DATA MINING 2

CRoss-Industry Standard Process for Data Mining

Riccardo Guidotti

a.a. 2019/2020

UNIVERSITÀ DI PISA

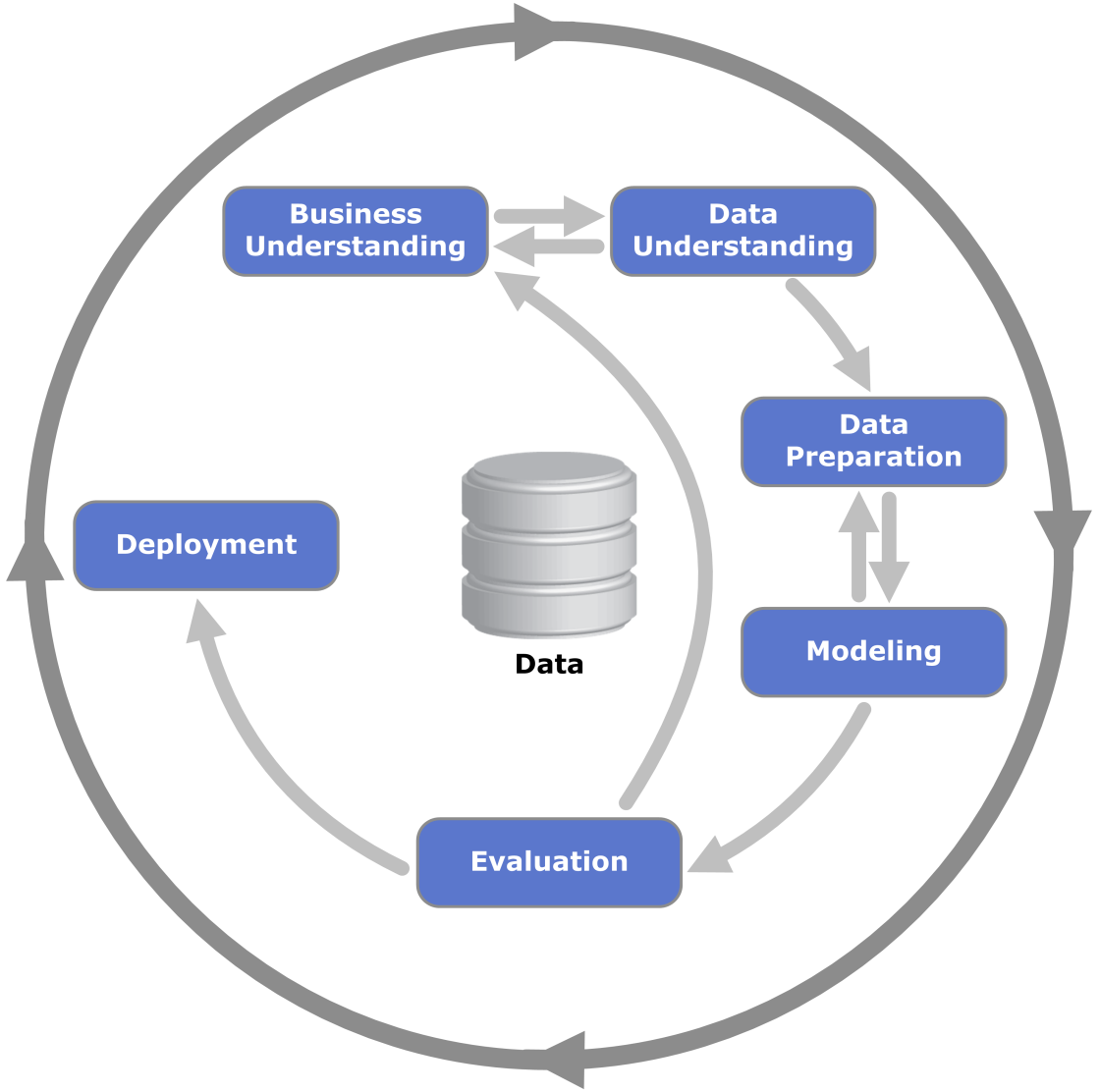
Why Should There be a Standard Process?

- The data mining process must be *reliable* and *repeatable* by people with little data mining background.
- 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
  - Reduces dependency on “stars”

CRISP-DM

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

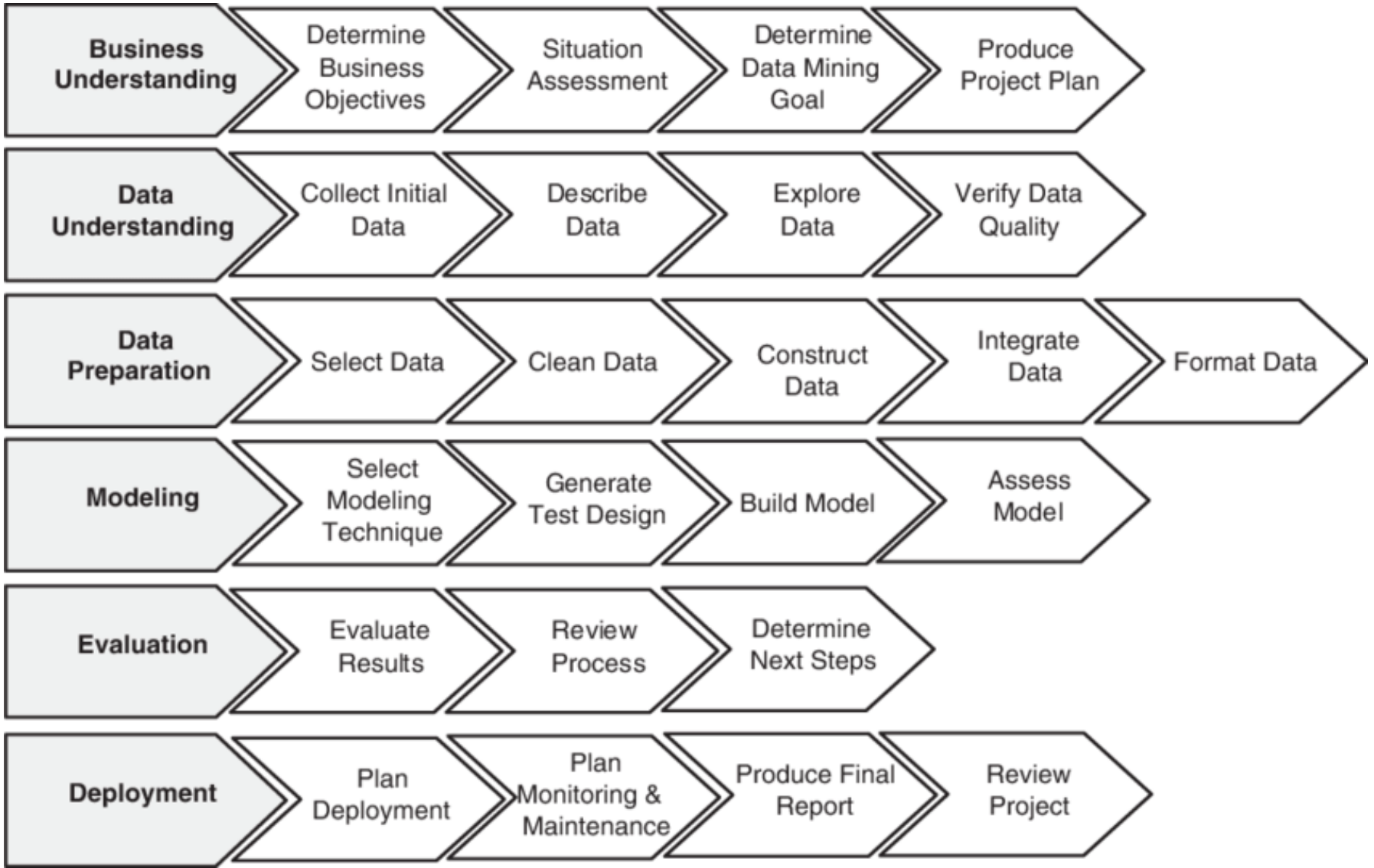
Overview


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

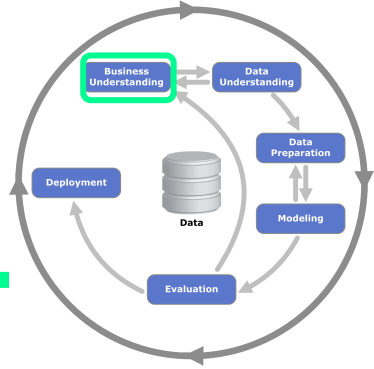
Phases

- **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

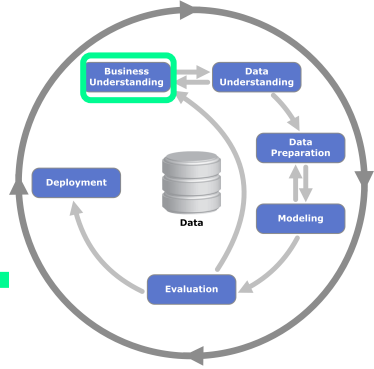


Business Understanding



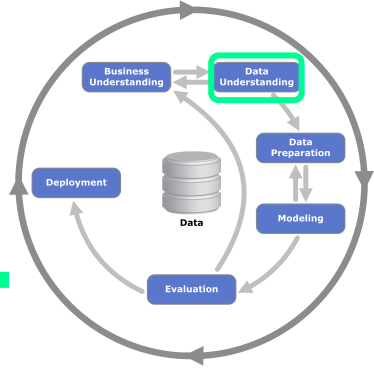
- **Determine business objectives**
  - thoroughly understand, from a business perspective, what the client really wants to accomplish
  - uncover important factors, at the beginning, that can influence the outcome of the project
  - neglecting this step is to expend a great deal of effort producing the right answers to the wrong questions
- **Assess situation**
  - more detailed fact-finding about all of the resources, constraints, assumptions and other factors that should be considered
  - flesh out the details

Business Understanding



- **Determine data mining 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.”
- **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

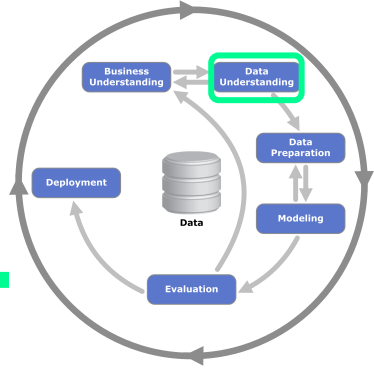
Data Understanding



- **Explore the Data**
- **Verify the Quality**
- **Find Outliers**

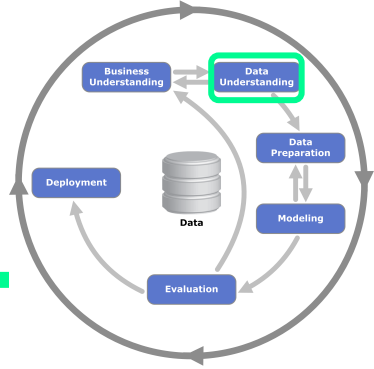
Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Data Understanding



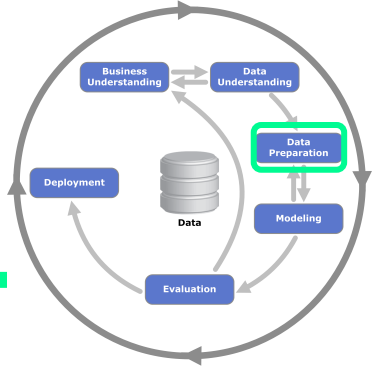
- **Collect initial data**
  - acquire within the project the data listed in the project resources
  - includes data loading if necessary for data understanding
  - possibly leads to initial data preparation steps
  - if acquiring multiple data sources, integration is an additional issue, either here or in the later data preparation phase
- **Describe data**
  - examine the “gross” or “surface” properties of the acquired data
  - report on the results

Data Understanding



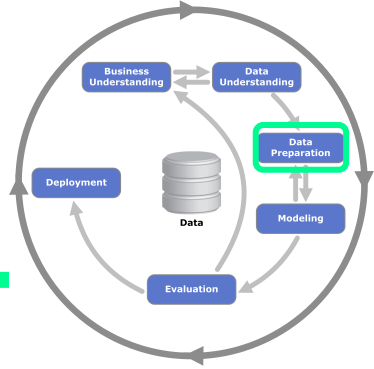
- **Explore data**
  - tackles the data mining questions, which can be addressed using querying, visualization and reporting including:
    - distribution of key attributes, through aggregations
    - relations between pairs of attributes
    - properties of significant sub-populations
  - may address directly the data mining goals
  - may contribute to data description and quality reports
  - may feed into the transformation and other data preparation needed
- **Verify data quality**
  - examine the quality of the data, addressing questions such as: “Is the data complete?”, Are there missing values in the data?”

Data Preparation



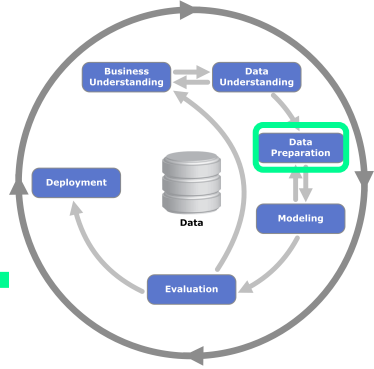
- Takes usually over 90% of the time
  - **Collection**
  - **Assessment**
  - **Consolidation and Cleaning**
  - **Data selection**
  - **Transformations**
- Covers all activities to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

Data Preparation



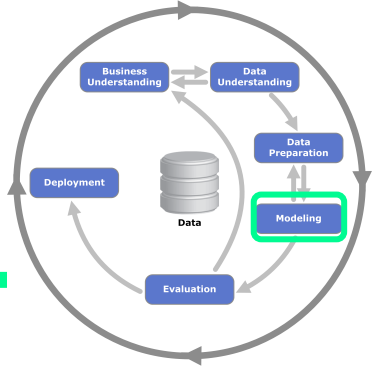
- **Select data**
  - decide on the data to be used for analysis
  - criteria include relevance to the data mining goals, quality and technical constraints such as limits on data volume or data types
  - covers selection of attributes as well as selection of records in a table
- **Clean data**
  - raise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

Data Preparation



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

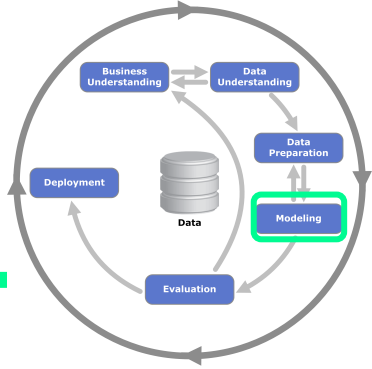
Modeling



- **Select the modeling technique**
  - (based upon data mining objectives)
- **Build model**
  - (Parameter settings)
- **Assess model**
  - (rank the models)

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*.

Modeling

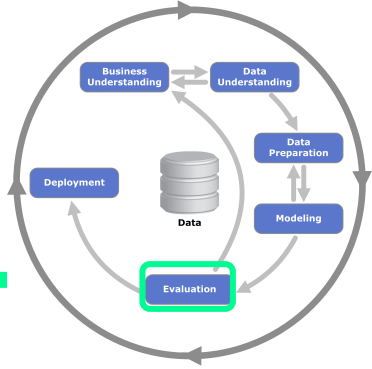


- **Select modeling technique**
  - select the actual modeling technique that is to be used ex) decision tree, neural network
  - if multiple techniques are applied, perform this task for each technique separately
- **Generate test design**
  - before actually building a model, generate a procedure or mechanism to test the model's quality and validity ex) 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

Modeling

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

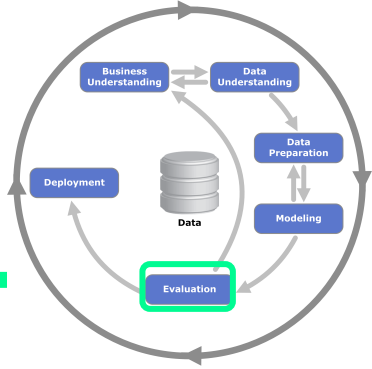
Evaluation



- **Evaluation of model**
  - how well it performed on test data
- **Methods and criteria**
  - depend on model type
- **Interpretation of model**
  - importance and hardness depend on the algorithm

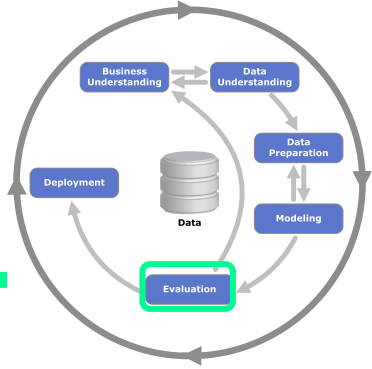
Thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is *to determine if there is some important business issue that has not been sufficiently considered*. At the end of this phase, a decision on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

Evaluation



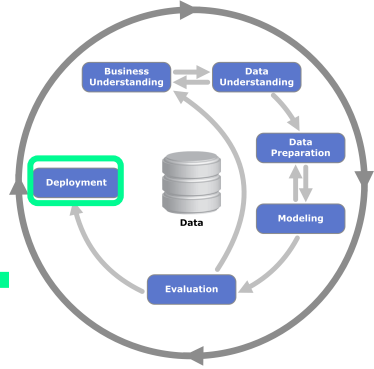
- **Evaluate results**
  - assesses the degree to which the model meets the business objectives
  - seeks to determine if there is some business reason why this model is deficient
  - test the model(s) on test applications in the real application if time and budget constraints permit
  - also assesses other data mining results generated
  - unveil additional challenges, information or hints for future directions

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 ex) “Did we correctly build the model?”
- **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

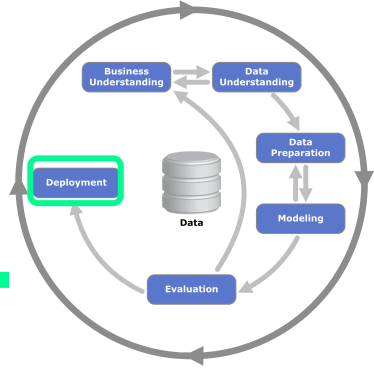
Deployment



- 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

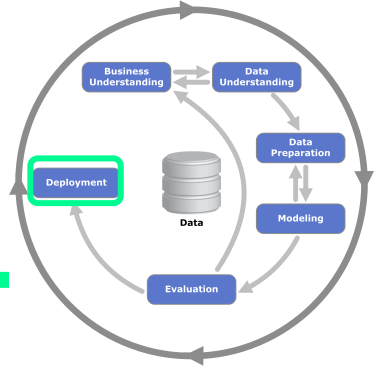
The knowledge gained will need to *be organized and presented in a way that the customer can use it*. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

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
- **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

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

Summary

Why CRISP-DM?

- 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

- Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz, (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler) “CRISP-DM 1.0 - Step-by-step data mining guide”

- Websites
  - http://www.crisp-dm.org/
  - http://www.spss.com/
  - http://www.kdnuggets.com/