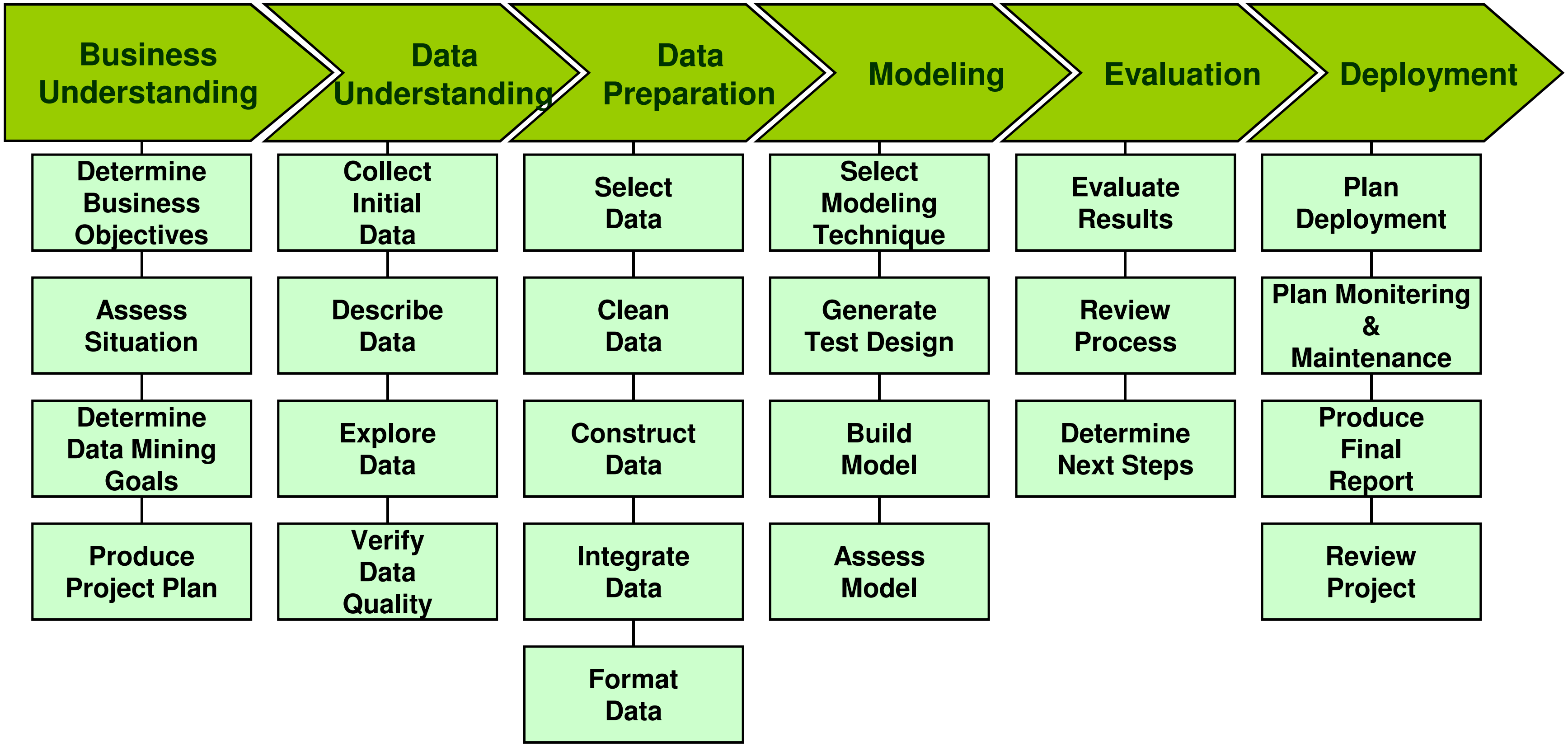
CRISP-DM: Overview



CRISP-DM: Phases

- Business Understanding
  - Project objectives and requirements understanding, Data mining problem definition
- Data Understanding
  - Initial data collection and familiarization, Data quality problems identification
- Data Preparation
  - Table, record and attribute selection, Data transformation and cleaning
- Modeling
  - Modeling techniques selection and application, Parameters calibration
- Evaluation
  - Business objectives & issues achievement evaluation
- Deployment
  - Result model deployment, Repeatable data mining process implementation

Phases and Tasks



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives	Collect Initial Data	Select Data	Select Modeling Technique	Evaluate Results	Plan Deployment
Assess Situation	Describe Data	Clean Data	Generate Test Design	Review Process	Plan Monitering & Maintenance
Determine Data Mining Goals	Explore Data	Construct Data	Build Model	Determine Next Steps	Produce Final Report
Produce Project Plan	Verify Data Quality	Integrate Data	Assess Model		Review Project
		Format Data			

Phase 1. Business Understanding

Phase 1. Business Understanding

- 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
- **Tasks:**
  - Determine business objectives
  - Assess situation
  - Determine data mining goals
  - Produce project plan

Phase 1. Business Understanding

- Determine business objectives
  - understanding the project objectives and requirements from a business perspective
  - thoroughly understand what the client really wants to accomplish
  - uncover important factors, at the beginning can influence the outcome of the project
  - neglecting this step is to expend a great deal of effort producing the right answers to the wrong questions

Phase 1. Business Understanding

- Assess situation
  - more detailed fact-finding about all of the resources, constraints, assumptions and other factors that should be considered

Phase 1. Business Understanding

- Determine data mining goals
  - a business goal states objectives in business terminology
  - a data mining goal states project objectives in technical terms
  - Example:
    - the 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.”

Phase 1. Business Understanding

- Produce project plan
  - describe the intended plan for achieving the data mining goals and 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

Phase 2. Data Understanding

Phase 2. Data Understanding

- Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
- **Tasks:**
  - Collect initial data
  - Describe data
  - Explore data
  - Verify data quality

Phase 2. Data Understanding

- Collect initial data
  - acquire within the project the data listed in the project resources
  - includes data loading if necessary for data understanding
  - possibly leads to initial data preparation steps
  - if acquiring multiple data sources, **integration** is an additional issue, either here or in the later data preparation phase

Phase 2. Data Understanding

- Describe data
  - examine the “gross” or “surface” properties of the acquired data
  - report on the results

Phase 2. Data Understanding

- Explore data
  - tackles the data mining questions, which can be addressed using **querying**, **visualization** and **reporting** including:
    - distribution of key attributes, results of simple aggregations
    - relations between pairs or small numbers of attributes
    - properties of significant sub-populations, simple statistical analyses
  - may address directly the data mining goals
  - may contribute to or refine the data description and quality reports
  - may feed into the transformation and other data preparation needed

Phase 2. Data Understanding

- Verify data quality
  - examine the quality of the data, addressing questions such as:
    - “Is the data complete?”, Are there missing values in the data?”

Phase 3. Data Preparation

Phase 3. Data Preparation

- Takes usually over 90% of the time
- Covers all activities to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order.
- **Tasks:**
  - Select data
  - Clean data
  - Construct data
  - Integrate data
  - Format data

Phase 3. Data Preparation

- Select data
  - decide on the data to be used for analysis
  - criteria include relevance to the data mining goals, quality and technical constraints such as limits on data volume or data types
  - covers selection of attributes as well as selection of records in a table

Phase 3. Data Preparation

- Clean data
  - raise the data quality to the level required by the selected analysis techniques
  - may involve selection of clean subsets of the data, the insertion of suitable defaults or more ambitious techniques such as the estimation of **missing data** by modeling

- Construct data
  - constructive data preparation operations such as the production of **derived attributes**, entire new records or **transformed values** for existing attributes

Phase 3. Data Preparation

- Integrate data
  - methods whereby information is combined from multiple tables or records to create new records or values

- Format data
  - formatting transformations refer to primarily syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool

Phase 4. Modeling

Phase 4. Modeling

- Various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary.
- **Tasks:**
  - Select modeling technique
  - Generate test design
  - Build model
  - Assess model

Phase 4. Modeling

- Select modeling technique
  - select the actual modeling technique that is to be used
  - Example: **decision tree**, **neural network**
  - if multiple techniques are applied, perform this task for each techniques separately

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

Phase 4. Modeling

- Build model
  - run the modeling tool on the prepared dataset to create one or more models

Phase 4. Modeling

- Assess model
  - interprets the models according to his domain knowledge, the data mining success criteria and the desired test design
  - judges the success of the application of modeling and discovery techniques more technically
  - contacts business analysts and domain experts later in order to discuss the data mining results in the business context
  - only consider models whereas the evaluation phase also takes into account all other results that were produced in the course of the project

Phase 5. Evaluation

Phase 5. Evaluation

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

Phase 5. Evaluation

- Evaluate results
  - assesses the degree to which the model meets the business objectives
  - 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
  - also assesses other data mining results generated
  - unveil additional challenges, information or hints for future directions

Phase 5. Evaluation

- Review process
  - do a more thorough review of the data mining engagement in order to determine if there is any important factor or task that has somehow been overlooked
  - review the quality assurance issues
  - Example: “Did we correctly build the model?”

Phase 5. Evaluation

- Determine next steps
  - decides how to proceed at this stage
  - decides whether to finish the project and move on to deployment if appropriate or whether to initiate further iterations or set up new data mining projects
  - include analyses of remaining resources and budget that influences the decisions

Phase 6. Deployment

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 by
  - Scoring a database, utilizing results as business rules, interactive scoring on-line
- 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

- **Tasks:**
  - Plan deployment
  - Plan monitoring and maintenance
  - Produce final report
  - Review project

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

Phase 6. Deployment

- Plan monitoring and maintenance
  - important if the data mining results become part of the day-to-day business and it environment
  - helps to avoid unnecessarily long periods of incorrect usage of data mining results
  - `needs a detailed on monitoring process
  - takes into account the specific type of deployment

Phase 6. Deployment

- Produce final report
  - the project leader and his team write up a final report
  - may be only a summary of the project and its experiences
  - may be a final and comprehensive presentation of the data mining result(s)

- Review project
  - assess what went right and what went wrong, what was done well and what needs to be improved

References

References

- Pete Chapman, et al. "**CRISP-DM 1.0 - Step-by-step Data Mining Guide**", 2000.

The end