Foreword

"If you torture the data long enough, Nature will confess," said 1991 Nobel-winning economist Ronald Coase. The statement is still true. However, achieving this lofty goal is not easy. First, "long enough" may, in practice, be "too long" in many applications and thus unacceptable. Second, to get "confession" from large data sets one needs to use state-of-the-art "torturing" tools. Third, Nature is very stubborn — not yielding easily or unwilling to reveal its secrets at all.

Fortunately, while being aware of the above facts, the reader (a data miner) will find several efficient data mining tools described in this excellent book. The book discusses various issues connecting the whole spectrum of approaches, methods, techniques and algorithms falling under the umbrella of data mining. It starts with data understanding and preprocessing, then goes through a set of methods for supervised and unsupervised learning, and concludes with model assessment, data security and privacy issues. It is this specific approach of using the knowledge discovery process that makes this book a rare one indeed, and thus an indispensable addition to many other books on data mining.

To be more precise, this is a book on knowledge discovery from data. As for the data sets, the easy-to-make statement is that there is no part of modern human activity left untouched by both the need and the desire to collect data. The consequence of such a state of affairs is obvious. We are surrounded by, or perhaps even immersed in, an ocean of all kinds of data (such as measurements, images, patterns, sounds, web pages, tunes, etc.) that are generated by various types of sensors, cameras, microphones, pieces of software and/or other human-made devices. Thus we are in dire need of automatically extracting as much information as possible from the data that we more or less wisely generate. We need to conquer the existing and develop new approaches, algorithms and procedures for knowledge discovery from data. This is exactly what the authors, world-leading experts on data mining in all its various disguises, have done. They present the reader with a large spectrum of data mining methods in a gracious and yet rigorous way.

To facilitate the book's use, I offer the following *roadmap* to help in:

a) reaching certain desired destinations without undesirable wandering, and
b) getting the basic idea of the breadth and depth of the book.

First, an overview: the volume is divided into seven parts (the last one being Appendices covering the basic mathematical concepts of Linear Algebra, Probability Theory, Lines and Planes in Space, and Sets). The main body of the book is as follows: Part 1, Data Mining and Knowledge Discovery Process (two Chapters), Part 2, Data Understanding (three Chapters), Part 3, Data Preprocessing (three Chapters), Part 4, Data Mining: Methods for Constructing Data Models (six Chapters), Part 5, Data Models Assessment (one Chapter), and Part 6, Data Security and Privacy Issues (one Chapter). Both the ordering of the sections and the amount of material devoted to each particular segment tells a lot about the authors' expertise and perfect control of the data mining field. Namely, unlike many other books that mainly focus on the modeling part, this volume discusses all the important — and elsewhere often neglected — parts before and after modeling. This breadth is one of the great characteristics of the book.

A dive into particular sections of the book unveils that Chapter 1 defines what data mining is about and stresses some of its unique features, while Chapter 2 introduces a Knowledge Discovery Process (KDP) as a process that seeks new knowledge about an application domain. Here, it is pointed out that Data Mining (DM) is just one step in the KDP. This Chapter also reminds us that the KDP consists of multiple steps that are executed in a sequence, where the next step is initiated upon successful completion of the previous one. It also stresses the fact that the KDP stretches between the task of understanding of the project domain and data, through data preparation and analysis, to evaluation, understanding and application of the generated knowledge. KDP is both highly iterative (there are many repetitions triggered by revision processes) and interactive. The main reason for introducing the process is to formalize knowledge discovery (KD) projects within a common framework, and emphasize independence of specific applications, tools, and vendors. Five KDP models are introduced and their strong and weak points are discussed. It is acknowledged that the data preparation step is by far the most time-consuming and important part of the KDP.

Chapter 3, which opens Part 2 of the book, tackles the underlying core subject of the book, namely, data and data sets. This includes an introduction of various data storage techniques and of the issues related to both the quality and quantity of data used for data mining purposes. The most important topics discussed in this Chapter are the different data types (numerical, symbolic, discrete, binary, nominal, ordinal and continuous). As for the organization of the data, they are organized into rectangular tables called data sets, where rows represent objects (samples, examples, patterns) and where columns represent features/attributes, i.e., the input dimension that describes the objects. Furthermore, there are sections on data storage using databases and data warehouses. The specialized data types — including transactional data, spatial data, hypertext, multimedia data, temporal data and the World Wide Web — are not forgotten either. Finally, the problems of scalability while faced with a large quantity of data, as well as the dynamic data and data quality problems (including imprecision, incompleteness, redundancy, missing values and noise) are also discussed. At the end of each and every Chapter, the reader can find good bibliographical notes, pointers to other electronic or written sources, and a list of relevant references.

Chapter 4 sets the stage for the core topics covered in the book, and in particular for Part 4, which deals with algorithms and tools for concepts introduced herein. Basic learning methods are introduced here (unsupervised, semi-supervised, supervised, reinforcement) together with the concepts of classification and regression.

Part 2 of the book ends with Chapter 5, which covers knowledge representation and its most commonly encountered schemes such as rules, graphs, networks, and their generalizations. The fundamental issue of abstraction of information captured by information granulation and resulting information granules is discussed in detail. An extended description is devoted to the concepts of fuzzy sets, granularity of data and granular concepts in general, and various other set representations, including shadow and rough sets. The authors show great care in warning the reader that the choice of a certain formalism in knowledge representation depends upon a number of factors and that while faced with an enormous diversity of data the data miner has to make prudent decisions about the underlying schemes of knowledge representation.

Part 3 of the book is devoted to *data preprocessing* and contains three Chapters. Readers interested in Databases (DB), Data Warehouses (DW) and On-Line Analytical Processing (OLAP) will find all the basics in Chapter 6, wherein the elementary concepts are introduced. The most important topics discussed in this Chapter are Relational DBMS (RDBMS), defined as a collection of interrelated data and a set of software programs to access those data; SQL, described as a declarative language for writing queries for a RDBMS; and three types of languages to retrieve and manipulate data: Data Manipulation Language (DML), Data Definition Language (DDL), and Data Control Language (DCL), which are implemented using SQL. DW is introduced as a subject-oriented, integrated, time-variant and non-volatile collection of data in support

of management's decision-making process. Three types of DW are distinguished: virtual data warehouse, data mart, and enterprise warehouse. DW is based on a multidimensional data model: the data is visualized using a multidimensional data cube, in contrast to the relational table that is used in the RDBMS. Finally, OLAP is discussed with great care to details. This Chapter is relatively unique, and thus enriching, among various data mining books that typically skip these topics.

If you are like the author of this Foreword, meaning that you love mathematics, your heart will start beating faster while opening Chapter 7 on *feature extraction (FE) and feature selection (FS) methods*. At this point, you can turn on your computer, and start implementing some of the many models nicely introduced and explained here. The titles of the topics covered reveal the depth and breadth of supervised and unsupervised techniques and approaches presented: Principal Component Analysis (PCA), Independent Component Analysis (ICA), Karhunen-Loeve Transformation, Fisher's linear discriminant, SVD, Vector quantization, Learning vector quantization, Fourier transform, Wavelets, Zernike moments, and several feature selection methods. Because FE and FS methods are so important in data preprocessing, this Chapter is quite extensive.

Chapter 8 deals with one of the most important, and often required, preprocessing methods, the overall goal of which is to reduce the complexity of the data for further data mining tasks. It introduces unsupervised and supervised discretization methods of continuous data attributes. It also outlines a dynamic discretization algorithm and includes a comparison between several state of the art algorithms.

Part 4, *Data Mining: Methods for Constructing Data Models,* is comprised of two Chapters on the basic types of unsupervised learning, namely, Clustering and Association Rules; three Chapters on supervised learning, namely Statistical Methods, Decision Trees and Rule Algorithms, and Neural Networks; and a Chapter on Text Mining. Part 4, along with Parts 3 and 6, forms the core algorithmic section of this great data mining volume. You may switch on your computer again and start implementing various data mining tools clearly explained here.

To show the main features of every Chapter in Part 4, let us start with Chapter 9, which covers clustering, a predominant technique used in unsupervised learning. A spectrum of clustering methods is introduced, elaborating on their conceptual properties, computational aspects and scalability. The treatment of huge databases through mechanisms of sampling and distributed clustering is discussed as well. The latter two approaches are essential for dealing with large data sets.

Chapter 10 introduces the other key unsupervised learning technique, namely, association rules. The topics discussed here are association rule mining, storing of items using transactions, the association rules categorization as single-dimensional and multidimensional, Boolean and quantitative, and single-level and multilevel, their measurement by using support, confidence, and correlation, and the association rules generation from frequent item sets (a priori algorithm and its modifications including: hashing, transaction removal, data set partitioning, sampling, and mining frequent item sets without generation of candidate item sets).

Chapter 11 constitutes a gentle encounter with *statistical methods* for *supervised learning*, which are based on exploitation of probabilistic knowledge about data. This becomes particularly visible in the case of Bayesian methods. The statistical classification schemes exploit concepts of conditional probabilities and prior probabilities — all of which encapsulate knowledge about statistical characteristics of the data. The Bayesian classifiers are shown to be optimal given known probabilistic characteristics of the underlying data. The role of effective estimation procedures is emphasized and estimation techniques are discussed in detail. Chapter 11 introduces regression models too, including both linear and nonlinear regression. Some of the most representative generalized regression models and augmented development schemes are covered in detail.

Chapter 12 continues along statistical lines as it describes main types of inductive machine learning algorithms: decision trees, rule algorithms, and their hybrids. Very detailed

description of these topics is given and the reader will be able to implement them easily or come up with their extensions and/or improvements. Comparative performances and discussion of the advantages and disadvantages of the methods on several data sets are also presented here.

The classical statistical approaches end here, and neural network models are presented in Chapter 13. This Chapter starts with presentation of biological neuron models: the spiking neuron model and a simple neuron model. This section leads to presentation of learning/plasticity rules used to update the weights between the interconnected neurons, both in networks utilizing the spiking and simple neuron models. Presentation of the most important neuron models and learning rules are unique characteristics of this Chapter. Popular neural network topologies are reviewed, followed by an introduction of a powerful Radial Basis Function (RBF) neural network that has been shown to be very useful in many data mining applications. Several aspects of the RBF are introduced, including its most important characteristic of being similar (almost practically equivalent) to the system of fuzzy rules.

In Chapter 14, concepts and methods related to text mining and information retrieval are presented. The most important topics discussed are information retrieval (IR) systems that concern an organization and retrieval of information from large collections of semi-structured or unstructured text-based databases and the World Wide Web, and how the IR system can be improved by latent semantic indexing and relevance feedback.

Part 5 of the book consists of Chapter 15, which discusses and explains several important and indispensable model selection and model assessment methods. The methods are divided into four broad categories: data re-use, heuristic, formal, and interestingness measures. The Chapter provides justification for why one should use methods from these different categories on the same data. The Akaike's information criterion and Bayesian information criterion methods are also discussed in order to show their relationship to the other methods covered.

The final part of the book, Part 6, and its sole Chapter 16, treats topics that are not usually found in other data mining books but which are very relevant and deserve to be presented to readers. Specifically, several issues of data privacy and security are raised and cast in the setting of data mining. Distinct ways of addressing them include data sanitation, data distortion, and cryptographic methods. In particular, the focus is on the role of information granularity as a vehicle for carrying out collaborative activities (such as clustering) while not releasing detailed numeric data. At this point, the roadmap is completed.

A few additional remarks are still due. The book comes with two important teaching tools that make it an excellent textbook. First, there is an *Exercises* section at the end of each and every Chapter expanding the volume beyond a great research monograph. The exercises are designed to augment the basic theory presented in each Chapter and help the reader to acquire practical skills and understanding of the algorithms and tools. This organization is suitable for both a textbook in a formal course and for self-study. The second teaching tool is a set of PowerPoint presentations, covering the material presented in all sixteen Chapters of the book.

All of the above makes this book a thoroughly enjoyable and solid read. I am sure that no data miner, scientist, engineer and/or interested layperson can afford to miss it.

Vojislav Kecman
University of Auckland
New Zeland

2

The Knowledge Discovery Process

In this Chapter, we describe the knowledge discovery process, present some models, and explain why and how these could be used for a successful data mining project.

1. Introduction

Before one attempts to extract useful knowledge from data, it is important to understand the overall approach. Simply knowing many algorithms used for data analysis is not sufficient for a successful data mining (DM) project. Therefore, this Chapter focuses on describing and explaining the **process** that leads to finding new knowledge. The process defines a sequence of steps (with eventual feedback loops) that should be followed to discover knowledge (e.g., patterns) in data. Each step is usually realized with the help of available commercial or open-source software tools.

To formalize the knowledge discovery processes (KDPs) within a common framework, we introduce the concept of a **process model**. The model helps organizations to better understand the KDP and provides a roadmap to follow while planning and executing the project. This in turn results in cost and time savings, better understanding, and acceptance of the results of such projects. We need to understand that such processes are nontrivial and involve multiple steps, reviews of partial results, possibly several iterations, and interactions with the data owners. There are several reasons to structure a KDP as a **standardized process model**:

1. *The end product must be useful for the user/owner of the data*. A blind, unstructured application of DM techniques to input data, called *data dredging*, frequently produces meaningless results/knowledge, i.e., knowledge that, while interesting, does not contribute to solving the user's problem. This result ultimately leads to the failure of the project. Only through the application of well-defined KDP models will the end product be valid, novel, useful, and understandable.
2. *A well-defined KDP model should have a logical, cohesive, well-thought-out structure and approach that can be presented to decision-makers who may have difficulty understanding the need, value, and mechanics behind a KDP*. Humans often fail to grasp the potential knowledge available in large amounts of untapped and possibly valuable data. They often do not want to devote significant time and resources to the pursuit of formal methods of knowledge extraction from the data, but rather prefer to rely heavily on the skills and experience of others (domain experts) as their source of information. However, because they are typically ultimately responsible for the decision(s) based on that information, they frequently want to understand (be comfortable with) the technology applied to those solution. A process model that is well structured and logical will do much to alleviate any misgivings they may have.

3. *Knowledge discovery projects require a significant project management effort that needs to be grounded in a solid framework*. Most knowledge discovery projects involve teamwork and thus require careful planning and scheduling. For most project management specialists, KDP and DM are not familiar terms. Therefore, these specialists need a definition of what such projects involve and how to carry them out in order to develop a sound project schedule.
4. *Knowledge discovery should follow the example of other engineering disciplines that already have established models*. A good example is the software engineering field, which is a relatively new and dynamic discipline that exhibits many characteristics that are pertinent to knowledge discovery. Software engineering has adopted several development models, including the waterfall and spiral models that have become well-known standards in this area.
5. *There is a widely recognized need for standardization of the KDP*. The challenge for modern data miners is to come up with widely accepted standards that will stimulate major industry growth. Standardization of the KDP model would enable the development of standardized methods and procedures, thereby enabling end users to deploy their projects more easily. It would lead directly to project performance that is faster, cheaper, more reliable, and more manageable. The standards would promote the development and delivery of solutions that use business terminology rather than the traditional language of algorithms, matrices, criterions, complexities, and the like, resulting in greater exposure and acceptability for the knowledge discovery field.

Below we define the KDP and its relevant terminology. We also provide a description of several key KDP models, discuss their applications, and make comparisons. Upon finishing this Chapter, the reader will know how to structure, plan, and execute a (successful) KD project.

2. What is the Knowledge Discovery Process?

Because there is some confusion about the terms data mining, knowledge discovery, and knowledge discovery in databases, we first define them. Note, however, that many researchers and practitioners use DM as a synonym for knowledge discovery; DM is also just one step of the KDP.

Data mining was defined in Chapter 1. Let us just add here that DM is also known under many other names, including *knowledge extraction*, *information discovery*, *information harvesting*, *data archeology*, and *data pattern processing*.

The **knowledge discovery process** (KDP), also called knowledge discovery in databases, seeks new knowledge in some application domain. It is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The process generalizes to nondatabase sources of data, although it emphasizes databases as a primary source of data. It consists of many steps (one of them is DM), each attempting to complete a particular discovery task and each accomplished by the application of a discovery method. Knowledge discovery concerns the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analyzing the application domain.

This book defines the term **knowledge extraction** in a narrow sense. While the authors acknowledge that extracting knowledge from data can be accomplished through a variety of methods — some not even requiring the use of a computer — this book uses the term to refer to knowledge obtained from a database or from textual data via the knowledge discovery process. Uses of the term outside this context will be identified as such.

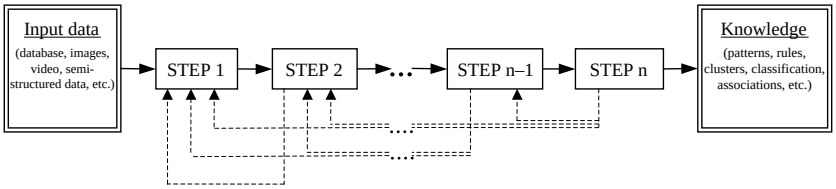

Figure 2.1. Sequential structure of the KDP model.

2.1. Overview of the Knowledge Discovery Process

The KDP model consists of a set of processing steps to be followed by practitioners when executing a knowledge discovery project. The model describes procedures that are performed in each of its steps. It is primarily used to plan, work through, and reduce the cost of any given project.

Since the 1990s, several different KDPs have been developed. The initial efforts were led by academic research but were quickly followed by industry. The first basic structure of the model was proposed by Fayyad et al. and later improved/modified by others. The process consists of multiple steps, that are executed in a sequence. Each subsequent step is initiated upon successful completion of the previous step, and requires the result generated by the previous step as its input. Another common feature of the proposed models is the range of activities covered, which stretches 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 the proposed models also emphasize the iterative nature of the model, in terms of many feedback loops that are triggered by a revision process. A schematic diagram is shown in Figure 2.1.

The main differences between the models described here lie in the number and scope of their specific steps. A common feature of all models is the definition of inputs and outputs. Typical inputs include data in various formats, such as numerical and nominal data stored in databases or flat files; images; video; semi-structured data, such as XML or HTML; etc. The output is the generated new knowledge — usually described in terms of rules, patterns, classification models, associations, trends, statistical analysis, etc.

3. Knowledge Discovery Process Models

Although the models usually emphasize independence from specific applications and tools, they can be broadly divided into those that take into account industrial issues and those that do not. However, the academic models, which usually are not concerned with industrial issues, can be made applicable relatively easily in the industrial setting and vice versa. We restrict our discussion to those models that have been popularized in the literature and have been used in real knowledge discovery projects.

3.1. Academic Research Models

The efforts to establish a KDP model were initiated in academia. In the mid-1990s, when the DM field was being shaped, researchers started defining multistep procedures to guide users of DM tools in the complex knowledge discovery world. The main emphasis was to provide a sequence of activities that would help to execute a KDP in an arbitrary domain. The two process models developed in 1996 and 1998 are the nine-step model by Fayyad et al. and the eight-step model by Anand and Buchner. Below we introduce the first of these, which is perceived as the leading research model. The second model is summarized in Sect. 2.3.4.

The Fayyad et al. KDP model consists of nine steps, which are outlined as follows:

1. *Developing and understanding the application domain*. This step includes learning the relevant prior knowledge and the goals of the end user of the discovered knowledge.
2. *Creating a target data set*. Here the data miner selects a subset of variables (attributes) and data points (examples) that will be used to perform discovery tasks. This step usually includes querying the existing data to select the desired subset.
3. *Data cleaning and preprocessing*. This step consists of removing outliers, dealing with noise and missing values in the data, and accounting for time sequence information and known changes.
4. *Data reduction and projection*. This step consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data.
5. *Choosing the data mining t*ask. Here the data miner matches the goals defined in Step 1 with a particular DM method, such as classification, regression, clustering, etc.
6. *Choosing the data mining al*gorithm. The data miner selects methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate.
7. *Data mining*. This step generates patterns in a particular representational form, such as classification rules, decision trees, regression models, trends, etc.
8. *Interpreting mined patterns*. Here the analyst performs visualization of the extracted patterns and models, and visualization of the data based on the extracted models.
9. *Consolidating discovered knowledge*. The final step consists of incorporating the discovered knowledge into the performance system, and documenting and reporting it to the interested parties. This step may also include checking and resolving potential conflicts with previously believed knowledge.

Notes: This process is iterative. The authors of this model declare that a number of loops between any two steps are usually executed, but they give no specific details. The model provides a detailed technical description with respect to data analysis but lacks a description of business aspects. This model has become a cornerstone of later models.

Major Applications: The nine-step model has been incorporated into a commercial knowledge discovery system called MineSet™ (for details, see Purple Insight Ltd. at http://www.purpleinsight.com). The model has been used in a number of different domains, including engineering, medicine, production, e-business, and software development.

3.2. Industrial Models

Industrial models quickly followed academic efforts. Several different approaches were undertaken, ranging from models proposed by individuals with extensive industrial experience to models proposed by large industrial consortiums. Two representative industrial models are the five-step model by Cabena et al., with support from IBM (see Sect. 2.3.4) and the industrial six-step CRISP-DM model, developed by a large consortium of European companies. The latter has become the leading industrial model, and is described in detail next.

The CRISP-DM (CRoss-Industry Standard Process for Data Mining) was first established in the late 1990s by four companies: Integral Solutions Ltd. (a provider of commercial data mining solutions), NCR (a database provider), DaimlerChrysler (an automobile manufacturer), and OHRA (an insurance company). The last two companies served as data and case study sources.

The development of this process model enjoys strong industrial support. It has also been supported by the ESPRIT program funded by the European Commission. The CRISP-DM Special Interest Group was created with the goal of supporting the developed process model. Currently, it includes over 300 users and tool and service providers.

The CRISP-DM KDP model (see Figure 2.2) consists of six steps, which are summarized below:

1. *Business understanding*. This step focuses on the understanding of objectives and requirements from a business perspective. It also converts these into a DM problem definition, and designs a preliminary project plan to achieve the objectives. It is further broken into several substeps, namely,

   – determination of business objectives,
   – assessment of the situation,
   – determination of DM goals, and
   – generation of a project plan.

2. *Data understanding*. This step starts with initial data collection and familiarization with the data. Specific aims include identification of data quality problems, initial insights into the data, and detection of interesting data subsets. Data understanding is further broken down into

   – collection of initial data,
   – description of data,
   – exploration of data, and
   – verification of data quality.

3. *Data preparation*. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into

   – selection of data,
   – cleansing of data,

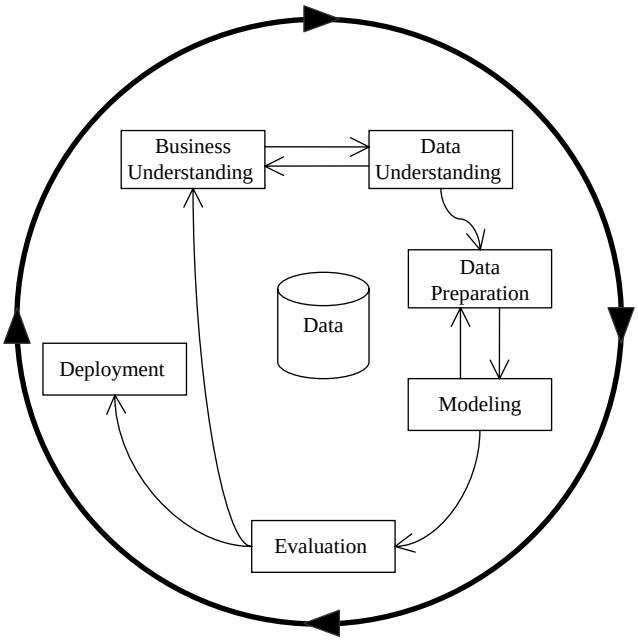

Figure 2.2. The CRISP-DM KD process model (source: http://www.crisp-dm.org/).

– construction of data,
– integration of data, and
– formatting of data substeps.

4. *Modeling*. At this point, various modeling techniques are selected and applied. Modeling usually involves the use of several methods for the same DM problem type and the calibration of their parameters to optimal values. Since some methods may require a specific format for input data, often reiteration into the previous step is necessary. This step is subdivided into

   – selection of modeling technique(s),
   – generation of test design,
   – creation of models, and
   – assessment of generated models.

5. *Evaluation*. After one or more models have been built that have high quality from a data analysis perspective, the model is evaluated from a business objective perspective. A review of the steps executed to construct the model is also performed. A key objective is to determine whether any important business issues have not been sufficiently considered. At the end of this phase, a decision about the use of the DM results should be reached. The key substeps in this step include

   – evaluation of the results,
   – process review, and
   – determination of the next step.

6. *Deployment*. Now the discovered knowledge must be organized and presented in a way that the customer can use. Depending on the requirements, this step can be as simple as generating a report or as complex as implementing a repeatable KDP. This step is further divided into

   – plan deployment,
   – plan monitoring and maintenance,
   – generation of final report, and
   – review of the process substeps.

Notes: The model is characterized by an easy-to-understand vocabulary and good documentation. It divides all steps into substeps that provide all necessary details. It also acknowledges the strong iterative nature of the process, with loops between several of the steps. In general, it is a very successful and extensively applied model, mainly due to its grounding in practical, industrial, real-world knowledge discovery experience.

Major Applications: The CRISP-DM model has been used in domains such as medicine, engineering, marketing, and sales. It has also been incorporated into a commercial knowledge discovery system called Clementine® (see SPSS Inc. at http://www.spss.com/clementine).

3.3. Hybrid Models

The development of academic and industrial models has led to the development of hybrid models, i.e., models that combine aspects of both. One such model is a six-step KDP model (see Figure 2.3) developed by Cios et al. It was developed based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include

– providing more general, research-oriented description of the steps,
– introducing a data mining step instead of the modeling step,

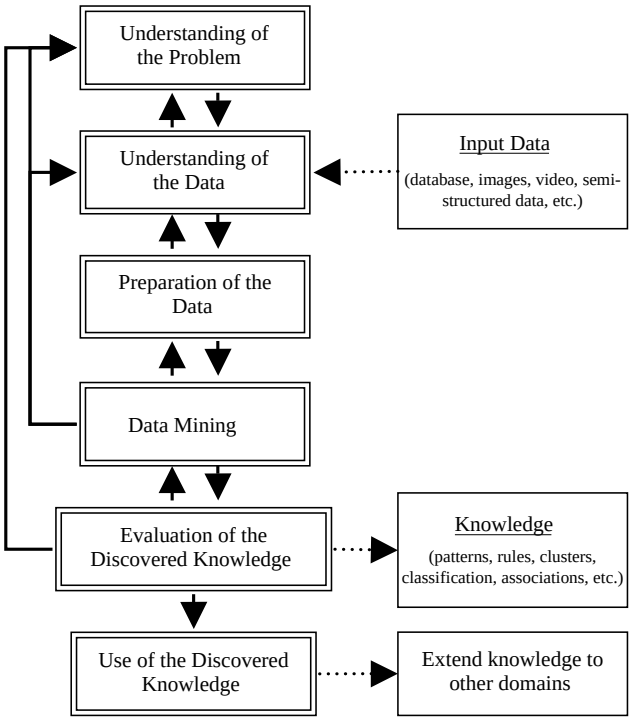

Figure 2.3. The six-step KDP model. Source: Pal, N.R., Jain, L.C., (Eds.) 2005. Advanced Techniques in Knowledge Discovery and Data Mining, Springer Verlag.

– introducing several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and
– modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains.

A description of the six steps follows

1. *Understanding of the problem domain*. This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology. A description of the problem, including its restrictions, is prepared. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.
2. *Understanding of the data*. This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.
3. *Preparation of the data*. This step concerns deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say, by discretization), and by summarization of data (data granularization). The end results are data that meet the specific input requirements for the DM tools selected in Step 1.

4. *Data mining*. Here the data miner uses various DM methods to derive knowledge from preprocessed data.
5. *Evaluation of the discovered knowledge*. Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared.
6. *Use of the discovered knowledge*. This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed.

Notes: The model emphasizes the iterative aspects of the process, drawing from the experience of users of previous models. It identifies and describes several explicit feedback loops:

– from *understanding of the data* to *understanding of the problem domain*. This loop is based on the need for additional domain knowledge to better understand the data.
– from *preparation of the data* to *understanding of the data*. This loop is caused by the need for additional or more specific information about the data in order to guide the choice of specific data preprocessing algorithms.
– from *data mining* to *understanding of the problem domain*. The reason for this loop could be unsatisfactory results generated by the selected DM methods, requiring modification of the project's goals.
– from *data mining* to *understanding of the data*. The most common reason for this loop is poor understanding of the data, which results in incorrect selection of a DM method and its subsequent failure, e.g., data were misrecognized as continuous and discretized in the *understanding of the data* step.
– from *data mining* to the *preparation of the data*. This loop is caused by the need to improve data preparation, which often results from the specific requirements of the DM method used, since these requirements may not have been known during the *preparation of the data* step.
– from *evaluation of the discovered knowledge* to the *understanding of the problem domain*. The most common cause for this loop is invalidity of the discovered knowledge. Several possible reasons include incorrect understanding or interpretation of the domain and incorrect design or understanding of problem restrictions, requirements, or goals. In these cases, the entire KD process must be repeated.
– from *evaluation of the discovered knowledge* to *data mining*. 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.

Awareness of the above common mistakes may help the user to avoid them by deploying some countermeasures.

Major Applications: The hybrid model has been used in medicine and software development areas. Example applications include development of computerized diagnostic systems for cardiac SPECT images and a grid data mining framework called GridMiner-Core. It has also been applied to analysis of data concerning intensive care, cystic fibrosis, and image-based classification of cells.

3.4. Comparison of the Models

To understand and interpret the KDP models described above, a direct, side-by-side comparison is shown in Table 2.1. It includes information about the domain of origin (academic or industry), the number of steps, a comparison of steps between the models, notes, and application domains.

Table 2.1. Comparison of the five KDP models. The double-lines group the corresponding steps.

Model	Fayyad et al.	Anand & Buchner	Cios et al.	Cabena et al.	CRISP-DM
Domain of origin	Academic	Academic	Hybrid academic/industry	Industry	Industry
# steps	9	8	6	5	6
Steps	1. Developing and Understanding the Application Domain	1. Human Resource Identification 2. Problem Specification	1. Understanding of the Problem Domain	1. Business Objectives Determination	1. Business Understanding
	2. Creating a Target Data Set	3. Data Prospecting 4. Domain Knowledge Elicitation	2. Understanding of the Data	2. Data Preparation	2. Data Understanding
	3. Data Cleaning and Preprocessing 4. Data Reduction and Projection 5. Choosing the Data Mining Task	6. Data Preprocessing	3. Preparation of the Data		3. Data Preparation
	6. Choosing the Data Mining Algorithm	5. Methodology Identification			
	7. Data Mining	7. Pattern Discovery	4. Data Mining	3. Data Mining	4. Modeling
	8. Interpreting Mined Patterns	8. Knowledge Post-Processing	5. Evaluation of the Discovered Knowledge	4. Analysis of Results	5. Evaluation

(Continued)

Table 2.1. *(Continued)*

Model	Fayyad et al.	Anand & Buchner	Cios et al.	Cabena et al.	CRISP-DM
	9. Consolidating Discovered Knowledge		6. Use of the Discovered Knowledge	5. Assimilation of Knowledge	6. Deployment
Notes	The most popular and most cited model; provides detailed technical description with respect to data analysis, but lacks business aspects	Provides detailed breakdown of the initial steps of the process; missing step concerned with application of the discovered knowledge and project documentation	Draws from both academic and industrial models and emphasizes iterative aspects; identifies and describes several explicit feedback loops	Business oriented and easy to comprehend by non-data-mining specialists; the model definition uses non-DM jargon	Uses easy-to-understand vocabulary; has good documentation; divides all steps into substeps that provide all necessary details
Supporting software	Commercial system MineSet™	N/A	N/A	N/A	Commercial system Clementine®
Reported application domains	Medicine, engineering, production, e-business, software	Marketing, sales	Medicine, software	Marketing, sales	Medicine, engineering, marketing, sales

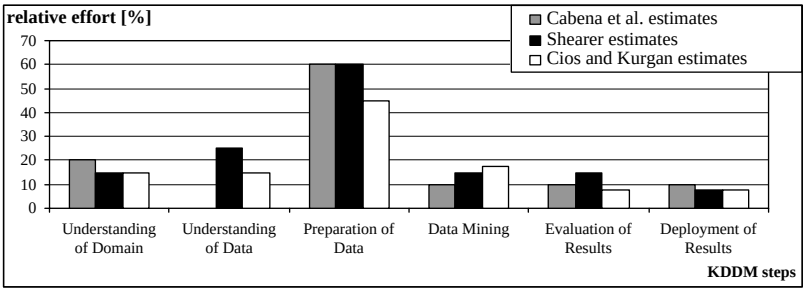

Figure 2.4. Relative effort spent on specific steps of the KD process. Source: Pal. N.R., Jain, L.C., (Eds.) 2005. Advanced Techniques in Knowledge Discovery and Data Mining, Springer Verlag.

Most models follow a similar sequence of steps, while the common steps between the five are domain understanding, data mining, and evaluation of the discovered knowledge. The nine-step model carries out the steps concerning the choice of DM tasks and algorithms late in the process. The other models do so before preprocessing of the data in order to obtain data that are correctly prepared for the DM step without having to repeat some of the earlier steps. In the case of Fayyad's model, the 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. The five-step model is very similar to the six-step models, except that it omits the data understanding step. The eight-step model gives a very detailed breakdown of steps in the early phases of the KDP, but it does not allow for a step concerned with applying the discovered knowledge. At the same time, it recognizes the important issue of human resource identification. This consideration is very important for any KDP, and we suggest that this step should be performed in all models.

We emphasize that there is no universally "best" KDP model. Each of the models has its strong and weak points based on the application domain and particular objectives. Further reading can be found in the Summary and Bibliographical Notes (Sect. 5).

A very important aspect of the KDP is the relative time spent in completing each of the steps. Evaluation of this effort enables precise scheduling. Several estimates have been proposed by researchers and practitioners alike. Figure 2.4 shows a comparison of these different estimates. We note that the numbers given are only estimates, which are used to quantify relative effort, and their sum may not equal 100%. The specific estimated values depend on many factors, such as existing knowledge about the considered project domain, the skill level of human resources, and the complexity of the problem at hand, to name just a few.

The common theme of all estimates is an acknowledgment that the data preparation step is by far the most time-consuming part of the KDP.

4. Research Issues

The ultimate goal of the KDP model is to achieve overall integration of the entire process through the use of industrial standards. Another important objective is to provide interoperability and compatibility between the different software systems and platforms used throughout the process. Integrated and interoperable models would serve the end user in automating, or more realistically semiautomating, work with knowledge discovery systems.

4.1. Metadata and the Knowledge Discovery Process

Our goal is to enable users to perform a KDP without possessing extensive background knowledge, without manual data manipulation, and without manual procedures to exchange data

and knowledge between different DM methods. This outcome requires the ability to store and exchange not only the data but also, most importantly, knowledge that is expressed in terms of data models, and meta-data that describes data and domain knowledge used in the process.

One of the technologies that can be used in achieving these goals is XML (eXtensible Markup Language), a standard proposed by the World Wide Web Consortium. XML allows the user to describe and store structured or semistructured data and to exchange data in a platform- and tool-independent way. From the KD perspective, XML helps to implement and standardize communication between diverse KD and database systems, to build standard data repositories for sharing data between different KD systems that work on different software platforms, and to provide a framework to integrate the entire KD process.

While XML by itself helps to solve some problems, metadata standards based on XML may provide a complete solution. Several such standards, such as PMML (Predictive Model Markup Language), have been identified that allow interoperability among different mining tools and that achieve integration with other applications, including database systems, spreadsheets, and decision support systems.

Both XML and PMML can be easily stored in most current database management systems. PMML, which is an XML-based language designed by the Data Mining Group, is used to describe data models (generated knowledge) and to share them between compliant applications. The Data Mining Group is an independent, vendor-led group that develops data mining standards. Its members include IBM, KXEN, Magnify Inc., Microsoft, MicroStrategy Inc., National Center for DM, Oracle, Prudential Systems Software GmbH, Salford Systems, SAS Inc., SPSS Inc., StatSoft Inc., and other companies (see http://www.dmg.org/). By using such a language, users can generate data models with one application, use another application to analyze these models, still another to evaluate them, and finally yet another to visualize the model. A PMML excerpt is shown in Figure 2.5.

XML and PMML standards can be used to integrate the KDP model in the following way. Information collected during the domain and data understanding steps can be stored as XML documents. These documents can be then used in the steps of data understanding, data preparation, and knowledge evaluation as a source of information that can be accessed automatically, across platforms, and across tools. In addition, knowledge extracted in the DM step and verified in the evaluation step, along with domain knowledge gathered in the domain understanding step, can be stored using PMML documents, which can then be stored and exchanged between different software tools. A sample architecture is shown in Figure 2.6.

5. Summary and Bibliographical Notes

In this Chapter we introduced the knowledge discovery process. The most important topics discussed are the following:

– **Knowledge discovery** is a **process** that seeks new knowledge about an application domain. It consists of many steps, one of which is data mining (DM), each aiming to complete a particular discovery task, and accomplished by the application of a discovery method.
– The KDP consists of **multiple steps** that are executed in a **sequence**. The subsequent step is initiated upon successful completion of the previous step and requires results generated by the previous step as its inputs.
– The KDP ranges from the task of understanding the project domain and data, through data preparation and analysis, to evaluation, understanding and application of the generated knowledge. It is highly **iterative**, and includes many feedback loops and repetitions, which are triggered by revision processes.

– The main reason for introducing **process models** is to formalize knowledge discovery projects within a common framework, a goal that will result in cost and time savings, and will improve understanding, success rates, and acceptance of such projects. The models emphasize **independence** from specific applications, tools, and vendors.
– Five KDP models, including the **nine-step model by Fayyad** et al., the **eight-step model by Anand and Buchner**, the **six-step model by Cios** et al., the **five-step model by Cabena** et al., and the **CRISP-DM model** were introduced. Each model has its strong and weak points, based on its application domain and particular business objectives.
– A very important consideration in the KDP is the relative time spent to complete each step. In general, we acknowledge that the **data preparation step is by far the most time-consuming part of the KDP**.
– The future of KDP models lies in achieving overall **integration** of the entire process through the use of popular industrial standards, such as XML and PMML.

The evolution of knowledge discovery systems has already undergone three distinct phases [16]:

– The **first-generation systems** provided only one data mining technique, such as a decision tree algorithm or a clustering algorithm, with very weak support for the overall process framework [11, 15, 18, 20, 21]. They were intended for expert users who already had an understanding of data mining techniques, the underlying data, and the knowledge being sought. Little attention was paid to providing support for the data analyst, and thus the first knowledge discovery systems had very limited commercial success [3]. The general research trend focused on the

```
<?xml version“1.0” encoding=“windows-1252”?>
<PMML version=“2.0”>
<DataDictionary numberOfFields=“4”>
   <DataField name=“PETALLEN” optype=“continuous” x-significance=“0.89”/>
   <DataField name=“PETALWID” optype=“continuous” x-significance=“0.39”/>
   <DataField name=“SEPALWID” optype=“continuous” x-significance=“0.92”/>
   <DataField name=“SPECIES” optype=“categorical” x-significance=“0.94”/>
   <DataField name=“SEPALLEN” optype=“continuous”/>
</DataDictionary>
<RegressionModel modelName=“...”functionName=“regression”
algorithmName=“polynomialRegression” modelType=“stepwisePolynomialRegression”
targetFieldName=“SEPALLEN”>
<MiningSchema>
   <MiningField name=“PETALLEN” usageType=“active”/>
   <MiningField name=“PETALWID” usageType=“active”/>
   ...
</MiningSchema>
<RegressionTable intercept=“−45534.5912666858”>
   <NumericPredictor name=“PETALLEN” exponent=“1” coefficient=“8.87” mean=“37.58”/>
   <NumericPredictor name=“PETALLEN” exponent=“2” coefficient=“−0.42” mean=“1722”/>
   ...
</RegressionTable>
</RegressionModel>
<Extension>
   <X-modelQuality x-rSquared=“0.8878700000000001”/>
   ...
</Extension>
</PMML>
```

Figure 2.5. A PMML excerpt that expresses the polynomial regression model for the popular iris dataset generated by the DB2 Intelligent Miner for Data V8.1. Source: http://www.dmg.org/.

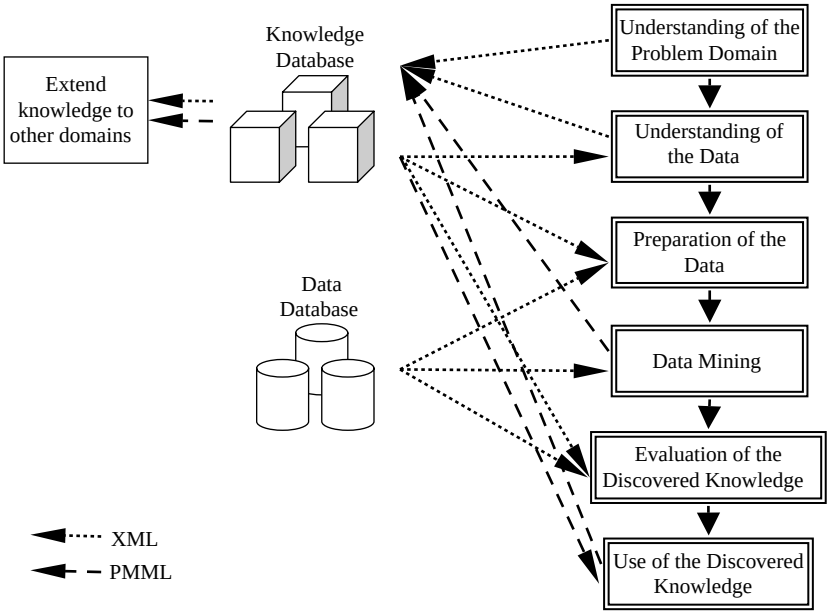

Figure 2.6. Application of PMML and XML standards in the framework of the KDP model.

development of new and improved data mining algorithms rather than on research to support other knowledge discovery activities.

– The **second-generation systems**, called *suites*, were developed in the mid-1990s. They provided multiple types of integrated data analysis methods, as well as support for data cleaning, preprocessing, and visualization. Examples include systems like SPSS's Clementine®, Silicon Graphics's MineSet™, IBM's Intelligent Miner, and SAS Institute's Enterprise Miner.
– The **third-generation systems** were developed in the late 1990s and introduced a vertical approach. These systems addressed specific business problems, such as fraud detection, and provided an interface designed to hide the internal complexity of data mining methods. Some of the suites also introduced knowledge discovery process models to guide the user's work. Examples include MineSet™, which uses the nine-step process model by Fayyad et al., and Clementine®, which uses the CRISP-DM process model.

The KDP model was first discussed during the inaugural workshop on Knowledge Discovery in Databases in 1989 [14]. The main driving factor in defining the model was acknowledgment of the fact that knowledge is the end product of a data-driven discovery process.

In 1996, the foundation for the process model was laid in a book entitled Advances in Knowledge Discovery and Data Mining [7]. The book presented a process model that had resulted from interactions between researchers and industrial data analysts. The model solved problems that were not connected with the details and use of particular data mining techniques but rather with providing support for the highly iterative and complex problem of overall knowledge generation process. The book also emphasized the close involvement of a human analyst in many, if not all, steps of the process [3].

The first KDP model was developed by Fayyad et al. [8 – 10]. Other KDP models discussed in this Chapter include those by Cabena et al. [4], Anand and Buchner [1, 2], Cios et al. [5, 6, 12], and the CRISP-DM model [17, 19]. A recent survey that includes a comprehensive comparison of several KDPs can be found in [13].

References

1. Anand, S., and Buchner, A. 1998. *Decision Support Using Data Mining*. Financial Times Pitman Publishers, London
2. Anand, S., Hughes, P., and Bell, D. 1998. A data mining methodology for cross-sales. *Knowledge Based Systems Journal*, 10:449–461
3. 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* 37–58, AAAI Press
4. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A. 1998. *Discovering Data Mining: From Concepts to Implementation*, Prentice Hall Saddle River, New Jersey
5. Cios, K., Teresinska, A., Konieczna, S., Potocka, J., and Sharma, S. 2000. Diagnosing myocardial perfusion from SPECT 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
6. Cios, K., and Kurgan, L. 2005. Trends in data mining and knowledge discovery. In Pal, N.R., and Jain L.C. (Eds.), Advanced Techniques in Knowledge Discovery and Data Mining, 1–26, Springer Verlag, London.
7. Fayyad, U., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.), 1996. *Advances in Knowledge Discovery and Data Mining*, AAAI Press, Cambridge
8. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. 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*, 1–34, AAAI Press, Cambridge
9. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34
10. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. Knowledge discovery and data mining: towards a unifying framework. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 82–88, Portland, Oregon
11. 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
12. 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*, 415–444, IEEE Press Piscataway, NJ
13. Kurgan, L., and Musilek, P. 2006. A survey of knowledge discovery and data mining process models. *Knowledge Engineering Review*, 21(1):1–24
14. Piatetsky-Shapiro, G. 1991. Knowledge discovery in real databases: a report on the IJCAI-89 workshop. *AI Magazine*, 11(5):68–70
15. Piatesky-Shapiro, G., and Matheus, C. 1992. Knowledge discovery workbench for exploring business databases. *International Journal of Intelligent Agents*, 7(7):675–686
16. Piatesky-Shapiro, G. 1999. The data mining industry coming to age. *IEEE Intelligent Systems*, 14(6): 32–33
17. Shearer, C. 2000. The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing*, 5(4):13–19
18. Simoudis, E., Livezey, B., and Kerber, R. 1994. Integrating inductive and deductive reasoning for data mining. *Proceedings of 1994 AAAI Workshop on Knowledge Discovery in Databases*, 37–48, Seattle, Washington, USA
19. Wirth, R., and Hipp, J. 2000. CRISP-DM: towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29–39, Manchester, UK
20. 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*, 89–101, Seattle, Washington
21. Zytow, J., and Baker, J. 1991. Interactive mining of regularities in databases. In Piatesky-Shapiro, G., and Frowley, W. (Eds.), *Knowledge Discovery in Databases*, 31–53, AAAI Press Cambridge

6. Exercises

1. Discuss why we need to standardize knowledge discovery process models.
2. Discuss the difference between terms *data mining* and *knowledge discovery process*. Which of these terms is broader?
3. Imagine that you are a chief data analyst responsible for deploying a knowledge discovery project related to mining data gathered by a major insurance company. The goal is to discover fraud patterns. The customer's data are stored in well-maintained data warehouse, and a team of data analysts who are familiar with the data are at your disposal. The management stresses the importance of analysis, documentation, and deployment of the developed solution(s). Which of the models presented in this Chapter would you choose to carry out the project and why? Also, provide a rationale as to why other models are less suitable in this case.
4. Provide a detailed description of the *Evaluation* and *Deployment* steps in the CRISP-DM process model. Your description should explain the details of the substeps in these two steps.
5. Compare side by side the six-step CRISP-DM and the eight-step model by Anand and Buchner. Discuss the main differences between the two models, and provide an example knowledge discovery application that is best suited for each of the models.
6. Find an industrial application for one of the models discussed in this Chapter. Provide details about the project that used the model, and discuss what benefits were achieved by deploying the model. (hint: see Hirji, K. 2001. Exploring data mining implementation. *Communications of the ACM*, 44(7), 87–93)
7. Provide a one-page summary of the PMML language standard. Your summary must include information about the newest release of the standard and which data mining models are supported by the standard.