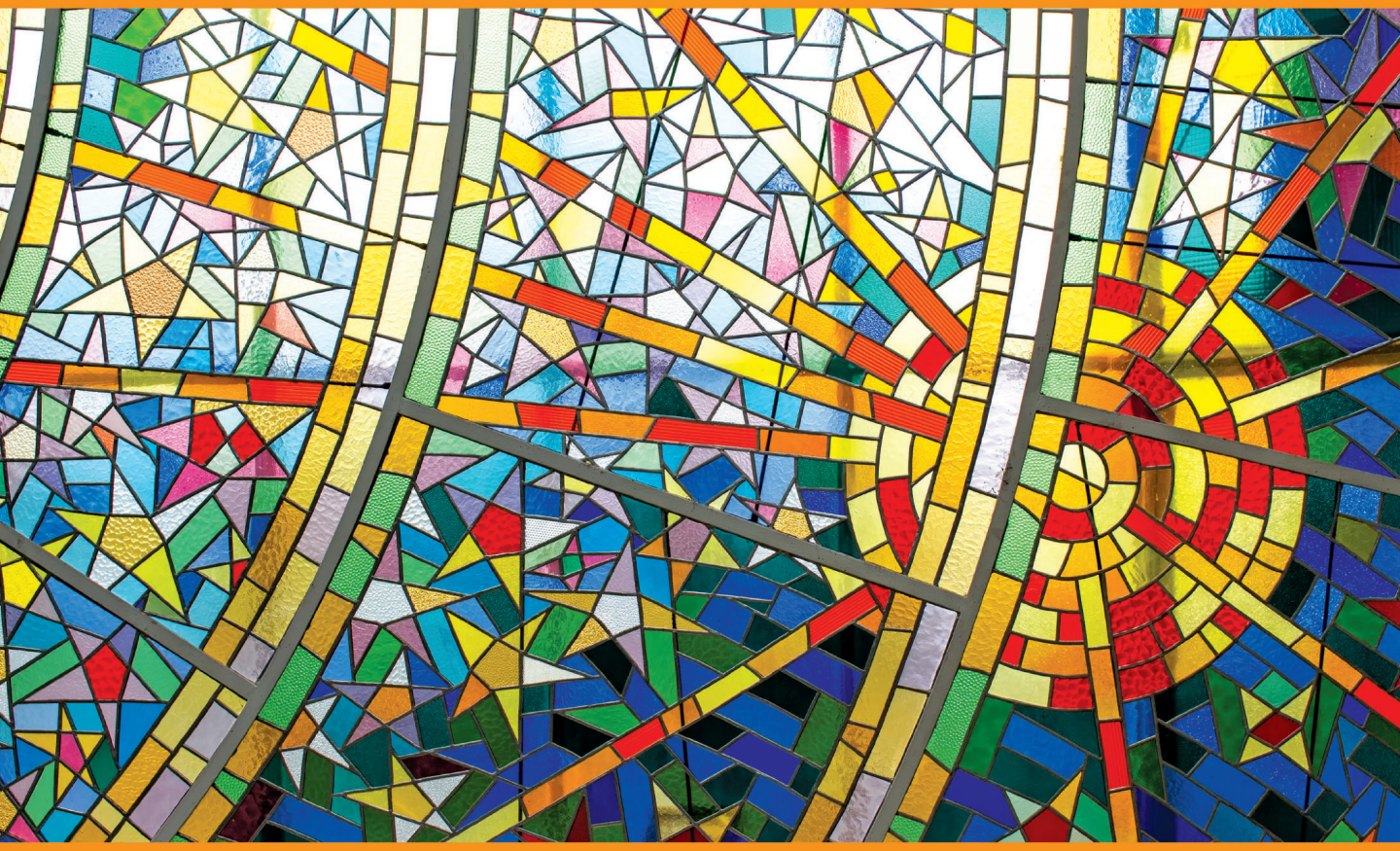
JIAWEI HAN ▪ JIAN PEI ▪ HANGHANG TONG



FOURTH EDITION

DATA MINING

CONCEPTS AND TECHNIQUES

MK
MORGAN KAUFMANN

Data Mining
Concepts and Techniques

This page intentionally left blank

Data Mining

Concepts and Techniques

Fourth Edition

Jiawei Han
Jian Pei
Hanghang Tong

MORGAN KAUFMANN PUBLISHERS
AN IMPRINT OF ELSEVIER

Morgan Kaufmann is an imprint of Elsevier
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-0-12-811760-6

Publisher: Katey Birtcher
Acquisitions Editor: Stephen Merken
Editorial Project Manager: Beth LoGiudice
Publishing Services Manager: Shereen Jameel
Production Project Manager: Gayathri S
Designer: Ryan Cook

Typeset by VTeX

Printed in the United States of America

Last digit is the print number: 9 8 7 6 5 4 3 2 1

To Dora and Lawrence for your love and encouragement

J.H.

To Jennifer, Jacqueline, and Jasmine for your never-failing care, encouragement, and support

J.P.

To Jingrui, Emma, and Nathaniel for your endless love and inspiration

H.T.

This page intentionally left blank

Contents

This page intentionally left blank

Foreword

Analyzing data is more important and prevalent than ever. Collecting and storing large datasets is easy; disks and "clouds" are well within budget of even small institutions. There is no excuse to not analyze the data to find patterns, trends, anomalies, and forecasts.

The 4th edition of *Data Mining: Concepts and Techniques* covers all the classics but adds significant material on recent developments. It has a whole chapter on deep learning, subchapters for vital topics like text mining (including one of my favorite algorithms, TopMine), frequent-subgraph discovery (covering gSpan and CloseGraph), and excellent summaries for explainability (LIME), genetic algorithms, reinforcement learning, misinformation detection, productivity and team science, causality, fairness, and social good.

The new appendix with mathematical background is extremely useful and convenient—it has all the fundamental formulas for data mining in one place, like gradient descent, Newton, and related material for optimization; SVD, eigenvalues and pseudo-inverse for matrix algebra; entropy and KL for information theory; and DFT and FFT for signal processing.

The book has an impressive, carefully chosen list of more than 800 citations, with more than 250 citations for papers after 2015. In short, this edition continues serving both as an excellent textbook and an encyclopedic reference book.

Christos Faloutsos
Carnegie Mellon University
Pittsburgh, June 2022

This page intentionally left blank

Foreword to second edition

We are deluged by data—scientific data, medical data, demographic data, financial data, and marketing data. People have no time to look at this data. Human attention has become the precious resource. So, we must find ways to automatically analyze the data, to automatically classify it, to automatically summarize it, to automatically discover and characterize trends in it, and to automatically flag anomalies. This is one of the most active and exciting areas of the database research community. Researchers in areas including statistics, visualization, artificial intelligence, and machine learning are contributing to this field. The breadth of the field makes it difficult to grasp the extraordinary progress over the last few decades.

Six years ago, Jiawei Han's and Micheline Kamber's seminal textbook organized and presented Data Mining. It heralded a golden age of innovation in the field. This revision of their book reflects that progress; more than half of the references and historical notes are to recent work. The field has matured with many new and improved algorithms, and has broadened to include many more datatypes: streams, sequences, graphs, time-series, geospatial, audio, images, and video. We are certainly not at the end of the golden age—indeed research and commercial interest in data mining continues to grow—but we are all fortunate to have this modern compendium.

The book gives quick introductions to database and data mining concepts with particular emphasis on data analysis. It then covers in a chapter-by-chapter tour the concepts and techniques that underlie classification, prediction, association, and clustering. These topics are presented with examples, a tour of the best algorithms for each problem class, and with pragmatic rules of thumb about when to apply each technique. The Socratic presentation style is both very readable and very informative. I certainly learned a lot from reading the first edition and got re-educated and updated in reading the second edition.

Jiawei Han and Micheline Kamber have been leading contributors to data mining research. This is the text they use with their students to bring them up to speed on the field. The field is evolving very rapidly, but this book is a quick way to learn the basic ideas and to understand where the field is today. I found it very informative and stimulating, and believe you will too.

Jim Gray
In his memory

This page intentionally left blank

Preface

The computerization of our society has substantially enhanced our capabilities for both generating and collecting data from diverse sources. A tremendous amount of data has flooded almost every aspect of our lives. This explosive growth in stored or transient data has generated an urgent need for new techniques and automated tools that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge. This has led to the generation of a promising and flourishing frontier in computer science called *data mining* and its various applications. Data mining, also popularly referred to as *knowledge discovery from data (KDD)*, is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories, or data streams.

This book explores the concepts and techniques of *knowledge discovery* and *data mining*. As a multidisciplinary field, data mining draws on work from areas including statistics, machine learning, pattern recognition, database technology, information retrieval, natural language processing, network science, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization. We focus on issues relating to the feasibility, usefulness, effectiveness, and scalability of techniques for the discovery of patterns hidden in *large data sets*. As a result, this book is not intended as an introduction to statistics, machine learning, database systems, or other such areas, although we do provide some background knowledge to facilitate the reader's comprehension of their respective roles in data mining. Rather, the book is a comprehensive introduction to data mining. It is useful for computer science students, application developers, and business professionals, as well as researchers involved in any of the disciplines listed above.

Data mining emerged during the late 1980s, made great strides during the 1990s, and continues to flourish into the new millennium. This book presents an overall picture of the field, introducing interesting data mining concepts and techniques and discussing applications and research directions. An important motivation for writing this book was the need to build an organized framework for the study of data mining—a challenging task, owing to the extensive multidisciplinary nature of this fast-developing field. We hope that this book will encourage people with different backgrounds and experiences to exchange their views regarding data mining to contribute toward the further promotion and shaping of this exciting and dynamic field.

Organization of the book

Since the publication of the first three editions of this book, great progress has been made in the field of data mining. Many new data mining methodologies, systems, and applications have been developed, especially for handling new kinds of data, including information networks, graphs, complex structures, and data streams, as well as text, Web, multimedia, time-series, and spatiotemporal data. Such fast development and rich, new technical contents make it difficult to cover the full spectrum of the field in a single book. Instead of continuously expanding the coverage of this book, we have decided to cover the core material in sufficient scope and depth, and leave the handling of complex data types and their applications to the books dedicated to those specific topics.

The 4th edition substantially revises the first three editions of the book, with numerous enhancements and a reorganization of the technical contents. The core technical material, which handles different mining methodologies on general data types, is expanded and substantially enhanced. To keep the book concise and up-to-date, we have done the following major revisions: (1) Two chapters in the 3rd edition, "Getting to Know You Data" and "Data Preprocessing" are merged into one chapter "Data, Measurements and Data Preprocessing," with the "Data Visualization" section removed since the concepts are easy to understand, the methods have been covered in multiple, dedicated data visualization books, and the software tools are widely available on the web; (2) two chapters in the 3rd edition, "Data Warehousing and Online Analytical Processing" and "Data Cube Technology" are merged into one chapter, with some not well-adopted data cube computation methods and data cube extensions omitted, but with a newer concept, "Data Lakes" introduced; (3) the key data mining method chapters in the 3rd edition on pattern discovery, classification, clustering and outlier analysis are retained with contents substantially enhanced and updated; (4) a new chapter "Deep Learning" is added, with a systematic introduction to neural network and deep learning methodologies; (5) the final chapter on "Data Mining Trends and Research Frontiers" is completely rewritten with many new advanced topics on data mining introduced in comprehensive and concise way; and finally, (6) a set of appendices that briefly introduce essential mathematical background needed to understand the contents of this book.

The chapters of this new edition are described briefly as follows, with emphasis on the new material.

Chapter 1 provides an *introduction* to the multidisciplinary field of data mining. It discusses the evolutionary path of information technology, which has led to the need for data mining, and the importance of its applications. It examines various kinds of data to be mined, and presents a general classification of data mining tasks, based on the kinds of knowledge to be mined, the kinds of technologies used, and the kinds of applications that are targeted. It shows that data mining is a confluence of multiple disciplines, with broad applications. Finally, it discusses how data mining may impact society.

Chapter 2 introduces the *data, measurements and data preprocessing*. It first discusses data objects and attribute types, and then introduces typical measures for basic statistical data descriptions. It also introduces ways to measure similarity and dissimilarity for various kinds of data. Then, the chapter moves to introduce techniques for data preprocessing. In particular, it introduces the concept of data quality and methods for data cleaning and data integration. It also discusses various methods for data transformation and dimensionality reduction.

Chapter 3 provides a comprehensive introduction to *datawarehouses* and online analytical processing (*OLAP*). The chapter starts with a well-accepted definition of data warehouse, an introduction to the architecture, and the concept of data lake. Then it studies the logical design of a data warehouse as a multidimensional data model, and looks at OLAP operations and how to index OLAP data for efficient analytics. The chapter includes an in-depth treatment of the techniques of building data cube as a way of implementing a data warehouse.

Chapters 4 and 5 present methods for *mining frequent patterns, associations*, and *correlations* in large data sets. **Chapter 4** introduces fundamental concepts, such as market basket analysis, with many techniques for frequent itemset mining presented in an organized way. These range from the basic Apriori algorithm and its variations to more advanced methods that improve efficiency, including the frequent pattern growth approach, frequent pattern mining with vertical data format, and mining closed and max frequent itemsets. The chapter also discusses pattern evaluation methods and introduces measures for mining correlated patterns. **Chapter 5** is on advanced pattern mining methods. It discusses methods for pattern mining in multilevel and multidimensional space, mining quantitative association

rules, mining high-dimensional data, mining rare and negative patterns, mining compressed or approximate patterns, and constraint-based pattern mining. It then moves to advanced methods for mining sequential patterns and subgraph patterns. It also presents applications of pattern mining, including phrase mining in text data and mining copy and paste bugs in software programs.

Chapters 6 and 7 describe methods for *data classification*. Due to the importance and diversity of classification methods, the contents are partitioned into two chapters. **Chapter 6** introduces basic concepts and methods for classification, including decision tree induction, Bayes classification, *k*-nearest neighbor classifiers, and linear classifiers. It also discusses model evaluation and selection methods and methods for improving classification accuracy, including ensemble methods and how to handle imbalanced data. **Chapter 7** discusses advanced methods for classification, including feature selection, Bayesian belief networks, support vector machines, rule-based and pattern-based classification. Additional topics include classification with weak supervision, classification with rich data type, multiclass classification, distant metric learning, interpretation of classification, genetic algorithms and reinforcement learning.

Cluster analysis forms the topic of Chapters 8 and 9. **Chapter 8** introduces the basic concepts and methods for data clustering, including an overview of basic cluster analysis methods, partitioning methods, hierarchical methods, density-based and grid-based methods. It also introduces methods for the evaluation of clustering. **Chapter 9** discusses advanced methods for clustering, including probabilistic model-based clustering, clustering high-dimensional data, clustering graph and network data, and semisupervised clustering.

Chapter 10 introduces *deep learning*, which is a powerful family of techniques based on artificial neural networks with broad applications in computer vision, natural language processing, machine translation, social network analysis, and so on. We start with the basic concepts and a foundational technique called backpropagation algorithm. Then, we introduce various techniques to improve the training of deep learning models, including responsive activation functions, adaptive learning rate, dropout, pretraining, cross-entropy, and autoencoder. We also introduce several commonly used deep learning architectures, ranging from feed-forward neural networks, convolutional neural networks, recurrent neural networks, and graph neural networks.

Chapter 11 is dedicated to *outlier detection*. It introduces the basic concepts of outliers and outlier analysis and discusses various outlier detection methods from the view of degree of supervision (i.e., supervised, semisupervised, and unsupervised methods), as well as from the view of approaches (i.e., statistical methods, proximity-based methods, reconstruction-based methods, clustering-based methods, and classification-based methods). It also discusses methods for mining contextual and collective outliers, and for outlier detection in high-dimensional data.

Finally, in **Chapter 12**, we discuss *future trends* and *research frontiers* in data mining. We start with a brief coverage of mining complex data types, including text data, graphs and networks, and spatiotemporal data. After that, we introduce a few data mining applications, ranging from sentiment and opinion analysis, truth discovery and misinformation identification, information and disease propagation, to productivity and team science. The chapter then moves ahead to cover other data mining methodologies, including structuring unstructured data, data augmentation, causality analysis, network-as-a-context, and auto-ML. Finally, it discusses social impacts of data mining, including privacy-preserving data mining, human-algorithm interaction, fairness, interpretability and robustness, and data mining for social good.

Throughout the text, *italic* font is used to emphasize terms that are defined, and **bold** font is used to highlight or summarize main ideas. Sans serif font is used for reserved words. Bold italic font is used to represent multidimensional quantities.

This book has several strong features that set it apart from other textbooks on data mining. It presents a very broad yet in-depth coverage of the principles of data mining. The chapters are written to be as self-contained as possible, so they may be read in order of interest by the reader. Advanced chapters offer a larger-scale view and may be considered optional for interested readers. All of the major methods of data mining are presented. The book presents important topics in data mining regarding multidimensional OLAP analysis, which is often overlooked or minimally treated in other data mining books. The book also maintains web sites with a number of online resources to aid instructors, students, and professionals in the field. These are described further in the following.

To the instructor

This book is designed to give a broad, yet detailed overview of the data mining field. First, it can be used to teach an introductory course on data mining at an advanced undergraduate level or at the first-year graduate level. Moreover, the book also provides essential materials for an advanced graduate course on data mining.

Depending on the length of the instruction period, the background of students, and your interests, you may select subsets of chapters to teach in various sequential orderings. For example, an introductory course may cover the following chapters.

- Chapter 1: Introduction
- Chapter 2: Data, measurements, and data preprocessing
- Chapter 3: Data warehousing and online analytical processing
- Chapter 4: Pattern mining: basic concepts and methods
- Chapter 6: Classification: basic concepts
- Chapter 8: Cluster analysis: basic concepts and methods

If time permits, some materials about deep learning (Chapter 10) or outlier detection (Chapter 11) may be chosen. In each chapter, the fundamental concepts should be covered, while some sections on advanced topics can be treated optionally.

As another example, for a place where a machine learning course is offered to cover supervised learning well, a data mining course can discuss in depth on clustering. Such a course can be based on the following chapters.

- Chapter 1: Introduction
- Chapter 2: Data, measurements, and data preprocessing
- Chapter 3: Data warehousing and online analytical processing
- Chapter 4: Pattern mining: basic concepts and methods
- Chapter 8: Cluster analysis: basic concepts and methods
- Chapter 9: Cluster analysis: advanced methods
- Chapter 11: Outlier detection

An instructor teaching an advanced data mining course may find Chapter 12 particularly informative, since it discusses an extensive spectrum of new topics in data mining that are under dynamic and fast development.

Alternatively, you may choose to teach the whole book in a two-course sequence that covers all of the chapters in the book, plus, when time permits, some advanced topics such as graph and network mining. Material for such advanced topics may be selected from the companion chapters available from the book's web site, accompanied with a set of selected research papers.

Individual chapters in this book can also be used for tutorials or for special topics in related courses, such as machine learning, pattern recognition, data warehousing, and intelligent data analysis.

Each chapter ends with a set of exercises, suitable as assigned homework. The exercises are either short questions that test basic mastery of the material covered, longer questions that require analytical thinking, or implementation projects. Some exercises can also be used as research discussion topics. The bibliographic notes at the end of each chapter can be used to find the research literature that contains the origin of the concepts and methods presented, in-depth treatment of related topics, and possible extensions.

To the student

We hope that this textbook will spark your interest in the young yet fast-evolving field of data mining. We have attempted to present the material in a clear manner, with careful explanation of the topics covered. Each chapter ends with a summary describing the main points. We have included many figures and illustrations throughout the text to make the book more enjoyable and reader-friendly. Although this book was designed as a textbook, we have tried to organize it so that it will also be useful to you as a reference book or handbook, should you later decide to perform in-depth research in the related fields or pursue a career in data mining.

What do you need to know to read this book?

- You should have some knowledge of the concepts and terminology associated with statistics, database systems, and machine learning. However, we do try to provide enough background of the basics, so that if you are not so familiar with these fields or your memory is a bit rusty, you will not have trouble following the discussions in the book.
- You should have some programming experience. In particular, you should be able to read pseudocode and understand simple data structures such as multidimensional arrays and structures.

To the professional

This book was designed to cover a wide range of topics in the data mining field. As a result, it is an excellent handbook on the subject. Because each chapter is designed to be as standalone as possible, you can focus on the topics that most interest you. The book can be used by application programmers, data scientists, and information service managers who wish to learn about the key ideas of data mining on their own. The book would also be useful for technical data analysis staff in banking, insurance, medicine, and retailing industries who are interested in applying data mining solutions to their businesses. Moreover, the book may serve as a comprehensive survey of the data mining field, which may also benefit researchers who would like to advance the state-of-the-art in data mining and extend the scope of data mining applications.

The techniques and algorithms presented are of practical utility. Rather than selecting algorithms that perform well on small "toy" data sets, the algorithms described in the book are geared for the discovery of patterns and knowledge hidden in large, real data sets. Algorithms presented in the book are illustrated in pseudocode. The pseudocode is similar to the C programming language, yet is designed so that it should be easy to follow by programmers unfamiliar with C or C++. If you wish to implement any of the algorithms, you should find the translation of our pseudocode into the programming language of your choice to be a fairly straightforward task.

Book web site with resources

The book has a website with Elsevier at *https://educate.elsevier.com/book/details/9780128117606*. This website contains many supplemental materials for readers of the book or anyone else with an interest in data mining. The resources include the following:

- **Slide presentations for each chapter**. Lecture notes in Microsoft PowerPoint slides are available for each chapter.
- **Instructors' manual**. This complete set of answers to the exercises in the book is available only to instructors from the publisher's web site.
- **Figures from the book**. This may help you to make your own slides for your classroom teaching.
- **Table of Contents** of the book in PDF format.
- **Errata on the different printings of the book**. We encourage you to point out any errors in this book. Once the error is confirmed, we will update the errata list and include acknowledgment of your contribution.

Interested readers may also like to check **Authors' course teaching websites**. All the authors are university professors in their respective universities. Please check their corresponding data mining course websites which may contain their undergraduate and/or graduate course materials for introductory and/or advanced courses on data mining, including updated course/chapter slides, syllabi, homeworks, programming assignments, research projects, errata, and other related information.

Acknowledgments

Fourth edition of the book

We would like to express our sincere thanks to Micheline Kamber, the co-author of the previous editions of this book. Micheline has contributed substantially to these editions. Due to her commitment of other duties, she will not be able to join us in this new edition. We really appreciate her long-term collaborations and contributions in the past many years.

We would also like to express our grateful thanks to the previous and current members, including faculty and students, of the Data and Information Systems (DAIS) Laboratory, the Data Mining Group, IDEA Lab and iSAIL Lab at UIUC and the Data Mining Group at SFU, and many friends and colleagues, whose constant support and encouragement have made our work on this edition a rewarding experience. Our thanks extend to the students and TAs in the many data mining courses we taught at UIUC and SFU, as well as those in summer schools and beyond, who carefully went through the early drafts and the early editions of this book, identified many errors, and suggested various improvements.

We also wish to thank Steve Merken and Beth LoGiudice at Elsevier, for their enthusiasm, patience, and support during our writing of this edition of the book. We thank Gayathri S, the Project Manager, and her team members, for keeping us on schedule. We are also grateful for the invaluable feedback from all of the reviewers.

We would like to thank the US National Science Foundation (NSF), US Defense Advanced Research Projects Agency (DARPA), US Army Research Laboratory (ARL), US National Institute of Health (NIH), US Defense Threat Reduction Agency (DTRA), and Natural Science and Engineering Research Council of Canada (NSERC), as well as Microsoft Research, Google Research, IBM Research, Amazon, Adobe, LinkedIn, Yahoo!, HP Labs, PayPal, Facebook, Visa Research, and other industry research labs for their support of our research in the form of research grants, contracts, and gifts. Such research support deepens our understanding of the subjects discussed in this book.

Finally, we thank our families for their wholehearted support throughout this project.

This page intentionally left blank

About the authors

Jiawei Han is a Michael Aiken Chair Professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign. He has received numerous awards for his contributions on research into knowledge discovery and data mining, including ACM SIGKDD Innovation Award (2004), IEEE Computer Society Technical Achievement Award (2005), and IEEE W. Wallace McDowell Award (2009). He is a Fellow of ACM and a Fellow of IEEE. He served as founding Editor-in-Chief of *ACM Transactions on Knowledge Discovery from Data* (2006–2011) and as an editorial board member of several journals, including *IEEE Transactions on Knowledge and Data Engineering* and *Data Mining and Knowledge Discovery*.

Jian Pei is currently Professor of Computer Science, Biostatistics and Bioinformatics, and Electrical and Computer Engineering at Duke University. He received a Ph.D. degree in computing science from Simon Fraser University in 2002 under Dr. Jiawei Han's supervision. He has published prolifically in the premier academic forums on data mining, databases, Web searching, and information retrieval and actively served the academic community. He is a fellow of the Royal Society of Canada, the Canadian Academy of Engineering, ACM, and IEEE. He received the 2017 ACM SIGKDD Innovation Award and the 2015 ACM SIGKDD Service Award.

Hanghang Tong is currently an associate professor at Department of Computer Science at University of Illinois at Urbana-Champaign. He received his Ph.D. degree from Carnegie Mellon University in 2009. He has published over 200 refereed articles. His research is recognized by several prestigious awards and thousands of citations. He is the Editor-in-Chief of SIGKDD Explorations (ACM) and an associate editor of several journals.

This page intentionally left blank

CHAPTER

1 Introduction

This book is an introduction to the young and fast-growing field of *data mining* (also known as *knowledge discovery from data*, or *KDD* for short). The book focuses on fundamental data mining concepts and techniques for discovering interesting patterns from data in various applications. In particular, we emphasize prominent techniques for developing effective, efficient, and scalable data mining tools.

This chapter is organized as follows. In Section 1.1, we learn what is data mining and why data mining is in high demand. Section 1.2 links data mining with the overall knowledge discovery process. Next, we learn about data mining from multiple aspects, such as the kinds of data that can be mined (Section 1.3), the kinds of knowledge to be mined (Section 1.4), the relationship between data mining and other disciplines (Section 1.5), and data mining applications (Section 1.6). Finally, we discuss the impact of data mining to society (Section 1.7).

1.1 What is data mining?

Necessity, who is the mother of invention.
– Plato

We live in a world where vast amounts of data are generated constantly and rapidly.

"*We are living in the information age*" is a popular saying; however, *we are actually living in the data age*. Terabytes or petabytes of data pour into our computer networks, the World Wide Web (WWW), and various kinds of devices every day from business, news agencies, society, science, engineering, medicine, and almost every other aspect of daily life. This explosive growth of available data volume is a result of the computerization of our society and the fast development of powerful computing, sensing, and data collection, storage, and publication tools.

Businesses worldwide generate gigantic data sets, including sales transactions, stock trading records, product descriptions, sales promotions, company profiles and performance, and customer feedback. Scientific and engineering practices generate high orders of petabytes of data in a continuous manner, from remote sensing, to process measuring, scientific experiments, system performance, engineering observations, and environment surveillance. Biomedical research and the health industry generate tremendous amounts of data from gene sequence machines, biomedical experiment and research reports, medical records, patient monitoring, and medical imaging. Billions of Web searches supported by search engines process tens of petabytes of data daily. Social media tools have become increasingly popular, producing a tremendous number of texts, pictures, and videos, generating various kinds of Web communities and social networks. The list of sources that generate huge amounts of data is endless.

Data Mining. https://doi.org/10.1016/B978-0-12-811760-6.00011-4

This explosively growing, widely available, and gigantic body of data makes our time truly *the data age*. Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge. This necessity has led to the birth of data mining.

Essentially, ***data mining*** *is the process of discovering interesting patterns, models, and other kinds of knowledge in large data sets*. The term, *data mining*, as a vivid view of searching for *gold nuggets* from data, appeared in 1990s. However, to refer to the mining of gold from rocks or sand, we say *gold mining* instead of rock or sand mining. Analogously, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. However, the shorter term, *knowledge mining* may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer carrying both "data" and "mining" became a popular choice. In addition, many other terms have a similar meaning to data mining—for example, *knowledge mining from data*, *KDD (i.e., Knowledge Discovery from Data)*, *pattern discovery*, *knowledge extraction*, *data archaeology*, *data analytics*, and *information harvesting*.

Data mining is a young, dynamic, and promising field. It has made and will continue to make great strides in our journey from the data age toward the coming information age.

Example 1.1. Data mining turns a large collection of data into knowledge. A search engine (e.g., Google) receives billions of queries every day. What novel and useful knowledge can a search engine learn from such a huge collection of queries collected from users over time? Interestingly, some patterns found in user search queries can disclose invaluable knowledge that cannot be obtained by reading individual data items alone. For example, Google's *Flu Trends* uses specific search terms as indicators of flu activity. It found a close relationship between the number of people who search for flu-related information and the number of people who actually have flu symptoms. A pattern emerges when all of the search queries related to flu are aggregated. Using aggregated Google search data, *Flu Trends* can estimate flu activity up to two weeks faster than what traditional systems can.[1] This example shows how data mining can turn a large collection of data into knowledge that can help meet a current global challenge. □

1.2 Data mining: an essential step in knowledge discovery

Many people treat data mining as a synonym for another popularly used term, **knowledge discovery from data**, or **KDD**, whereas others view data mining as merely an essential step in the overall process of knowledge discovery. The overall knowledge discovery process is shown in Fig. 1.1 as an iterative sequence of the following steps:

1. **Data preparation**
   a. **Data cleaning** (to remove noise and inconsistent data)
   b. **Data integration** (where multiple data sources may be combined)[2]

[1] This is reported in [GMP+09]. The *Flu Trend* reporting stopped in 2015.

[2] A popular trend in the information industry is to perform data cleaning and data integration as a preprocessing step, where the resulting data are stored in a data warehouse.

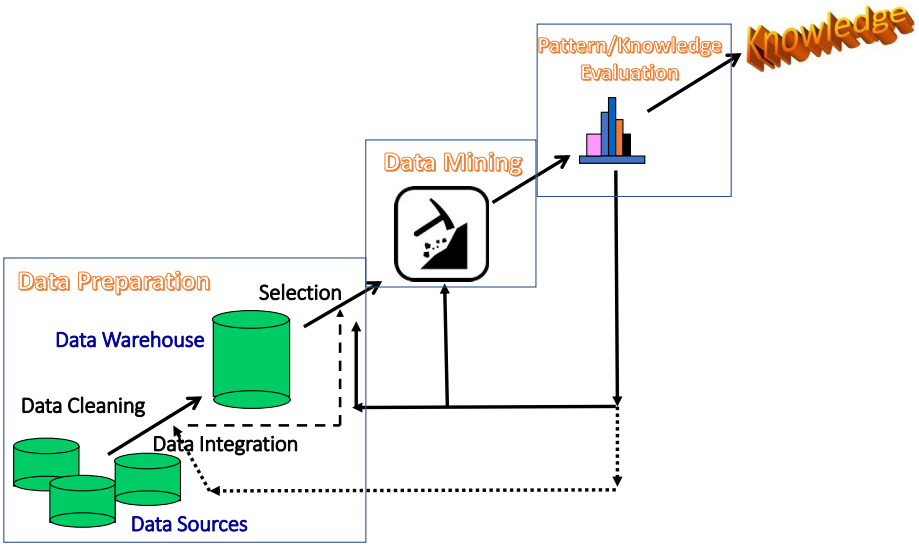


FIGURE 1.1

Data mining: An essential step in the process of knowledge discovery.

c. **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)[3]
d. **Data selection** (where data relevant to the analysis task are retrieved from the database)

2. **Data mining** (an essential process where intelligent methods are applied to extract patterns or construct models)
3. **Pattern/model evaluation** (to identify the truly interesting patterns or models representing knowledge based on *interestingness measures*—see Section 1.4.7)
4. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)

Steps 1(a) through 1(d) are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with a user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

The preceding view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns or models for evaluation. However, in industry, in media, and in the research milieu, the term *data mining* is often used to refer to the entire knowledge discovery process (perhaps because the term is shorter than *knowledge discovery from data*). Therefore, we adopt a broad view of data mining functionality: *Data mining is the* process *of discovering inter-*

[3] Data transformation and consolidation are often performed before the data selection process, particularly in the case of data warehousing. *Data reduction* may also be performed to obtain a smaller representation of the original data without sacrificing its integrity.

esting patterns and knowledge from large *amounts of data*. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

1.3 Diversity of data types for data mining

As a general technology, data mining can be applied to any kind of data as long as the data are meaningful for a target application. However, different kinds of data may need rather different data mining methodologies, from simple to rather sophisticated, making data mining a rich and diverse field.

Structured vs. unstructured data

Based on whether data have clear structures, we can categorize data as *structured* vs. *unstructured data*.

Data stored in *relational databases*, *data cubes*, *data matrices*, and many *data warehouses* have uniform, record- or table-like structures, defined by their data dictionaries, with a fixed set of attributes (or fields, columns), each with a fixed set of value ranges and semantic meaning. These data sets are typical examples of highly structured data. In many real applications, such strict structural requirement can be relaxed in multiple ways to accommodate *semistructured* nature of the data, such as to allow a data object to contain a set value, a small set of heterogeneous typed values, or nested structures or to allow the structure of objects or subobjects to be defined flexibly and dynamically (e.g., XML structures).

There are many data sets that may not be as structured as relational tables or data matrices. However, they do have certain structures with clearly defined semantic meaning. For example, a *transactional data set* may contain a large set of transactions each containing a set of items. A *sequence data set* may contain a large set of sequences each containing an ordered set of elements that can in turn contain a set of items. Many application data sets, such as shopping transaction data, time-series data, gene or protein data, or Weblog data, belong to this category.

A more sophisticated type of semistructured data set is *graph or network data*, where a set of nodes are connected by a set of edges (also called links); and each node/link may have its own semantic description or substructures.

Each of such categories of structured and semistructured data sets may have special kinds of patterns or knowledge to be mined and many dedicated data mining methods, such as sequential pattern mining, graph pattern mining, and information network mining methods, have been developed to analyze such data sets.

Beyond such structured or semistructured data, there are also large amounts of unstructured data, such as text data and multimedia (e.g., audio, image, video) data. Although some studies treat them as one-dimensional or multidimensional byte streams, they do carry a lot of interesting semantics. Domain-specific methods have been developed to analyze such data in the fields of natural language understanding, text mining, computer vision, and pattern recognition. Moreover, recent advances on deep learning have made tremendous progress on processing text, image, and video data. Nevertheless, mining hidden structures from unstructured data may greatly help understand and make good use of such data.

The real-world data can often be a mixture of structured data, semistructured data, and unstructured data. For example, an online shopping website may host information for a large set of products, which

can be essentially structured data stored in a relational database, with a fixed set of fields on product name, price, specifications, and so on. However, some fields may essentially be text, image, and video data, such as product introduction, expert or user reviews, product images, and advertisement videos. Data mining methods are often developed for mining some particular type of data, and their results can be integrated and coordinated to serve the overall goal.

Data associated with different applications

Different applications may generate or need to handle very different data sets and require rather different data analysis methods. Thus when categorizing data sets for data mining, we should take specific applications into consideration.

Take sequence data as an example. *Biological sequences* such as DNA or protein sequences may have very different semantic meaning from *shopping transaction sequences* or *Web click streams*, calling for rather different sequence mining methods. A special kind of sequence data is time-series data where a *time-series* may contain an ordered set of numerical values with equal time interval, which is also rather different from shopping transaction sequences, which may not have fixed time gaps (a customer may shop at anytime she likes).

Data in some applications can be associated with spatial information, time information, or both, forming *spatial, temporal, and spatiotemporal data*, respectively. Special data mining methods, such as spatial data mining, temporal data mining, spatiotemporal data mining, or trajectory pattern mining, should be developed for mining such data sets as well.

For graph and network data, different applications may also need rather different data mining methods. For example, social networks (e.g., Facebook or LinkedIn data), computer communication networks, biological networks, and information networks (e.g., authors linking with keywords) may carry rather different semantics and require different mining methods.

Even for the same data set, finding different kinds of patterns or knowledge may require different data mining methods. For example, for the same set of software (source) programs, finding plagiarized subprogram modules or finding copy-and-paste bugs may need rather different data mining techniques.

Rich data types and diverse application requirements call for very diverse data mining methods. Thus data mining is a rich and fascinating research domain, with lots of new methods waiting to be studied and developed.

Stored vs. streaming data

Usually, data mining handles finite, stored data sets, such as those stored in various kinds of large data repositories. However, in some applications such as video surveillance or remote sensing, data may stream in dynamically and constantly, as infinite *data streams*. Mining stream data will require rather different methods than stored data, which may form another interesting theme in our study.

1.4 Mining various kinds of knowledge

Different kinds of patterns and knowledge can be uncovered via data mining. In general, data mining tasks can be put into two categories: **descriptive data mining** and **predictive data mining**. Descriptive mining characterizes properties of the interested set of data, whereas predictive mining performs induction on the data set in order to make predictions.

In this section, we introduce different data mining tasks. These include multidimensional data summarization (Section 1.4.1); the mining of frequent patterns, associations, and correlations (Section 1.4.2); classification and regression (Section 1.4.3); cluster analysis (Section 1.4.4); and outlier analysis (Section 1.4.6). Different data mining functionalities generate different kinds of results that are often called patterns, models, or knowledge. In Section 1.4.7, we will also introduce the interestingness of a pattern or a model. In many cases, only interesting patterns or models will be considered as *knowledge*.

1.4.1 Multidimensional data summarization

It is often tedious for a user to go over the details of a large set of data. Thus it is desirable to automatically summarize an interested set of data and compare it with the contrasting sets at some high levels. Such summaritive description of an *interested set of data* is called **data summarization**. Data summarization can often be conducted in a multidimensional space. If the multidimensional space is well defined and frequently used, such as product category, producer, location, or time, massive amounts of data can be aggregated in the form of **data cubes** to facilitate user's drill-down or roll-up of the summarization space with mouse clicking. The output of such multidimensional summarization can be presented in various forms, such as **pie charts**, **bar charts**, **curves**, **multidimensional data cubes**, and **multidimensional tables**, including crosstabs.

For structured data, multidimensional aggregation methods have been developed to facilitate such precomputation or online computation of multidimensional aggregations using data cube technology, which will be discussed in Chapter 3. For unstructured data, such as text, this task becomes challenging. We will give a brief discussion of such research frontiers in our last chapter.

1.4.2 Mining frequent patterns, associations, and correlations

Frequent patterns, as the name suggests, are patterns that occur frequently in data. There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures. A *frequent itemset* typically refers to a set of items that often appear together in a transactional data set—for example, milk and bread, which are frequently bought together in grocery stores by many customers. A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a computer bag, and then other accessories, is a (*frequent*) *sequential pattern*. A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (*frequent*) *structured pattern*. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

Example 1.2. Association analysis. Suppose that, a webstore manager wants to know which items are frequently purchased together (i.e., in the same transaction). An example of such a rule, mined from the transactional database, is

$$buys(X, \text{“}computer\text{”}) \Rightarrow buys(X, \text{“}webcam\text{”}) \quad [support = 1\%, confidence = 50\%],$$

where X is a variable representing a customer. A **confidence**, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy webcam as well. A 1% **support** means

that 1% of all the transactions under analysis show that computer and webcam are purchased together. This association rule involves a single attribute or predicate (i.e., *buys*) that repeats. Association rules that contain a single predicate are referred to as **single-dimensional association rules**. Dropping the predicate notation, the rule can be written simply as "*computer* $\Rightarrow$ *webcam* [1%, 50%]."

Suppose, mining the same database generates another association rule:

$$age(X, \text{“20..29”}) \wedge income(X, \text{“40K..49K”}) \Rightarrow buys(X, \text{“laptop”})$$
$$[support = 0.5\%, confidence = 60\%].$$

The rule indicates that of all its customers under study, 0.5% are 20 to 29 years old with an income of $40,000 to $49,000 and have purchased a laptop (computer). There is a 60% probability that a customer in this age and income group will purchase a laptop. Note that this is an association involving more than one attribute or predicate (i.e., *age, income*, and *buys*). Adopting the terminology used in multidimensional databases, where each attribute is referred to as a dimension, the above rule can be referred to as a **multidimensional association rule**. □

Typically, association rules are discarded as uninteresting if they do not satisfy both a **minimum support threshold** and a **minimum confidence threshold**. Additional analysis can be performed to uncover interesting statistical **correlations** between associated attribute–value pairs.

Frequent itemset mining is a fundamental form of frequent pattern mining. Mining frequent itemsets, associations, and correlations will be discussed in Chapter 4. Mining diverse kinds of frequent pattern, as well as mining sequential patterns and structured patterns, will be covered in Chapter 5.

1.4.3 Classification and regression for predictive analysis

Classification is the process of finding a **model** (or function) that describes and distinguishes data classes or concepts. The model is derived based on the analysis of a set of **training data** (i.e., data objects for which the class labels are known). The model is used to predict the class labels of objects for which the class labels are unknown.

Depending on the classification methods, a derived model can be in various forms, such as a set of *classification rules* (i.e., *IF-THEN rules*), a *decision tree*, a *mathematical formula*, or a learned *neural network* (Fig. 1.2). A **decision tree** is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can easily be converted to classification rules. A **neural network**, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units. There are many other methods for constructing classification models, such as naïve Bayesian classification, support vector machines, and k-nearest-neighbor classification.

Whereas classification predicts categorical (discrete, unordered) labels, **regression** models continuous-valued functions. That is, regression is used to predict missing or unavailable *numerical data values* rather than (discrete) class labels. The term *prediction* refers to both numeric prediction and class label prediction. **Regression analysis** is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution *trends* based on the available data.

Classification and regression may need to be preceded by **feature selection** or **relevance analysis**, which attempts to identify attributes (often called *features*) that are significantly relevant to the clas-

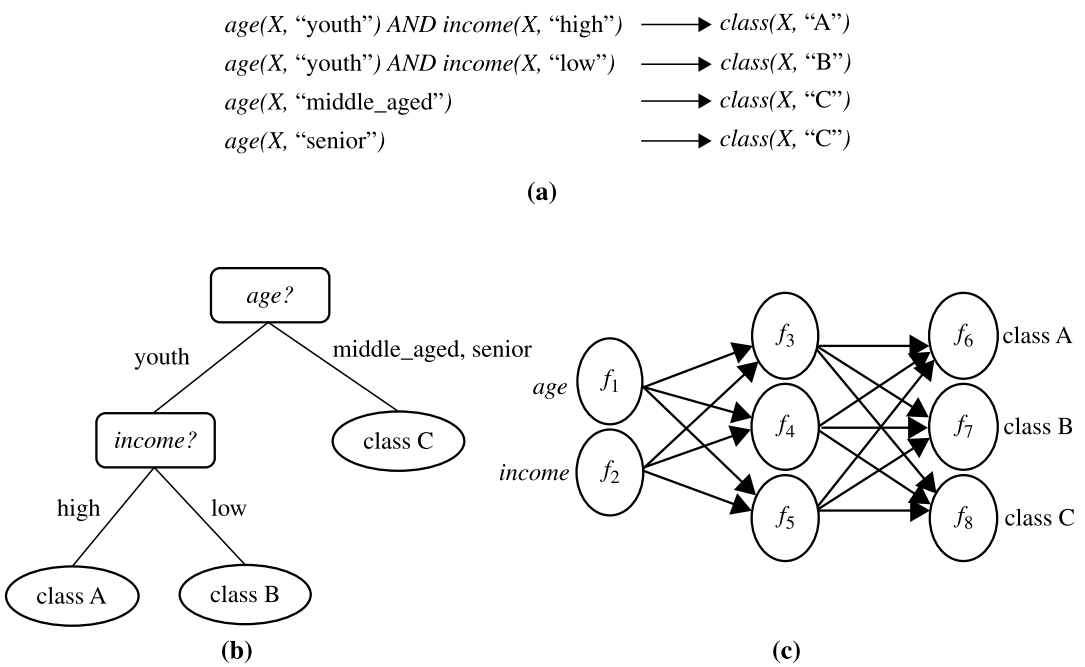


FIGURE 1.2

A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

sification and regression process. Such attributes will be selected for the classification and regression process. Other attributes, which are irrelevant, can then be excluded from consideration.

Example 1.3. Classification and regression. Suppose a webstore sales manager wants to classify a large set of items in the store, based on three kinds of responses to a sales campaign: *good response*, *mild response*, and *no response*. You want to derive a model for each of these three classes based on the descriptive features of the items, such as *price*, *brand*, *place_made*, *type*, and *category*. The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set.

Suppose that the resulting classification is expressed as a decision tree. The decision tree, for instance, may identify *price* as being the first important factor that best distinguishes the three classes. Other features that help further distinguish objects of each class from one another include *brand* and *place_made*. Such a decision tree may help the manager understand the impact of the given sales campaign and design a more effective campaign in the future.

Suppose instead, that rather than predicting categorical response labels for each store item, you would like to predict the amount of revenue that each item will generate during an upcoming sale, based on the previous sales data. This is an example of regression analysis because the regression model constructed will predict a continuous function (or ordered value.) □

Chapters 6 and 7 discuss classification in further detail. Regression analysis is covered lightly in these chapters since it is typically introduced in statistics courses. Sources for further information are given in the bibliographic notes.

1.4.4 Cluster analysis

Unlike classification and regression, which analyze class-labeled (training) data sets, **cluster analysis** (also called **clustering**) groups data objects without consulting class labels. In many cases, class-labeled data may simply not exist at the beginning. Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of *maximizing the intraclass similarity and minimizing the interclass similarity*. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters. Each cluster so formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate **taxonomy formation**, that is, the organization of observations into a hierarchy of classes that group similar events together.

Example 1.4. Cluster analysis. Cluster analysis can be performed on the webstore customer data to identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing. Fig. 1.3 shows a 2-D plot of customers with respect to customer locations in a city. Three clusters of data points are evident. □

Cluster analysis forms the topic of Chapters 8 and 9.

1.4.5 Deep learning

For many data mining tasks, such as classification and clustering, a key step often lies in finding "good features," which is a vector representation of each input data tuple. For example, in order to predict

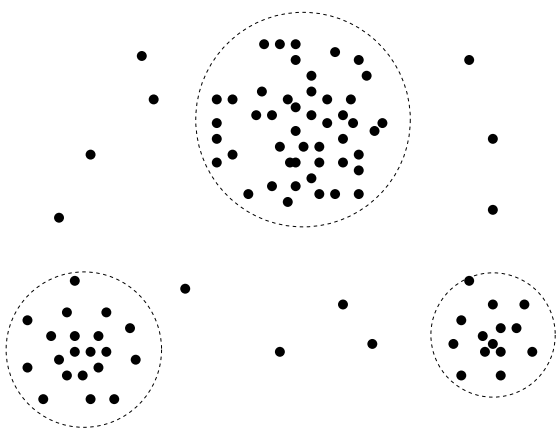


FIGURE 1.3

A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.