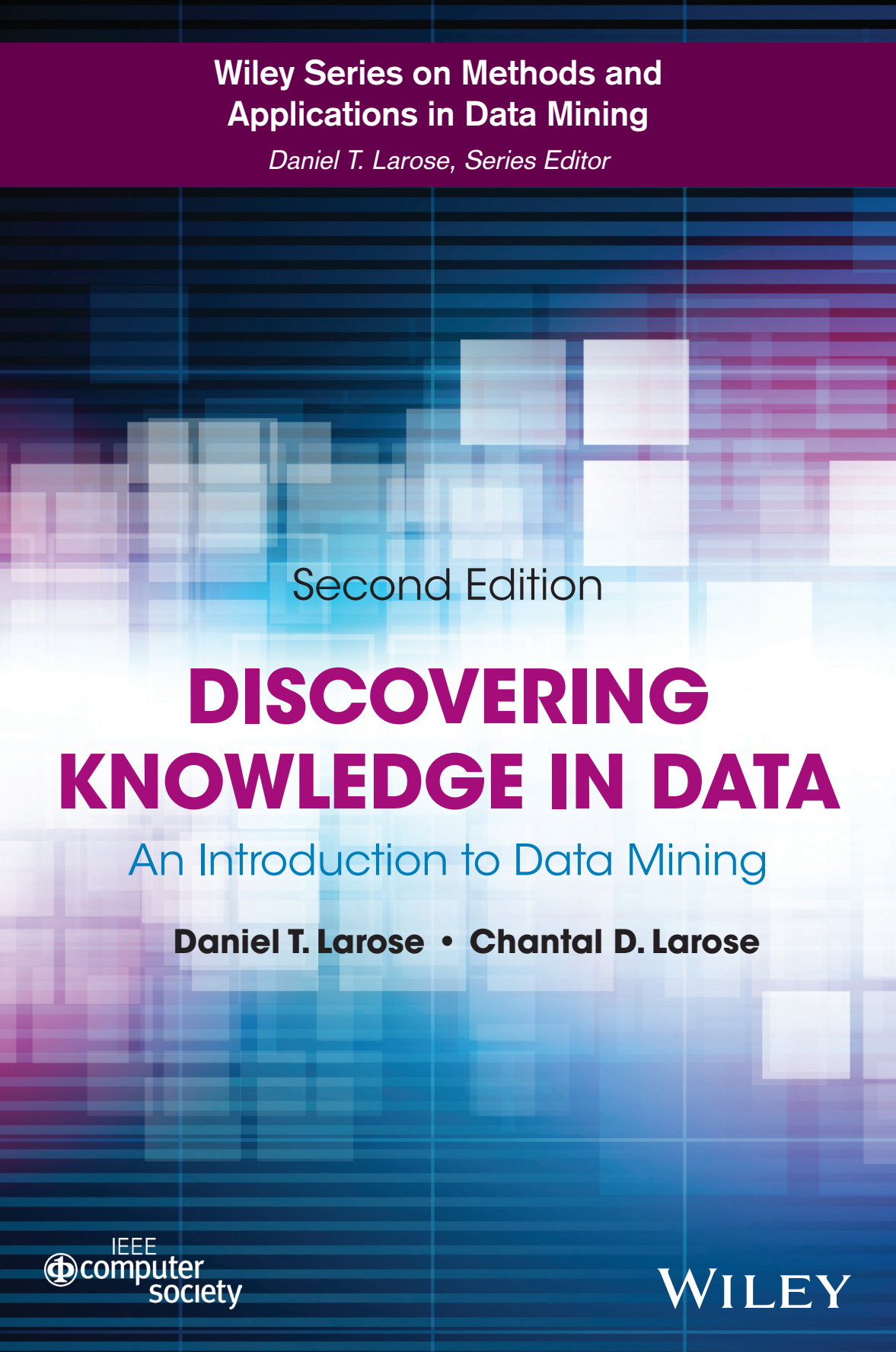
Wiley Series on Methods and Applications in Data Mining

Daniel T. Larose, Series Editor



Second Edition

DISCOVERING KNOWLEDGE IN DATA

An Introduction to Data Mining

Daniel T. Larose • Chantal D. Larose

IEEE computer society

WILEY

DISCOVERING KNOWLEDGE IN DATA

WILEY SERIES ON METHODS AND APPLICATIONS IN DATA MINING

Series Editor: **Daniel T. Larose**

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition • Daniel T. Larose and Chantal D. Larose

Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data • Darius M. Dziuda

Knowledge Discovery with Support Vector Machines • Lutz Hamel

Data-Mining on the Web: Uncovering Patterns in Web Content, Structure, and Usage • Zdravko Markov and Daniel Larose

Data Mining Methods and Models • Daniel Larose

Practical Text Mining with Perl • Roger Bilisoly

SECOND EDITION

DISCOVERING KNOWLEDGE IN DATA

An Introduction to Data Mining

DANIEL T. LAROSE
CHANTAL D. LAROSE

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Larose, Daniel T.
Discovering knowledge in data : an introduction to data mining / Daniel T. Larose and Chantal D. Larose. – Second edition.
pages cm
Includes index.
ISBN 978-0-470-90874-7 (hardback)
1. Data mining. I. Larose, Chantal D. II. Title.
QA76.9.D343L38 2014
006.3′12–dc23 2013046021

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

WHAT IS DATA MINING?

According to the Gartner Group,

> Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.

Today, there are a variety of terms used to describe this process, including *analytics*, *predictive analytics*, *big data*, *machine learning*, and *knowledge discovery in databases*. But these terms all share in common the objective of mining actionable nuggets of knowledge from large data sets. We shall therefore use the term *data mining* to represent this process throughout this text.

WHY IS THIS BOOK NEEDED?

Humans are inundated with data in most fields. Unfortunately, these valuable data, which cost firms millions to collect and collate, are languishing in warehouses and repositories. *The problem is that there are not enough trained human analysts available who are skilled at translating all of these data into knowledge*, and thence up the taxonomy tree into wisdom. This is why this book is needed.

The McKinsey Global Institute reports:[1]

> There will be a shortage of talent necessary for organizations to take advantage of big data. A significant constraint on realizing value from big data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning, and the managers and analysts who know how to operate companies by using insights from big data . . . . We project that demand for deep analytical positions in a big data world could exceed the supply being produced on current trends by 140,000 to 190,000 positions. . . . In addition, we project a need for 1.5 million additional managers and analysts in the United States who can ask the right questions and consume the results of the analysis of big data effectively.

This book is an attempt to help alleviate this critical shortage of data analysts. *Discovering Knowledge in Data: An Introduction to Data Mining* provides readers with:

- The models and techniques to uncover hidden nuggets of information,

[1] *Big data: The next frontier for innovation, competition, and productivity*, by James Manyika et al., Mckinsey Global Institute, www.mckinsey.com, May, 2011. Last accessed March 16, 2014.

- The insight into how the data mining algorithms really work, and
- The experience of actually performing data mining on large data sets.

Data mining is becoming more widespread everyday, because it empowers companies to uncover profitable patterns and trends from their existing databases. Companies and institutions have spent millions of dollars to collect megabytes and terabytes of data, but are not taking advantage of the valuable and actionable information hidden deep within their data repositories. However, as the practice of data mining becomes more widespread, companies which do not apply these techniques are in danger of falling behind, and losing market share, because their competitors are applying data mining, and thereby gaining the competitive edge.

In *Discovering Knowledge in Data*, the step-by-step, hands-on solutions of real-world business problems, using widely available data mining techniques applied to real-world data sets, will appeal to managers, CIOs, CEOs, CFOs, and others who need to keep abreast of the latest methods for enhancing return-on-investment.

WHAT'S NEW FOR THE SECOND EDITION?

The second edition of *Discovery Knowledge in Data* is enhanced with an abundance of new material and useful features, including:

- Nearly 100 pages of new material.
- Three new chapters:
  - Chapter 5: *Multivariate Statistical Analysis* covers the hypothesis tests used for verifying whether data partitions are valid, along with analysis of variance, multiple regression, and other topics.
  - Chapter 6: *Preparing to Model the Data* introduces a new formula for balancing the training data set, and examines the importance of establishing baseline performance, among other topics.
  - Chapter 13: *Imputation of Missing Data* addresses one of the most overlooked issues in data analysis, and shows how to impute missing values for continuous variables and for categorical variables, as well as how to handle patterns in missingness.
- *The R Zone*. In most chapters of this book, the reader will find *The R Zone*, which provides the actual R code needed to obtain the results shown in the chapter, along with screen shots of some of the output, using *R Studio*.
- A host of new topics not covered in the first edition. Here is a sample of these new topics, chapter by chapter:
  - Chapter 2: *Data Preprocessing*. Decimal scaling; Transformations to achieve normality; Flag variables; Transforming categorical variables into numerical variables; Binning numerical variables; Reclassifying categorical variables; Adding an index field; Removal of duplicate records.

- Chapter 3: *Exploratory Data Analysis*. Binning based on predictive value; Deriving new variables: Flag variables; Deriving new variables: Numerical variables; Using EDA to investigate correlated predictor variables.
- Chapter 4: *Univariate Statistical Analysis*. How to reduce the margin of error; Confidence interval estimation of the proportion; Hypothesis testing for the mean; Assessing the strength of evidence against the null hypothesis; Using confidence intervals to perform hypothesis tests; Hypothesis testing for the proportion.
- Chapter 5: *Multivariate Statistics*. Two-sample test for difference in means; Two-sample test for difference in proportions; Test for homogeneity of proportions; Chi-square test for goodness of fit of multinomial data; Analysis of variance; Hypothesis testing in regression; Measuring the quality of a regression model.
- Chapter 6: *Preparing to Model the Data*. Balancing the training data set; Establishing baseline performance.
- Chapter 7: *k-Nearest Neighbor Algorithm*. Application of *k*-nearest neighbor algorithm using IBM/SPSS Modeler.
- Chapter 10: *Hierarchical and k-Means Clustering*. Behavior of MSB, MSE, and pseudo-*F* as the *k*-means algorithm proceeds.
- Chapter 12: *Association Rules*. How can we measure the usefulness of association rules?
- Chapter 13: *Imputation of Missing Data*. Need for imputation of missing data; Imputation of missing data for continuous variables; Imputation of missing data for categorical variables; Handling patterns in missingness.
- Chapter 14: *Model Evaluation Techniques*. Sensitivity and Specificity.

- An *Appendix* on *Data Summarization and Visualization*. Readers who may be a bit rusty on introductory statistics may find this new feature helpful. Definitions and illustrative examples of introductory statistical concepts are provided here, along with many graphs and tables, as follows:
  - Part 1: *Summarization 1: Building Blocks of Data Analysis*
  - Part 2: *Visualization: Graphs and Tables for Summarizing and Organizing Data*
  - Part 3: *Summarization 2: Measures of Center, Variability, and Position*
  - Part 4: *Summarization and Visualization of Bivariate Relationships*
- New Exercises. There are over 100 new chapter exercises in the second edition.

DANGER! DATA MINING IS EASY TO DO BADLY

The plethora of new off-the-shelf software platforms for performing data mining has kindled a new kind of danger. The ease with which these graphical user interface (GUI)-based applications can manipulate data, combined with the power of the

formidable data mining algorithms embedded in the black box software currently available, makes their misuse proportionally more hazardous.

Just as with any new information technology, *data mining is easy to do badly*. A little knowledge is especially dangerous when it comes to applying powerful models based on large data sets. For example, analyses carried out on unpreprocessed data can lead to erroneous conclusions, or inappropriate analysis may be applied to data sets that call for a completely different approach, or models may be derived that are built upon wholly specious assumptions. These errors in analysis can lead to very expensive failures, if deployed.

"WHITE BOX" APPROACH: UNDERSTANDING THE UNDERLYING ALGORITHMIC AND MODEL STRUCTURES

The best way to avoid these costly errors, which stem from a blind black-box approach to data mining, is to instead apply a "white-box" methodology, which emphasizes an understanding of the algorithmic and statistical model structures underlying the software.

***Discovering Knowledge in Data* applies this white-box approach by:**

- Walking the reader through the various algorithms;
- Providing examples of the operation of the algorithm on actual large data sets;
- Testing the reader's level of understanding of the concepts and algorithms;
- Providing an opportunity for the reader to do some real data mining on large data sets; and
- Supplying the reader with the actual R code used to achieve these data mining results, in *The R Zone*.

Algorithm Walk-Throughs

Discovering Knowledge in Data walks the reader through the operations and nuances of the various algorithms, using small sample data sets, so that the reader gets a true appreciation of what is really going on inside the algorithm. For example, in Chapter 10, *Hierarchical and K-Means Clustering*, we see the updated cluster centers being updated, moving toward the center of their respective clusters. Also, in Chapter 11, *Kohonen Networks*, we see just which kind of network weights will result in a particular network node "winning" a particular record.

Applications of the Algorithms to Large Data Sets

Discovering Knowledge in Data provides examples of the application of the various algorithms on actual large data sets. For example, in Chapter 9, *Neural Networks*, a classification problem is attacked using a neural network model on a real-world data set. The resulting neural network topology is examined, along with the network connection weights, as reported by the software. These data sets are included on the

data disk, so that the reader may follow the analytical steps on their own, using data mining software of their choice.

Chapter Exercises: Check Your Understanding

Discovering Knowledge in Data includes over 260 chapter exercises, which allow readers to assess their depth of understanding of the material, as well as have a little fun playing with numbers and data. These include conceptual exercises, which help to clarify some of the more challenging concepts in data mining, and "Tiny data set" exercises, which challenge the reader to apply the particular data mining algorithm to a small data set, and, step-by-step, to arrive at a computationally sound solution. For example, in Chapter 8, *Decision Trees*, readers are provided with a small data set and asked to construct—by hand, using the methods shown in the chapter—a *C4.5* decision tree model, as well as a *classification and regression tree* model, and to compare the benefits and drawbacks of each.

Hands-On Analysis: Learn Data Mining by Doing Data Mining

Most chapters provide *hands-on analysis problems*, representing an opportunity for the reader to apply newly-acquired data mining expertise to solving real problems using large data sets. Many people learn by doing. This book provides a framework where the reader can learn data mining by doing data mining.

The intention is to mirror the real-world data mining scenario. In the real world, dirty data sets need to be cleaned; raw data needs to be normalized; outliers need to be checked. So it is with *Discovering Knowledge in Data*, where about 100 hands-on analysis problems are provided. The reader can "ramp up" quickly, and be "up and running" data mining analyses in a short time.

For example, in Chapter 12, *Association Rules*, readers are challenged to uncover high confidence, high support rules for predicting which customer will be leaving the company's service. In Chapter 14, *Model Evaluation Techniques*, readers are asked to produce lift charts and gains charts for a set of classification models using a large data set, so that the best model may be identified.

The R Zone

R is a powerful, open-source language for exploring and analyzing data sets (www.r-project.org). Analysts using *R* can take advantage of many freely available packages, routines, and GUIs, to tackle most data analysis problems. In most chapters of this book, the reader will find *The R Zone*, which provides the actual R code needed to obtain the results shown in the chapter, along with screen shots of some of the output. *The R Zone* is written by Chantal D. Larose (Ph.D. candidate in Statistics, University of Connecticut, Storrs), daughter of the author, and *R* expert, who uses *R* extensively in her research, including research on multiple imputation of missing data, with her dissertation advisors, Dr. Dipak Dey and Dr. Ofer Harel.

DATA MINING AS A PROCESS

One of the fallacies associated with data mining implementations is that data mining somehow represents an isolated set of tools, to be applied by an aloof analysis department, and marginally related to the mainstream business or research endeavor. Organizations which attempt to implement data mining in this way will see their chances of success much reduced. Data mining should be viewed as a *process*.

Discovering Knowledge in Data presents data mining as a well-structured *standard process*, intimately connected with managers, decision makers, and those involved in deploying the results. Thus, this book is not only for analysts, but for managers as well, who need to communicate in the language of data mining.

The standard process used is the *CRISP-DM* framework: the *Cross-Industry Standard Process for Data Mining*. *CRISP-DM* demands that data mining be seen as an entire process, from communication of the business problem, through data collection and management, data preprocessing, model building, model evaluation, and, finally, model deployment. Therefore, this book is not only for analysts and managers, but also for data management professionals, database analysts, and decision makers.

GRAPHICAL APPROACH, EMPHASIZING EXPLORATORY DATA ANALYSIS

Discovering Knowledge in Data emphasizes a graphical approach to data analysis. There are more than 170 screen shots of computer output throughout the text, and 40 other figures. Exploratory data analysis (EDA) represents an interesting and fun way to "feel your way" through large data sets. Using graphical and numerical summaries, the analyst gradually sheds light on the complex relationships hidden within the data. *Discovering Knowledge in Data* emphasizes an EDA approach to data mining, which goes hand-in-hand with the overall graphical approach.

HOW THE BOOK IS STRUCTURED

Discovering Knowledge in Data: An Introduction to Data Mining provides a comprehensive introduction to the field. Common myths about data mining are debunked, and common pitfalls are flagged, so that new data miners do not have to learn these lessons themselves. The first three chapters introduce and follow the *CRISP-DM* standard process, especially the data preparation phase and data understanding phase. The next nine chapters represent the heart of the book, and are associated with the *CRISP-DM* modeling phase. Each chapter presents data mining methods and techniques for a specific data mining task.

- Chapters 4 and 5 examine univariate and multivariate statistical analyses, respectively, and exemplify the *estimation* and *prediction* tasks, for example, using multiple regression.

- Chapters 7–9 relate to the *classification* task, examining *k*-nearest neighbor (Chapter 7), decision trees (Chapter 8), and neural network (Chapter 9) algorithms.
- Chapters 10 and 11 investigate the *clustering* task, with hierarchical and *k*-means clustering (Chapter 10) and Kohonen networks (Chapter 11) algorithms.
- Chapter 12 handles the *association* task, examining association rules through the *a priori* and *GRI* algorithms.
- Finally, Chapter 14 considers model evaluation techniques, which belong to the *CRISP-DM* evaluation phase.

Discovering Knowledge in Data as a Textbook

Discovering Knowledge in Data: An Introduction to Data Mining naturally fits the role of textbook for an introductory course in data mining. Instructors may appreciate:

- The presentation of data mining as a *process*
- The "White-box" approach, emphasizing an understanding of the underlying algorithmic structures
  - Algorithm walk-throughs
  - Application of the algorithms to large data sets
  - Chapter exercises
  - Hands-on analysis, and
  - *The R Zone*
- The graphical approach, emphasizing exploratory data analysis, and
- The logical presentation, flowing naturally from the *CRISP-DM* standard process and the set of data mining tasks.

Discovering Knowledge in Data is appropriate for advanced undergraduate or graduate-level courses. Except for one section in the neural networks chapter, no calculus is required. An introductory statistics course would be nice, but is not required. No computer programming or database expertise is required.

ACKNOWLEDGMENTS

I first wish to thank my mentor Dr. Dipak K. Dey, Distinguished Professor of Statistics, and Associate Dean of the College of Liberal Arts and Sciences at the University of Connecticut, as well as Dr. John Judge, Professor of Statistics in the Department of Mathematics at Westfield State College. My debt to the two of you is boundless, and now extends beyond one lifetime. Also, I wish to thank my colleagues in the data mining programs at Central Connecticut State University: Dr. Chun Jin, Dr. Daniel S. Miller, Dr. Roger Bilisoly, Dr. Darius Dziuda, and Dr. Krishna Saha. Thanks to my daughter Chantal Danielle Larose, for her important contribution to this book, as well as for her cheerful affection and gentle insanity. Thanks to my twin children

Tristan Spring and Ravel Renaissance for providing perspective on what life is really about. Finally, I would like to thank my wonderful wife, Debra J. Larose, for our life together.

Daniel T. Larose, Ph.D.

Professor of Statistics and Data Mining
Director, Data Mining@CCSU
www.math.ccsu.edu/larose

I would first like to thank my PhD advisors, Dr. Dipak Dey, Distinguished Professor and Associate Dean, and Dr. Ofer Harel, Associate Professor, both from the Department of Statistics at the University of Connecticut. Their insight and understanding have framed and sculpted our exciting research program, including my PhD dissertation, Model-Based Clustering of Incomplete Data. Thanks also to my father Daniel for kindling my enduring love of data analysis, and to my mother Debra for her care and patience through many statistics-filled conversations. Finally thanks to my siblings, Ravel and Tristan, for perspective, music, and friendship.

Chantal D. Larose, MS

Department of Statistics
University of Connecticut

> Let us settle ourselves, and work, and wedge our feet downwards through the mud and slush of opinion and tradition and prejudice and appearance and delusion, . . . till we come to a hard bottom with rocks in place which we can call *reality* and say, "This is, and no mistake."
>
> Henry David Thoreau

CHAPTER 1

AN INTRODUCTION TO DATA MINING

1.1 WHAT IS DATA MINING?

The McKinsey Global Institute (MGI) reports [1] that most American companies with more than 1000 employees had an average of at least 200 terabytes of stored data. MGI projects that the amount of data generated worldwide will increase by 40% annually, creating profitable opportunities for companies to leverage their data to reduce costs and increase their bottom line. For example, retailers harnessing this "big data" to best advantage could expect to realize an increase in their operating margin of more than 60%, according to the MGI report. And healthcare providers and health maintenance organizations (HMOs) that properly leverage their data storehouses could achieve $300 in cost savings annually, through improved efficiency and quality.

The MIT Technology Review reports [2] that it was the Obama campaign's effective use of data mining that helped President Obama win the 2012 presidential election over Mitt Romney. They first identified likely Obama voters using a data mining model, and then made sure that these voters actually got to the polls. The campaign also used a separate data mining model to predict the polling outcomes

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition.
By Daniel T. Larose and Chantal D. Larose.

county-by-county. In the important swing county of Hamilton County, Ohio, the model predicted that Obama would receive 56.4% of the vote; the Obama share of the actual vote was 56.6%, so that the prediction was off by only 0.02%. Such precise predictive power allowed the campaign staff to allocate scarce resources more efficiently.

About 13 million customers per month contact the West Coast customer service call center of the Bank of America, as reported by *CIO Magazine* [3]. In the past, each caller would have listened to the same marketing advertisement, whether or not it was relevant to the caller's interests. However, "rather than pitch the product of the week, we want to be as relevant as possible to each customer," states Chris Kelly, vice president and director of database marketing at Bank of America in San Francisco. Thus Bank of America's customer service representatives have access to individual customer profiles, so that the customer can be informed of new products or services that may be of greatest interest to him or her. This is an example of mining customer data to help identify the type of marketing approach for a particular customer, based on customer's individual profile.

So, what is data mining?

Data mining is the process of discovering useful patterns and trends in large data sets.

While waiting in line at a large supermarket, have you ever just closed your eyes and listened? You might hear the beep, beep, beep, of the supermarket scanners, reading the bar codes on the grocery items, ringing up on the register, and storing the data on company servers. Each beep indicates a new row in the database, a new "observation" in the information being collected about the shopping habits of your family, and the other families who are checking out.

Clearly, a lot of data is being collected. However, what is being learned from all this data? What knowledge are we gaining from all this information? Probably not as much as you might think, because there is a serious shortage of skilled data analysts.

1.2 WANTED: DATA MINERS

As early as 1984, in his book *Megatrends* [4], John Naisbitt observed that "We are drowning in information but starved for knowledge." The problem today is not that there is not enough data and information streaming in. We are in fact inundated with data in most fields. Rather, the problem is that there are not enough trained *human* analysts available who are skilled at translating all of these data into knowledge, and thence up the taxonomy tree into wisdom.

The ongoing remarkable growth in the field of data mining and knowledge discovery has been fueled by a fortunate confluence of a variety of factors:

- The explosive growth in data collection, as exemplified by the supermarket scanners above,

- The storing of the data in data warehouses, so that the entire enterprise has access to a reliable, current database,
- The availability of increased access to data from web navigation and intranets,
- The competitive pressure to increase market share in a globalized economy,
- The development of "off-the-shelf" commercial data mining software suites,
- The tremendous growth in computing power and storage capacity.

Unfortunately, according to the McKinsey report [1],

> There will be a shortage of talent necessary for organizations to take advantage of big data. A significant constraint on realizing value from big data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning, and the managers and analysts who know how to operate companies by using insights from big data .... We project that demand for deep analytical positions in a big data world could exceed the supply being produced on current trends by 140,000 to 190,000 positions.... In addition, we project a need for 1.5 million additional managers and analysts in the United States who can ask the right questions and consume the results of the analysis of big data effectively.

This book is an attempt to help alleviate this critical shortage of data analysts.

1.3 THE NEED FOR HUMAN DIRECTION OF DATA MINING

Many software vendors market their analytical software as being a plug-and-play, out-of-the-box application that will provide solutions to otherwise intractable problems, without the need for human supervision or interaction. Some early definitions of data mining followed this focus on automation. For example, Berry and Linoff, in their book *Data Mining Techniques for Marketing, Sales and Customer Support* [5] gave the following definition for data mining: "Data mining is the process of exploration and analysis, *by automatic or semi-automatic means*, of large quantities of data in order to discover meaningful patterns and rules" [emphasis added]. Three years later, in their sequel *Mastering Data Mining* [6], the authors revisit their definition of data mining, and mention that, "If there is anything we regret, it is the phrase 'by automatic or semi-automatic means' ... because we feel there has come to be too much focus on the automatic techniques and not enough on the exploration and analysis. This has misled many people into believing that data mining is a product that can be bought rather than a discipline that must be mastered."

Very well stated! Automation is no substitute for human input. Humans need to be actively involved at every phase of the data mining process. Rather than asking where humans fit into data mining, we should instead inquire about how we may design data mining into the very human process of problem solving.

Further, the very power of the formidable data mining algorithms embedded in the black box software currently available makes their misuse proportionally more dangerous. Just as with any new information technology, *data mining is easy to do badly*. Researchers may apply inappropriate analysis to data sets that call for a

completely different approach, for example, or models may be derived that are built upon wholly specious assumptions. Therefore, an understanding of the statistical and mathematical model structures underlying the software is required.

1.4 THE CROSS-INDUSTRY STANDARD PRACTICE FOR DATA MINING

There is a temptation in some companies, due to departmental inertia and compartmentalization, to approach data mining haphazardly, to re-invent the wheel and duplicate effort. A cross-industry standard was clearly required, that is industry-neutral, tool-neutral, and application-neutral. The Cross-Industry Standard Process for Data Mining (CRISP-DM) [7] was developed by analysts representing Daimler-Chrysler, SPSS, and NCR. CRISP provides a nonproprietary and freely available standard process for fitting data mining into the general problem solving strategy of a business or research unit.

According to CRISP-DM, a given data mining project has a life cycle consisting of six phases, as illustrated in Figure 1.1. Note that the phase-sequence is *adaptive*.

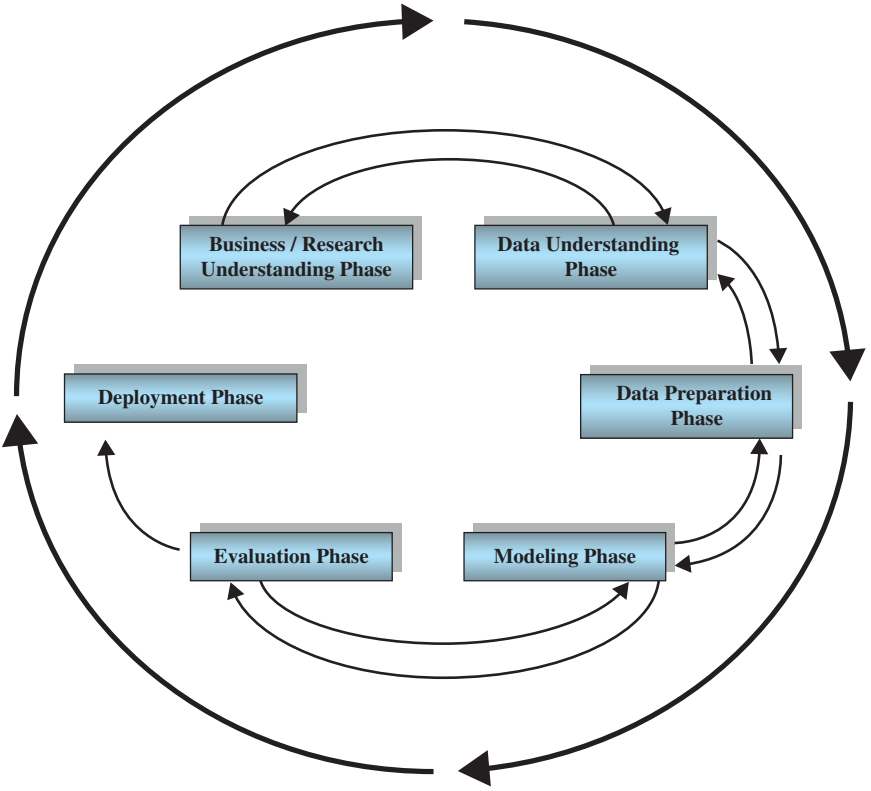


Figure 1.1 CRISP-DM is an iterative, adaptive process.

That is, the next phase in the sequence often depends on the outcomes associated with the previous phase. The most significant dependencies between phases are indicated by the arrows. For example, suppose we are in the modeling phase. Depending on the behavior and characteristics of the model, we may have to return to the data preparation phase for further refinement before moving forward to the model evaluation phase.

The iterative nature of CRISP is symbolized by the outer circle in Figure 1.1. Often, the solution to a particular business or research problem leads to further questions of interest, which may then be attacked using the same general process as before. Lessons learned from past projects should always be brought to bear as input into new projects. Here is an outline of each phase.

Issues encountered during the evaluation phase can conceivably send the analyst back to any of the previous phases for amelioration.

1.4.1 Crisp-DM: The Six Phases

1. **Business/Research Understanding Phase**
   - **a.** First, clearly enunciate the project objectives and requirements in terms of the business or research unit as a whole.
   - **b.** Then, translate these goals and restrictions into the formulation of a data mining problem definition.
   - **c.** Finally, prepare a preliminary strategy for achieving these objectives.
2. **Data Understanding Phase**
   - **a.** First, collect the data.
   - **b.** Then, use exploratory data analysis to familiarize yourself with the data, and discover initial insights.
   - **c.** Evaluate the quality of the data.
   - **d.** Finally, if desired, select interesting subsets that may contain actionable patterns.
3. **Data Preparation Phase**
   - **a.** This labor-intensive phase covers all aspects of preparing the final data set, which shall be used for subsequent phases, from the initial, raw, dirty data.
   - **b.** Select the cases and variables you want to analyze, and that are appropriate for your analysis.
   - **c.** Perform transformations on certain variables, if needed.
   - **d.** Clean the raw data so that it is ready for the modeling tools.
4. **Modeling Phase**
   - **a.** Select and apply appropriate modeling techniques.
   - **b.** Calibrate model settings to optimize results.

   c. Often, several different techniques may be applied for the same data mining problem.
   d. May require looping back to data preparation phase, in order to bring the form of the data into line with the specific requirements of a particular data mining technique.

5. **Evaluation Phase**
   a. The modeling phase has delivered one or more models. These models must be evaluated for quality and effectiveness, before we deploy them for use in the field.
   b. Also, determine whether the model in fact achieves the objectives set for it in Phase 1.
   c. Establish whether some important facet of the business or research problem has not been sufficiently accounted for.
   d. Finally, come to a decision regarding the use of the data mining results.
6. **Deployment Phase**
   a. Model creation does not signify the completion of the project. Need to make use of created models.
   b. Example of a simple deployment: Generate a report.
   c. Example of a more complex deployment: Implement a parallel data mining process in another department.
   d. For businesses, the customer often carries out the deployment based on your model.

This book broadly follows CRISP-DM, with some modifications. For example, we prefer to clean the data (Chapter 2) before performing exploratory data analysis (Chapter 3).

1.5 FALLACIES OF DATA MINING

Speaking before the US House of Representatives SubCommittee on Technology, Information Policy, Intergovernmental Relations, and Census, Jen Que Louie, President of Nautilus Systems, Inc. described four fallacies of data mining [8]. Two of these fallacies parallel the warnings we have described above.

- **Fallacy 1.** There are data mining tools that we can turn loose on our data repositories, and find answers to our problems.
  - *Reality.* There are no automatic data mining tools, which will mechanically solve your problems "while you wait." Rather data mining is a process. CRISP-DM is one method for fitting the data mining process into the overall business or research plan of action.

- **Fallacy 2.** The data mining process is autonomous, requiring little or no human oversight.
  - *Reality.* Data mining is not magic. Without skilled human supervision, blind use of data mining software will only provide you with the wrong answer to the wrong question applied to the wrong type of data. Further, the wrong analysis is worse than no analysis, since it leads to policy recommendations that will probably turn out to be expensive failures. Even after the model is deployed, the introduction of new data often requires an updating of the model. Continuous quality monitoring and other evaluative measures must be assessed, by human analysts.
- **Fallacy 3.** Data mining pays for itself quite quickly.
  - *Reality*. The return rates vary, depending on the start-up costs, analysis personnel costs, data warehousing preparation costs, and so on.
- **Fallacy 4.** Data mining software packages are intuitive and easy to use.
  - *Reality.* Again, ease of use varies. However, regardless of what some software vendor advertisements may claim, you cannot just purchase some data mining software, install it, sit back, and watch it solve all your problems. For example, the algorithms require specific data formats, which may require substantial preprocessing. Data analysts must combine subject matter knowledge with an analytical mind, and a familiarity with the overall business or research model.

To the above list, we add three further common fallacies:

- **Fallacy 5.** Data mining will identify the causes of our business or research problems.
  - *Reality.* The knowledge discovery process will help you to uncover patterns of behavior. Again, it is up to the humans to identify the causes.
- **Fallacy 6.** Data mining will automatically clean up our messy database.
  - *Reality.* Well, not automatically. As a preliminary phase in the data mining process, data preparation often deals with data that has not been examined or used in years. Therefore, organizations beginning a new data mining operation will often be confronted with the problem of data that has been lying around for years, is stale, and needs considerable updating.
- **Fallacy 7.** Data mining always provides positive results.
  - *Reality.* There is no guarantee of positive results when mining data for actionable knowledge. Data mining is not a panacea for solving business problems. But, used properly, by people who understand the models involved, the data requirements, and the overall project objectives, data mining can indeed provide actionable and highly profitable results.

The above discussion may have been termed, *what data mining cannot or should not do*. Next we turn to a discussion of what data mining can do.

1.6 WHAT TASKS CAN DATA MINING ACCOMPLISH?

The following list shows the most common data mining tasks.

Data Mining Tasks

Description
Estimation
Prediction
Classification
Clustering
Association

1.6.1 Description

Sometimes researchers and analysts are simply trying to find ways to *describe* patterns and trends lying within the data. For example, a pollster may uncover evidence that those who have been laid off are less likely to support the present incumbent in the presidential election. Descriptions of patterns and trends often suggest possible explanations for such patterns and trends. For example, those who are laid off are now less well off financially than before the incumbent was elected, and so would tend to prefer an alternative.

Data mining models should be as *transparent* as possible. That is, the results of the data mining model should describe clear patterns that are amenable to intuitive interpretation and explanation. Some data mining methods are more suited to transparent interpretation than others. For example, decision trees provide an intuitive and human-friendly explanation of their results. On the other hand, neural networks are comparatively opaque to nonspecialists, due to the nonlinearity and complexity of the model.

High quality description can often be accomplished with *exploratory data analysis*, a graphical method of exploring the data in search of patterns and trends. We look at exploratory data analysis in Chapter 3.

1.6.2 Estimation

In estimation, we approximate the value of a numeric target variable using a set of numeric and/or categorical predictor variables. Models are built using "complete" records, which provide the value of the target variable, as well as the predictors. Then, for new observations, estimates of the value of the target variable are made, based on the values of the predictors.

For example, we might be interested in estimating the systolic blood pressure reading of a hospital patient, based on the patient's age, gender, body-mass index, and blood sodium levels. The relationship between systolic blood pressure and the predictor variables in the training set would provide us with an estimation model. We can then apply that model to new cases.

Examples of estimation tasks in business and research include

- Estimating the amount of money a randomly chosen family of four will spend for back-to-school shopping this fall
- Estimating the percentage decrease in rotary movement sustained by a football running back with a knee injury
- Estimating the number of points per game, LeBron James will score when double-teamed in the playoffs
- Estimating the grade point average (GPA) of a graduate student, based on that student's undergraduate GPA.

Consider Figure 1.2, where we have a scatter plot of the graduate GPAs against the undergraduate GPAs for 1000 students. Simple linear regression allows us to find the line that best approximates the relationship between these two variables, according to the least squares criterion. The regression line, indicated as a straight line increasing from left to right in Figure 1.2 may then be used to estimate the graduate GPA of a student, given that student's undergraduate GPA.

Here, the equation of the regression line (as produced by the statistical package *Minitab*, which also produced the graph) is $\hat{y} = 1.24 + 0.67x$. This tells us that the estimated graduate GPA $\hat{y}$ equals 1.24 plus 0.67 times the student's undergrad GPA. For example, if your undergrad GPA is 3.0, then your estimated graduate GPA is $\hat{y} = 1.24 + 0.67\,(3) = 3.25$. Note that this point ($x = 3.0$, $\hat{y} = 3.25$) lies precisely on the regression line, as do all of the linear regression predictions.

The field of statistical analysis supplies several venerable and widely used estimation methods. These include, point estimation and confidence interval estimations, simple linear regression and correlation, and multiple regression. We examine these

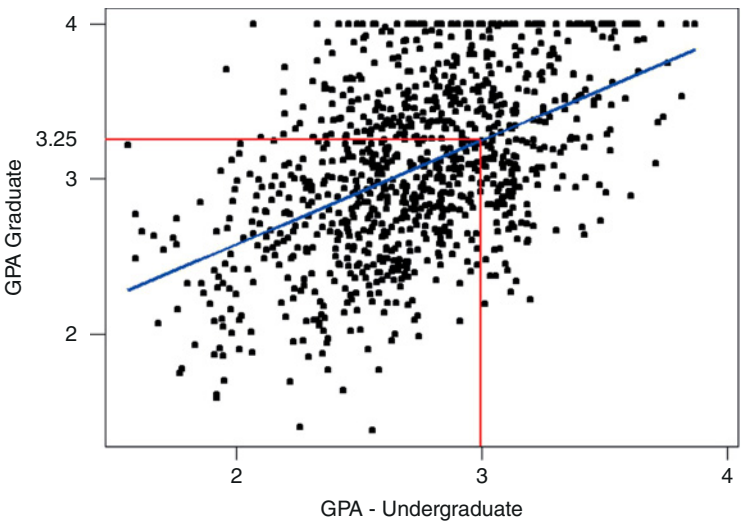


Figure 1.2 Regression estimates lie on the regression line.

methods and more in Chapters 4 and 5. Neural networks (Chapter 9) may also be used for estimation.

1.6.3 Prediction

Prediction is similar to classification and estimation, except that for prediction, the results lie in the future. Examples of prediction tasks in business and research include

- Predicting the price of a stock 3 months into the future.
- Predicting the percentage increase in traffic deaths next year if the speed limit is increased.
- Predicting the winner of this fall's World Series, based on a comparison of the team statistics.
- Predicting whether a particular molecule in drug discovery will lead to a profitable new drug for a pharmaceutical company.

Any of the methods and techniques used for classification and estimation may also be used, under appropriate circumstances, for prediction. These include the traditional statistical methods of point estimation and confidence interval estimations, simple linear regression and correlation, and multiple regression, investigated in Chapters 4 and 5, as well as data mining and knowledge discovery methods like *k*-nearest neighbor methods (Chapter 7), decision trees (Chapter 8), and neural networks (Chapter 9).

1.6.4 Classification

Classification is similar to estimation, except that the target variable is categorical rather than numeric. In classification, there is a target categorical variable, such as *income bracket*, which, for example, could be partitioned into three classes or categories: high income, middle income, and low income. The data mining model examines a large set of records, each record containing information on the target variable as well as a set of input or predictor variables. For example, consider the excerpt from a data set shown in Table 1.1.

Suppose the researcher would like to be able to *classify* the income bracket of new individuals, not currently in the above database, based on the other characteristics associated with that individual, such as age, gender, and occupation. This task is a classification task, very nicely suited to data mining methods and techniques.

TABLE 1.1 Excerpt from data set for classifying income

Subject	Age	Gender	Occupation	Income Bracket
001	47	F	Software Engineer	High
002	28	M	Marketing Consultant	Middle
003	35	M	Unemployed	Low
...	...	...	...	...