The Knowledge Engineering Review, Vol. 25:2, 137–166. © Cambridge University Press, 2010
doi:10.1017/S0269888910000032

A survey of data mining and knowledge discovery process models and methodologies

GONZALO MARISCAL[1], ÓSCAR MARBÁN[2] and COVADONGA FERNÁNDEZ[2]

[1]*Universidad Europea de Madrid, C/Tajo, S/N. 28670 - Villaciciosa de Odon, Madrid, Spain;* [2]*Facultad de Informatica, Universidad Politecnica de Madrid, Campus de Montegancedo, 28660 - Boadilla del Monte, Madrid, Spain;*
e-mails: gonzalo.mariscal@uem.es, omarban@fi.upm.es, cfbaizan@fi.upm.es

Abstract

Up to now, many data mining and knowledge discovery methodologies and process models have been developed, with varying degrees of success. In this paper, we describe the most used (in industrial and academic projects) and cited (in scientific literature) data mining and knowledge discovery methodologies and process models, providing an overview of its evolution along data mining and knowledge discovery history and setting down the state of the art in this topic. For every approach, we have provided a brief description of the proposed knowledge discovery in databases (KDD) process, discussing about special features, outstanding advantages and disadvantages of every approach. Apart from that, a global comparative of all presented data mining approaches is provided, focusing on the different steps and tasks in which every approach interprets the whole KDD process. As a result of the comparison, we propose a new data mining and knowledge discovery process named *refined data mining process* for developing any kind of data mining and knowledge discovery project. The refined data mining process is built on specific steps taken from analyzed approaches.

1 Introduction

Data mining (DM), knowledge discovery in databases (KDD), knowledge discovery, and *data mining and knowledge discovery* (DM & KD) are terms used to refer to results of research, techniques and tools used to extract useful information from large volumes of data (Agrawal & Shafer, 1996). The complete process of extracting information is known as KDD process (Piatetsky-Shaphiro & Frawley, 1991; Fayyad *et al*., 1996c). Data mining is just one step in the whole KDD process.

In general, *data mining* is used by many researchers as a synonym of the KDD process (Cabena *et al*., 1997; Chapman *et al*., 2000; Piatetsky-Shapiro, 2000; Kurgan & Musilek, 2006). In general, in industrial and press worlds, *data mining* is used to refer to the whole KDD process. Therefore, both terms can be used indistinctly when referring to this area.

Lately, the term *data mining and knowledge discovery* has been proposed as the most adequate name for the overall process of KDD (Reinartz, 2002; Cios & Kurgan, 2005; Kurgan & Musilek, 2006; SpringerLink, 2008).

In the early 90s, when the KDD term was first coined (Piatetsky-Shapiro, 1991), there was a rush to develop data mining algorithms that were capable of solving all the problems related to the search for useful knowledge in large volumes of data. Apart from developing algorithms, some specific tools, such as: clementine (ISL, 1995; Khabaza & Shearer, 1995; Shearer, 1996), IBM Intelligent Miner (Tkach, 1998; IBM, 1999), Weka (Witten & Frank, 2005), and DBMiner (The Data Mining Research Group, 1997) were also developed to simplify the application of data mining algorithms and provide some kind of support for all the activities involved in the KDD.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

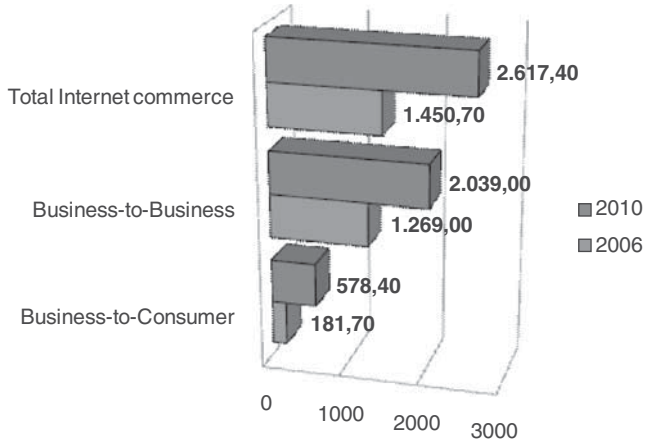


Figure 1 Internet commerce in Europe, 2006–2010, in billion Euro (EITO (European Information Technology Observatory), 2007)

From the viewpoint of data mining methodologies and process models, the year 2000 marked the most important milestone: Cross-Industry Standard Process for Data Mining (CRISP-DM) was first proposed (Chapman *et al.*, 2000). CRISP-DM is the most widely used methodology for developing data mining projects (KdNuggets.Com, 2007b). It is considered the *de facto* standard[1] (Chapman *et al.*, 2000).

This model describes the activities that must been done to develop a data mining project. Every activity is composed of tasks. For every task, generated outputs and needed inputs are detailed. CRISP-DM comes up to resolve the problems that existed in data mining project developments. Their main objectives are listed below (Presutti, 1999):

- Ensure quality of data mining projects results
- Reduce skills required for data mining
- Capture experience for reuse
- General purpose (i.e., widely stable across varying applications)
- Robust (i.e., insensitive to changes in the environment)
- Tool and technique independent
- Tool supportable

The number of applied projects in the Data Mining area is expanding rapidly (KdNuggets.Com, 2007a; Kriegel *et al.*, 2007). This growth is confirmed by the annual reports by the Gartner Group (McDonald *et al.*, 2006; Gartner, Inc., 2008a, 2008b). They claim that business intelligence (BI)[2] is the area in which companies are investing the most since 2006. Data Mining investment grew by 4.8% from 2005 to 2006 (McDonald *et al.*, 2006), and by 11.2% from 2007 to 2008 (Gartner, Inc., 2008b).

On the other hand, Figure 1 shows that a great growth in the BI area (business to consumer commerce) is expected from 2006 to 2010.

Having seen data, it can be said that data mining is continuously growing. However, not all data mining results are positive. While it is true that a lot of data mining projects are being developed, neither all the project results are in use (Eisenfeld *et al.*, 2003a, 2003b; Zornes, 2003;

[1] *De facto standards* are those that have come into existence without any formal plan by any of the standard organizations. Rather, they are developed through the industry's acceptance of a specific vendor's standard, which is placed in the public domain (*De facto* is Latin for *from the fact*) (Gallo & Hancock, 2001).

[2] BI is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions (SearchDataManagement.com, 2008). Data mining is an important component of BI (Cabena *et al.*, 1997).

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

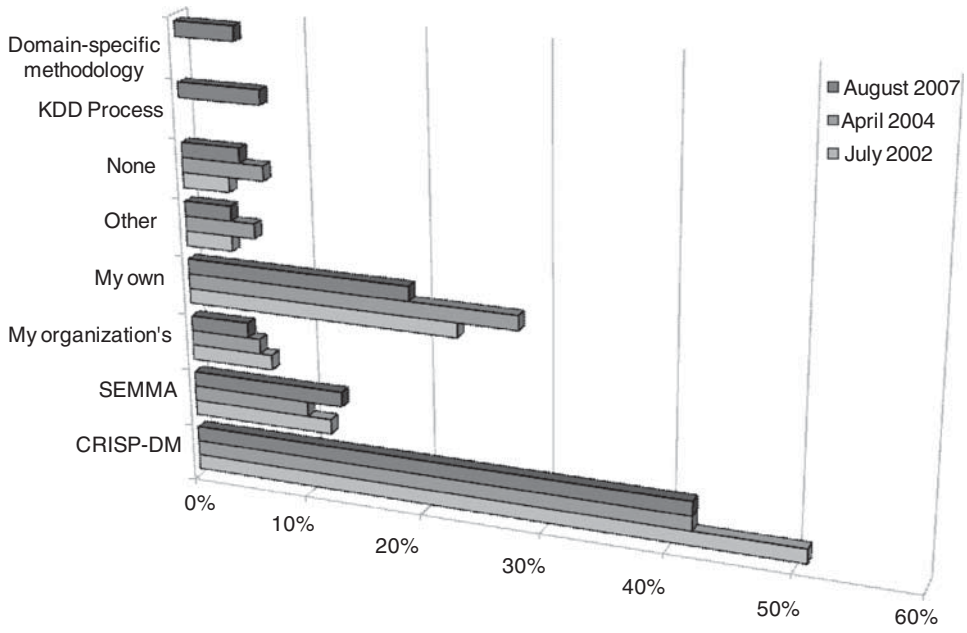


Figure 2 Use of data mining methodologies (KdNuggets.Com, 2002, 2004, 2007b)

KdNuggets.Com, 2008), nor all projects end successfully (Edelstein & Edelstein, 1997; Strand, 2000; Gartner, Inc., 2005; Gondar, 2005; KdNuggets.Com, 2008; McMurchy, 2008).

CRISP-DM is the most commonly used methodology for developing data mining projects (KdNuggets.Com, 2002, 2004, 2007b). However, its use is not becoming any more widespread due to rivalry with other, in-house methodologies developed by work teams and sample, explore, modify, model, assess (SEMMA) methodology (see Figure 2).

This decrease in the use of CRISP-DM is due to the fact that it just defines *what to do* and not *how to do*. Because of that, work teams have started to use its own methodologies. Another inconvenience is that CRISP-DM does not include project management activities such as quality management or change management (Marbán *et al*., 2008). On the other hand, the use of SEMMA methodology has lightly increased because of the growth in the use of its data mining support tool, Enterprise Miner, developed by SAS and based in SEMMA methodology. SAS is a leader company in BI and it has the most comprehensive BI platform in the industry with the most advanced analysis capabilities. It is corroborated by the Magic Quadrant for BI platforms published by Gartner (Richardson *et al*., 2008), where SAS has been placed in the Leaders Quadrant. It confirms that SAS is the leader in BI and analytical software and services.

In the study by Yang and Wu (2006), one of the 10 challenging problems to be solved in data mining research is the need to build a new methodology to help users avoid many common data mining mistakes, by improving the automation of the KDD process. As it will be seen later, there are a lot of proposed methodologies for data mining projects, but a correct and complete methodology, which complies with the *methodology* definition detailed in the next section, has not been developed yet.

Previously, some surveys about process models and methodologies have been published. It is fitting to point out (Kurgan & Musilek, 2006), it presents a historical overview, description and future directions concerning a standard for a data mining process model.

It presents a comparison of five data mining process models and methodologies: KDD process (Fayyad *et al*., 1996b), Cabena *et al*. (1997), Anand and Buchner (1998), CRISP-DM (Chapman *et al*., 2000); and proposes a six-step generic model based on the five surveyed models (Application domain understanding, data understanding, data preparation and identification of data mining technology, data mining, evaluation and knowledge consolidation and deployment). The survey

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

also proposes to enhance existing models by embedding other current standards to enable automation and interoperability of the entire process.

The five approaches analyzed in the study by Kurgan and Musilek (2006) are included in this survey (and nine more). In later sections, we also propose a new approach based on the comparison of the 14 analyzed approaches. And we agree with Kurgan and Musilek (2006) that standardization of data mining process models should be an essential research line in present and future of data mining and knowledge discovery.

This paper presents a review of most used and cited process models and methodologies for data mining and knowledge discovery projects. First of all, some technical terms referenced along the paper are defined. After describing most important models and methodologies, a comparative study between all of them is built. As a result, a new *refined data mining process*, including specific steps taken out of studied approaches, is presented. Finally, some conclusions are listed.

2 Basic terms about process models

Process models, paradigms, methodologies, techniques and tools are essential strategic, tactical and technical elements for developing an engineering project. In the next sections every element is described:

- Process model
  A process model can be defined as the set of framework activities and tasks to get the job done, including inputs and outputs in every task (Pressman, 2005). The final objective of a process model is to do it manageable, repeatable and measurable. A good process model should comply the following characteristics (McCall *et al*., 1977; Tyrrell, 2000).
  - o Effective. An effective process must help us produce the right product.
  - o Maintainable. So we can quickly and easily find and remedy faults or work out where to make changes.
  - o Predictable. Any new product development needs to be planned, and those plans are used as the basis for allocating resources: both time and people. A good process will help us do this. The process helps lay out the steps of development.
  - o Repeatable. If a process is discovered to work, it should be replicated in future projects. *Ad hoc* processes are rarely repeatable unless the same team is working on the new project. Even with the same team, it is difficult to keep things exactly the same.
  - o Quality. Quality in this case may be defined as the product's fitness for its purpose.
  - o Improvable. No one would expect his or her process to reach perfection and need no further improvement itself. Even if we were as good as we could be now, both development environments and requested products are changing so quickly that our processes will always be running to catch up.
  - o Traceable. A defined process should allow the project staff to follow the status of a project.
- Paradigm
  In a study by Harman (1970) paradigm is defined as a basic way of perceiving, thinking, valuing and doing things that are associated with a particular vision of reality. On the other hand, in Barker (1992) and Capra (1996) paradigm is defined as a set of rules that define limits, and state what is necessary to be successful within those limits.
- Methodology
  Methodology can be defined as a process model instance, in which not only tasks, inputs and outputs must be specified but also the way in which the tasks must be carried out. It has to be taken into account that a methodology can be an instance of different process models, and vice versa, such as shown in Figure 3. Tasks are executed by using techniques that defines how to do them (Pressman, 2005). After selecting the techniques, a tool can be used to support the carrying out of the established tasks. These tools apply the techniques, and tasks are made easier. To sum up, process models define *what to do*, methodologies define *how to do*.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

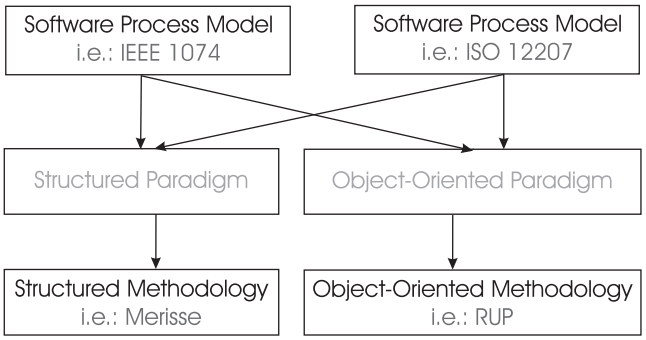


Figure 3 Process models with different methodologies

- Life cycle
  Life cycle determines the order in which activities will be done (McConnell, 1997). A life cycle model is the description of different ways of developing a project. The main functions of a life cycle are:
  - o To establish the order of phases and processes involved in project development.
  - o To establish the criteria to step forward from a phase to the next (intermediate deliverable). It includes the criteria to validate a phase and select the next one.

Life cycles provide an order guide (phases, activities, prototypes, validations) in which the most important activities of the project should be done. Project success depends on the selected life cycle, since it can help to assure every step takes us to achieve the objective. A wrong choice can be a source of delays and unnecessary work.

As an example, if we take a software project development, the process model defines tasks to carry out, such as requirements specification, implementation, maintenance, management and support activities. As process model we could use ISO 12207 (ISO, 1995). The paradigm would show the approach of developing the software, for instance with object-oriented techniques, structured techniques, real-time techniques. The methodology must be consistent with techniques. For example, if we choose an object-oriented paradigm, the methodology would be object-oriented. An object-oriented methodology could be rational unified process (RUP) (Jacobson *et al*., 1999), widespread and accepted by software industry. Techniques would be requirement elicitation, test techniques and analysis and design techniques for object-oriented software. About tools, we can use CASE tools like rational rose that supports analysis and design of software tests. About life cycle, if RUP is selected as methodology, life cycle would be iterative and incremental.

3 Data mining and knowledge discovery process models and methodologies

In this section, there will be presented the evolution of data mining and knowledge discovery process models and methodologies to provide a review of them.

Figure 4 shows the evolution of 14 data mining process models and methodologies. We can point out KDD as the initial approach, and CRISP-DM as the central approach of the evolution diagram. Most of the approaches are based on them.

The two main approaches, KDD and CRISP-DM, are described in depth. The rest of the approaches are briefly described as they are based on KDD or CRISP-DM. Apart from that, contributions and disadvantages of every approach are shown.

The description of the 14 existing data mining and knowledge discovery methodology and process model approaches are divided into three subsections:

- *KDD related approaches*: this subsection includes a detailed description of the original KDD process proposed by Fayyad *et al*. (1996a) and a brief description of the approaches directly derived from the original KDD process.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

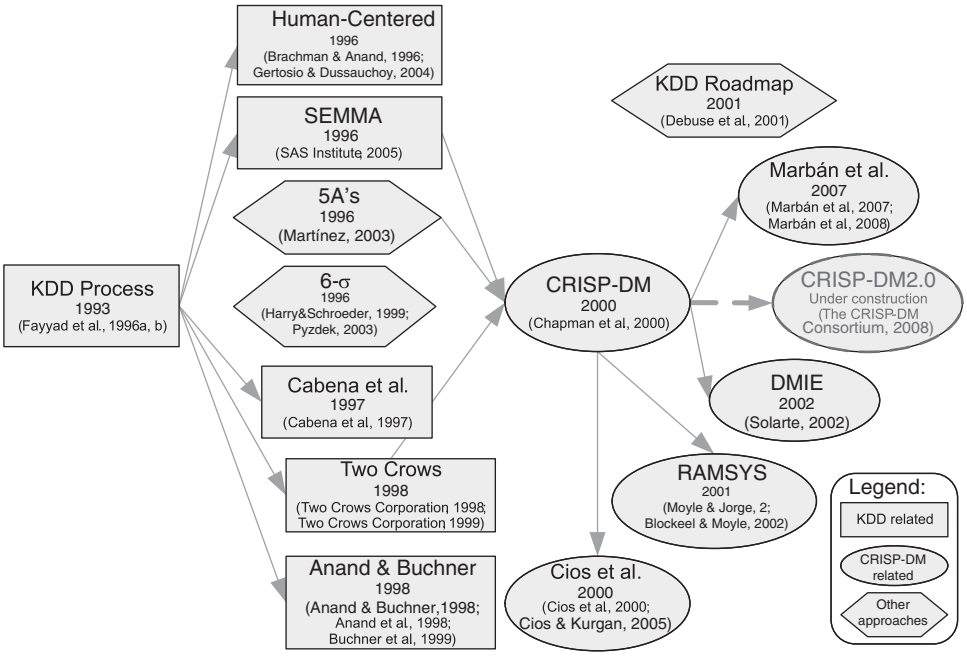


Figure 4 Evolution of data mining process models and methodologies

- *CRISP-DM related approaches*: this subsection includes a detailed description of CRISP-DM (Chapman *et al*., 2000) and a brief description of the approaches directly derived from CRISP-DM.
- *Other approaches*: this subsection includes a brief description of other independent approaches.

3.1 KDD related approaches

3.1.1 KDD process

The term KDD, appeared first time in latest 1980s (Piatetsky-Shapiro, 1991) to emphasize that knowledge is the product of a discovery process guided by data, and it is a joint point of different research areas focused on data analysis and knowledge extraction from different points of view, such as data bases, statistics, mathematics, logic or artificial intelligence.

KDD is defined as the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data (Fayyad *et al*., 1996c). Here, *data* are a set of facts, and *pattern* is an expression in some language describing a subset of the data or a model applicable to the subset. Extracting a pattern also designates fitting a model to data. The discovered patterns should be valid on new data with some degree of certainty. We also want patterns to be novel (at least to the system and preferably to the user) and potentially useful, that is, lead to some benefit to the user or task. Finally, the patterns should be understandable, if not immediately then after some postprocessing.

The term *process* implies that KDD comprises many steps, which involve data preparation, search for patterns, knowledge evaluation and refinement, all repeated in multiple iterations. By *non-trivial*, authors mean that some search or inference is involved.

According to the aforementioned, KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling and projections of the data before the data mining step.

Data mining step refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

The KDD process is outlined in Figure 5 from the data viewpoint.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

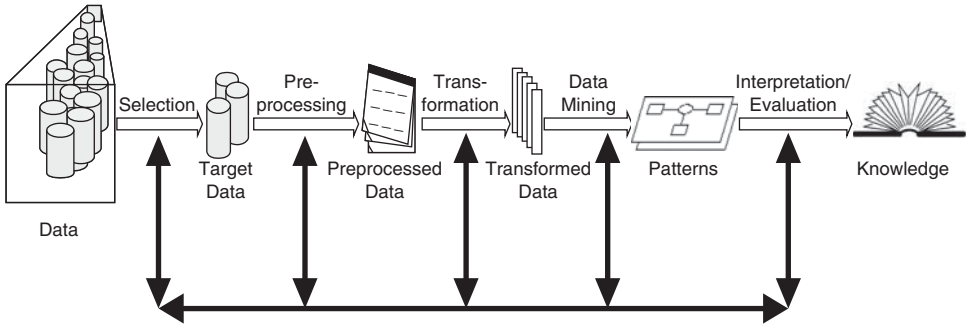


Figure 5 Overview of the steps constituting the knowledge discovery in databases (KDD) process (Fayyad *et al.*, 1996b)

The KDD process is interactive and iterative (with many decisions made by the user), involving nine steps, described from the practical viewpoint as:

- Learning the application domain
  It includes developing an understanding of the relevant prior knowledge and the goals of the application.
- Creating a target data set
  It includes selecting a data set or focusing on a subset of variables or data samples on which discovery is to be performed.
- Data cleaning and preprocessing
  It includes basic operations, such as removing noise or outliers if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes, as well as deciding data base management system issues, such as data types, schema and mapping of missing and unknown values.
- Data reduction and projection
  It includes finding useful features to represent the data, depending on the goal of the task, and using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.
- Choosing the function of data mining
  It includes deciding the purpose of the model derived by the data mining algorithm (e.g., summarization, classification, regression and clustering).
- Choosing the data mining algorithm
  It includes selecting method(s) to be used for searching for patterns in the data, such as deciding which models and parameters may be appropriate and matching a particular data mining method with the overall criteria of the KDD process.
- Data mining
  It includes searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, clustering, sequence modeling, dependency, association rules and line analysis.
- Interpretation
  It includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns and translating the useful ones into terms understandable by users.
- Using discovered knowledge
  It includes incorporating this knowledge into the performance system, taking actions based on the knowledge or simply documenting it and reporting it to interested parties, as well as checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

3.1.2 *Human-centered approach of data mining*

Brachman and Anand (1996) and Gertosio and Dussauchoy (2004) gave a practical view of the KDD process, emphasizing the interactive nature of the process. The human-centered model emphasized the interactive involvement of a data analyst (data miner) during the process.

Its basic steps are shown in Figure 6: task discovery, data discovery, data cleaning, model development, data analysis and output generation.

These six steps cover the same tasks that are included in Fayyad *et al.* (1996c) KDD process.

The main difference between both approaches is that human-centered process is focused in the tasks from the *data miner* viewpoint, while KDD process is more focused in data transformations. Human-centered model shows in a clearer way which decisions the user has to make.

3.1.3 *SEMMA*

SAS Institute defines SEMMA as a logical organization of the functional tool set of SAS enterprise miner for carrying out the core tasks of data mining (SAS Institute, 2005). Enterprise miner can be used as part of any iterative data mining methodology adopted by the client. SEMMA is focused on the model development aspects of data mining. Figure 7 shows SEMMA steps.

The main difference between the original KDD process and SEMMA is that SEMMA is integrated into SAS tools such as Enterprise Miner and it's unlikely to use SEMMA methodology out of them, while KDD is an open process and it can be applied in very different environments.

There are other two important differences between SEMMA and the original KDD process. On the one hand, SEMMA skips the first step of KDD process, learning the application domain, and starts directly with sample step. On the other hand, SEMMA does not include an explicit step to use the discovered knowledge, while KDD includes using discovered knowledge step. These two steps are considered essential to carry out a data mining project with success.

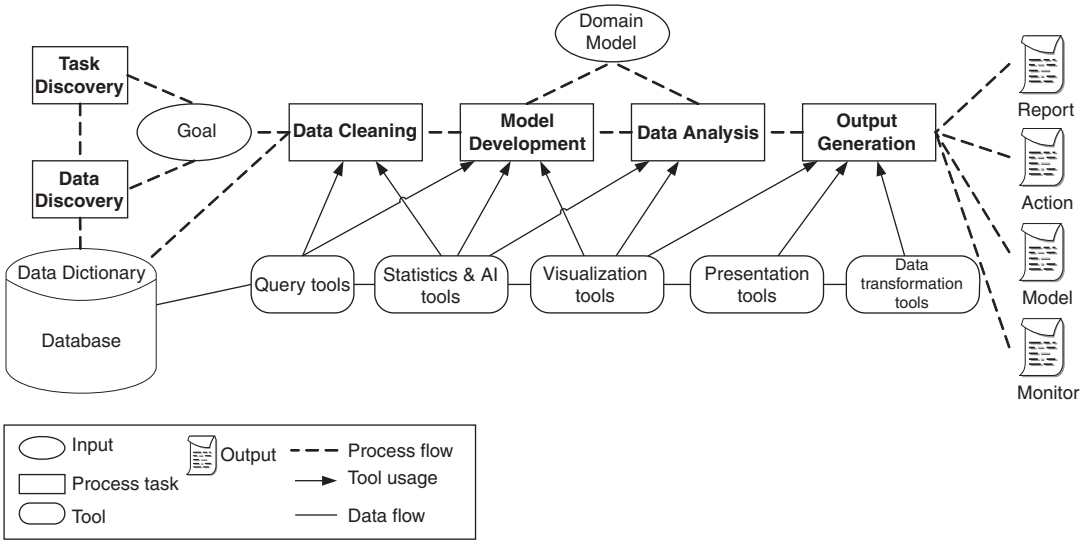


Figure 6 Human-centered process

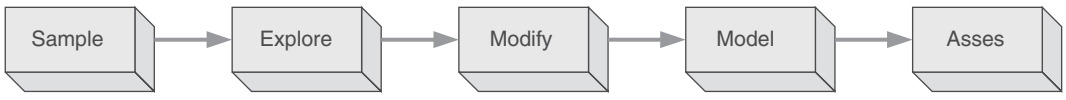


Figure 7 Sample, explore, modify, model, assess (SEMMA) methodology steps

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

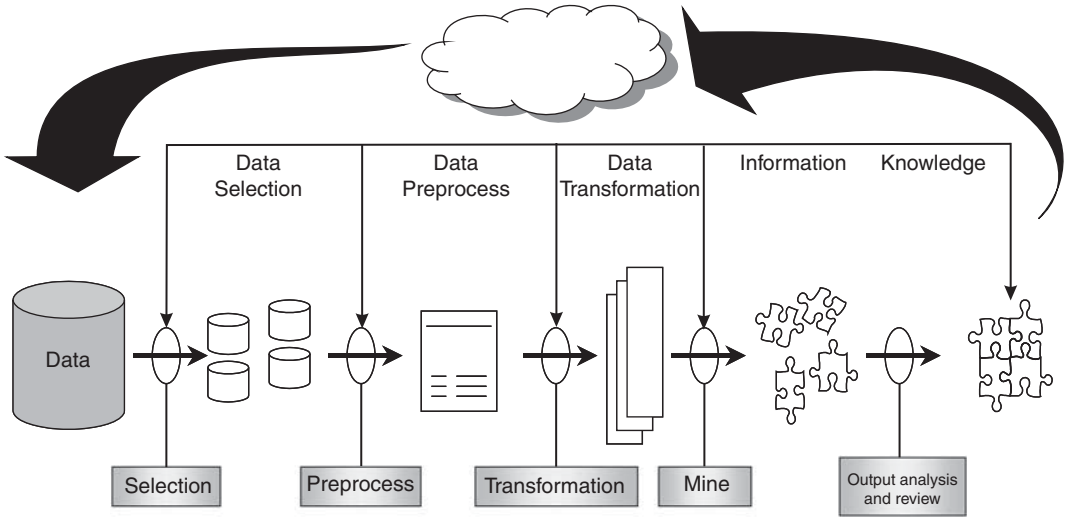


Figure 8 Data mining process according to (Cabena *et al*., 1997)

3.1.4 Cabena et al.

Cabena *et al*. define in (Cabena *et al*., 1997) data mining (referring to the complete KDD process) as the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions.

Figure 8 shows the steps of data mining process according to Cabena *et al*: business objectives determination, data preparation (that includes data selection, data preprocessing and data transformation), data mining, analysis of results, and assimilation of knowledge.

There are not big differences between the data mining tasks proposed by the original KDD process and Cabena *et al*. approach, although they structure the process in a different number of steps.

3.1.5 Two Crows

The Two Crows data mining process model is proposed by Two Crows Corporation (1999). This model is derived from the previous edition of the Two Crows process model (Two Crows Corporation, 1998), and also takes advantage of some insights from first versions of CRISP-DM (before CRISP-DM 1.0 is released).

While the steps appear in a list, the data mining process is not linear you will inevitably need to loop back to previous steps.

Figure 9 shows the basic steps of Two Crows: define business problem, build data mining data base, explore data, prepare data for modeling, build model, evaluate model, and deploy model and results.

Two Crows approach is very closed to the original KDD process, although they use different names for similar steps.

3.1.6 Anand and Buchner

Anand and Buchner (Buchner *et al*., 1999) have proposed a model covering the entire life cycle of an online customer, the available operational and materialized data, as well as the incorporation of marketing knowledge. The web-enabled knowledge discovery process, also known as internet-enabled knowledge discovery process, is an adoption of a generic process defined in earlier work (Anand & Buchner, 1998; Anand *et al*., 1998) adapted to web mining projects in this case.

As shown in Figure 10, the model consists of eight steps: human resource identification, problem specification, problem specification, data prospecting, methodology identification, data preprocessing, pattern discovery, and knowledge postprocessing. While it is true that Anand and

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

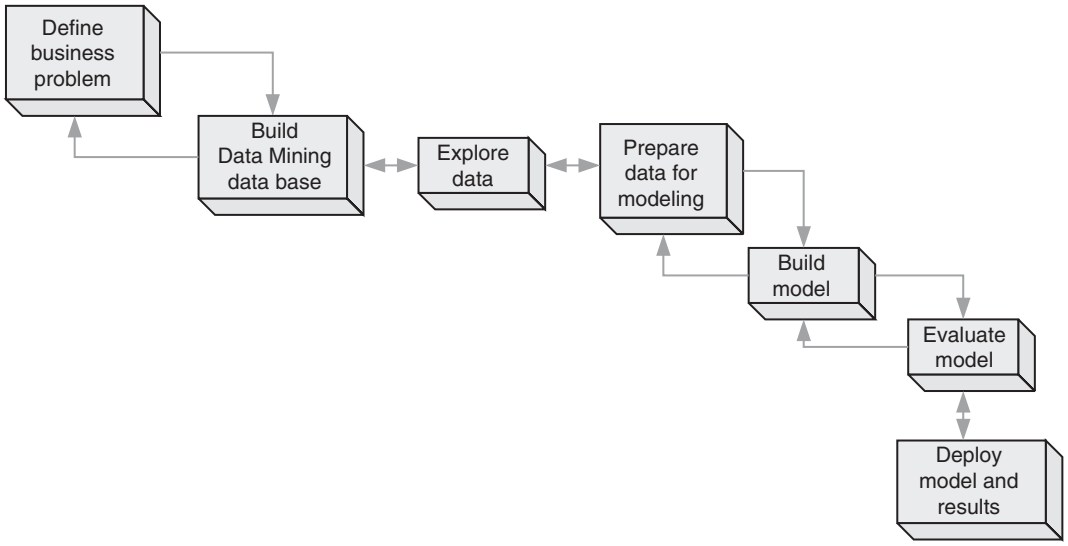


Figure 9 Two crows data mining process model

Buchner (1998) provide a detailed analysis of initial steps of the process, unfortunately, it does not include the needed activities to use the discovered knowledge.

3.2 CRISP-DM related approaches

3.2.1 CRISP-DM

In response to common issues and needs in data mining project in the mid 90's, a group of organizations involved in data mining (Teradata, SPSS -ISL-, Daimler-Chrysler and OHRA) proposed a reference guide to develop data mining projects, named CRISP-DM (*CRoss Industry Standard Process for Data Mining*) (Chapman *et al*., 2000). CRISP-DM is considered the *de facto* standard for developing data mining and knowledge discovery projects. One important factor of CRISP-DM success is the fact that CRISP-DM is industry-, tool- and application-neutral.

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 (from general to specific) (see Figure 11). At the top level, the data mining process is organized into a number of phases; each phase consists of several second-level generic tasks. This second level is called generic, because it is intended to be general enough to cover all possible data mining situations. The third level, the specialized task level, is the place to describe how actions in the generic tasks should be carried out in certain specific situations. The fourth level, the process instance, is a record of the actions, decisions and results of an actual data mining engagement.

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.

CRISP-DM distinguishes between four different dimensions of data mining contexts:

- The application domain is the specific area in which the data mining project takes place.
- The data mining problem type describes the specific classes of objectives that the data mining project deals with.
- The technical aspect covers specific issues in data mining that describe different (technical) challenges that usually occur during data mining.
- The tool and technique dimension specifies, which data mining tool(s) and/or techniques are applied during the data mining project.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

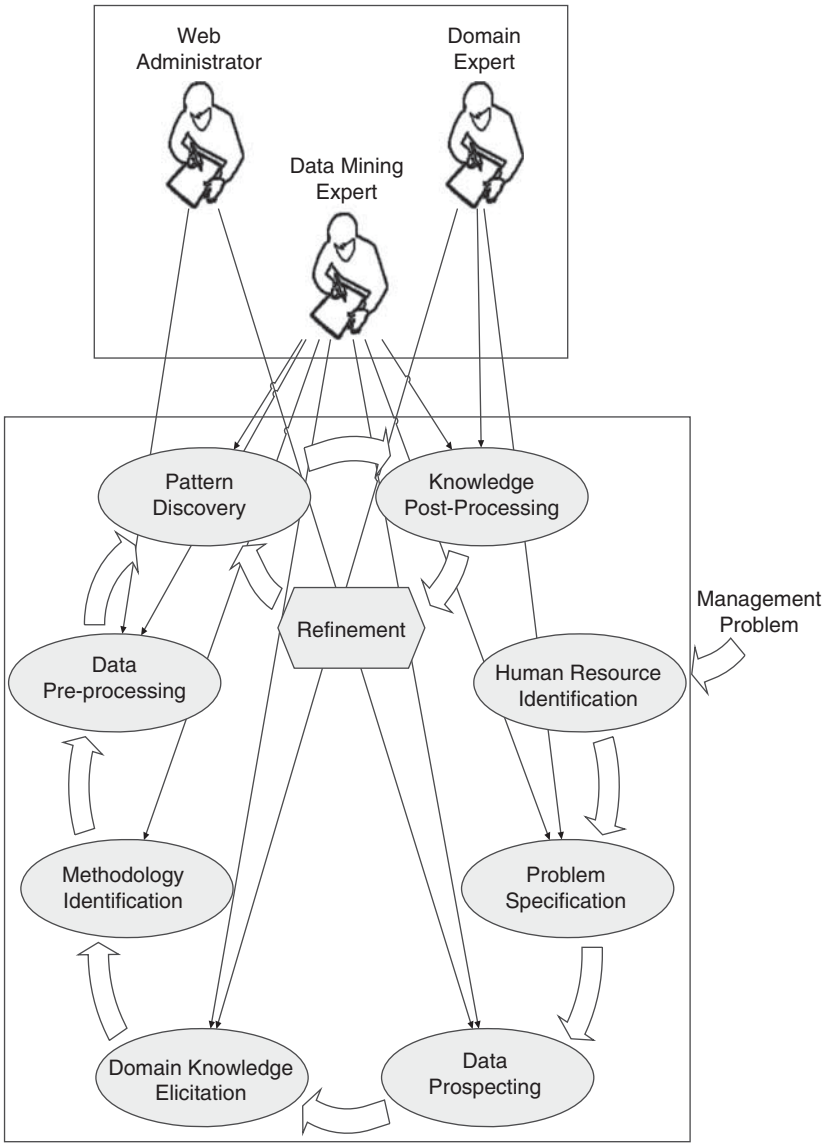


Figure 10 Anand and Buchner process model (Buchner *et al.*, 1999)

The CRISP-DM process model for data mining[3] 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.

The life cycle of a data mining project according to CRISP-DM consists of six phases (Figure 12). The sequence of the phases is not strict. Moving back and forth between different phases is always required. It depends on the outcome of each phase, which phase or which particular task of a phase, that has to be performed next. The arrows indicate the most important and frequent dependencies between phases.

[3] Actually, although authors refer to CRISP-DM as a process model, it is really an instanced process model because it establishes a waterfall life cycle (CRISP-DM states which tasks have to be carried out to successfully complete a data mining project and its order). Therefore, it must not be considered as a process model. It must not be considered as a pure methodology either, because it does not describe how to do all the tasks. It can be considered as a mixing between both terms.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

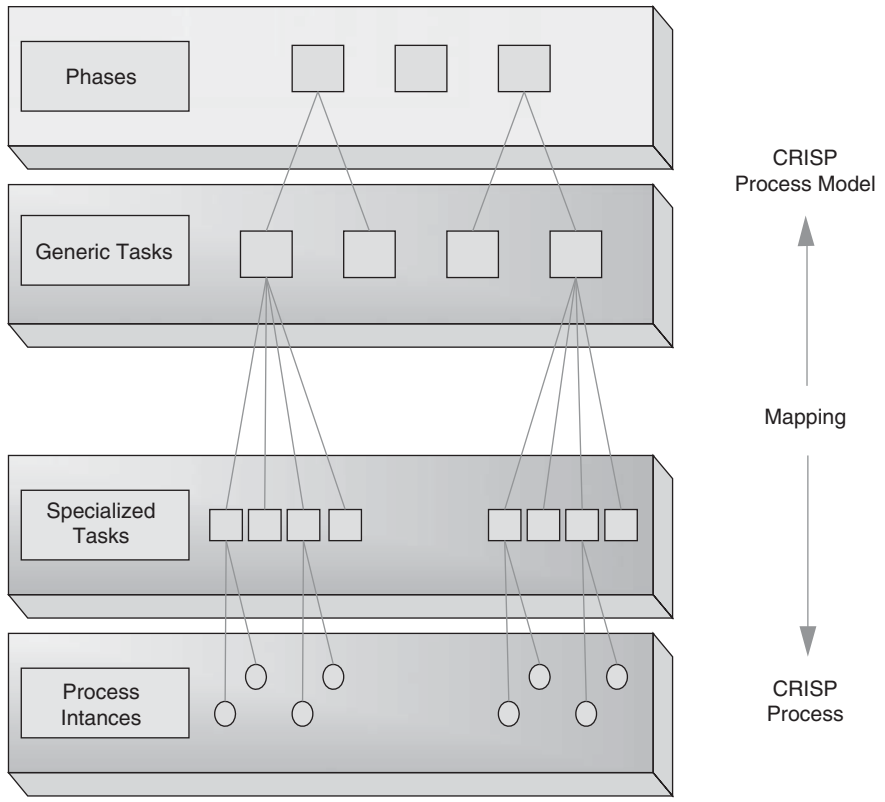


Figure 11 Four-level breakdown of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology

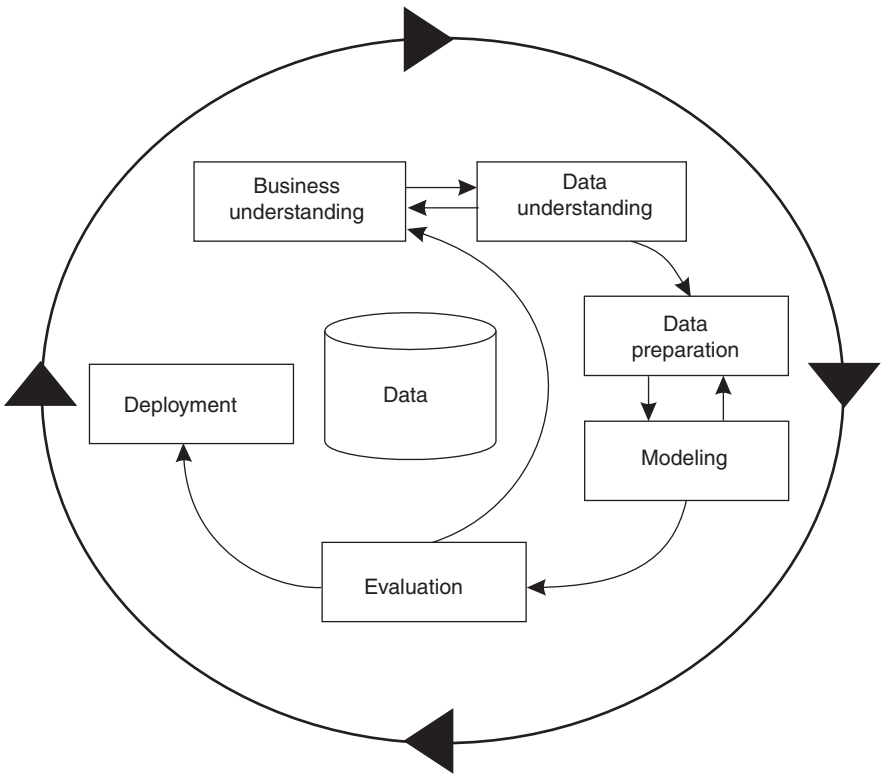


Figure 12 Cross-Industry Standard Process for Data Mining (CRISP-DM) process model (Chapman *et al.*, 2000)

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

Below follows a brief outline of the phases:

- 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.
- Data understanding
  The data understanding 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 preparation
  The data preparation phase covers all activities to construct the final data set (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. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.
- Modeling
  In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.
- Evaluation
  At this stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be sure 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.
- Deployment
  Generally, the creation of the model is not the end of the project. Even if the purpose of the model is to increase knowledge of the data, it will be necessary to organize the knowledge extracted, as well as to present it in a useful way to 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. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions will need to be carried out in order to actually make use of the created models.

3.2.2 CRISP-DM 2.0

Many changes have occurred in the business application of data mining since CRISP-DM 1.0 was published (CRISP-DM, 2007). Emerging issues and requirements include:

- The availability of new types of data (e.g., text, web and attitudinal data) along with new techniques for preprocessing, analyzing and combining them with related case data.
- Integration and deployment of results with operational systems such as call centers and web sites.
- Far more demanding requirements for scalability and for deployment into real-time environments.
- The need to package analytical tasks for non-analytical end users and integrate these tasks in business workflows.
- The need to seamlessly integrate the deployment of results and closed-loop feedback with existing business processes.
- The need to mine large-scale databases *in situ*, rather than exporting an analytical data set.
- Organizations' increasing reliance on teams, making it important to educate greater numbers of people on the processes and best practices associated with data mining and predictive analytics.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

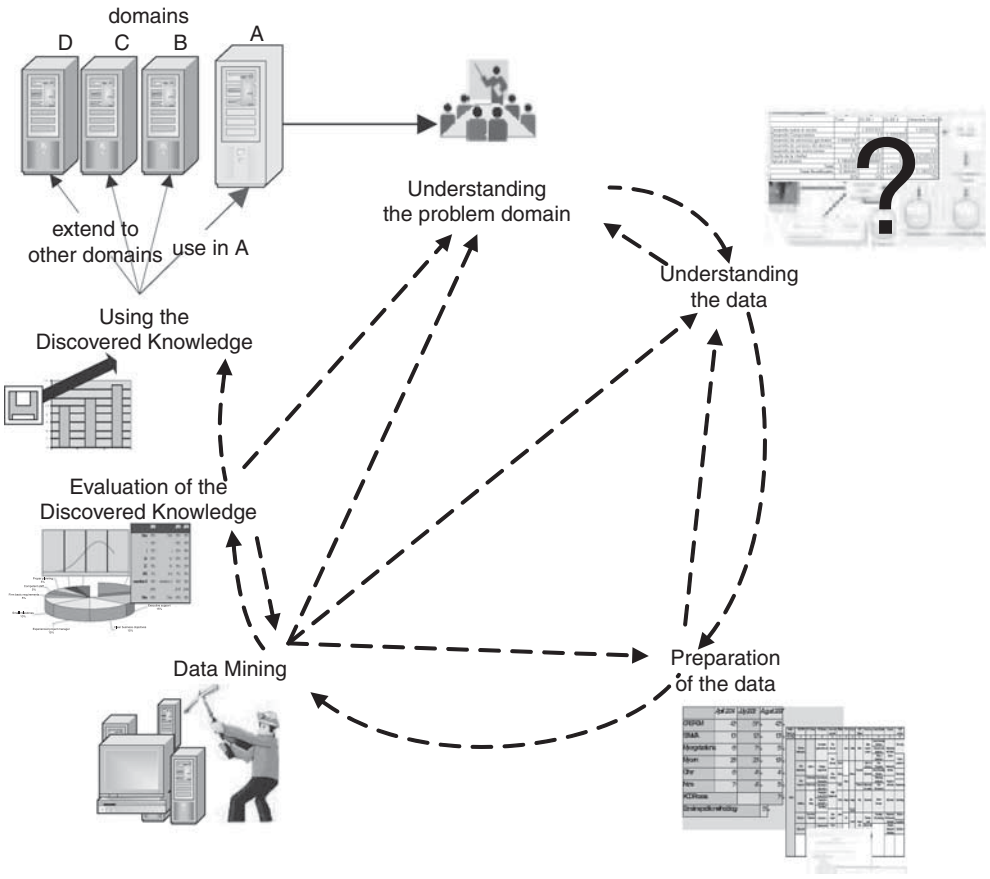


Figure 13 Cios *et al.*'s process model (Cios & Kurgan, 2005)

In response to evolving business needs, the CRISP-DM methodology is to be updated. For this purpose, it has been established the CRISP-DM 2.0 special interest group (SIG) to work in the new methodology. Apart from members of the original consortium established for CRISP-DM 1.0, vendors, service providers, researchers and end-users are being sought to join them in the consortium. Although there is not a definitive model yet, the CRISP-DM 2.0 SIG is working on it. This new methodology is expected to include changes such as adding new steps, renaming existing phases and/or deleting any old phase (The CRISP-DM Consortium, 2008).

3.2.3 Cios et al.

The process model of Cios *et al.* (Cios *et al.*, 2000; Cios & Kurgan, 2005) was first proposed in 2000 by adapting the CRISP-DM model to the needs of academic research community. The main extensions of the latter model include a more general, research-oriented description of the steps, introduction of several explicit feedback mechanisms and a modification of the description of the last step, which emphasizes that knowledge discovered for a particular domain may be applied in other domains. Cios *et al.* process model is based on technologies like XML, PMML, SOAP, UDDI and OLE BD-DM.

The model, as shown in Figure 13, consists of six steps: understanding the problem domain, understanding the data, preparation of the data, data mining, evaluation of the discovered knowledge and using the discovered knowledge.

Figure 13 shows that the process is iterative and interactive. Since any changes and decisions made in one of the steps can result in changes in later steps, the feedback loops are necessary.

3.2.4 Rapid collaborative data mining system (RAMSYS)

In Moyle and Jorge (2001) and Blockeel and Moyle (2002) RAMSYS methodology is described. RAMSYS is a methodology for performing data mining work where different groups are

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

geographically dispersed, but work together on the same problem in a collaborative way. The RAMSYS methodology attempts to achieve the combination of a problem solving methodology, knowledge sharing and ease of communication. It is guided by the following principles: it should enable light management, it should allow collaborators to start and stop any time and leave them problem-solving freedom, it should provide efficient knowledge sharing and security.

The RAMSYS methodology is based on CRISP-DM and keeps the same steps and generic tasks (see Figure 14). It can be regarded as a refinement of CRISP-DM, where some of the generic tasks have to be carried out in a collaborative mode, while accounting for the possibility of the 'remote' restriction. In the RAMSYS Methodology, the current best understanding of the problem is kept in the information vault, where information is shared between the different groups.

The RAMSYS methodology proposes a new task called model submission into the modeling step, where the current best models from each of the nodes will be selected, evaluated and delivered.

3.2.5 Data mining for industrial engineering (DMIE)

In a study by Solarte (2002) a methodology is presented for data mining projects oriented to industrial engineering domain. This methodology is based on CRISP-DM, but DMIE methodology establishes different phases (see Figure 15). DMIE consists of five steps: analyze the

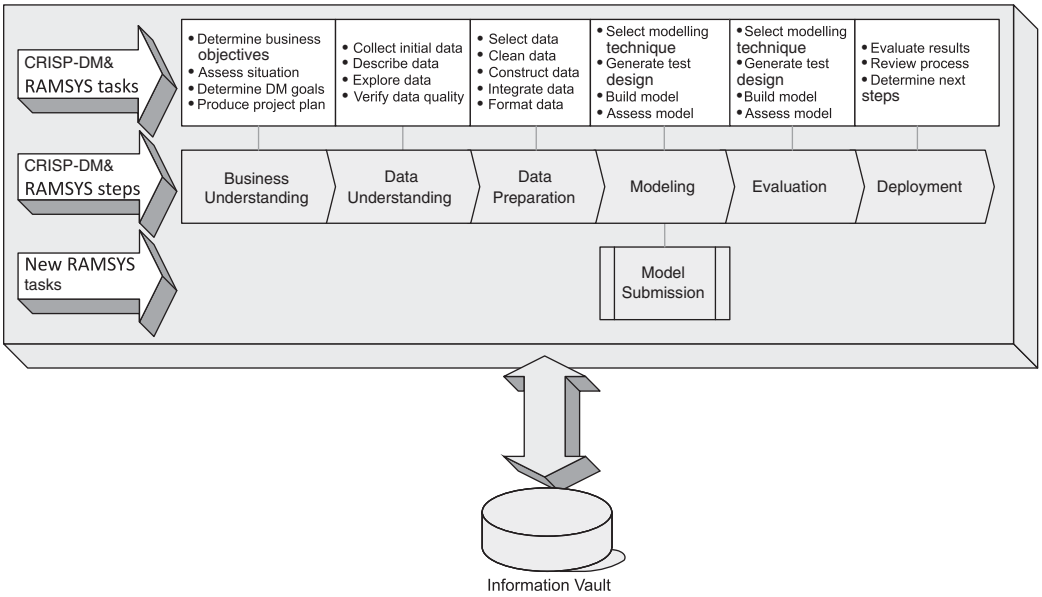


Figure 14 Rapid collaborative data mining system (RAMSYS) methodology

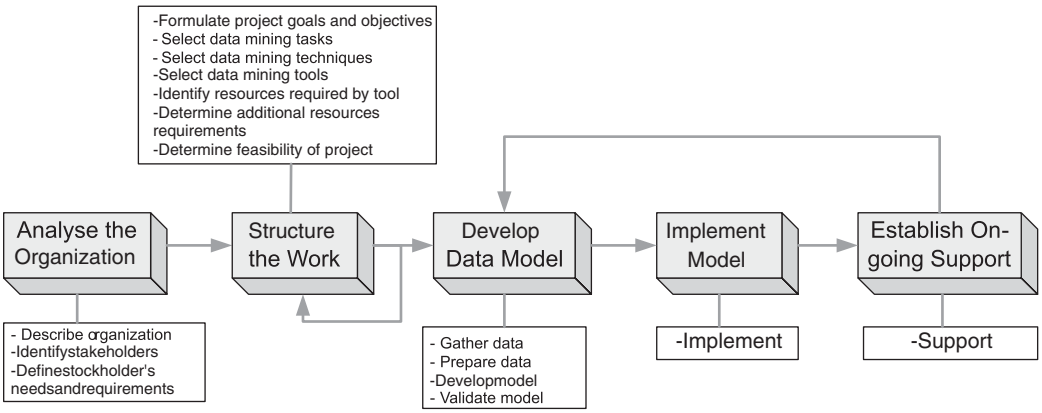


Figure 15 Data mining process for industrial engineering (DMIE) (Solarte, 2002)

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

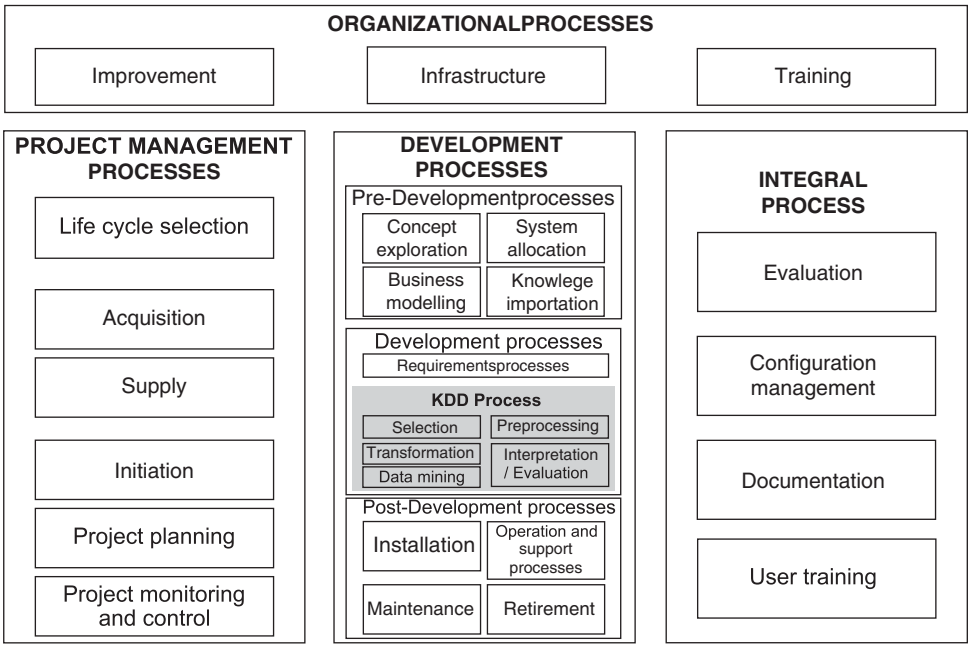


Figure 16 Process model for data mining engineering (Marbán *et al.*, 2007)

organization, structure the work, develop the data model, implement the model, establish on-going support.

The main contribution of the DMIE approach is a phase that is included only in it (until the moment DMIE was published), named on-going support. It consists of a support and maintenance phase, involving data backups, data maintenance, data mining model updates and software updates when needed.

3.2.6 Marbán et al.

This approach (Marbán *et al.*, 2007, 2008) is based on the idea that data mining problems are taking on the dimensions of an engineering problem. Therefore, the processes to be applied should include all the activities and tasks required in an engineering process, tasks that CRISP-DM (the most currently used data mining model) does not cover. (Marbán *et al.*, 2007) enhances CRISP-DM by embedding other current standards, as suggested in (Piatetsky-Shapiro, 2000; Kurgan & Musilek, 2006), inspired by the work done recently in software engineering derived from other branches of engineering and from developer experience. This approach proposes a data mining engineering process model that covers the above points, making a distinction between a process model and a methodology and life cycle.

Figure 16 shows a general scheme of Marbán *et al.* process model including most relevant subprocesses, pointing out the KDD process (based on Fayyad *et al.* (1996b) KDD process and CRISP-DM) as the core of the development processes. The rest of management and development processes are based on two software engineering standard process models: IEEE 1074 (IEEE, 1991) and ISO 12207 (ISO, 1995).

3.3 Other approaches

3.3.1 The five A's

SPSS defines the five A's methodology (de Pisón Ascacibar, 2003; SPSS, 2007) (assess, access, analyze, act, automate) as a general vision of data analysis and data mining processes. Five A's is closer to a trend in data mining project development than a process model, so that it does not describe how to develop a data mining project.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

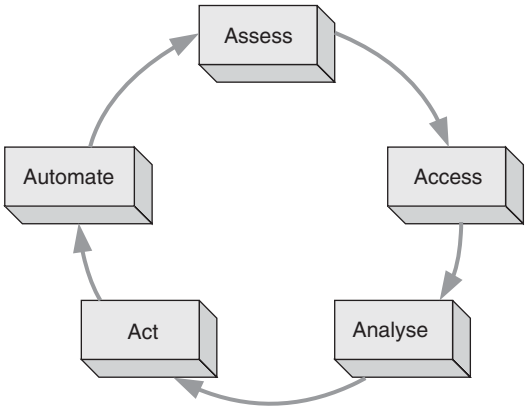


Figure 17 5 A's methodology phases

Figure 17 shows five A's steps.

The main contribution of five A's approach is the automate step. It is based on the interesting idea of the automation of data mining process in order to let non-expert data mining users to apply previous obtained models to new data. The resulting tool can be very useful for supporting non-expert data mining users to get new knowledge from new data in an easy way.

The negative aspect of five A's is that it does not establish other alternative ways of applying the built model or discovered knowledge, that is why automate step of five A's is only partially equivalent to deployment step of five A's.

Another important disadvantage of five A's is that five A's does not include data understanding step that is considered an essential step in CRISP-DM to understand the data and test its quality to prevent possible problems during the project development.

The five A's approach was abandoned by SPSS in 1999 when it joined CRISP-DM consortium to develop the CRISP-DM process model.

3.3.2 The six sigma (6-σ)

In mid-1996, Motorola developed the 6-σ method (Harry & Schroeder, 1999; Pyzdek, 2003). The 6-σ is a paradigm for quality and excellence in management. In other words, it defines how to improve quality and customer's satisfaction and, at the same time, reduce production costs.

On account of Motorola's success in applying 6-σ method, other companies like Texas Instrument, IBM, Kodak, General Electric, Ford, Microsoft or American Express have decided to apply this method in its production process (Arranz, 2007).

The 6-σ method has also been applied in data mining projects (StatSoft, 2005). It is based on DMAIC (define, measure, analyze, improve and control). The steps and tasks of 6-σ are shown in Figure 18.

3.3.3 KDD Roadmap

KDD Roadmap (Debuse *et al.,* 2001) is a data mining methodology used in *Witness Miner* toolkit (Group, 2006). As shown in Figure 19, KDD roadmap is an iterative methodology and it consists of eight steps: problem specification, resourcing, data cleansing, preprocessing, data mining, evaluation, interpretation and exploitation.

The main contribution of KDD roadmap is the resourcing task. It is considered an essential task in the knowledge discovery process as it becomes an independent step in KDD roadmap.

4 Data mining and knowledge discovery methodologies and process models comparative analysis

This section is intended to provide a detailed comparison between all methodologies and process models described above, focusing on phases and tasks included in every approach.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

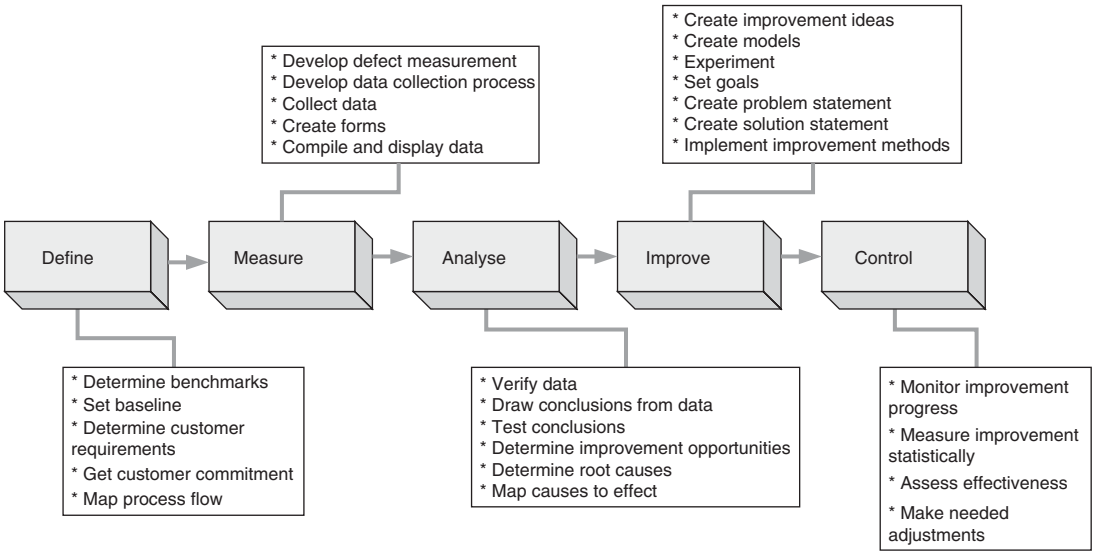


Figure 18 6-σ paradigm

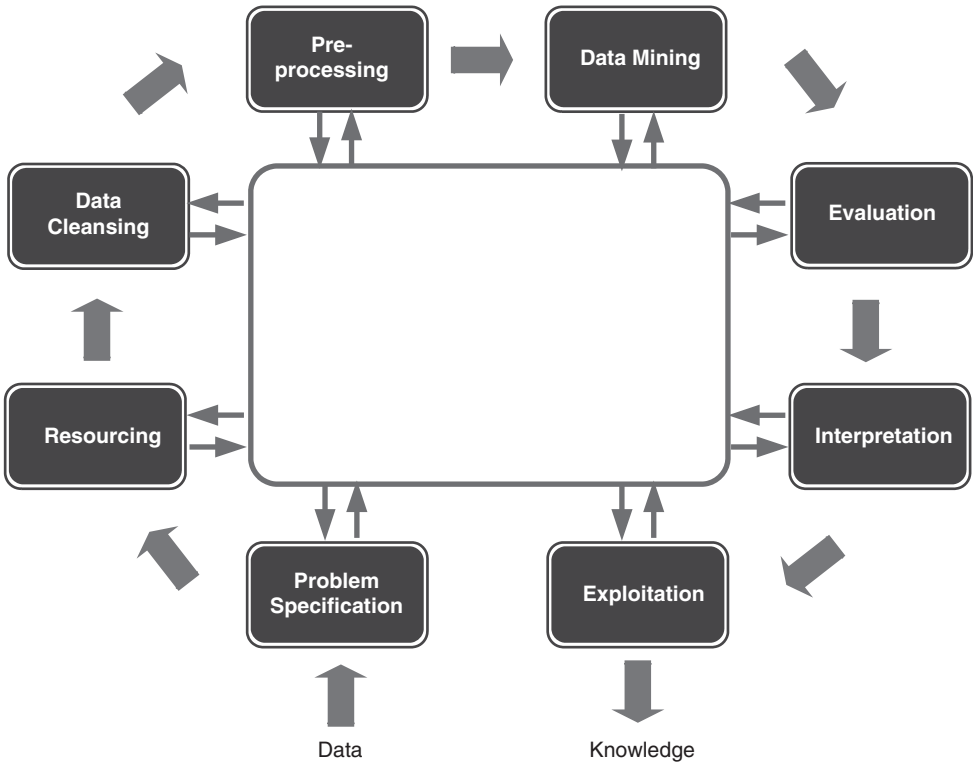


Figure 19 Knowledge discovery in databases (KDD) Roadmap

First of all, detailed and outlined descriptions of KDD process will be compared with CRISP-DM. In earlier sections of this paper, it was shown that CRISP-DM is the central approach of the evolution of data mining methodologies and process models. Moreover, CRISP-DM is the most used data mining methodology and it is considered the *de facto* standard. For all these reasons, we choose CRISP-DM as the reference model to be compared with the rest of the approaches in order to detect specific differences in the steps of the discovery process proposed by each approach in comparison with CRISP-DM.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

<table>
<tr><th>Methodology</th><th colspan="6">Phases</th></tr>
<tr><td>CRISP-DM</td><td>Business understanding</td><td>Data understanding</td><td>Data preparation</td><td>Modeling</td><td>Evalution</td><td>Deployment</td></tr>
<tr><td rowspan="3">KDD-Detailed</td><td rowspan="3">Learning the application domain</td><td rowspan="3">Creating a target data set</td><td>Data cleaning and pre-processing</td><td>Choosing the function of DM</td><td rowspan="3">Interpretation</td><td rowspan="3">Using discovered knowledge</td></tr>
<tr><td rowspan="2">Data reduction and projection</td><td>Choosing the DM algorithm</td></tr>
<tr><td>Data mining</td></tr>
<tr><td rowspan="2">KDD-Outlined</td><td rowspan="2"></td><td>Selection</td><td>Pre-processing</td><td>Modeling</td><td rowspan="2">Interpretation / Evalution</td><td rowspan="2"></td></tr>
<tr><td>Pre-processing</td><td>Transformation</td><td>Data mining</td></tr>
</table>

Figure 20 Knowledge discovery in databases (KDD) process vs. Cross-Industry Standard Process for Data Mining (CRISP-DM)

Therefore, the comparison is divided in three different mappings:

- KDD vs. CRISP-DM: To start with, the two main approaches will be compared. A mapping of the steps of every approach is shown, including KDD-detailed, KDD-outlined and CRISP-DM.
- CRISP-DM vs. KDD related approaches: The second mapping includes a comparison between CRISP-DM and KDD related approaches.
- CRISP-DM vs. CRISP-DM related approaches: The third mapping includes a comparison between CRISP-DM and CRISP-DM related approaches.
- CRISP-DM vs. other approaches: The last mapping includes a comparison between CRISP-DM and other independent approaches that are five A's, 6-σ and KDD Roadmap.

4.1 KDD process vs. CRISP-DM

Figure 20 shows the mapping between KDD process (KDD-outlined and KDD-detailed) steps and CRISP-DM steps.

The mapping focuses on phases and tasks included in every approach. When two different steps of two different approaches carry out a common task, it will be noticed with a match between both steps.

In the first two rows of Figure 20, we compare the six steps shown in Figure 12 (CRISP-DM) with the KDD nine steps (KDD-detailed). Both of them cover the same tasks, although KDD-detailed proposes more specific phases in this high-level analysis (six steps in CRISP-DM vs. nine steps in KDD). To be precise, data preparation step of CRISP-DM is split up into two steps in KDD-detailed: on the one hand, data cleaning and preprocessing, on the other hand, data reduction and projection. In addition, modeling step of CRISP-DM is separated into three steps in KDD-detailed: choosing the data mining function, choosing the data mining algorithm and data mining.

On the other hand, according to KDD-outlined vs. CRISP-DM comparison, it is shown that KDD-outlined is incomplete, as it does not include business understanding or deployment steps.

4.2 CRISP-DM vs. KDD related approaches

Figure 21 shows the mapping between CRISP-DM and KDD related approaches. We consider it is interesting to analyze if they propose contributions to CRISP-DM process (the currently most used process model).

In the first two rows, human-centered and CRISP-DM steps are compared. In general, both solutions cover the same tasks. For instance, task discovery step of human-centered sets the business objectives and it analyzes the data sources and its understanding, while CRISP-DM sets the business objectives in the business understanding step, and it analyzes the data sources and its understanding in the subsequent data understanding step.

On the other hand, in CRISP-DM, the data preprocessing step includes the data segmentation task that is carried out in the model development step in the human-centered approach.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

<table>
<tr><th>Methodology</th><th colspan="6">Phases</th></tr>
<tr><td>CRISP-DM</td><td>Business understanding</td><td>Data understanding</td><td>Data preparation</td><td>Modeling</td><td>Evaluation</td><td>Deployment</td></tr>
<tr><td rowspan="3">Human-Centered</td><td rowspan="3">Data discovery</td><td>Task discovery</td><td>Data cleaning</td><td rowspan="2">Model development</td><td rowspan="3">Data analysis</td><td rowspan="3">Output generation</td></tr>
<tr><td>Data discovery</td><td rowspan="2">Model development</td></tr>
<tr><td>Data cleaning</td><td>Data analysis</td></tr>
<tr><td rowspan="2">SEMMA</td><td rowspan="2"></td><td>Sample</td><td>Explore</td><td>Model</td><td rowspan="2">Assess</td><td rowspan="2"></td></tr>
<tr><td>Explore</td><td>Modify</td><td>Assess</td></tr>
<tr><td>Cabena et al.</td><td>Select</td><td>Preprocess</td><td>Pre-process
Transform</td><td>Mine</td><td>Analyse & assimilate</td><td>Analyse & assimilate</td></tr>
<tr><td>Two Crows</td><td>Define business problem</td><td>Build DM data base
Explore data</td><td>Explore data for modeling
Prepare data</td><td>Build model</td><td>Evaluate model</td><td>Deploy model and results</td></tr>
<tr><td rowspan="3">Anand & Buchner</td><td>Domain knowledge elicitation</td><td>Domain knowledge elicitation</td><td>Methodology identification</td><td>Methodology identification</td><td rowspan="3">Knowledge post-processing</td><td rowspan="3"></td></tr>
<tr><td>Human resource identification</td><td rowspan="2">Data prospecting</td><td rowspan="2">Data pre-processing</td><td rowspan="2">Pattern discovery</td></tr>
<tr><td>Problem specification</td></tr>
</table>

Figure 21 Cross-Industry Standard Process for Data Mining (CRISP-DM) vs. knowledge discovery in databases (KDD) related approaches

Moreover, the modeling step of CRISP-DM is separated into two steps in the human-centered approach. The first one is the model development step and it includes selecting the model, parameters and tools to be used. The second step is data analyzing, in which the model is performed, validated and refined, whereas in CRISP-DM the model is validated in the evaluation step.

Focusing on the mapping between the steps proposed by SEMMA and CRISP-DM approaches, SEMMA and CRISP-DM methodologies share the same philosophy because they structure the data mining project in steps connected among themselves. This is why the data mining process becomes iterative and interactive.

SEMMA methodology is more focused on the technical characteristics involved in the process development, whereas CRISP-DM methodology has a broader perspective regarding the business objectives of the project. This difference is shown from the beginning in the first data mining step, where SEMMA methodology starts with a data sampling, while CRISP-DM starts analyzing the business problem in order to turn it into a technical problem, as shown in Figure 21. From this viewpoint, looking at the project as a whole, it can be considered that CRISP-DM methodology is closer to the real project concept. Moreover, SEMMA does not include an explicit step to use the discovered knowledge, while CRISP-DM includes the deployment step.

Focusing on the mapping between (Cabena *et al.*, 1997) and CRISP-DM steps. Both approaches cover practically the same tasks to complete the whole KDD process. There are some overlaps between tasks of each process. For instance, preprocess step of Cabena *et al.* apart from detecting mistakes in data, prepares a set of suited data for the later analysis, while data understanding step of CRISP-DM just selects data sources and analyze data quality without preparing data for the next step.

Apart from that, we can point out that Cabena *et al.* model joins evaluation and deployment CRISP-DM steps in analyze and assimilate step. Although analyze and assimilate step in Cabena *et al.* model carries out equivalent tasks to two named steps of CRISP-DM, it does not include reports or documents related to obtained knowledge.

Focusing on the mapping between Two Crows data mining process model steps and CRISP-DM steps, Figure 21 shows that both models are quite similar, although there is a clear difference. Data understanding and data preparation steps of CRISP-DM that are distributed in Two Crows

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

in three steps: build data mining data base, explore data and prepare data for modeling. Explore data step of Two Crows consists of tasks extracted from both data understanding and data preparation CRISP-DM steps.

In Figure 21, Anand and Buchner (1998) steps are compared with CRISP-DM steps. As shown, Anand and Buchner (1998) does not include deployment step of CRISP-DM to use discovered knowledge. Anand and Buchner approach is the only one that proposes carrying out the algorithm selection (in methodology identification step) before data preprocessing. In CRISP-DM, algorithm selection task is included in modeling step, after preparing the data.

4.3 CRISP-DM vs. CRISP-DM related approaches

Figure 22 shows a comparison between CRISP-DM and CRISP-DM related approaches. It is worthwhile examining the possible contributions of these approaches to CRISP-DM.

The first two rows of the figure shows the mapping between Cios *et al.* (2000) approach steps and CRISP-DM steps, Figure 22 shows that both approaches are quite similar from a high-level viewpoint. To find differences, it is necessary focusing on process iterations. For example, Cios approach let the process go back from data preparation step to data understanding step, while CRISP-DM does not show it as a common transition.

The RAMSYS methodology is based on CRISP-DM and keeps the same steps and generic tasks (Figure 22). It can be regarded as a refinement of CRISP-DM, where some of the generic tasks have to be carried out in a collaborative mode, while accounting for the possibility of the 'remote' restriction. As RAMSYS only introduce some changes at task level and it does not change the CRISP-DM steps, the mapping in figure shows the same steps.

Figure 22 also shows the mapping between DMIE (Solarte, 2002) steps and CRISP-DM steps. There are clear differences between both models as shown in that figure. First, business understanding step of CRISP-DM is considered by DMIE as two different steps: analyze the organization and structure the work. According to this, DMIE separates tasks related to organization from tasks related to project management. However, DMIE joins develop data model, data understanding and data preparation steps. DMIE also joins modeling, evaluation and deployment steps of CRISP-DM in implement model step. The main contribution of DMIE approach is a phase that only it includes (until the moment DMIE was published), named on-going support. It consists of a support and maintenance phase, involving data backups, data maintenance, data mining model updates and software updates when needed.

About Marbán *et al.* (2008), it is a process model based on software engineering standard process models, CRISP-DM model and KDD process.

It classifies processes in three large groups: project management processes (including life cycle selection processes), development processes (including predevelopment processes, development processes itself, and postdevelopment processes) and integral processes.

Figure 22 shows the mapping between process groups (or subgroups) included in Marbán *et al.* and CRISP-DM steps.

<table>
<tr><th>Methodology</th><th colspan="8">Phases</th></tr>
<tr><td>CRISP-DM</td><td></td><td>Business understanding</td><td>Data understanding</td><td>Data Preparation</td><td>Modeling</td><td>Evalution</td><td>Deployment</td><td></td></tr>
<tr><td>Cios et al.</td><td></td><td>Understanding the problem domain</td><td>Understanding the data</td><td>Preparation of the data</td><td>Build model</td><td>Evalution of the discovered knowledge</td><td>Using the discovered knowledge</td><td></td></tr>
<tr><td>RAMSYS</td><td></td><td>Business understanding</td><td>Data understanding</td><td>Data preparation</td><td>Modeling</td><td>Evalution</td><td>Deployment</td><td></td></tr>
<tr><td>DMIE</td><td></td><td>Analyse the organization
Structure the work</td><td>Develop data model</td><td>Develop data model</td><td>Implement model</td><td>Implement model</td><td>Implement model</td><td>Establish on-going support</td></tr>
<tr><td>Marbán et al.</td><td>Life cycle selection</td><td>Project management
Pre-development
Development</td><td>Pre-development
Development</td><td>Development</td><td>Development
Integral processes</td><td>Development
Integral processes</td><td>Project management
Integral processes
Post-development</td><td></td></tr>
</table>

Figure 22 Cross-Industry Standard Process for Data Mining (CRISP-DM) vs. CRISP-DM-related approaches

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

<table>
<tr><th>Methodology</th><th colspan="7">Phases</th></tr>
<tr><td>CRISP-DM</td><td>Business understanding</td><td>Data understanding</td><td>Data preparation</td><td>Modeling</td><td>Evaluation</td><td>Deployment</td><td></td></tr>
<tr><td>5 A's</td><td>Assess</td><td></td><td>Access</td><td>Analyse</td><td>Act</td><td></td><td>Automate</td></tr>
<tr><td>6-sigma</td><td>Define</td><td>Measure</td><td>Measure</td><td>Analyse
Improve</td><td>Control</td><td>Control</td><td></td></tr>
<tr><td rowspan="2">KDD Roadmap</td><td>Problem specification</td><td>Problem specification</td><td>Data cleaning</td><td rowspan="2">Data mining</td><td>Evaluation</td><td rowspan="2">Exploitation</td><td rowspan="2"></td></tr>
<tr><td>Resourcing</td><td>Resourcing
Data cleaning</td><td>Pre-processing</td><td>Interpretation</td></tr>
</table>

Figure 23 Cross-Industry Standard Process for Data Mining (CRISP-DM) vs. other approaches

From the comparison of CRISP-DM with a software engineering process model (Marbán *et al.*, 2008), it is found that many of the processes defined in software engineering that are very important for developing any type of DM engineering project are missing from CRISP-DM. This could be the reason why CRISP-DM is not as effective as it should be.

The activities missing from CRISP-DM are primarily project management processes, integral processes (that assure project function completeness and quality) and organizational processes (that help to achieve a more effective organization).

4.4 CRISP-DM vs. other approaches

Figure 23 shows the mapping between CRISP-DM and other independent approaches: five A's, 6-σ and KDD Roadmap. Five A's and 6-σ approaches are early approaches that were proposed after KDD and before CRISP-DM. In contrast, KDD Roadmap was proposed after CRISP-DM.

Focusing on the comparison between the steps proposed by five A's (de Pisón Ascacibar, 2003) and CRISP-DM, Figure 21 shows that there are two main differences between them. First, five A's does not include data understanding step that is considered an essential step in CRISP-DM to understand the data and test its quality to prevent possible problems during the project development.

The main contribution of five A's approach is the automate step. The result is a useful tool for supporting non-expert data mining users to get new knowledge from new data in an easy way.

Analyzing CRISP-DM vs. 6-σ comparison in Figure 21, 6-σ joins data understanding and data preparation steps of CRISP-DM in measure step of 6-σ.

However, 6-σ considers important to separately analyze and improve steps in order to build the model.

These tasks are joint in the modeling step of CRISP-DM, although CRISP-DM points out the need of carrying out both tasks in an iterative way before going forward with the next step. According to this, CRISP-DM tries out different models before ending the modeling step, whereas 6-σ does not go back after analyzing and improving the steps just once.

Taking into account the comparison between KDD Roadmap (Debuse *et al.*, 2001) steps and CRISP-DM steps, although both approaches are quite similar, a clear difference is shown: KDD Roadmap includes one phase more than CRISP-DM. To be precise, KDD Roadmap considers that the business understanding step of CRISP-DM must be divided into problem specification and resourcing steps. According to this, KDD Roadmap emphasize the importance of the resourcing task by creating an independent step.

5 Refined data mining process

In this section, a global mapping of all approaches will be shown. After that, a new *refined data mining process* is built, based on this global comparative analysis.

Figure 24 shows a global comparison of steps of every described approaches. In addition, a generic data mining process model is proposed and detailed in last two columns of the table in that figure. We have named it *Refined Data Mining Process*.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

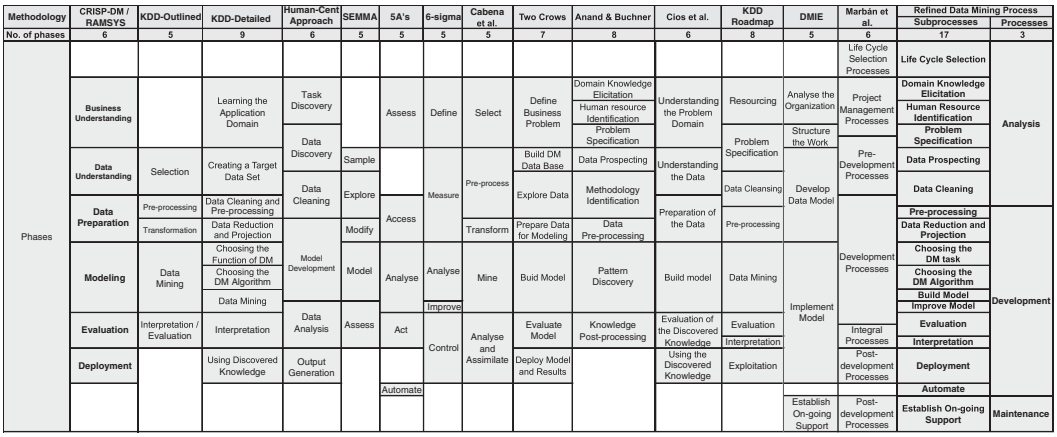


Figure 24 Cross-Industry Standard Process for Data Mining (CRISP-DM) vs. rest of the approaches and refined data mining process.

We consider that a new process model for developing data mining and knowledge discovery projects is necessary as any of the existing approaches is complete. Some steps are included only in some specific proposals; that is, life cycle selection in Marbán *et al.* (2007) model, on-going support in DMIE (Solarte, 2002), or automate in five A's (de Pisón Ascacibar, 2003). Hence, new important phases are appearing in new approaches.

CRISP-DM is the most widely used methodology for developing data mining projects, but it does not include some project management activities. Currently, not all data mining results are positive.

On the other hand, the proposed *Refined Data Mining Process* includes more specific phases than other approaches. One of the reasons of the problems in data mining and knowledge discovery is the numerous dependencies between the various phases and tasks of a data mining project (Sharma & Osei-Bryson, 2009). Therefore, it will be easier to follow the *Refined Data Mining Process* and identify dependencies between phases. Anyway, this new approach is just a proposal and it has not been tested yet in real projects.

The refined data mining process consists of three large processes: analysis, development and maintenance. At the same time, these processes are divided into subprocesses. The refined data mining process is composed of 17 subprocesses extracted from the methodologies and process models described above.

The names of the three large processes (analysis, development and maintenance) have been chosen based on terminology commonly used in project development methodologies in any engineering area. Every subprocess of the refined data mining process are included in one of the three main processes.

The names of the 17 subprocesses have been chosen based on terminology used in described approaches. The established criteria to select steps from the described approaches is based on concreteness and relevance of phases defined in any approach. Each approach considers different tasks and steps as the most important and relevant in the KDD process. After building the global comparative table (see Figure 24) we have selected the most specific steps avoiding overlaps and covering every task included in any data mining and knowledge discovery project development.

According to that selection of steps, we propose the refined data mining process as a new theoretical process model for developing data mining and knowledge discovery projects. We will focus on the description of subprocesses, but we will not establish a concrete life cycle for the refined data mining process because we consider that a process model defines *what to do*, and a methodology defines *how to do it* including the order in which tasks had to be carried out. The definition of a methodology (with a concrete life cycle) based on the refined data mining process is proposed as a future challenge.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

The 17 subprocess of the refined data mining process are described below:

- Analysis process
  - o Life cycle selection

    This subprocess has been extracted from Marban *et al*. process model (Marbán *et al*., 2007), and it consists of three essential tasks to be carried out in any engineering project: acquisition, supply and life cycle selection.

    The purpose of this subprocess is to identify and select a life cycle for the project that is to be developed. Possible life cycle models are identified and analyzed based on the type of project to be developed and its requirements. Then a model that provides proper support for the project is selected. This set of processes also extends to third party resource acquisition and supply. These two processes cover all the tasks related to supply or acquisition management.

    CRISP-DM does not include any of the acquisition or supply processes at all. Authors own experience in DM project development suggests that acquisition and supply processes may be considered necessary and third parties engaged to develop or create DM models for projects of some size or complexity. Their management should therefore be specified as processes.
- Domain knowledge elicitation

  This subprocess is extracted from Anand and Buchner (1998), and it must be considered as an important one because domain knowledge elicitation can be a complex and useful task to have a better understanding of the data and the problem. The main objective of domain knowledge elicitation and later incorporation at the pattern discovery stage is to constrain the learning algorithms search space and to reduce the number of patterns discovered. Marketing knowledge is a type of domain expertise, obtained internally or externally has usually been formulated by (human or artificial) marketing experts.
- Human resource identification

  This subprocess is extracted from Anand and Buchner (1998) too, and it is an essential task to define a proper project team and project participants.

  After a problem has been identified at the management level of a virtual enterprise, human resource identification is the first stage of the knowledge discovery process, which requires domain, data and data mining expertise. The synergy of these human resources as early as possible within any project is imperative to its success.
- Problem specification

  This subprocess is extracted from Anand and Buchner (1998). DMIE (Solarte, 2002) also includes it as *Structure the work*. This subprocess is important to establish concrete data mining objectives. Otherwise, we could make wrong efforts answering the wrong questions during the KDD process and it would cause delays.

  At the problem specification stage, a better understanding of the problem is developed by the human resources identified in the previous stage. Project objectives and goals are defined from the data mining viewpoint. These objectives and goals have to be clear, concrete and understandable by every participant in the project. At the same time, it has to be considered that objectives and goals are dynamic due to changes in the requirements of the organization or business, and they must be reviewed and updated regularly. The problem is decomposed into tasks and those tasks that can be solved using a knowledge discovery approach are identified. Each of these tasks is associated with a particular goal.

  In addition, tasks, techniques and tools to be used in the project must be selected in this step, taking into account objectives and goals.
- Data prospecting

  This subprocess is extracted from Anand and Buchner (1998), Two Crows (Two Crows Corporation, 1999) (where named *Build Data Mining data base*), and SEMMA (SAS Institute, 2005) (where named *Sample*).

  This subprocess needs to be included in refined data mining process because it includes complex and important tasks. This subprocess is usually the first of the four subprocesses that constitute the data preparation. All four together take more time and effort than all the other subprocesses combined.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

In order to perform data prospecting, the data mined should be collected. In addition, you may want to bring data from outside your company or you may want to add new fields computed from existing fields. Apart from that, a model is selected to research the efficacy of data mining.

- Data cleaning

  This subprocess is extracted from KDD Roadmap methodology (Debuse *et al.*, 2001), where it is named *data cleansing*. Human-centered (Brachman & Anand, 1996) also includes this subprocess as *data cleaning*. This task is very important in KDD process to get cleaned data to work with later.

  The aim of the data cleaning stage is to prepare and clean the data for subsequent tasks of the process. The cleaning stage involves operations such as searching for and removal of errors, sampling, dealing with outliers, missing and unreliable values and possibly balancing.

- Development process
  - o Preprocessing

    This subprocess is included in KDD (Fayyad *et al.*, 1996b), KDD Roadmap (Debuse *et al.*, 2001), (Cabena *et al.*, 1997), and Anand and Buchner (1998) approaches.

    Although the entire process is iterative, the preprocessing is one of the main subprocesses that is applied a number of times; second only to the data mining subprocess. For which reason, we include the preprocessing subprocess in the refined data mining process.

    Preprocessing is usually the first place where learning takes place and potentially useful patterns can be found.
  - o Data reduction and projection

    This subprocess is extracted from the original KDD process Fayyad *et al.* (1996b). It includes essential tasks to complete the data preparation.

    It includes finding useful features to represent the data, depending on the goal of the task, and using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.
  - o Choosing the data mining function

    According to KDD process by Fayyad *et al.* (1996b), this subprocess involves deciding the purpose of the model derived by the data mining algorithm and deciding the data mining techniques to apply (e.g., summarization, classification, regression and clustering).
  - o Choosing the data mining algorithm

    According to KDD process by Fayyad *et al.* (1996b), it includes selecting method(s) to be used for searching for patterns in the data, such as deciding which models and parameters may be appropriate and matching a particular data mining method with the overall criteria of the KDD process.
  - o Build model

    All approaches include this subprocess. This step is also called *data mining*, as some approaches refer to it. In this subprocess, the data mining model or models are applied to the data.
  - o Improve model

    This subprocess is only included in 6-σ (Harry & Schroeder, 1999). Most of the approaches include this subprocess as part of Build model subprocess. Six Sigma is focused in quality aspects and we agree that Improve model must be considered as an independent subprocess including concrete tasks to do it.

    It involves an iterative process repeated as many times as necessary. It includes reviewing a model that was applied in Build model subprocess and trying to improve it.
  - o Evaluation

    Most of the approaches include this subprocess, although there is a range of approaches, which can be used for evaluating the results of the Build model subprocess. The decision as to which method to use will depend on the data mining goal and the interest measure chosen to compare results. This subprocess includes evaluating the performance of the discovered knowledge with previously unseen examples (a test database is needed).

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

- Interpretation

  This subprocess is only included in KDD Roadmap (Debuse *et al.*, 2001) as an independent task out of evaluation. Interpreting the results is an advanced task of evaluation that is performed by domain experts.

  The crucial test for discovered knowledge is to satisfy the domain experts who can justify the results using their much deeper knowledge of the problem domain. Patterns that differ greatly from the knowledge of the domain expert should be carefully analyzed to explain such an anomaly.
- Deployment

  This subprocess is extracted from CRISP-DM (Chapman *et al.*, 2000). Other approaches includes this subprocess with other names as using discovered knowledge (KDD; Fayyad *et al.*, 1996b), Output generation (human-centered; Brachman & Anand, 1996), deploy model and results (Two Crows; Two Crows Corporation, 1999), Using the discovered knowledge (Cios *et al.*, 2000), and exploitation (KDD Roadmap; Debuse *et al.*, 2001). This is an essential subprocess that consists of applying the discovered knowledge. Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained need to be organized and presented in a way that the customer can use it. 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. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models.
- Automate

  This subprocess is only included in five A's methodology (SPSS, 2007; de Pisón Ascacibar, 2003). It is based on the interesting idea of the automation of data mining process in order to let non-expert data mining users to apply previous obtained models to new data.

  We include this subprocess because we consider that the resulting tool can be very useful for supporting non-expert data mining users to get new knowledge from new data and existing models in an easy way.

- Maintenance process
  - Establish on-going support

    This subprocess is only included as an independent subprocess in DMIE (Solarte, 2002), although Marbán *et al.* proposal also includes maintenance tasks. We totally agree that maintenance tasks are essential in any engineering project, and it is not different in data mining projects. Data mining projects need a support and maintenance subprocess.

    This maintenance involves data backups, data maintenance, data mining model updates (because new data appears and data mining models can change their behavior), and software updates when needed.

6 Conclusions

To date, many data mining and knowledge discovery methodologies and process models have been developed, with varying degrees of success. In this paper, we have described the most used (in industrial and academic projects) and cited (in scientific literature) data mining and knowledge discovery methodologies and process models, providing an overview of its evolution along with data mining and knowledge discovery history and setting down the state of the art in this topic.

For every approach, we have provided a detailed description of the proposed KDD process, discussing about special features, outstanding advantages and disadvantages of every approach.

Currently, the most widely used data mining model is CRISP-DM (Chapman *et al.*, 2000), considered as the *de facto* standard. CRISP-DM does not cover many tasks related to project management, organization and quality in the way required by the increasing complexity of the

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

more recent data mining and knowledge discovery projects. These projects not only involve examining huge volumes of data, but also managing and organizing big interdisciplinary human teams. All the above goes to show that while CRISP-DM was an improvement on the earlier state of affairs, the process model is not perhaps yet mature enough to deal with the complexity of the problems it has to address. And this detracts from the effectiveness of its deployment, as it does not produce the expected results.

As explained in Piatetsky-Shapiro (2000) and Kurgan and Musilek (2006), the challenge for the 21st century data miners is to develop and popularize widely accepted standards that, if adopted, will stimulate major industry growth and interest. Perhaps it is time to take a look at other mature engineering fields as software engineering, with over 40 years of experience, and taking up widely accepted standards from it and adapt them to data mining. We found that many of the processes defined in software engineering that are very important for developing any type of DM engineering project are missing from CRISP-DM. The activities missing from CRISP-DM are primarily project management processes, integral processes (that assure project function completeness and quality) and organizational processes (that help to achieve a more effective organization). This could be the reason why CRISP-DM is not as effective as it should be.

The idea behind Marbán *et al.* approach (Marbán *et al.*, 2007), is that data mining problems are taking on the dimensions of an engineering problem. Therefore, the processes to be applied should include all the activities and tasks required by an engineering process, that are not covered by CRISP-DM. Marbán *et al.* (2007) enhances CRISP-DM by embedding other current standards. This approach proposes a data mining engineering process model that covers CRISP-DM mistakes, making a distinction between a *process model* and a *methodology* and *life cycle*. Marbán *et al.* process model is based on the current data mining *de facto* standard CRISP-DM, and the two most used software engineering standard process models: IEEE 1074 (IEEE, 1991) and ISO 12207 (ISO, 1995).

From a formal and strict engineering viewpoint, only Marban *et al.* approach (Marbán *et al.*, 2007) can be considered as a true process model according to the definition given in Section 2. Although some authors classify its proposals as process models (Fayyad *et al.*, 1996b; Cabena *et al.*, 1997; Two Crows Corporation, 1999; Chapman *et al.*, 2000), they cannot be considered correctly as it. They must be better classified as methodologies that establish tasks to do with a concrete life cycle that sets the order in which tasks must be done.

From the standardization viewpoint in data mining and knowledge discovery area, the year 2000 marked the most important milestone when CRISP-DM 1.0 was launched in response to common issues and needs in data mining project in the mid 90s by a group of organizations involved in data mining (Teradata, SPSS -ISL-, Daimler, Chrysler and OHRA). These organizations established a consortium called SIG.

In addition, the future methodology CRISP-DM 2.0 (CRISP-DM, 2007), currently under development, seems to follow the right steps for standardization. In this case, apart from members of the original consortium established for CRISP-DM 1.0, vendors, service providers, researchers and end-users are being sought to join them in the CRISP-DM 2.0 SIG consortium. The CRISP-DM 2.0 methodology is to be developed in response to evolving business needs.

Apart from the approaches description and comments, a global comparative of all the described data mining approaches has been provided, focusing on the different steps in which every approach interprets the whole KDD process. Previously, some surveys about process models and methodologies have been published. It is fitting to point out (Kurgan & Musilek, 2006), but it focuses in the comparison of just five process models and methodologies, and it does not go into how they work in depth as it has been done in this paper.

As a result of the presented comparison, we have proposed a new data mining and knowledge discovery process model named *Refined Data Mining Process* for developing any kind of data mining and knowledge discovery project. The refined data mining process has been built based on specific steps taken out of analyzed approaches.

The refined data mining process is composed of three large processes: analysis, development and maintenance. At the same time, these processes are divided into subprocesses. The refined data

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

mining process is composed of 17 subprocesses extracted from analyzed methodologies and process models: life cycle selection, domain knowledge elicitation, human resource identification, problem specification, data prospecting, data cleaning, preprocessing, data reduction and projection, choosing the data mining function, choosing the data mining algorithm, Build model, Improve model, evaluation, interpretation, deployment, automate, and establish on-going support.

The established criteria to select steps from the described approaches were based on concreteness and relevance of phases defined in any approach. We have selected the most specific steps avoiding overlaps and covering every task included in any data mining and knowledge discovery project development. We have focused on the description of subprocesses, but we have not established a concrete life cycle for the refined data mining process. The definition of a concrete methodology (with a concrete life cycle) based on the refined data mining process is proposed as a future challenge.

Acknowledgement

This work has been partially funded by the project no. TIN 2008-05924/TIN of the Ministry of Science and Innovation of Spain.

References

Agrawal, R. & Shafer, J. C. 1996. Parallel mining of association rules. *IEEE Engineering in Medicine and Biology Magazine Trans. On Knowledge and Data Engineering* **8**, 962–969.

Anand, S. & Buchner, A. 1998. *Decision Support Using Data Mining*. Financial Times Management, 184.

Anand, S. S., Patrick, A. R., Hughes, J. G. & Bell, D. A. 1998. A data mining methodology for cross sales. *Knowledge-based System Journal* **10**(7), 449–461.

Arranz, C. 2007. *6 sigma desde la praxis. Experiencias concretas de empresas españnolas*, AEC (Asociación Española para la Calidad), chapter ¿Qué Es En Realidad Six-Sigma? 36–46. Morgan Kaufmann.

Barker, J. 1992. *Paradigms: The Business of Discovering the Future*. HarperBusiness.

Blockeel, H. & Moyle S. 2002. Collaborative data mining needs centralised model evaluation. In *Proceedings of ICML'02 Workshop on Data Mining: Lessons Learned*, T. Fawcett (ed.), 21–28. `citeseer.ist.psu.edu/568060.html`.

Brachman, R. J. & Anand, T. 1996. The process of knowledge discovery in databases. *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, 37–57.

Buchner, A. G., Mulvenna, M. D., Anand, S. S. & Hughes, J. G. 1999. *An Internet-enabled Knowledge Discovery Process*, 13–27. `citeseer.ist.psu.edu/290505.html`.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. & Zanasi, A. 1997. *Discovery Data Mining. From Concept to Implementation*. Prentice Hall.

Capra, F. 1996. *The Web of Life: A New Scientific Understanding of Living Systems*. Anchor Books.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. 2000. *CRISP-DM 1.0 Step-by-Step Data Mining Guide*. Technical report, CRISP-DM.

Cios, K. J. & Kurgan, L. A. 2005. Trends in data mining and knowledge discovery. In *Advanced Techniques in Knowledge Discovery and Data Mining*, Pal, L. C. Jain, N. (eds), Advanced Information and Knowledge Processing. Springer, 1–26.

Cios, K., Teresinska, A., Konieczna, S., Potocka, J. & Sharma, S. 2000. Diagnosing myocardial perfusion from pect bull's-eye maps — a knowledge discovery approach. *IEEE Engineering in Medicine and Biology Magazine* **19**, 17–25.

CRISP-DM 2007. `http://www.crisp-dm.org/new.htm`.

de Pisón Ascacibar, F. M. 2003. *Optimización Mediante Técnicas de Minería de Datos Del Ciclo de Recocido de Una Línea de Galvanizado*. PhD thesis, Univeridad de la Rioja.

Debuse, J. C. W., de la Iglesia, B., Howard, C. & Rayward-Smith, V. 2001. Building the KDD Roadmap: A Methodology for Knowledge Discovery. *Industrial Knowledge Management*. Springer-Verlag, 179–196.

Edelstein, H. A. & Edelstein, H. C. 1997. *Building, Using, and Managing the Data Warehouse*, Data Warehousing Institute, 1st edition. Prentice Hall PTR.

Eisenfeld, B., Kolsky, E. & Topolinski, T. 2003a. *42 percent of crm Software Goes Unused.* `http://www.gartner.com`.

Eisenfeld, B., Kolsky, E., Topolinski, T., Hagemeyer, D. & Grigg, J. 2003b. *Unused CRM Software Increases TCO and Decreases ROI*. `http://www.gartner.com`.

EITO (European Information Technology Observatory) 2007. Eito report 2007.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. 1996a. From data mining to knowledge discovery: an overview, *Advances in Knowledge Discovery and Data Mining*, 1–34. American Association for Artificial Intelligence.
Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. 1996b. The KDD PROCESS for extracting useful knowledge from volumes of data. *Communication of the ACM* **39**, 27–34. citeseer.ist.psu.edu/fayyad96kdd.html.
Fayyad, U., Piatetsky-Shapiro, G., Smith, P. & Uthurusamy, R. 1996c. *Advances in Knowledge Discovey and Data Mining*. AAAI/MIT Press.
Gallo, M. A. & Hancock, W. M. 2001. *Networking Explained*. Butterworth-Heinemann.
Gartner, Inc. 2005. Gartner says more than 50 percent of data warehouse projects will have limited acceptance or will be failures through 2007. http://www.gartner.com.
Gartner, Inc. 2008a. Gartner exp survey of more than 1,400 cios shows cios must create leverage to remain relevant to the business.
Gartner, Inc. 2008b. Gartner exp worldwide survey of 1,500 cios shows 85 percent of cios expect *Significant Change* over next three years. http://www.gartner.com/it/page.jsp?id=587309.
Gertosio, C. & Dussauchoy, A. 2004. Knowledge discovery from industrial databases. *Journal of Intelligent Manufacturing* **15**, 29–37.
Gondar, J. E. 2005. *Metodología Del Data Mining*. Data Mining Institute S. L.
Group, L. 2006. http://www.witnessminer.com/witminerwebhelp.htm.
Harman, W. 1970. *An Incomplete Guide to the Future*. W. W. Norton.
Harry, M. & Schroeder, R. 1999. *Six Sigma, the Breakthrough Management Strategy Revolutionizing the World's Top Corporations*. Currency.
IBM 1999. *Application Programming Interface and Utility Reference*. IBM DB2 Intelligent Miner for Data, IBM.
IEEE 1991. *Standard for Developing Software Life Cycle Processes. IEEE Std. 1074-1991*. IEEE Computer Society.
ISL 1995. *Clementine User Guide*, Version 5, ISL, Integral Solutions Limited.
ISO 1995. *ISO/IEC Standard 12207:1995. Software Life Cycle Processes*. International Organization for Standarization.
Jacobson, I., Booch, G. & Rumbaugh, J. 1999. *The Unified Software Development Process*. Addison Wesley Longman Inc.
KdNuggets.Com 2002. *Data Mining Methodology*. http://www.kdnuggets.com/polls/2002/methodology.htm.
KdNuggets.Com 2004. *Data Mining Methodology*. http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm.
KdNuggets.Com 2007a. *Data Mining Activity in 2007 vs 2006*. http://www.kdnuggets.com/polls/2007/data_mining_2007_vs_2006.htm.
KdNuggets.Com 2007b. *Data Mining Methodology*. http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm.
KdNuggets.Com 2008. *Data Mining Roi*. http://www.kdnuggets.com/polls/2008/roi-data-mining.htm.
Khabaza, T. & Shearer, C. 1995. *Data Mining with Clementine* **16**(2), 1–5. London.
Kriegel, H.-P., Borgwardt, K. M., Kröger, P., Pryakhin, A., Schubert, M. & Zimek, A. 2007. Future trends in data mining. *Data Mining Knowledge Discovery* **15**(1), 87–97.
Kurgan, L. A. & Musilek P. 2006. A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review* **21**(1), 1–24.
Marbán, O., Mariscal, G., Menasalvas, E. & Segovia, F. J. 2007. An engineering approach to data mining projects. Lecture Notes in Computer Science **4881**, 578–588. Springer.
Marbán, O., Segovia, J., Menasalvas, E. & Fernandez-Baizan, C. 2008. Towards data mining engineering: a software engineering approach. *Information Systems Journal*.
McCall, J., Richards, P. & Walters, G. 1977. Factors in software quality. *NTIS AD-A049-014* **015**(055).
McConnell, S. 1997. *Desarrollo y gestión de proyectos informáticos*. McGraw-Hill.
McDonald, M., Blosch, M., Jaffarian, T., Mok, L. & Stevens, S. 2006. *Growing It's Contribution: The 2006 Cio Agenda*. http://www.gartner.com.
McMurchy, N. 2008. *Toolkit Tactical Guideline: Five Success Factors for Effective Bi Initiatives*. http://www.gartner.com.
Moyle, S. & Jorge, A . 2001. Ramsys—a methodology for supporting rapid remote collaborative data mining projects, *ECML/PKDD 2001 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning: Internal SolEuNet Session*, 20–31.
Piatetsky-Shaphiro, G. & Frawley, W. 1991. *Knowledge Discovery in Databases*. AAAI/MIT Press.
Piatetsky-Shapiro, G. 1991. *Report on the AAAI-91 Workshop on Knowledge Discovery in Databases*. Technical report 6, IEEE Expert.
Piatetsky-Shapiro, G. 2000. Knowledge discovery in databases: 10 years after. *SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining* **1**(2), 59–61.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press

Pressman, R. S. 2005. *Software Engineering: A Practitioner's Approach*, 6th edition. McGraw-Hill Science.
Presutti, G. D. 1999. CRoss industry standard process for data mining: CRISP-DM, *4th CRISP-DM Special Interest Group (SIG) Meeting*. http://www.crisp-dm.org, Brussels.
Pyzdek, T. 2003. *The Six Sigma Handbook*, 2nd edition. McGraw-Hill.
Reinartz, T. 2002. *Stages of the Discovery Process*. Oxford University Press, Inc., 185–192.
Richardson, J., Schlegel, K., Hostmann, B. & McMurchy, N. 2008. *Magic Quadrant for Business Intelligence Platforms, 2008*. http://www.gartner.com.
SAS Institute 2005. *Semma Data Mining Methodology*. http://www.sas.com/technologies/analytics/datamining/miner/semma.html.
SearchDataManagement.com 2008. *What is Business Intelligence*? http://SearchDataManagement.com.
Sharma, S. & Osei-Bryson, K.-M. 2009. Framework for formal implementation of the business understanding phase of data mining projects. *Expert Systems with Applications* **36**(2), 4114–4124.
Shearer, C. 1996. User driven data mining. *Unicom Data Mining Conference*. London.
Solarte, J. 2002. *A Proposed Data Mining Methodoloy and Its Aplication to Industrial Engineering*, Master's thesis, University of Tennessee, Knoxville.
SpringerLink 2008. *Data Mining and Knowledge Discovery*. http://www.springerlink.com/content/100254/.
SPSS 2007. *Spss Website*. http://www.spss.com.
StatSoft, I. 2005. *Data Mining Techniques*. http://www.statsoftinc.com/textbook/stathome.html.
Strand, M. 2000. *The Business Value of Data Warehouses–Opportunities, Pitfalls and Future Directions*. PhD thesis, Department of Computer Science, University of Skövde.
The CRISP-DM Consortium 2008. *The crisp-dm Blog*. http://crispdm.wordpress.com.
The Data Mining Research Group 1997. *DBMiner User Manual*. Simnon Fraser University, Intelligent Database Systems Laboratory.
Tkach, D. 1998. *Information Mining with the IBM Intelligent Miner Family*. IBM Software Solutions White Paper.
Two Crows Corporation 1998. *Introduction to Data Mining and Knowledge Discovery*, 2nd edition. Two Crows Corporation. ISBN 892095-00-0.
Two Crows Corporation 1999. *Introduction to Data Mining and Knowledge Discovery*, 3rd edition. Two Crows Corporation. ISBN 1-892095-02-5.
Tyrrell, S. 2000. The many dimensions of the software process. *ACM Crossroads* **6**(4), 22–26.
Witten, I. H. & Frank, E. 2005. *Data Mining: Practical Machine Learning Tools with Java Implementations*, 2nd edition. Morgan Kaufmann.
Yang, Q. & Wu, X. 2006. 10 challenging problems in data mining research. *International Journal of Information Technology and Decision Making* **5**(4), 597–604.
Zornes, A. 2003. The top 5 global 3000 data mining trends for 2003/04. *META Group Research-Delta Summary* **2061**, 1–20.

https://doi.org/10.1017/S0269888910000032 Published online by Cambridge University Press