Minería de Datos

U1 Introducción

Héctor Maravillo

Section 1

The CRISP-DM Process Model

CRISP-DM

CRISP-DM, which stands for **Cross-Industry Standard Process for Data Mining**, is a robust and versatile framework for data mining projects.

- As a **methodology**, it includes descriptions of the normal **phases** of a project, the **tasks** required in each phase, and an explanation of the **relationships** between tasks.
- As a **process model**, CRISP-DM provides an overview of the **data mining life cycle**.

CRISP-DM History

- CRISP-DM was conceived in 1996 and became an **European Union project** under the European Strategic Programme on Research in Information Technology (ESPRIT) funding initiative in 1997.
- The project was led by **five companies**: Integral Solutions Ltd ISL (England), NCR Systems Engineering Copenhagen (Denmark), DaimlerChrysler AG (Germany), Teradata (American) and OHRA Verzekeringen en Bak Groep B.V. (The Netherlands).
- The first version of the methodology was presented at the 4th CRISP-DM SIG Workshop in Brussels in March **1999**, and published as a step-by-step data mining guide later that year.
- ISL was acquired and merged into SPSS, which was later acquired by **IBM**. Currently, IBM is the primary corporation that uses the CRISP-DM process model.

The CRISP-DM Methodology

The CRISP-DM data mining methodology is described in terms of a **hierarchical process model**, consisting of sets of tasks described at four levels of abstraction: **phases**, **generic tasks**, **specialized tasks** and, **process instances**.

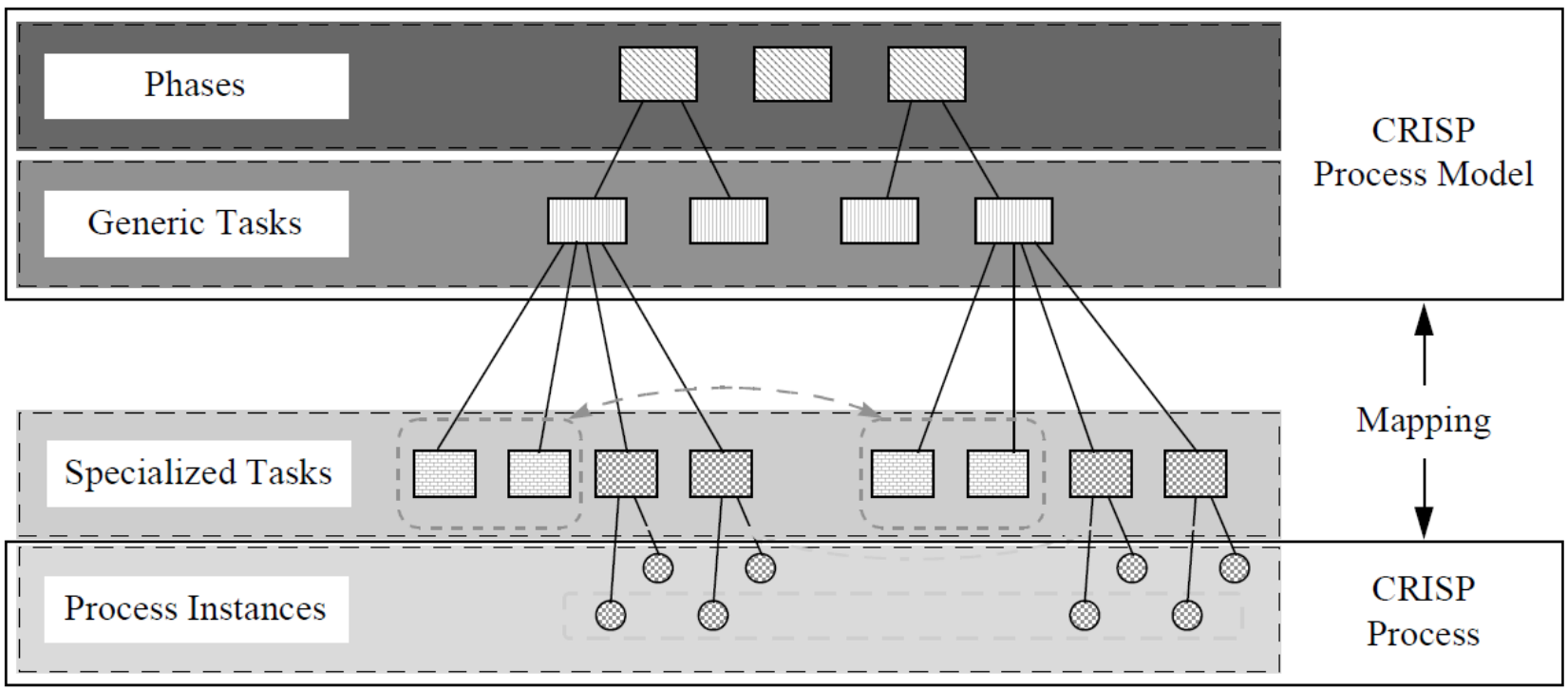

Figure: Four Levels of the CRISP-DM Methodology

The CRISP-DM Methodology

- **Phase**: The top level of abstraction.
- **Generic task**: It is intended to be general enough to **cover** all possible data mining situations. These tasks must be complete and stable as possible.
  - **Complete**, means **covering** both the whole process of data mining and all possible data mining applications.
  - **Stable** means that the model should be **valid** for yet unforeseen developments like new modelling techniques.
- **Specialized task**: Describe **how** action in the generic task should be achieved in certain **specific situation**.
- **Process instance**: It is a **record** of the **actions**, **decision**, and **results** of an actual data mining project. A process instance is organized according to the tasks defined at the higher levels but represents **what actually happened** in a particular engagement.

Reference model and user guide

Horizontally, the CRISP-DM methodology distinguishes between the Reference Model and the User Guide.

- The **Reference Model** presents a quick overview of phases, tasks, and their outputs, and describes **what to do** in a data mining project.
- The **User Guide** gives more detailed tips and hints for each phase and each task within a phase and depicts **how to do** a data mining project.

The Reference Model

The **CRISP-DM Reference Model** provides an overview of the **life cycle** of a data mining project.

It contains the corresponding phases of a project, their respective tasks, and relationships between these tasks.

Figure: Phases of the CRISP-DM Reference Model

Business Understanding

- **Business Understanding**. This initial phase focuses on understanding the **project objectives** and **requirements** from a business perspective, and then converting this knowledge into a **data mining problem definition**, and a **preliminary plan** designed to achieve the objectives.

Business Understanding

- Business understanding tasks:
  - **Determine business objectives**: Identify and understand a client's core goals and desired outcomes from a business perspective, while also considering and balancing any competing objectives and constraints.
  - **Assess situation**: Investigate and evaluate all resources, constraints, assumptions, and other relevant factors to develop the project plan.
  - **Determine data mining goals**: State project objective in technical terms.
  - **Produce project plan**: Describe the intended plan for achieving the data mining goals, including an initial selection of tools and techniques.

Business Understanding

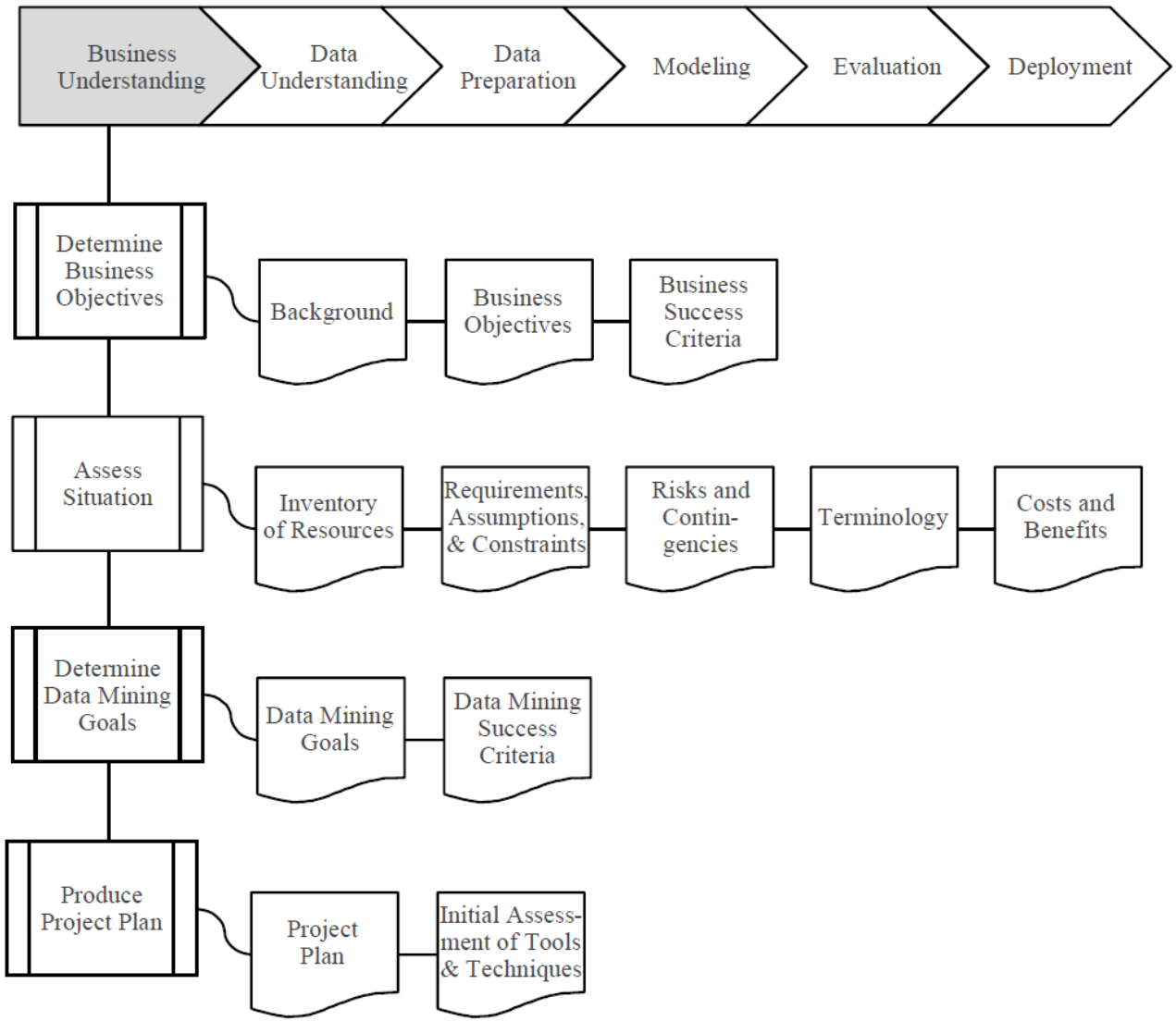

Figure: Business Understanding

Data Understanding

- **Data Understanding**. This phase 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

- Data understanding tasks
  - **Collect initial data**: Acquire within the project the data (or access to the data) listed in the project resources. This task may include data integration or loading. This task may include data integration or loading.
  - **Describe data**: Examine the *surface* properties of the acquired data and report on the results.
  - **Explore data**: Analyze data using querying, visualization, and reporting to address data mining questions. This includes examining attribute distributions, relationships between attributes, simple aggregations, and statistical properties of sub-populations.
  - **Verify data quality**: Examine the quality of the data, addressing questions such as: Is the data complete (does it cover all the cases required)? Is it correct or does it contain errors? Are there missing values?

Data Understanding

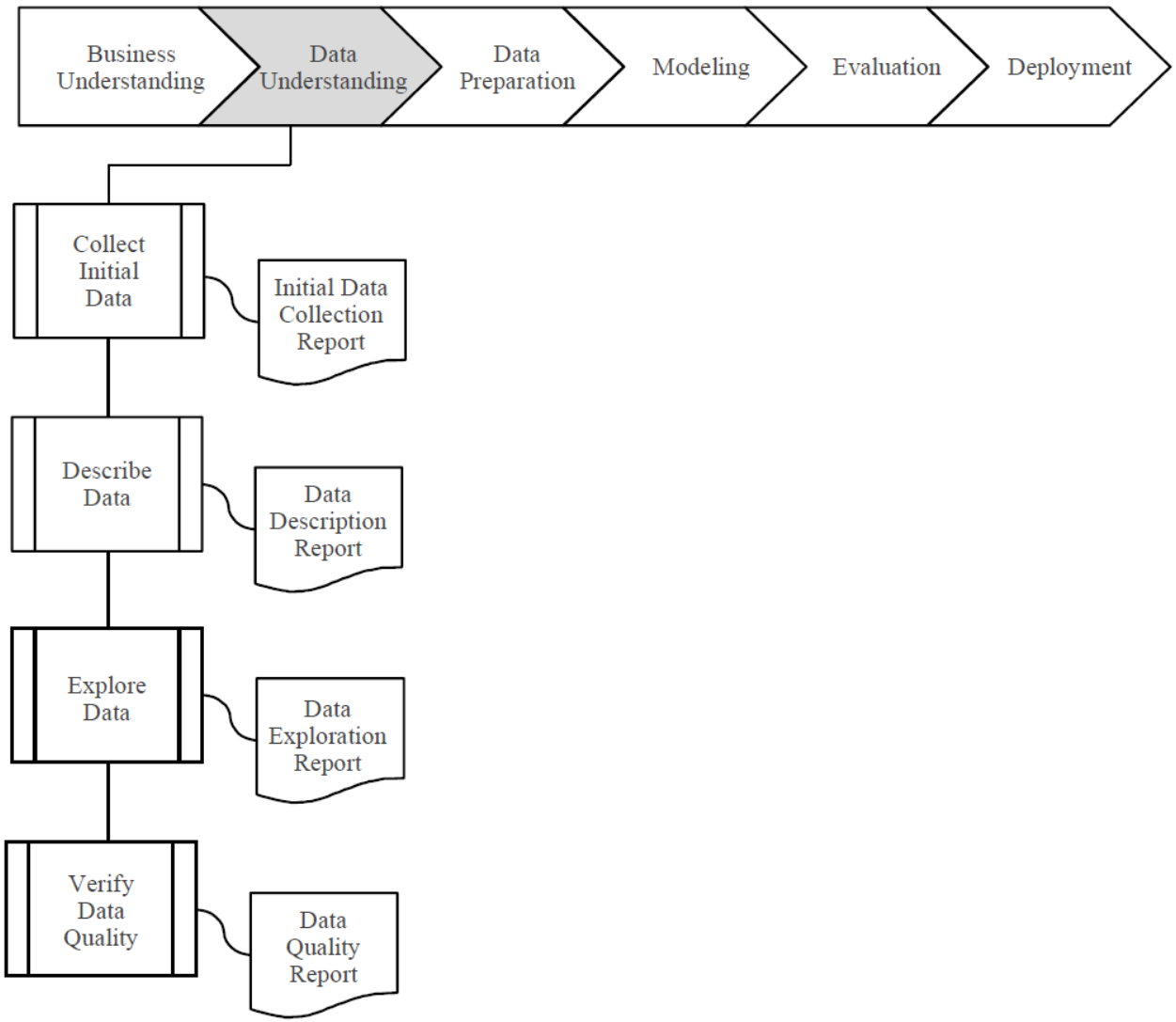

Figure: Data Understanding

Data Preparation

- **Data Preparation**. This phase covers all activities to **construct the final dataset** (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed **multiple times**, and not in any prescribed order.

  Outputs:
  - Data set.
  - Data set description.

Data Preparation

- Data preparation tasks:
  - **Select data**: Decide on the data to be used for analysis. This task covers selection of attributes (columns) as well as selection of records (rows) in a table.
  - **Clean data**: Raise the data quality to the level required by the selected analysis techniques.
  - **Construct data**: This task includes constructive data preparation operation such as the production of derived attributes, entire new records, or transformed values for existing attributes.
  - **Integrate data**: These are methods by which information from several tables or records is combined to create new records or values.
  - **Format data**: It refers mainly to syntactic modifications made to data that do not change its meaning, but which may be required by the modeling tool.

Data Preparation

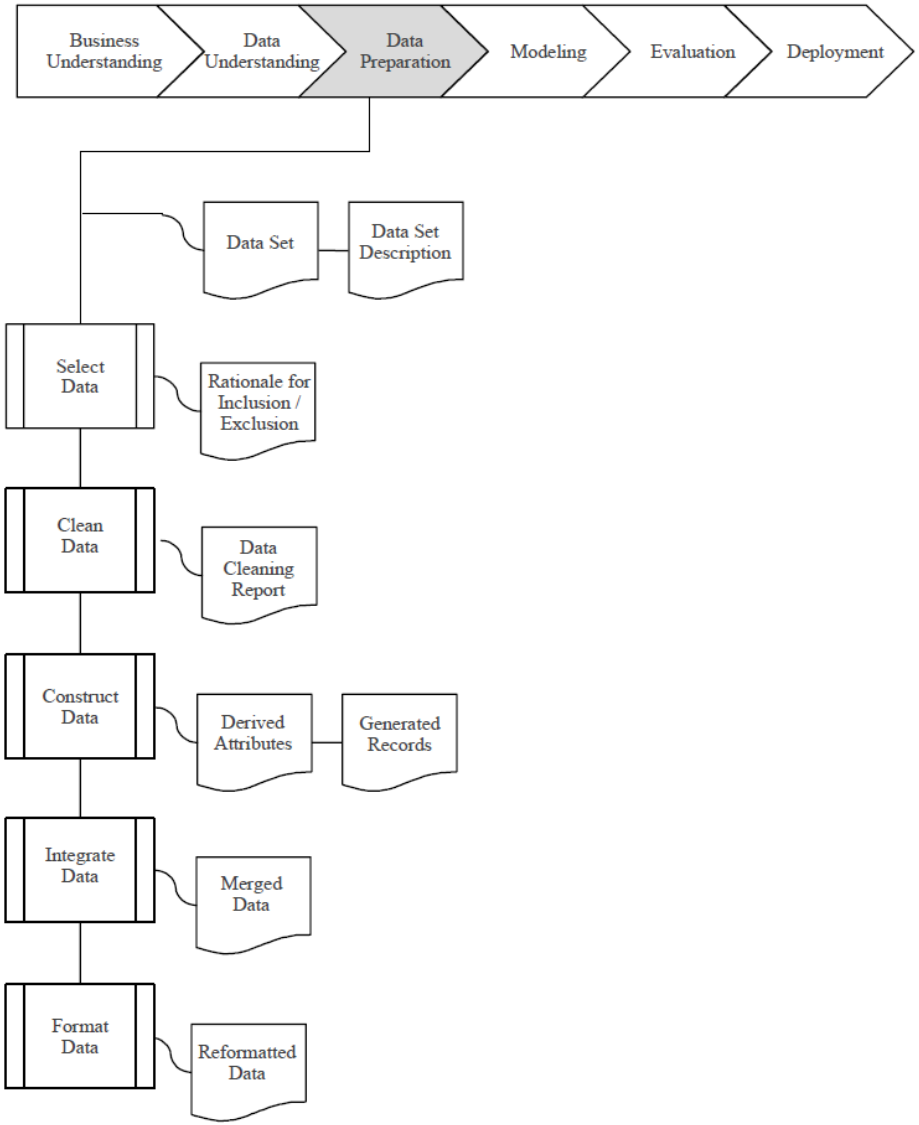

Figure: Data Preparation

Modeling

- **Modeling**. In this phase, various **modeling techniques** are selected and applied, and their **parameters** are calibrated to optimal values.

Modeling

- Modeling tasks:
  - **Select modeling technique**: select the actual and specific modeling technique which is used now. If multiple techniques are applied, perform this task for each technique separately.
  - **Generate test design**: Generate a procedure or mechanism to test the model's quality and validity. For example, split the data set into train and test set.
  - **Build model**: Run the modeling tool on the prepared data set to create one or more models.
  - **Assess model**: The data mining engineer evaluates and ranks models based on technical criteria, domain knowledge, and business objectives, typically assessing the model's accuracy and generality.

Modeling

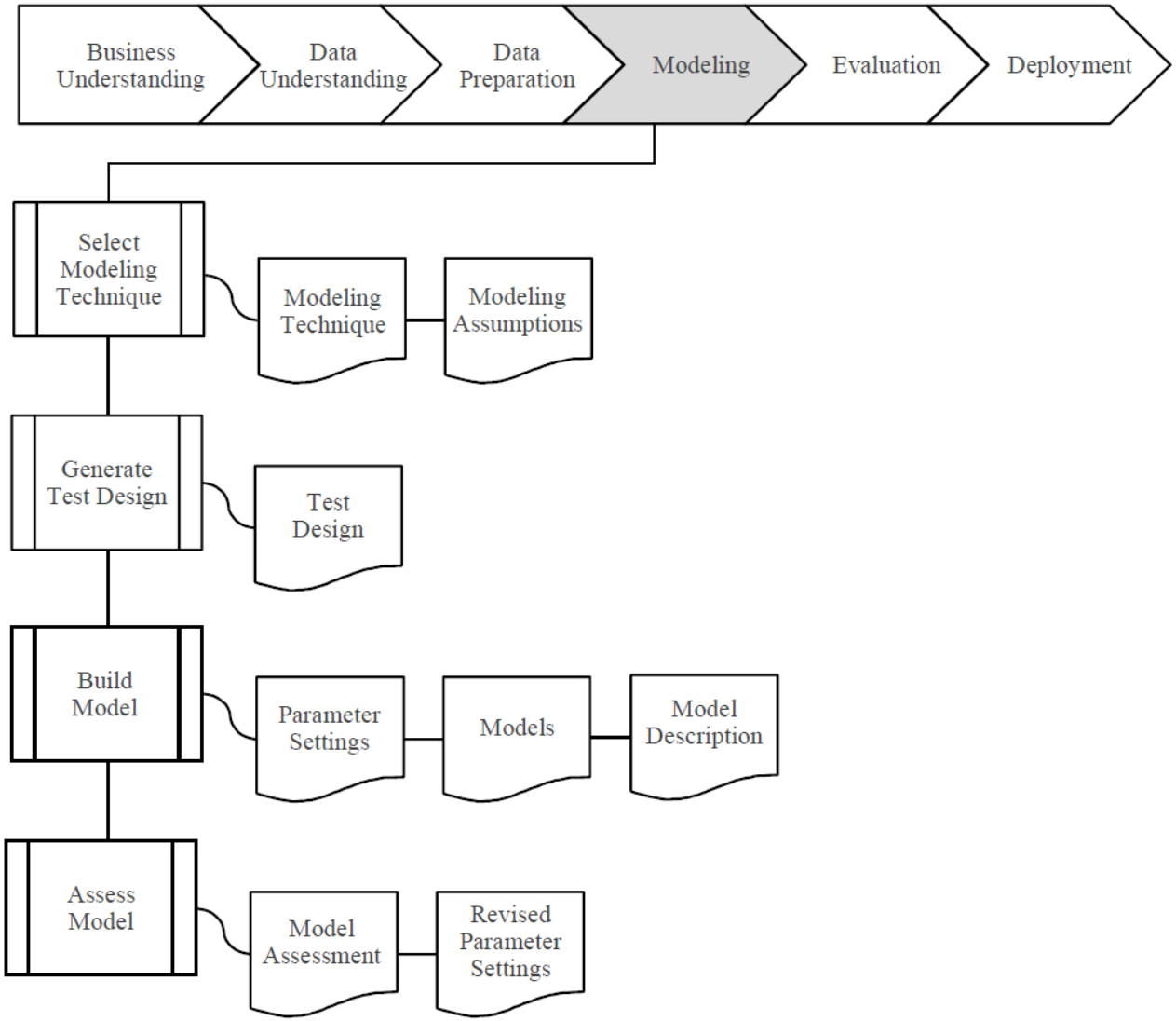

Figure: Modeling

Evaluation

- **Evaluation**: Thoroughly assess the model, and **review the steps** executed to construct the model to ensure that it adequately **achieves business objectives**. A key objective is to determine if there is some important business issue that has not been sufficiently considered.

Evaluation

- Tasks for evaluating results
  - **Evaluate results**: This step assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why this model is deficient.
  - **Review process**: Conduct a thorough evaluation of the data mining project in order to determine if there is any important factor or task that has somehow been overlooked.
  - **Determine next steps**: Decide whether to finish this project and move on to deployment if appropriate, or whether to initiate further iterations or set up new data mining projects.

Evaluation

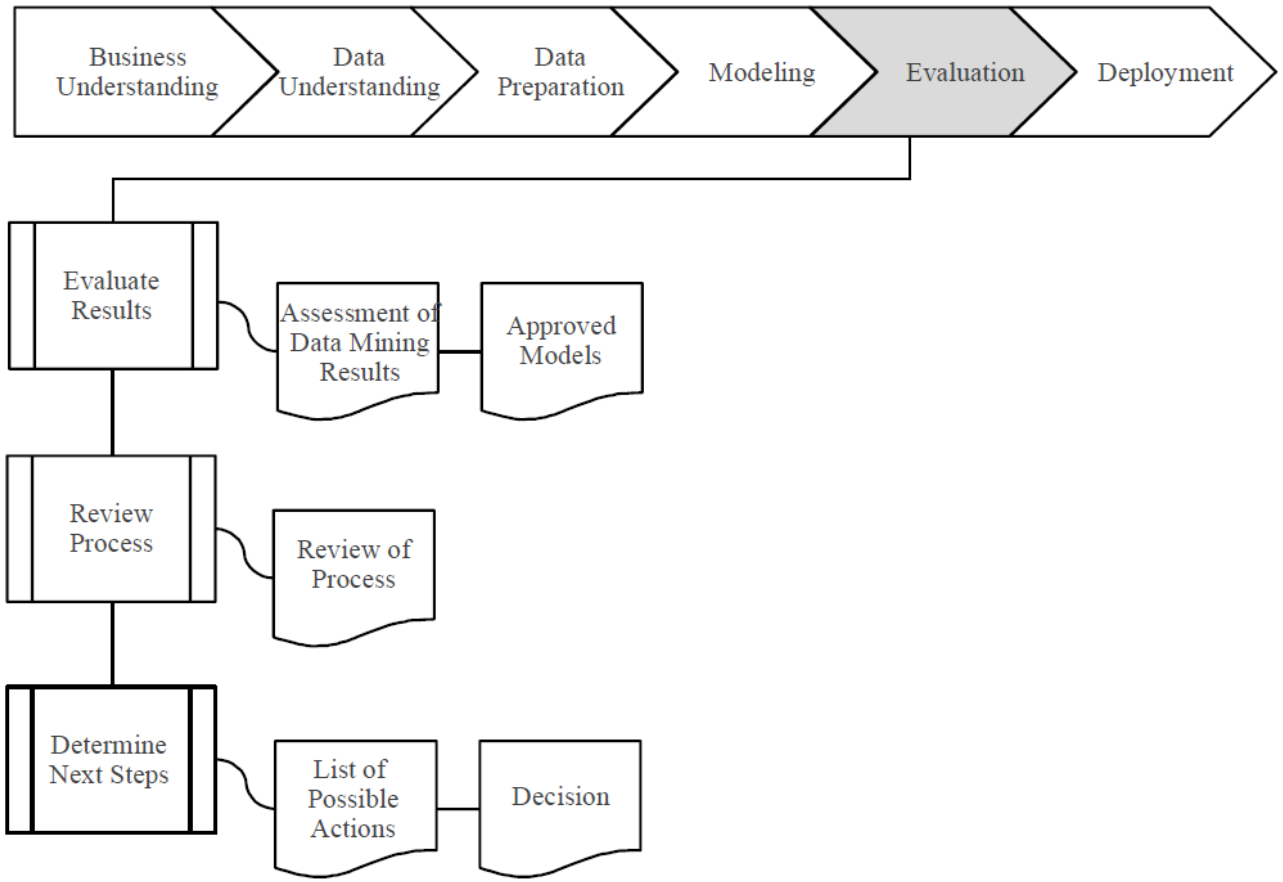

Figure: Evaluation

Deployment

- **Deployment.** Organize and **present** the knowledge gained in a **usable format** for the **customer**. 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.

Deployment

- Deployment tasks:
  - **Plan deployment**: In order to deploy the data mining result(s) into the business, this task takes the evaluation results and concludes a strategy for deployment.
  - **Plan monitoring and maintenance**: Develop a detailed strategy for monitoring and maintaining data mining results to ensure their correct and timely use in daily business operations.
  - **Produce final report**: At the end of the project, the project leader and his team write up a final report.
  - **Review project**: Assess what went right and what went wrong, what was done well and what needs to be improved.

Deployment

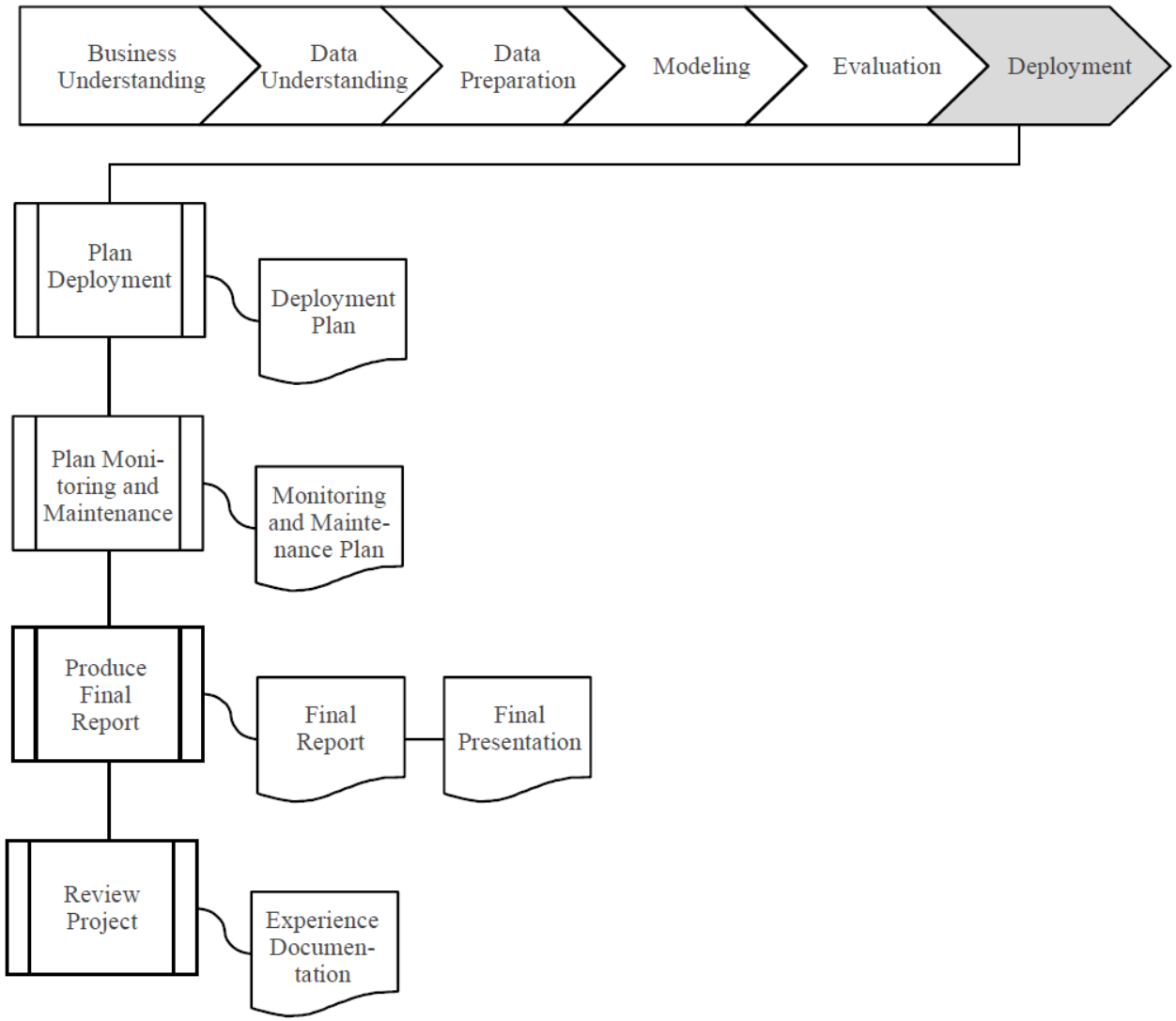

Figure: Deployment

References

- Chapman, P., Kerber, R., Clinton, J., Khabaza, T., Reinartz, T., & Wirth, R. *The CRISP-DM Process Model*, NCR Systems Engineering Copenhagen - DaimlerChrysler AG - Integral Solutions Ltd. - OHRA Verzekeringen en Bank Groep B.V., 1999.