The Knowledge Engineering Review, Vol. 21:1, 1–24. © 2006, Cambridge University Press
doi:10.1017/S0269888906000737 Printed in the United Kingdom

A survey of Knowledge Discovery and Data Mining process models

LUKASZ A. KURGAN[1] and PETR MUSILEK[2]

[1]*Department of Electrical and Computer Engineering, University of Alberta, ECERF 2nd Floor, 9107 116 Street, Edmonton, Alberta, T6G 2V4, Canada;*
e-mail: lkurgan@ece.ualberta.ca
[2]*Department of Electrical and Computer Engineering, University of Alberta, ECERF 2nd Floor, 9107 116 Street, Edmonton, Alberta, T6G 2V4, Canada;*
e-mail: musilek@ece.ualberta.ca

Abstract

Knowledge Discovery and Data Mining is a very dynamic research and development area that is reaching maturity. As such, it requires stable and well-defined foundations, which are well understood and popularized throughout the community. This survey presents a historical overview, description and future directions concerning a standard for a Knowledge Discovery and Data Mining process model. It presents a motivation for use and a comprehensive comparison of several leading process models, and discusses their applications to both academic and industrial problems. The main goal of this review is the consolidation of the research in this area. The survey also proposes to enhance existing models by embedding other current standards to enable automation and interoperability of the entire process.

1 Introduction

> '. . . Knowledge Discovery is the most desirable end-product of computing. Finding new phenomena or enhancing our knowledge about them has a greater long-range value than optimizing production processes or inventories, and is second only to task that preserve our world and our environment. It is not surprising that it is also one of the most difficult computing challenges to do well . . .' Gio Wiederhold (1996).

Current technological progress permits the storage and access of large amounts of data at virtually no cost. Although many times preached, the main problem in a current information-centric world remains to properly put the collected raw data to use. The true value is not in storing the data, but rather in our ability to extract useful reports and to find interesting trends and correlations, through the use of statistical analysis and inference, to support decisions and policies made by scientists and businesses (Fayyad *et al.*, 1996c).

Before any attempt can be made to perform the extraction of this useful knowledge, an overall approach that describes how to extract knowledge needs to be established. Therefore, the focus of this paper is not on describing the methods that can be used to extract knowledge from data, but rather on discussing the methodology that supports the process that leads to finding this knowledge. The main reason for establishing and using process models is to organize the Knowledge Discovery and Data Mining (KDDM) projects within a common framework. The models help organizations to understand the Knowledge Discovery process and provide a road map to follow while planning and carrying out the projects. This in turn results in time and cost savings, and in a better understanding and acceptance of such projects. The first step is to understand that such processes

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

are not trivial, but rather involve multiple steps, reviews and iterations. To date, there have been several attempts made to develop such models, with varying degrees of success. This paper summarizes the state-of-the-art in this subject area, and discusses future research directions.

The main motivation for this paper is a lack of a comprehensive overview and comparison of KDDM models. Although several models have been developed that have received broad attention of both research and industrial communities, they have been usually discussed separately, making their comparison and selection of the most suitable model a daunting task.

This survey is organized as follows. First, basic definitions concerning the Knowledge Discovery domain, motivation for the existence of process models, and a historical overview are provided in Section 2. Next, in Section 3, several leading models are reviewed and discussed. A formal comparison of the models and their applications in both research and industrial context are presented in Section 4. Finally, future trends in this area are discussed and conclusions are provided in Sections 5 and 6, respectively.

2 KDDM process models

2.1 Terminology

There is a common confusion in understanding the terms of Data Mining (DM), Knowledge Discovery (KD), and Knowledge Discovery in Databases (KDD). For this reason, the meanings of these terms are first explained with references to definitions published in scientific literature. For many researchers the term DM is used as a synonym for KD besides being used to describe one of the steps of the KD process.

Data Mining concerns application, under human control, of low-level DM methods, which in turn are defined as algorithms designed to analyze data, or to extract patterns in specific categories from data (Klosgen & Zytkow, 1996). DM is also known under many other names, including knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing (Fayyad *et al*., 1996c).

Knowledge Discovery is a process that seeks new knowledge about an application domain. It consists of many steps, one of them being DM, each aimed at completion of a particular discovery task, and accomplished by the application of a discovery method (Klosgen & Zytkow, 1996).

Knowledge Discovery in Databases concerns the knowledge discovery process applied to databases (Klosgen & Zytkow, 1996). It is also defined as a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad *et al*., 1996b). This definition is the most popular among the KD community, developed by revising the original definition published by Frawley *et al*. (1991). It generalizes the application of the process to non-database sources, although it emphasizes them as a primary source of data.

Knowledge Discovery and Data Mining concerns the KD process applied to any data source. The term KDDM has been proposed as the most appropriate name for the overall process of KD (Reinartz, 2002; Cios & Kurgan, 2005).

KDDM concerns the entire knowledge extraction process, including how the data is stored and accessed, how to develop efficient and scalable algorithms that can be used to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine (Fayyad *et al*., 1996c). It also concerns the support for learning and analyzing the application domain. DM is always included as one of the steps in the KDDM process.

2.2 Motivation

The main motivation factor to formally structure the KDDM as a process results from an observation of problems associated with a blind application of DM methods to input data. Such activity, called ‘data dredging’ in the statistical literature, can lead to discovery of meaningless

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

knowledge (Fayyad *et al.*, 1996c). Therefore, the main reason for defining and implementing KDDM process models is to ensure that the end product will be useful to the user (Fayyad *et al.*, 1996d). This is why the definition of the process emphasizes validity, novelty, usefulness, and understandability of the results. Only by using well-defined and formal development methods can such desirable properties be successfully achieved.

The second motivating factor is associated with the understanding of the process itself, and understanding of the concerns and needs of the end-users of process models. There are several common human behaviors associated with the knowledge-searching task. Humans very often lack perception of large amounts of untapped and potentially valuable data. In addition, they are usually not willing to devote time and resources toward formal methods of knowledge seeking, but rather heavily rely on other people, such as domain experts, as a source of valuable information (Rouse, 2002). One of the reasons for such behavior may be uncertainty of new technology and processes that needs to be applied to provide the solution (Rouse, 2002). This calls for popularization and standardization of solutions developed in this area.

Another important factor, which is very often underestimated by researchers, is associated with providing support for management problems. Some KDDM projects involve relatively large teams of people working together, and thus require careful planning and scheduling. For most project management specialists, KD and DM are unfamiliar terms. Therefore, they need a definition of what such projects involve, and how to carry them out. Such definition is a basis for developing a sound project schedule, typically based on milestones. A milestone is defined as a concrete, specific, measurable event used to define completion of particular phases of the overall project (Brooks, 1995). They can be properly defined only in the context of a well-defined larger framework. Other engineering disciplines have already established development models and used them for many years. A good example is software engineering, which is also a relatively new and dynamic discipline that exhibits many characteristics similar to KDDM. Software engineering adopted Waterfall (Roy, 1970) and Spiral (Boehm, 1988) models that became well-known standards in this area. Other examples are models proposed in fields intimately related to the KDDM, such as Statistics (Hand, 1994), and Machine Learning (Brodley & Smith, 1997).

Lastly, there is a widely recognized need for the standardization of KDDM processes to provide a unified view on existing process descriptions and to allow an appropriate usage of technology to solve current business problems in practice (Reinartz, 2002). This trend can be followed by looking at the number of papers on this subject, and the strong industrial support for such initiatives, both demonstrated throughout this paper.

2.3 History

The concept of a KDDM process model was originally discussed during the first workshop on KDD in 1989 (Piatesky-Shapiro, 1991). The main driving factor to define the model was acknowledgement of the fact that knowledge is the end product of a data-driven discovery process. One of the outcomes of the workshop was also the acknowledgement of the need to develop interactive systems that would provide visual and perceptual tools for data analysis.

Following this seminal event, the idea of a process model was iteratively developed by the KDD community over the several years that followed. Initial KD systems provided only a single DM technique, such as a decision tree or clustering algorithm, with a very weak support for the overall process framework (Zytow & Baker, 1991; Klosgen, 1992; Piatesky-Shapiro & Matheus, 1992; Ziarko *et al.*, 1993; Simoudis *et al.*, 1994). Such systems were intended for expert users who had understanding of DM techniques, the underlying data, and the knowledge sought. There was very little attention focused on the support for the layman data analyst, and thus the first KD systems had minimal commercial success (Brachman & Anand, 1996). The general research trends were concentrated on the development of new and improved DM algorithms rather than on the support for other KD activities.

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

In 1996, the foundation of the process model was laid down with the release of *Advances in Knowledge Discovery and Data Mining* (Fayyad *et al*., 1996a). This book presented a process model that resulted from interactions between researchers and industrial data analysts. The model did not address particular DM techniques, but rather provided support for the complicated and highly iterative process of knowledge generation. It also emphasized the close involvement of a human analyst in the majority of steps of the process (Brachman & Anand, 1996).

Research presented in the book resulted in proposing two major types of process models. The human-centric model emphasized the interactive involvement of a data analyst during the process, and the data-centric model emphasized the iterative and interactive nature of the data analysis tasks (Fayyad *et al*., 1996a). The human-centric process was defined as a series of knowledge-intensive tasks consisting of complex interactions, protracted over time, between a human and a (large) database, possibly supported by a heterogeneous suite of tools (Brachman & Anand, 1996). The structure of the process included three main tasks: model selection and execution, data analysis, and output generation. These tasks were further broken down into sub-tasks. The first task was divided into data segmentation, model selection, and parameter selection. The data analysis task consisted of model specification, model fitting, model evaluation, and model refinement. Finally, the output generation task included the generation of reports, and development of so called monitor implementing the obtained results into the original problem domain. This model was further discussed by Reinartz (1999). Another human-centric model was proposed by Gupta *et al*. (2000). It was based on a data-centric model with extensions to provide support for experimentation and monitoring activities.

Since the data-centric model has become dominant in industrial and research settings, the remaining part of the survey concentrates only on models of this type. In general, data-centric models are structured as sequences of steps that focus on performing manipulation and analysis of data and information surrounding the data, such as domain knowledge and extracted results. Such models are usually defined as a fixed sequence of predefined steps. The user's role is to assure that specific objectives for each step are met, which is usually carried out by supervising and guiding the data processing tasks.

Despite the differences in understanding their basic structure, models of both above mentioned types bring to attention many similar issues. They define the process as being highly interactive and complex. They also suggest that KDDM process may use, or at least should consider the use of, a set of DM technologies, while admitting that the DM step constitutes only a small portion of the overall process (Brachman & Anand, 1996; Fayyad *et al*., 1996).

3 The KDDM process models

A KDDM process model consists of a set of processing steps to be followed by practitioners when executing KDDM projects. Such a model describes procedures that are performed in each of the steps, primarily used to plan, work through, and reduce the cost of any given project. The basic structure of the model was proposed by Fayyad *et al*. (1996a). Since then, several different KDDM models have been developed in both academia and industry.

All process models consist of multiple steps executed in a sequence, which often includes loops and iterations. Each subsequent step is initiated upon the successful completion of a previous step, and requires a result generated by the previous step as its inputs. Another common feature of the proposed models is the span of covered activities. It ranges from the task of understanding the project domain and data, through data preparation and analysis, to evaluation, understanding, and application of the generated results. All proposed models also emphasize the iterative nature of the model, in terms of many feedback loops and repetitions, which are triggered by a revision process. This can be contrasted to the Exploratory Data Analysis (EDA) that is typically executed as a unidirectional process intended to gain insight into the data without the usual assumptions about what kind of model the data follow (Tukey, 1977). Instead, EDA represents a more direct approach

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

of allowing the data itself to reveal its underlying structure and model (NIST, 2005). The absence of feedback loops is the main cause of general inefficiency of EDA methods on contemporary databases characterized by large size and complexity.

The main difference among the surveyed models is in the proposed number and scope of their specific steps. Although the models usually emphasize independence of specific applications, tools, and vendors, they can be broadly divided into two groups depending on whether they take into account industrial aspects of KDDM projects or not. Compared with academic projects, industrial KDDM projects are usually concerned with different types of data, have more complex application scenarios, and are associated with different burdens and pitfalls.

We will now introduce several KDDM models in chronological order. The discussion is restricted to the models popularized in scientific publications that are subject to peer-review evaluation and thus present unbiased information.

3.1 Overview of the major KDDM process models

The development of the standard KDDM model was initiated several years ago. The first reported KDDM model consists of nine steps and was developed by Fayyad *et al.* in the mid-1990s (Fayyad *et al.*, 1996b, c; d, e). The next model, by Cabena *et al.*, consists of five steps and was introduced in 1998 (Cabena *et al.*, 1998). The third model, which consists of eight steps, was developed by Anand & Buchner at about the same time (Anand & Buchner, 1998; Anand *et al.*, 1998). The CRISP-DM (CRoss-Industry Standard Process for DM) process model that includes six steps was first proposed in early 1996 by a consortium of four companies: SPSS (a provider of commercial DM solutions), NCR (a database provider), Daimler Chrysler, and OHRA (an insurance company). The last two companies served as sources of data and case studies. The model was officially released (version 1.0) in 2000 (Shearer, 2000; Wirth & Hipp, 2000) and it continues to enjoy a strong industrial support. It has been also supported by the ESPRIT program funded by the European Commission. The CRISP-DM Special Interest Group was created with the goal of supporting the model. Currently it includes over 300 DM users and tool and service providers (CRISP-DM, 2003). Finally, the six-step process model of Cios *et al.* was first proposed in 2000 (Cios *et al.*, 2000; Cios & Kurgan, 2005), by adopting the CRISP-DM model to the needs of academic research community. The main extensions of the latter model include providing 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. These five models constitute a group that made a substantial impact in terms of their development process, background, academic and industrial involvement, and the number of projects that have already applied the models. Each of these models was applied in at least several real KDDM projects in either industrial, research, or both settings.

Over the last ten years, the research efforts have been focused on proposing new models, rather than improving design of a single model or proposing a generic unifying model. Despite the fact that most models have been developed in isolation, a significant progress has been made. The subsequent models provide more generic and appropriate descriptions. Most of them are not tied specifically to academic or industrial needs, but rather provide a model that is independent of a particular tool, vendor, or application. Fayyad's nine-step model is best geared towards specific academic research features, while omitting several important business issues. The CRISP-DM model, on the other hand, is very industry-oriented. The remaining three models occupy the middle ground, mixing both academic and industrial aspects of KDDM. To facilitate the understanding and interpretation of the described KDDM models, a direct, side-by-side comparison is shown in Table 1. The table aligns different models by maximizing overlap between the steps from different models. In some isolated cases the scope of the steps is different between the models, but in general similar steps are used, although they may be named differently, for example, DM, Modeling and

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

Table 1 Side-by-side comparison of the major existing KDDM models

Model	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*	Generic model
Area	Academic	Industrial	Academic	Industrial	Academic	N/A
No of steps	9	5	8	6	6	6
Refs	(Fayyad *et al.*, 1996d)	(Cabena *et al.*, 1998)	(Anand & Buchner, 1998)	(Shearer, 2000)	(Cios *et al.*, 2000)	N/A
Steps	1 Developing and Understanding of the Application Domain	1 Business Objectives Determination	1 Human Resource Identification	1 Business Understanding	1 Understanding the Problem Domain	1 Application Domain Understanding
			2 Problem Specification			
	2 Creating a Target Data Set	2 Data Preparation	3 Data Prospecting	2 Data Understanding	2 Understanding the Data	2 Data Understanding
			4 Domain Knowledge Elicitation			
	3 Data Cleaning and Preprocessing		5 Methodology Identification	3 Data Preparation	3 Preparation of the Data	3 Data Preparation and Identification of DM Technology
	4 Data Reduction and Projection		6 Data Preprocessing			
	5 Choosing the DM Task					
	6 Choosing the DM Algorithm					
	7 DM	3 DM	7 Pattern Discovery	4 Modeling	4 DM	4 DM
	8 Interpreting Mined Patterns	4 Domain Knowledge Elicitation	8 Knowledge Post-processing	5 Evaluation	5 Evaluation of the Discovered Knowledge	5 Evaluation
	9 Consolidating Discovered Knowledge	5 Assimilation of Knowledge		6 Deployment	6 Using the Discovered Knowledge	6 Knowledge Consolidation and Deployment

Pattern Discovery. The table provides information about the developing party (academic of industry), number of steps, main reference, and a comparison of steps across all models.

There are several features common for all KDDM process models, which have not changed over time. Most of the existing process models follow the same sequence of steps and they often use similar steps. Most models involve complex and time-consuming data preparation tasks (Brachman & Anand, 1996). The processes have iterative nature that can involve significant iterations and loops between most, if not all, of its steps (Fayyad *et al*., 1996b). Following these observations a generic model is proposed based on the five surveyed models. This generic model, described in the last column of Table 1, provides a consolidated view based on information accumulated among the five models. It consists of six steps and in general follows the lines of the Cios *et al*. and CRISP-DM models. This choice was motivated by two main factors: (1) these two models were developed based on experiences coming from the older models, and (2) all five models fit the six steps structure with a minor modification that combines several original steps into a major step. For instance, Steps 2, 3, 4 and 5 of Fayyad's model constitute a set of sub-steps of step 3, of the generic model. There are several exceptions from these rules. For instance, Anand & Buchner's model is missing the last step of knowledge consolidation and system deployment. Also, data preprocessing and selection of algorithms used for DM are performed in different order, i.e., in the case of Fayyad's model, preprocessing is performed first, while Anand & Buchner's model suggests that preprocessing should be performed after the selection of algorithms.

3.2 Other KDDM process models

There are several other process models that made a less significant impact and thus are not discussed in detail in this survey.

- A six-step KDDM process model by Adriaans & Zantinge (1996), which consists of Data Selection, Cleaning, Enrichment, Coding, DM, and Reporting.
- A four-step model by Berry & Linoff (1997), which consists of Identifying the Problem, Analyzing the Problem, Taking Action, and Measuring the Outcome.
- A five-step SEMMA model by the SAS Institute Inc. (1997), which consists of steps named Sample, Explore, Modify, Model, and Assess. This model was incorporated into commercial KD software platform SAS Enterprise Miner™.
- A seven-step model by Han & Kamber (2001), which consists of Learning the Application domain, Creating a Target Data Set, Data Cleaning and Preprocessing, Data Reduction and Transformation, Choosing Functions of DM, Choosing the Mining Algorithm(s), DM, Pattern Evaluation and Knowledge Presentation, and Use of Discovered Knowledge.
- A five-step model by Edelstein (2001), which consists of Identifying the Problem, Preparing the Data, Building the Model, Using the Model, and Monitoring the Model.
- A seven-step model by Klosgen & Zytkow (2002), which consist of Definition and Analysis of Business Problems, Understanding and Preparation of Data, Setup of the Search for Knowledge, Search for Knowledge, Knowledge Refinement, Application of Knowledge in Solving the Business Problems, and Deployment and Practical Evaluation of the Solutions.
- A seven-step model by Haglin *et al*. (2005), which consists of Goal Identification, Target Data Creation, Data Preprocessing, Data Transformation, DM, Evaluation and Interpretation, and Take Action steps.

In general, these models are either too new for evaluation in application settings (in particular, the model by Haglin *et al*.), or have not been widely cited in professional literature, perhaps indicating a low user interest. A more detailed description and comparison of the five models is provided in the following section.

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

4 Analysis of KDDM process models

This section starts with a detailed side-by-side comparison of the five KDDM models introduced in Section 3.1. Next, applications of the five models are surveyed, and a direct comparison of impact and quality of the models is performed. Finally, a discussion about the relative effort required to complete each step is presented.

4.1 Detailed description of KDDM models

Table 2 shows a side-by-side comparison of individual steps of the models. The description of the steps is based on the original papers that introduced the models, and uses a unified terminology to facilitate the comparison. The scope and description of the steps of the generic model can be inferred based on the description of the corresponding steps of the five major models.

Most models compared in Table 2 follow a similar sequence of steps. The common steps among the five models are: Domain Understanding, Data Preparation, DM, and evaluation of the DK. The main difference is in Fayyad's nine-step model, which performs activities related to the choice of DM task and algorithm relatively late in the process. The other models perform this step before preprocessing the data. This way, the data are correctly prepared for the DM step without the need to repeat some of the earlier steps (Cios & Kurgan, 2005). In the case of Fayyad's model, prepared data may not be suitable for the tool of choice, and thus a loop back to the second, third or fourth step may be required. Cabena's model is very similar to that of Cios and the CRISP-DM model, however it omits the Data Understanding step. This incompleteness of Cabena's model was pointed out by Hirji, who used this model in a business project and concluded that adding one more step between Data Preparation and DM, which he called Data Audit, was necessary (Hirji, 2001). The eight-step model by Anand & Buchner provides a very detailed breakdown of steps in the early phases of the KDDM process. Unfortunately, it does not include activities necessary for putting the discovered knowledge to work. The nine-step and eight-step models were developed from an academic perspective. Cabena's model was developed from an industrial perspective, with support from IBM. CRISP-DM was also developed based on a significant industrial input, and involved several major companies covering all development aspects, together with academic and governmental support. It is a very mature model that has been thoroughly documented and tested in many applications. The six-step model by Cios *et al.* draws significantly from the CRISP-DM model, but emphasizes academic aspects of the process. It is also the only model that provides detailed guidelines concerning possible loops, rather than just mentioning their presence (see footnote of Table 2).

4.2 Applications and impact of KDDM models

4.2.1 Applications of KDDM models

To complement the description of existing KDDM models, their applications to a variety of research and industrial domains are briefly summarized. In general, research applications are perceived as easier to use than industrial applications. This is mainly due to the fact that research users typically know the data much better than industrial users; they have better understanding of novel technologies, and are better trained to organize intuitions into computerized procedures (Fayyad *et al.*, 1996f). This observation only corroborates the need for standard process models to support planning and execution of KDDM projects in industrial communities. The following summary shows a real interest to use KDDM process models, in both industrial and research communities. Applications are grouped by the model used.

The nine-step model by Fayyad *et al.* is the most cited model in the professional literature to date. It has been incorporated into an industrial DM software system called MineSet™ (Brunk *et al.*, 1997), and applied in a number of KDDM projects (all projects, except the last, are predominantly research oriented):

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

Table 2 Detailed description of individual steps of the major existing KDDM models

Model	Generic	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*
Steps	STEP 1. Application Domain Understanding	1 Learning goals of the end-user and relevant prior knowledge	1 Understanding the business problem and defining business objectives, which are later redefined into DM goals	1 Identification of human resources and their roles 2 Partitioning of the project into smaller tasks that can be solved using a particular DM method	1 Understanding of business objectives and requirements, which are converted into a DM problem definition	1 Defining project goals, identifying key people, learning current solutions and domain terminology, translation of project goals into DM goals, and selection of DM methods for Step 4
	STEP 2 Data Understanding	2. Selection of a subset of variables and sampling of the data to be used in later steps	2 Identification of internal and external data sources, selection of subset of data relevant to a given DM task. It also includes verifying and improving data quality, such as noise and missing data. Determination of DM methods that will be used in the next step and transformation of the data into analytical model required by selected DM methods	3 Analysis of accessibility and availability of data, selection of relevant attributes and a storage model 4 Elicitation of the project domain knowledge	2 Identification of data quality problems, data exploration, and selection of interesting data subsets	2 Collecting the data, verification of data completeness, redundancy, missing values, plausibility, and usefulness of the data with respect to the DM goals

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

Table 2 *Continued*

Model	Generic	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*
Steps	STEP 3 Data Preparation and Identification of DM Technology	3 Preprocessing of noise, outliers, missing values, etc, and accounting for time sequence information 4 Selection of useful attributes by dimension reduction and transformation, development of invariant data representation 5 Goals from Step 1 are matched with a particular DM method, i.e. classification, regression, etc. 6 Selection of particular data model(s), method(s), and method's parameters		5 Selection of the most appropriate DM method, or a combination of DM methods 6 Preprocessing of the data, including removal of outliers, dealing with missing and noisy data, dimensionality reduction, data quantization, transformation and coding, and resolution of heterogeneity issues	3 Preparation of the final dataset, which will be fed into DM tool(s), and includes data and attribute selection, cleaning, construction of new attributes, and data transformations	3 Preprocessing via sampling, correlation and significance tests, cleaning, feature selection and extraction, derivation of new attributes, and data summarization. The end result is a data set that meets specific input requirements for the selected DM methods
	STEP 4 Data Mining	7 Generation of knowledge (patterns) from data, for example classification rules, regression model, etc.	3 Application of the selected DM methods to the prepared data	7 Automated pattern discovery from the preprocessed data	4 Calibration and application of DM methods to the prepared data	4 Application of the selected DM methods to the prepared data, and testing of the generated knowledge

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

Table 2 *Continued*

Model	Generic	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*
Steps	STEP 5 Evaluation	8 Interpretation of the model(s) based on visualization of the model(s) and the data based on the model(s)	4 Interpretation and analysis of DM results; usually visualization technique(s) are used	8 Filtering out trivial and obsolete patterns, validation and visualization of the discovered knowledge	5 Evaluation of the generated knowledge from the business perspective	5 Interpretation of the results, assessing impact, novelty and interestingness of the discovered knowledge. Revisiting the process to identify which alternative actions could have been taken to improve the results
	STEP 6 Knowledge Consolidation and Deployment	9 Incorporation of the discovered knowledge into a final system, creation of documentation and reports, checking and resolving potential conflicts with previously held knowledge	5 Presentation of the generated knowledge in a business-oriented way, formulation of how the knowledge can be exploited, and incorporation of the knowledge into organization's systems		6 Presentation of the discovered knowledge in a customer-oriented way. Performing deployment, monitoring, maintenance, and writing final report	6 Deployment of the discovered knowledge. Creation of a plan to monitor the implementation of the discovered knowledge, documenting the project, extending the application area from the current to other possible domains

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

Table 2 *Continued*

Model	Generic	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*
Notes	Unified set of steps. Each step's scope can be inferred from the corresponding steps of other models	Significant iterations by looping between any two steps are possible, but no details are given. This model became a cornerstone for the future models, and is currently the most cited model in the scientific literature	The first business-oriented model, which is easy to comprehend by the layman. Emphasizes iterative nature of the model, but no details are given. Authors note that DM is often performed together with the Step 5	Provides a detailed breakdown of the initial steps. Emphasizes iterative nature of the model, where experts examine the knowledge after the last step and may decide to refine and rerun part or the entire process. Lacks step where the discovered knowledge is applied	Uses easy to understand vocabulary, and has good documentation. Divides steps into sub-steps that provide all necessary details. Acknowledges strong iterative nature of the process, but without details	Emphasizes and explicitly describes iterative and interactive aspects of the process*

* The specific feedback loops described in Cios *et al.* model include (Cios & Kurgan, 2005).

- From Step 2 to Step 1: execution of this loop is triggered by the need for additional domain knowledge to improve data understanding.
- From Step 3 to Step 2: execution of this loop is triggered by the need for additional or more specific information about the data to guide choice of specific data preprocessing.
- From Step 4 to Step 1: the loop is performed if results generated by selected DM methods are not satisfactory and modification of project's goals is required.
- From Step 4 to Step 2: the most common reason is poor understanding of the data, which results in incorrect selection of DM method(s) and its subsequent failure (e.g. data was misclassified as continuous and discretized in Understanding the Data step).
- From Step 4 to Step 3: the loop is motivated by the need to improve data preparation; this is often caused by specific requirements of used DM method, which may have been unknown during Step 3.
- From Step 5 to Step 1: the most common cause is invalidity of the discovered knowledge; there are several possible reasons including misunderstanding or misinterpretation of the domain, incorrect design or misunderstanding of problem restrictions, requirements, or goals. In these cases the entire KDDM process needs to be repeated.
- From Step 5 to Step 4: this loop is executed when the discovered knowledge is not novel, interesting, or useful; the least expensive solution is to choose a different DM tool and repeat the DM step.

The importance of these feedback mechanisms has been confirmed by several research application of the model (Cios *et al.*, 2000; Sacha *et al.*, 2000; Kurgan *et al.*, 2001; Maruster *et al.*, 2002; Kurgan *et al.*, 2005). Introduction and detailed description of these mechanisms and their triggers is important as it increases awareness and helps the user of the process to avoid similar problems by deploying appropriate countermeasures.

– the analysis of alarm correlation systems data (Klemettinen *et al.*, 1997);
– the development of a learning system to schedule workshop production (Morello *et al.*, 2001);
– the development of KDB2000, a support tool that provides integrated set of techniques for entire KDDM process (Appice *et al.*, 2002);
– the evaluation of different aspects of recommender systems based on Web usage mining (Geyer-Schulz & Hahsler, 2002);
– the discovery of conversation models from sequences of messages in a multi-agent systems (Mounier *et al.*, 2003);
– the development of a high-throughput screening, assay design, and detection system (Vaschetto *et al.*, 2003);
– the extension of Fayyad's basic model with addition of case studies (Li & Ruan, 2004);
– the development of an architecture for defining and validation of KDDM processes for e-business (Castellano *et al.*, 2004);
– the development of a software platform, called VidaMine, for support visualization when performing tasks related to KDDM processes (Kimani *et al.*, 2004).

The five-step model by Cabena *et al.* has been applied in an industrial project analyzing data from a Canadian fast-food retailer (Hirji, 2001).

The eight-step model by Anand & Buchner has been applied in the following projects:

– an industrial project concerning customer cross sales (Anand *et al.*, 1998);
– a research project concerning analysis of marketing Internet data (Buchner *et al.*, 1999).

The CRISP-DM model has been used in the following research projects:

– performance evaluation of heating, ventilation, and air-conditioning (HVAC) systems (Buchheit *et al.*, 2000);
– analysis of thrombosis data (Jensen, 2001);
– analysis of retail store data (Butler, 2002);
– development of a new methodology for collaborative KD projects through providing support for distributed teams (Blockeel & Moyle, 2002);
– text mining (Silva *et al.*, 2002);

It has also been used in the following industrial projects:

– analysis of warranty claims at Daimler Chrysler (Hipp & Linder, 1999);
– automotive direct marketing (Gersten *et al.*, 2000);
– analysis of data concerning construction of large and tall buildings (Moyle *et al.*, 2002);
– control and improvement of air quality in Taiwan with the goal of identifying national pollutant distribution, with data retrieved from 71 monitoring stations (Li & Shue, 2004);
– development of new tinplate quality diagnostic models (De Abajo *et al.*, 2004).

In addition, this model has been recently incorporated into a commercial KD software platform called Clementine 8.0 (SPSS, 2003).

The six-step model by Cios *et al.* has been used in several research projects:

– the development of a computerized system for diagnoses of SPECT bull-eye images (Cios *et al.*, 2000);
– creating and mining a database of cardiac SPECT images (Sacha *et al.*, 2000);
– the development of an automated diagnostic system for cardiac SPECT images (Kurgan *et al.*, 2001);

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

Table 3 Number of citations to five major KDDM models (references between different indices may overlap)

Model	Paper	Indexing Service			Total
		SCOPUS	Citeseer	ISI Web of Science	
Fayyad *et al.* (9 step)	Fayyad *et al.* (1996b)	253	244	—	
	Fayyad *et al.* (1996c)	105	99	91	748
	Fayyad *et al.* (1996d)	—	47	—	
Cabena *et al.* (5 step)	Cabena *et al.* (1998)	—	16	—	16
Anand & Buchner (8 step)	Anand & Buchner (1998)	—	—	—	13
	Anand *et al.* (1998)	8	—	5	
CRISP-DM	Wirth & Hipp (2000)	4	3	—	
	Shearer (2000)	—	—	—	19
	http://www.crisp-dm.org	—	12	—	
Cios *et al.* (6 step)	Cios *et al.* (2000)	7	3	7	20
	Cios & Kurgan (2005)	—	3	—	

– the development of a diabetic retinal image screening system (Goh *et al.*, 2001), clustering and visualization of epidemiological pathology data (Shalvi & DeClaris, 2001) and about ten other medical application described in (Cios, 2001);
– the development of a logistic-based patient grouping for multi-disciplinary treatment (Maruster *et al.*, 2002);
– analysis of respiratory pressure–volume curves in intensive care medicine (Ganzert *et al.*, 2002);
– image based analysis and classification of cells (Perner *et al.*, 2002);
– the development of a Grid DM framework GridMiner-Core (Hofer & Brezany, 2004);
– analysis of clinical data related to cystic fibrosis disease (Kurgan *et al.*, 2005).

The above application list was compiled through a comprehensive literature search using several leading indexing services, including SCOPUS (Elsevier citation index that covers 14 000 peer-reviewed titles from more than 4000 international publishers, see, http://www.scopus.com/), Citeseer (a scientific index search engine that focuses primarily on computer and information science that includes over 700 000 articles, see Lawrence *et al.*, 1999), and ISI Web of Science (a portal providing simultaneous access to the Science Citaton Index©, Social Sciences Citation Index©, Arts & Humanities Citation Index©, Index Chemicus©, and Current Chemical Reactions©, see http://www.isinet.com/isihome/products/citation/wos/). Although the presented list of applications is not guaranteed to be complete, it shows general application trends. CRISP-DM is currently the most popular and broadly adopted model. This model, and its derivative by Cios *et al.*, have been already acknowledged and relatively widely used in both research and industrial communities. CRISP-DM model is used not only in many real development projects, but also in many 'request for proposal' documents (Shearer, 2000). In fact it has already been assessed as meeting industrial needs (Piatesky-Shapiro, 1999b).

4.2.2 Evaluation of the impact of KDDM models

Most research and industrial projects based on standard models have appeared only recently, and are expected to grow in number. However, in order to evaluate the relative rate of adoption of the models, the total number of references to papers introducing individual models has been extracted from the three leading indexing services. The results of this literature search are reported in Table 3.

The nine-step model is the most cited. The large number of the citations is mainly due to the fact that the papers that described the model also introduced the most commonly used definitions of

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

DM and KDD. In addition, the number of citation should be normalized by the number of years since the model's introduction. Therefore, the total number of citations is 83.1 per year for the nine-step model, 2.3 for the five-step model, 1.9 for the eight-step model, 3.8 for CRISP-DM, and 4 for the six-step model.

Additional information about the usage of the models has been compiled using two recent polls conducted by KDnuggets (`http://www.kdnuggets.com/`),which is a leading Web resource on DM. The poll from July 2002, which included 189 respondents, shows that 51% of respondents used the CRISP-DM model, 23% used their own model, 12% used SEMMA, 7% used their organization's specific model, and 4% used some other model or no model. The second poll from April 2004, which included 170 respondents, shows 42% for CRISP-DM, 28% for respondent's own model, 10% for SEMMA, 6% for organization's specific, 7% for no model, and 6% for other models. The poll specified only the above six choices and named only two models, i.e., CRISP-DM and SEMMA. The results clearly indicate the CRISP-DM received significant attention. They also show that a significant number of people uses their own model to perform KDDM. Information on applications and citations of the models and the poll results are used as a basis for comparison of the five models in the following section.

4.3 Comparison of the KDDM models

Finally, an overall comparison of the five models, which includes both quantitative and qualitative aspects, is shown in Table 4. For convenience, columns 1, 2 and 4 repeat some important facts from Table 3, namely the year when the model was introduced, the total number of citations per year, and the average results of the KDnuggets polls. Columns 3a-d include the number of applications in academia and industry, the total number of applications, and the total number of applications with applications by model author(s) discounted. Each paper is categorized as either an academic or an industrial application based on the application's context, and papers that are authored by one of the model's authors are counted. Column 2 shows the total number of citations per year based on the performed citation analysis study, while columns 3a-d include only papers that applied given model to perform a KDDM project. Column 2 may also include some repetition of citations (due to using multiple indexing services), as well as citations from papers that just acknowledge, rather than use a given KDDM model. In the case of the Fayyad *et al.* model, the high citation count is due to citations of the definition DM and KDD, which are included in this paper. Remaining columns contain nominal and qualitative evaluation of the models. First, areas in which the models have been applied are listed. Next, the level of industrial involvement in development of the models is estimated using an ordinal scale, from 0 (no industrial involvement) to 5 (strong industrial involvement), followed by information on the existence of software tools supporting the models (if a model is supported, the name of the software support tool is listed). Next, information about documentation is given, followed by ease of use estimated for each tool. Finally, the main drawbacks of each model are enumerated. Please note that the evaluations included in columns 5–9 of the table provide subjective view of the authors.

The motivation behind the model comparison shown in Table 4 is to provide existing and future users of the models with information necessary to perform an independent assessment of benefits and drawbacks of the individual models. In general, CRISP-DM is the most suitable for novice data miners and especially miners working on industrial projects, due to the easy to read documentation and intuitive, industry-applications-focused description. It is a very successful and extensively applied model. This stems from grounding its development on practical, industrial, real-world KD experience (Shearer, 2000). The Cios's model, on the other hand, is geared towards users that work on research projects and projects that require feedback loops, and have good prior understanding of DM terminology and concepts. Fayyad's model is the most suitable for projects requiring extensive data preprocessing, while Cabena's model is suitable for applications where data is virtually ready for mining before the project starts. The detailed assessment and choice of a

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

Table 4 Comparison of quantitative and qualitative aspects of the five major KDDM models (1 – year when the model was introduced, 2 – total number of citations per year, 3 – number of applications of the model: 3a – in academia, 3b – in industry, 3c – total number of applications, 3d – total number of applications with applications by model authors discounted, 4 – average KDnuggets poll results, 5 – application areas, 6 – industry involvement (0 none – 5 strong), 7 – software tool support, 8 – documentation, 9 – ease of use (0 novice – 5 expert), 10 – main drawbacks)

KDDM model	1	2	3				4	5	6	7	8	9	10
			a	b	c	d							
Fayyad *et al.* (9 step)	1996	83	8	2	10	10	Not listed	Medicine, engineering, production, e-business, software	0	Yes MineSet™	No Web site, description based on research papers	4 Requires background in DM	— Prepared data may not be suitable for the tool of choice, and thus unnecessary loop back previous steps may be required —limited discussion of feedback loops
Cabena *et al.* (5 step)	1998	2	0	1	1	1	Not listed	Marketing and sales	2 (one company)	No	No Web site, description based on a book	2 Requires some knowledge of DM terminology	—Omits the data understanding step —Limited discussion of feedback loops
Anand & Buchner (8 step)	1998	2	1	1	2	0	Not listed	Marketing and sales	0	No	No Web site, description based on research papers	4 Requires background in DM	—Too detailed breakdown of steps in the early phases of the KDDM process —Does not accommodate for a step that is concerned with putting the discovered knowledge to work —Limited discussion of feedback loops
CRISP-DM	2000	4	5	6	11	9	46%	Medicine, engineering, marketing and sales, environment	5 (consortium of companies)	Yes Clementine™	Has Web site, description based on research and white papers	1 Easy to understand, in lay words	—limited discussion of feedback loops
Cios *et al.* (6 step)	2000	4	21	0	21	16	Not listed	Medicine, software	0	No	No Web site, description based on research papers	3 Requires knowledge of DM terminology	—Popularized and geared towards research applications

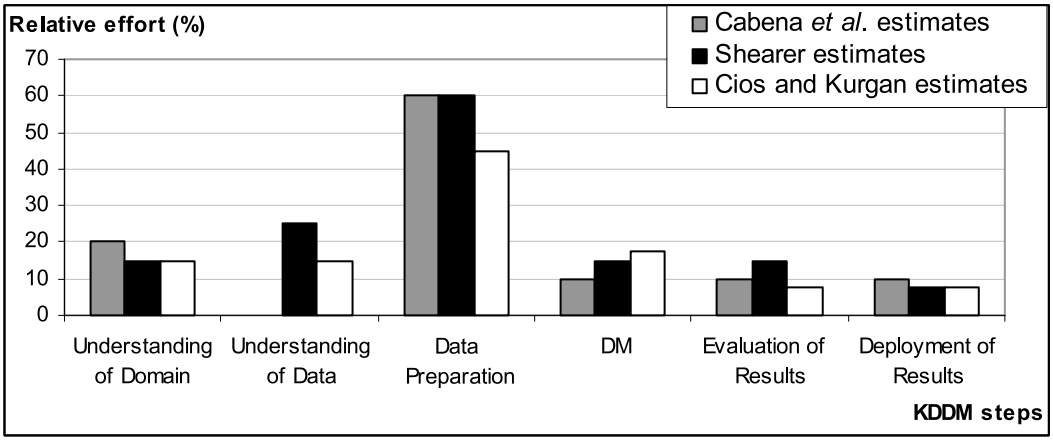


Figure 1 Relative effort spent on specific steps in the KDDM process

particular model is left for the readers since it strongly depends on application domain, user's background, individual preferences, and familiarity with technologies and tools.

4.4 Evaluation of relative effort required for particular steps

An important aspect of a KDDM process is the relative time spent to complete its individual steps. Evaluation of such effort would enable more precise scheduling and thus significantly increase the usability of the model. There have been several estimates proposed by various researchers. This section summarizes the findings.

Cabena *et al.* estimate that about 20% of effort is spent on Business Objective Determination, about 60% on Data Preparation and about 10% on DM, Analysis of Results and Knowledge Assimilation steps (Cabena *et al.*, 1998). Based on experience with industrial applications, Shearer estimates that about 50–70% of time is spent on Data Preparation, 20–30% on Data Understanding, 10–20% on Modeling, Evaluation, and Business Understanding steps, and 5–10% on the Deployment step (Shearer, 2000). Cios & Kurgan estimate that 10–20% of effort is spent of Understanding of the Domain and Understanding of the Data, 30–60% on Data Preparation, 10–25% on DM and 5–10% on Evaluation and Using the Discovered Knowledge (Cios & Kurgan, 2005). On the other hand, Brachman & Anand assert that larger amount of time, about 15–25%, is spent on the DM step (Brachman & Anand, 1996). This may be related to the limited availability of easy to use DM tools at the time of writing. Figure 1 shows a comparison of these estimates in form of a chart. An average value is plotted for interval-based estimates. Note that these numbers are estimates quantifying relative effort, and thus may not sum up to 100%.

These estimates have several characteristics in common. First, it is usually assumed that about half of the project effort is spent on Data Preparation. Second, the DM step usually takes very little time, comparable with the effort required for each of the remaining steps.

There are several reasons why the Data Preprocessing step requires so much time. Enterprise companies often do not collect all the necessary data, and the collected data often contain about 1–5% errors, and can be redundant and inconsistent (Redman, 1998). Many domains have their specific problems, such as medicine where problems are often encountered with a physician's interpretations that are written in unstructured free-text English (Cios & Moore, 2002). Preprocessing of data very often requires a significant amount of manual work involving data manipulation that is difficult to automate (Kurgan *et al.*, 2005). Such common and serious data quality problems contribute to the extent of the Data Preprocessing step.

On the other hand, the DM step is assumed to require a relatively small amount of effort. The main reason is that it uses automated or semi-automated methods on already prepared data (Cabena *et al.*, 1998). Many years of research in the DM field have shown that there is no

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

universally best mining algorithm, but rather a set of equivalently good algorithms for any particular domain (e.g. transactional data, temporal data, etc.). Thus, there is relatively little effort required to select and use a DM method in practice since users do not have to spent time optimizing algorithmic details (Fayyad *et al*., 1996c). Data analysts can use many off-the-shelf commercial DM tools to perform this step. See Goebel & Gruenwald (1999) for a survey of 43 existing software implementations. An overview of several modern commercial KDDM systems that provide extensive DM support can be found in Klosgen & Zytkow (2002).

Of course, there may be deviations from the above estimates. Hirji has reported on a KDDM application in an industrial domain in which the DM step took about 45% of the total project's effort compared to only 30% spent on data preparation (Hirji, 2001). The main reason for this variation was availability of a well-maintained data warehouse that was used as a data source.

4.5 Interestingness of knowledge generated during a KDDM process

Every KDDM project heavily relies on a concept of interestingness. Knowledge discovered today may no longer be interesting tomorrow, and data miners should be able to not only recognize this fact, but also accommodate for it during the process.

The assessment of knowledge generated during the KDDM process is usually approached in one of the following two ways. In the first approach, a domain expert manually analyzes the generated knowledge, and judges its usefulness and interestingness according to their own knowledge and established project goals. The other approach performs a more formal evaluation, usually involving statistical tests via cross-validation or more advanced test schemas. This evaluation usually involves different measures, which mainly depend on applied learning paradigms and a predefined goal. For example, in the case of predictive tasks, common measures include predictive accuracy tests, sometimes also supported by specificity and sensitivity values (Cios & Moore, 2002).

There are also more general measures, such as interestingness, which provides an overall measure of pattern values combining novelty, usefulness, and simplicity (Piatesky-Shapiro & Matheus, 1994; Silbershatz & Tuzhilin, 1995). An excellent review of methods used to characterize interestingness of discovered patterns is provided in (Hilderman & Hamilton, 1999). A similar overview, but concentrating on the evaluation of rules, is given by Mitra *et al*. (2002). It includes multitude of measures, such as accuracy, fidelity, confusion, coverage, confidence, complexity, etc. The evaluation of results is executed most commonly by combining both manual (by an expert) and automated statistical approaches. Usually a large number of patterns are mined, and the formal evaluation is used to sort out all irrelevant or obvious cases before a human expert performs the manual assessment (Padmanaphan & Tuzhilin, 1998).

5 Evolution and future KDDM process models

5.1 History of knowledge discovery systems

The evolution of KD systems has already undergone three distinct periods (Piatesky-Shapiro, 1999a). The first generation systems appeared in the 1980s and included research tools focused on individual DM tasks, such as building a classifier using a decision-tree or a neural network. Such tools addressed specific data-analysis problems, and required technically sophisticated users. The main difficulty was to use more than one tool on the same data, which often required significant transformation of data and metadata. The second-generation systems, called suites, were developed in the mid-1990s. They provided multiple types of integrated data analysis methods and support for data cleaning, preprocessing, and visualization. Examples include systems such as SPSS's Clementine©, Purple Insight's MineSet™, IBM's Intelligent Miner™, and SAS Institute's Enterprise Miner™. The third generation systems were developed in late the 1990s and introduced a vertical approach. These systems addressed specific business problems, such as fraud detection,

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

and provided interface that was designed to hide the internal complexity of DM methodologies. Some of these suites also used KDDM process models to guide the execution of projects. Examples include Purple Insight's MineSet™ that uses the nine-step process model by Fayyad *et al*. SPSS's Clementine© that uses the CRISP-DM process model, and SAS's Enterprise Miner™ that uses the SEMMA process model.

5.2 New trends in KDDM process models

The future of KDDM process models is in achieving overall integration of the entire process through the use of other popular industrial standards. Another currently very important issue is to provide interoperability and compatibility between different software systems and platforms, which also concerns KDDM models. Such systems would serve end-users in automating, or more realistically semi-automating, work with KD systems (Cios & Kurgan, 2005).

The interactive nature that is inherent to KDDM process (Brachman & Anand, 1996) is currently one of the major reasons why solutions provided by the KDDM community have only a limited impact compared to potentially very large industrial interest. A current goal is to enable users to carry out KDDM projects without possessing extensive background knowledge, without manual data manipulation, and with manual procedures to exchange data and knowledge between different DM methods. This requires the ability to store and exchange not only the data, but most importantly knowledge that is expressed in terms of data models generated by the KDDM process, and meta-data that describes data and domain knowledge used in the process. A technology, which can help in achieving these goals, is XML (eXtensible Markup Language), a standard proposed and maintained by the World Wide Web Consortium (Bray *et al*., 2000). XML and other standards that are developed based on XML to support the integration of KDDM process are discussed next. This is followed by a brief summary of research in non-XML based integration.

5.2.1 XML-based integration

XML permits the description and storage of structured or semi-structured data, and to exchange data in a platform- and tool-independent way. From the KD perspective, XML can help to achieve multiple goals (Cios & Kurgan, 2005):

– to implement and standardize communication between diverse KD and database systems;
– to build standard repositories for sharing data between KD systems that work on different software platforms;
– to provide a framework for integrating the entire KDDM process.

While XML itself helps to solve only some problems connected with the support for consolidation of the entire KDDM process, metadata standards based on XML can provide a complete solution (Clifton & Thuraisingham, 2001). Some metadata standards, such as Predictive Model Markup Language (PMML) (PMML, 2001), were identified to allow interoperability among different DM tools and to achieve integration with other applications, including database systems, spreadsheets, and decision support systems (Piatesky-Shapiro, 1999a; Apps, 2000).

PMML is an XML-based language designed by the DM Group. (DMG is an independent, vendor-led group, developing DM standards. DMG members include IBM, Microsoft, Oracle, SPSS Inc., Angoss, MineIt Software Ltd., and about 15 other industrial companies.) PMML describes the input to DM models, the transformations used to prepare data for DM, and the parameters that define the models themselves (DMG, 2005). Currently in version 3.0, it is used to describe data models (generated knowledge) and share them between compliant applications. By using PMML, users can generate data models using one application, use another application to analyze them, another to evaluate them, and finally yet another application to visualize the model. In effect, PMML brings DM closer to domain experts that understand the business process but not

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

necessarily the DM techniques used to analyze business data (Swoyer, 2005). The language has already been used as a standard to support the Evaluation step from the KDDM process. A tool called VizWiz was developed for visualization of models expressed using PMML (Wettschereck *et al.*, 2003). Another tool, called PEAR, was developed for exploration and visualization of association rules expressed in PMML (Jorge *et al.*, 2002).

Both XML and PMML can be easily stored in most current database management systems. A model scenario of XML/PMML-based integration of KDDM process contains the following activities:

– information collected during Domain and Data Understanding steps is stored as XML documents;
– these documents are used in Data Preparation and DM steps as a source of information that can be accessed automatically, across platforms and across tools;
– knowledge extracted in DM step and domain knowledge gathered in the Domain Understanding step can be stored using PMML documents that can be exchanged among different software tools.

In short, integration and interoperability of modern KDDM models may be achieved by application of modern industrial standards, such as XML and PMML. Such synergic application of several well-known standards brings additional exposure to the KDDM industry. This way, new users following XML standards will be exposed to, and attracted by, KDDM applications despite their lack of direct knowledge of KDDM.

5.2.2 Other approaches to integration

KDDM process integration is also approached using methods that are not based on XML. Several recent examples include KDB2000 (Appice *et al.*, 2002), RuleViz (Han & Cercone, 2000), VidaMine (Kimani *et al.*, 2004), and Intelligent Discovery Assistant (Bernstein *et al.*, 2005).

KDB2000 is a support tool that provides integrated set of techniques for entire the KDDM process, which includes database access, data preprocessing and transformation techniques, a full range of DM methods, and pattern validation and visualization (Appice *et al.*, 2002). Another integration support endeavor is RuleViz, a system for visualizing the entire process of KDDM. The system consists of five components corresponding to the main constituents of a KDDM process: original data visualization, visual data reduction, visual data preprocessing, visual rule discovery, and rule visualization. The aim is to help users navigate through enormous search spaces, recognize their intentions, help them gain a better insight into multi-dimensional data, understand intermediate results, and interpret the discovered patterns (Han & Cercone, 2000). Another similar tool is VidaMine (Kimani *et al.*, 2004). Other approaches to integration include the development of an ontology-driven Intelligent Discovery Assistant, which helps to automate the process of selecting the most beneficial processing steps for execution of a valid KDDM process. This system has been applied to a problem of cost sensitive classification for data from KDDCUP 1998. The process is built as a combination between the Fayyad and CRISP-DM processes (Bernstein *et al.*, 2005). Finally, a documentation infrastructure model and prototype tool has been developed to support project management activities related to KDDM projects. This tool allows for organization and contextualization of various artifacts used or generated by the process in terms of process activities, maintenance of process history, and support for process/task re-execution, restructuring, or project continuation (Becker & Ghedini, 2005). The authors refer to CRISP-DM, Cabena's and Fayyad's models.

6 Conclusions and summary

The KDDM industry is on the verge of possessing one of the most successful business technologies. What stands in the way to the success is the inaccessibility of the related applications to broad

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

scientific and industrial communities. This shortcoming can be overcome only by moving beyond its algorithm-centric roots (Apps, 2000). 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 (Piatesky-Shapiro, 1999a). The fuel for the growth is the strong economic and social need for solutions provided by the KDDM community (Fayyad *et al*., 1996c).

This survey has provided an overview of the state-of-the-art in developing one of the most important standards, the KDDM process model. The description and comprehensive comparison of several main models has been provided, along with discussion of issues associated with their implementation. The goal of this survey has been to consolidate research in this area, to inform users about different models and how to select the appropriate model, and to develop improved models that are based on previous experiences. Repeated reference and the introduction of KDDM and other standard methodologies significantly contributes to establishing *de facto* industrial standards. There are many commonly recognized advantages of introducing standards. Standardization of a KDDM process model will enable standard methods and procedures to be developed, resulting in making it easier for the end users to deploy related projects (Clifton & Thuraisingham, 2001). It will directly lead to performing projects faster, cheaper, more manageably, and more reliably. The standards will promote development and delivery of solutions that use business language, rather than the traditional language of algorithms, matrices, criterions, complexities, and the like, which will result in a greater exposure and acceptance of the KDDM industry. This, in turn, will be a significant factor in pushing the industry beyond the edge, and into the mainstream.

Acknowledgements

The authors would like to thank Dr Marek Reformat for useful comments and suggestions on earlier versions of this article and the anonymous reviewers for their constructive criticism. Dr. Kurgan also gratefully acknowledges contributions of Dr. Krzysztof Cios who first introduced him to the subject of KDDM process models. Research presented in this paper was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Adriaans, P and Zantinge, D, 1996, *Data Mining*. Reading, MA: Addison-Wesley.

Anand, S and Buchner, A, 1998, *Decision Support Using Data Mining*. Financial Time Management, London.

Anand, S, Patrick, A, Hughes, J and Bell, D, 1998, A data mining methodology for cross-sales. *Knowledge Based Systems Journal* **10**, 449–461.

Appice, A, Ceci, M and Malerba, D, 2002, KDB2000: an integrated knowledge discovery tool. *Management Information Systems* **6**, 531–540.

Apps, E, 2000, New mining industry standards: moving from monks to the mainstream. *PC AI* **14**(6), 46–50.

Becker, K and Ghedini, C, 2005, A documentation infrastructure for the management of data mining projects. *Information and Software Technology* **47**(2), 95–111.

Bernstein, A, Provost, F and Hill, S, 2005, Toward intelligent assistance for a data mining process: an ontology-based approach for cost-sensitive classification. *IEEE Transactions on Knowledge and Data Engineering* **17**(4), 503–518.

Berry, M and Linoff, G, 1997, *Data Mining Techniques for Marketing, Sales, and Customer Support*. Wiley.

Blockeel, H and Moyle, S, 2002, Collaborative data mining needs centralized model evaluation. In *Proceedings of the ICML-2002 Workshop on Data Mining Lessons Learned*, pp. 21–28.

Boehm, B, 1998, A spiral model of software development and enhancement. *IEEE Computer* **21**(5), 61–72.

Brachman, R and Anand, T, 1996, The process of knowledge discovery in databases: a human-centered approach. In Fayyad, U, Piatetsky-Shapiro, G, Smyth, P and Uthurusamy, R (eds) *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press, pp. 37–58.

Bray, T, Paoli, J and Maler, E, 2000, Extensible Markup Language (XML) 1.0 (2nd edition). W3C Recommendation, `http://www.w3.org/TR/2000/REC-xml-20001006`.

Brodley, C and Smyth, P, 1997, Applying classification algorithms in practice. *Statistics and Computing* **7**, 45–56.

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

Brooks, F, 1995, *The Mythical Man-Month*, anniversary edn. Reading, MA: Addison-Wesley, pp. 153–160.
Brunk, C, Kelly, J and Kohavi, R, 1997, MineSet: an integrated system for data mining. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*. AAAI Press, pp. 135–138.
Buchheit, RB, Garrett, JH, Jr, Lee, SR and Brahme, R, 2000, A knowledge discovery framework for civil infrastructure: a case study of the intelligent workplace. *Engineering with Computers* **16**(3–4), 264–274.
Buchner, A, Mulvenna, M, Anand, S and Hughes, J, 1999, An internet-enabled knowledge discovery process. In *Proceedings of the 9th International Database Conference, Hong Kong*, pp. 13–27.
Butler, S, 2002, An investigation into the relative abilities of three alternative data mining methods to derive information of business value from retail store-based transaction data. BSc thesis, School of Computing and Mathematics, Deakin University, Australia.
Cabena, P, Hadjinian, P, Stadler, R, Verhees, J and Zanasi, A, 1998, *Discovering Data Mining: From Concepts to Implementation*. Prentice Hall.
Castellano, M, Pastore, N, Arcieri, F, Summo, V and De Grecis, GB, 2004, A model-view-controller architecture for knowledge discovery. *Management Information Systems* **10**, 383–392.
Cios, K, Teresinska, A, Konieczna, S, Potocka, J and Sharma, S, 2000, Diagnosing myocardial perfusion from PECT bull's-eye maps—a knowledge discovery approach. *IEEE Engineering in Medicine and Biology Magazine*, Special issue on Medical Data Mining and Knowledge Discovery **19**(4), 17–25.
Cios, K (ed.) 2001, *Medical Data Mining and Knowledge Discovery*. Springer-Verlag.
Cios, K and Moore, G, 2002, Uniqueness of medical data mining. *Artificial Intelligence in Medicine* **26**(1–2), 1–24.
Cios, K and Kurgan, L, 2005, Trends in data mining and knowledge discovery. In Pal, N and Jain, L (eds) *Advanced Techniques in Knowledge Discovery and Data Mining*. Springer, pp. 1–26.
Clifton, C and Thuraisingham, B, 2001, Emerging standards for data mining. *Computer Standards and Interfaces* **23**, 187–193.
CRISP-DM, 2003, CRoss Industry Standard Process for Data Mining, `http://www.crisp-dm.org`.
De Abajo, N, Lobato, V, Diez, AB and Cuesta, SR, 2004, ANN quality diagnostic models for packaging manufacturing: an industrial Data Mining case study. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 799–804.
DMG, 2005, The Data Mining Group, `http://www.dmg.org/`.
Edelstein, H, 1998, Data mining: let's get practical. *DB2 Magazine* **3**(2), summer.
Fayyad, U, Piatesky-Shapiro, G, Smyth, P and Uthurusamy, R (eds), 1996a, *Advances in Knowledge Discovery and Data Mining*. AAAI Press.
Fayyad, U, Piatetsky-Shapiro, G and Smyth, P, 1996b, From data mining to knowledge discovery: an overview. In Fayyad, U, Piatetsky-Shapiro, G, Smyth, P and Uthurusamy, R (eds) *Advances in Knowledge Discovery and Data Mining*. AAAI Press, pp. 1–34.
Fayyad, U, Piatetsky-Shapiro, G, & Smyth, P, 1996c, The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM* **39**(11), 27–34.
Fayyad, U, Piatetsky-Shapiro, G and Smyth, P, 1996d, Knowledge discovery and data mining: towards a unifying framework. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR*, pp. 82–88.
Fayyad, U, Piatetsky-Shapiro, G and Smyth, P, 1996e, From data mining to knowledge discovery in databases. *AI Magazine* **17**(3), 37–54.
Fayyad, U, Haussler, D and Stolorz, P, 1996f, KDD for science data analysis: issues and examples. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR*, pp. 50–56.
Frawley, W, Piatesky-Shapiro, G and Matheus, C, 1991, Knowledge discovery in databases: an overview. In Piatesky-Shapiro, G and Frowley, W (eds) *Knowledge Discovery in Databases*. AAAI/MIT Press, pp. 1–27.
Ganzert, S. Guttmann, J, Kersting, K, Kuhlen, R, Putensen, C, Sydow, M and Kramer, S, 2002, Analysis of respiratory pressure–volume curves in intensive care medicine using inductive machine learning. *Artificial Intelligence in Medicine* **26**(1–2), 69–86.
Gersten, W, Wirth, R and Arndt D, 2000, Predictive modeling in automotive direct marketing: tools, experiences and open issues. In *Proceeding of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 398–406.
Geyer-Schulz, A and Hahsler, M, 2002, Comparing two recommender algorithms with the help of recommendations by peers. In *Proceedings of the 4th International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles, Edmonton, Canada.*
Goebel, M and Gruenwald, L, 1999, A survey of data mining software tools. *SIGKDD Explorations* **1**(1), 20–33.
Goh, KG, Hsu, W, Lee, ML and Wang, H, 2001, ADRIS: an automatic diabetic retinal image screening system. In Cios, K (ed.) *Medical Data Mining and Knowledge Discovery*, pp. 181–207.

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

Gupta, S, Bhatnagar, N, Wasan, S and Somayajulu, D, 2000, Intension mining: a new paradigm in knowledge discovery. Technical Report No. IITD/CSE/TR2000/001, Department of Computer Science and Engineering, Indian Institute of Technology.

Haglin, D, Roiger, R, Hakkila, J and Giblin, T, 2005, A tool for public analysis of scientific data. *Data Science Journal* **4**(30), 39–53.

Han, J and Cercone, N, 2000, RuleViz: a model for visualizing knowledge discovery process. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discover and Data Mining*, pp. 244–253.

Han, J and Kamber, M, 2001, *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Hand, D, 1994, Deconstructing statistical questions. *Journal of Royal Statistical Society*, 317–356.

Hilderman, R and Hamilton, H, 1999, Knowledge discovery and interestingness measures: a survey. Technical Report CS 99–04, University of Regina, Regina, Saskatchewan, Canada.

Hipp, J and Lindner, G, 1999, Analyzing warranty claims of automobiles. An application description following the CRISP-DM data mining process. In *Proceedings of 5th International Computer Science Conference, Hong Kong, China*, pp. 31–40.

Hirji, K, 2001, Exploring data mining implementation. *Communications of the ACM* **44**(7), 87–93.

Hofer, J and Brezany P, 2004, Distributed Decision Tree Induction within the Grid Data Mining Framework GridMiner-Core. GridMiner TR2004–04, Institute for Software Science, University of Vienna.

Jensen, S, 2001, Mining medical data for predictive and sequential patterns: PKDD 2001. In *Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD2001 Discovery Challenge on Thrombosis Data.

Jorge, A, Pocas, J and Azevedo, P, 2002, Post-processing operators for browsing large sets of association rules. In *Proceedings of the ECML/PKDD'02 workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, pp. 53–64.

Kimani, S, Lodi, S, Catarci, T, Santucci, G and Sartori, C, 2004, VidaMine: a visual data mining environment. *Journal of Visual Languages and Computing* **15**(1), 37–67.

Klemettinen, M, Mannila, H and Toivonen, H, 1997, A data mining methodology and its application to semi-automatic knowledge acquisition. In *Proceedings of the 8th International Conference and Workshop on Database and Expert Systems Applications (DEXA'97), Toulouse, France*, pp. 670–677.

Klosgen, W, 1992, Problems for knowledge discovery in databases and their treatment in the statistics interpreter explora. *Journal of Intelligent Systems* **7**(7), 649–673.

Klosgen, W and Zytkow, J, 1996, Knowledge discovery in databases terminology. In Fayyad, U, Piatetsky-Shapiro, G, Smyth, P and Uthurusamy, R (eds) *Advances in Knowledge Discovery and Data Mining*. AAAI Press, pp. 573–592.

Klosgen, W and Zytkow, J (eds), 2002, *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press.

Klosgen, W and Zytkow, J, 2002, The knowledge discovery process. In Klosgen, W and Zytkow, J (eds) *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press, pp. 10–21.

Kurgan, L, Cios, K, Tadeusiewicz, R, Ogiela, M and Goodenday, L, 2001, Knowledge discovery approach to automated cardiac SPECT diagnosis. *Artificial Intelligence in Medicine* **23**(2), 149–169.

Kurgan, L, Cios, K, Sontag, M and Accurso, F, 2005, Mining the cystic fibrosis data. In Zurada, J and Kantardzic, M (eds) *Next Generation of Data-Mining Applications*. IEEE Press and Wiley, pp. 415–444.

Lawrence, S, Giles, C and Bollacker, K, 1999, Digital libraries and autonomous citation indexing. *IEEE Computer* 32(6), 67–71.

Li, S-T and Shue, L-Y, 2004, Data mining to aid policy making in air pollution management. *Expert Systems with Applications* **27**(3), 331–340.

Li, T and Ruan, D, 2004, An extended process model of knowledge discovery in databases. In *Applied Computational Intelligence—Proceedings of the 6th International FLINS Conference*, pp. 185–188.

Maruster, L, Weijters, T, De Vries, G, Van den Bosch, A and Daelemans, W, 2002, Logistic-based patient grouping for multi-disciplinary treatment. *Artificial Intelligence in Medicine* **26**(1–2), 87–107.

Mitra, S, Pal, S and Mitra, P, 2002, Data mining in soft computing framework: a survey. *IEEE Transactions on Neural Networks* **13**(1), 3–14.

Morello, BC, Michaut, D and Baptiste, P, 2001, A knowledge discovery process for a flexible manufacturing system. In *IEEE Symposium on Emerging Technologies and Factory Automation*, pp. 651–658.

Mounier, A, Boissier, O and Jacquenet, F, 2003, How to learn to interact? In *Proceedings of the International Conference on Autonomous Agents*, pp. 1072–1073.

Moyle, S, Bohanec, M and Ostrowski, E, 2002, Large and tall buildings: a case study in the application of decision support and data mining. In *Proceedings of the ECML/PKDD'02 workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, pp. 88–99.

NIST, 2005, NIST/SEMATECH e-Handbook of Statistical Methods, `http://www.itl.nist.gov/div898/handbook/`.

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press

Padmanaphan, B and Tuzhilin, A, 1998, A belief-driven method for discovering unexpected patterns. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pp. 94–100.

Perner, P, Perner, H and Muller, B, 2002, Mining knowledge for HEp-2 cell image classification. *Artificial Intelligence in Medicine* **26**(1–2), 161–173.

Piatetsky-Shapiro, G, 1991, Knowledge discovery in real databases: a report on the IJCAI-89 workshop. *AI Magazine* **11**(5), 68–70.

Piatesky-Shapiro, G and Matheus, C, 1992, Knowledge discovery workbench for exploring business databases. *International Journal of Intelligent Agents* **7**(7), 675–686.

Piatesky-Shapiro, G and Matheus, C, 1994, The interestingness of deviations. In *Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases, Seattle, Washington*, pp. 25–36.

Piatesky-Shapiro, G, 1999a, The data mining industry coming to age. *IEEE Intelligent Systems* **14**(6), 32–33.

Piatesky-Shapiro, G, 1999b, CRISP-DM: a proposed global standard for data mining. *On-Line Executive Journal for Data-Intensive Decision Support* **3**(15).

PMML, 2001, *Second Annual Workshop on the Predictive Model Markup Language, San Francisco, CA, August 2001*.

Redman, T, 1998, The impact of poor data quality on the typical enterprise. *Communications of the ACM* **41**(2), 79–81.

Reinartz, T, 1999, *Focusing Solutions for Data Mining: Analytical Studies and Experimental Results in Real-World Domains (Lecture Notes in Computer Science, 1623)*. Springer.

Reinartz, T, 2002, Stages of the discovery process. In Klosgen, W and Zytkow, J (eds) *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press, pp. 185–192.

Rouse, W, 2002, Need to know—information, knowledge, and decision making. *IEEE Transactions on Systems, Man, and Cybernetics, Part C* **32**(4), 282–292.

Royce, W, 1970, Managing the development of large software system concepts and techniques. In *Proceedings of the WESCON*. IEEE, pp. 1–9.

Sacha, J, Cios, K and Goodenday, L, 2000, Issues in automating cardiac SPECT diagnosis. *IEEE Engineering in Medicine and Biology Magazine*, Special issue on Medical Data Mining and Knowledge Discovery **19**(4), 78–88.

SAS, 1997, SAS Institute Inc, *From Data to Business Advantage: Data Mining, SEMMA Methodology and the SAS System* (White Paper).

Shalvi, D and DeClaris, N, 2001, A data clustering and visualization methodology for epidemiological pathology discoveries. In Cios, K (ed.) *Medical Data Mining and Knowledge Discovery*, pp. 129–151.

Shearer, C, 2000, The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing* **l5**(4), 13–19.

Silberschatz, A and Tuzhilin, A, 1995, On subjective measures of interestingness in knowledge discovery. In *Proceedings of KDD-95, the 1st International Conference on Knowledge Discovery and Data Mining, Menlo Park, CA*, pp. 275–281.

Silva, EM, Do Prado, HA and Ferneda, E, 2002, Text mining: crossing the chasm between the academy and the industry. *Management Information Systems* **6**, 351–361.

Simoudis, E, Livezey, B and Kerber, R, 1994, Integrating inductive and deductive reasoning in data mining. In *Proceedings of 1994 AAAI Workshop on Knowledge Discovery in Databases*, pp. 37–48.

SPSS, 2003, Clementine 8.0, `http://www.spss.com/spssbi/clementine/`.

Swoyer, S, 2005, PMML: data mining for the masses? *Enterprise Strategies Newsletters*, `http://esj.com/`.

Tukey, JW, 1977, *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Vaschetto, M, Weissbrod, T, Bodle, D and Guner, O, 2003, Enabling high-throughput discovery. *Current Opinion in Drug Discovery and Development* **6**(3), 377–383.

Wettschereck, D, Jorge, A and Moyle, S, 2003, Visualization and evaluation support of knowledge discovery through the predictive model markup language. In *Proceedings of 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*. Springer, pp. 493–501.

Wirth, R and Hipp, J, 2000, CRISP-DM: towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK*, pp. 29–39.

Wiederhold, G, 1996, Foreword: on the barriers and future of knowledge discovery. In Fayyad, U, Piatetsky-Shapiro, G, Smyth, P and Uthurusamy, R (eds) *Advances in Knowledge Discovery and Data Mining*. AAAI Press.

Ziarko, R, Golan, R and Edwards, D, 1993, An application of Datalogic/R knowledge discovery tool to identify strong predictive rules in stock market data. Working notes from the *Workshop on Knowledge Discovery in Databases, Seattle, Washington*, pp. 89–101.

Zytow, J and Baker, J, 1991, Interactive mining of regularities in databases. In Piatesky-Shapiro, G and Frowley, WJ (eds) *Knowledge Discovery in Databases*. AAAI Press, pp. 31–53.

https://doi.org/10.1017/S0269888906000737 Published online by Cambridge University Press