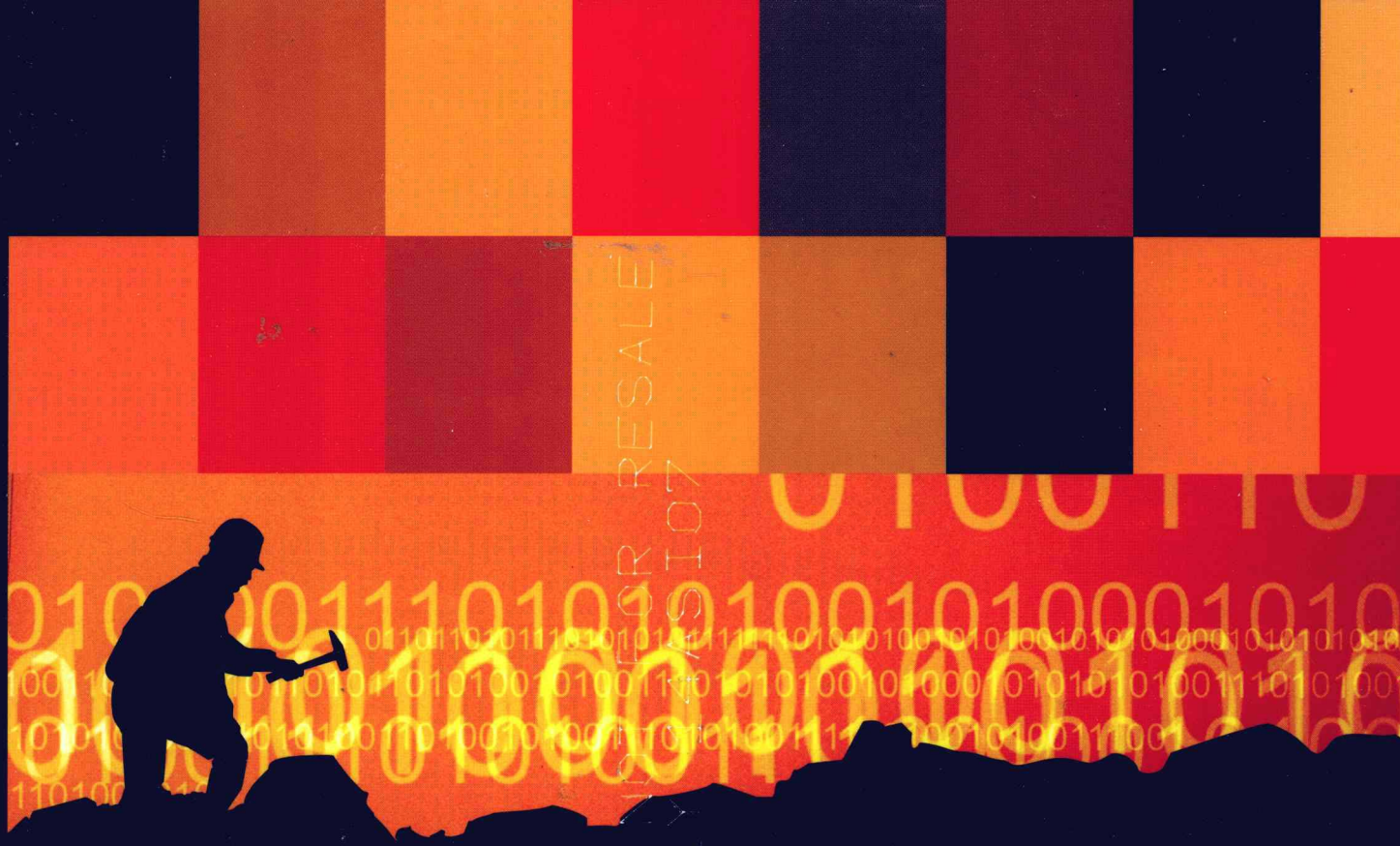



# Introduction to Business Data Mining

DAVID OLSON

YONG SHI

# Introduction to Business Data Mining

**David Olson**
*University of Nebraska–Lincoln*

**Yong Shi**
*Graduate University of the Chinese Academy of Sciences*
*University of Nebraska–Omaha*


**McGraw-Hill Irwin**

Boston Burr Ridge, IL Dubuque, IA Madison, WI New York San Francisco St. Louis
Bangkok Bogotá Caracas Kuala Lumpur Lisbon London Madrid Mexico City
Milan Montreal New Delhi Santiago Seoul Singapore Sydney Taipei Toronto

INTRODUCTION TO BUSINESS DATA MINING



This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 0 CCW/CCW 0 9 8 7 6 5

ISBN-13: 978-0-07-295971-0
ISBN-10: 0-07-295971-1

Editorial director: *Brent Gordon*
Executive editor: *Scott Isenberg*
Editorial coordinator: *Lee Stone*
Senior marketing manager: *Douglas Reiner*
Senior media producer: *Victor Chiu*
Project manager: *Harvey Yep*
Senior production supervisor: *Sesha Bolisetty*
Designer: *Cara David*
Lead media project manager: *Brian Nacik*
Cover design: *Chris Bowyer*
Typeface: *10/12 Palatino*
Compositor: *International Typesetting & Composition*
Printer: *Courier Westford*

**Library of Congress Cataloging-in-Publication Data**

Olson, David Louis.
Introduction to business data mining / David Olson, Yong Shi.
p. cm.
Includes index.
ISBN-13: 978-0-07-295971-0 (alk. paper)
ISBN-10: 0-07-295971-1 (alk. paper)
1. Data mining. 2. Business—Data processing. I. Shi, Yong, 1956– II. Title.
HF5548.2.O46 2007
006.3'12—dc22 2005044525

www.mhhe.com

To my family.

*David Olson*

This book is dedicated to my parents,
Li Guihua and Shi Yuanqing for their constant support and understanding of my academic career.

*Yong Shi*

# About the Authors

## David L. Olson

*University of Nebraska–Lincoln*

**David Olson** is the James & H. K. Stuart Professor in MIS and Othmer Professor at the University of Nebraska. He received his Ph.D. in Business from the University of Nebraska in 1981. He has published research in over eighty refereed journal articles, primarily on the topic of multiple objective decision making. He teaches in the management information systems, management science, and operations management areas. He has authored the books *Decision Aids for Selection Problems, Introduction to Information Systems Project Management,* and *Managerial Issues of Enterprise Resource Planning Systems,* and co-authored the books *Decision Support Models and Expert Systems; Introduction to Management Science; Introduction to Simulation and Risk Analysis; Business Statistics: Quality Information for Decision Analysis; Statistics, Decision Analysis, and Decision Modeling; Multiple Criteria Analysis in Strategic Siting Problems;* and *Introduction to Business Data Mining.* He has made over one hundred presentations at international and national conferences on research topics. He is a member of the Association for Information Systems, the Decision Sciences Institute, the Institute for Operations Research and Management Sciences, and the Multiple Criteria Decision-Making Society. He has coordinated the Decision Sciences Institute Dissertation Competition and the Innovative Education Competition, chaired the Doctoral Affairs Committee, and served as nationally elected vice president three times and as National Program Chair. He was with Texas A&M University from 1981 through 2001—the last two years as Lowry Mays Professor of Business in the Department of Information and Operations Management. He received a Research Fellow Award from the College of Business and Graduate School of Business at Texas A&M University and held the Business Analysis Faculty Excellence Fellowship for two years. He is a Fellow of the Decision Sciences Institute.

## Yong Shi

*Graduate University of the Chinese Academy of Sciences*
*University of Nebraska–Omaha*

**Professor Yong Shi** currently is the director of Chinese Academy of Sciences Research Center on Data Technology & Knowledge Economy and Assistant President of the Graduate University of Chinese Academy of Sciences. He has been the Charles W. and Margre H. Durham Distinguished Professor of Information Technology, College of Information Science and Technology, University of Nebraska at Omaha since 1999. Dr. Shi's research interests cover data mining, information overload, multiple criteria decision making, and telecommunications management. He has published seven books, more than sixty papers in various journals, and numerous conferences/proceedings papers. He is the Editor-in-Chief of *International Journal of Information Technology and Decision Making (SCI),* an Area Editor of *International Journal of Operations and*

*Quantitative Management,* and an Editorial Board Member for a number of academic journals. Dr. Shi has received many distinguished awards including Outstanding Young Scientist Award, National Natural Science Foundation of China, 2001; Member of Overseas Assessor for the Chinese Academy of Sciences, May 2000; and Speaker of Distinguished Visitors Program (DVP) for 1997–2000, IEEE Computer Society. He has consulted for a number of famous companies in data mining and knowledge management projects.

# Preface

The intent of this book is to serve advanced undergraduate and graduate classes presenting data mining. Data mining is a very useful topic, applying quantitative analysis to large-scale data made available through recently developed information technology. Each of us has taught such material, and we both have extensive experience in quantitative analysis in business. Yong Shi also has extensive real experience in commercial data mining analysis. We want to take this opportunity to acknowledge the graduate students at the University of Nebraska at Omaha, Gang Kou, Nian Yan, and Wei Zhuang, who helped us prepare the data mining reports by using computer software for this book.

## Book Concept

Our intent is to cover the fundamental concepts of data mining, to demonstrate the potential of gathering large sets of data and analyzing these data sets to gain useful business understanding. We have organized the material into four parts. Part I introduces concepts. Part II describes and demonstrates basic data mining algorithms. Part III focuses on business applications of data mining. Part IV presents developing areas, including web mining, text mining, and ethical aspects of data mining. Part I is overview material. Part II contains chapters on a number of different techniques often used in data mining. Not all of these chapters need to be covered, and their sequence can be varied according to instructor need. Part III covers applications, and while we feel that these chapters contain the most interesting and important material, instructors who wish to focus on techniques might not wish to cover these chapters. Conversely, instructors more interested in business applications can cover Part III before reviewing content as needed in Part II. This approach would work especially well if data mining software is available to do the modeling. Part IV contains material we feel is important now and is growing in importance. However, again, coverage and sequence is up to the instructor.

The book includes short vignettes of how specific concepts have been applied in real business. A series of representative data sets are generated to demonstrate specific methods and concepts. References to data mining software and sites such as www.kdnuggets.com are provided.

Supplements accompanying this text include (1) an instructor's CD-ROM, containing a solutions guide, PowerPoint slides, and the data set; (2) a student's CD-ROM, containing PowerPoint slides and the data set; and (3) an online learning center.

### Part I: Introduction

Chapter 1 gives an overview of data mining and provides a description of the data mining process. An overview of useful business applications is provided. Chapter 2 presents the data mining process in more detail. It demonstrates this process with a typical set of data. Visualization of data through data mining software is addressed. Chapter 3 presents database support to data mining. Different software tools are described, from data warehouse products through data marts to online analytic

processing. Data quality is addressed. Again, different concepts are demonstrated through prototypical data.

## Part II: Data Mining Methods as Tools

Chapter 4 provides an overview of data mining techniques and functions. Chapter 5 describes and demonstrates clustering algorithms. Software product support available is reviewed. Chapter 6 reviews various forms of regression tools to identify the best fit over given data sets. Chapter 7 discusses neural networks, a popular application of artificial intelligence suitable for many data mining applications. Chapter 8 reviews decision tree algorithms. The basic algorithm is described, along with descriptions of tree structure, machine learning, and fuzzy set aspects of decision trees. Software products are reviewed, and See5 is demonstrated. Chapter 9 presents linear programming-based methods of fitting data. Real data mining applications are described and demonstrated.

## Part III: Business Applications

Chapter 10 reviews the major applications of data mining in business, focusing on the value of these analyses to business decision making. This includes the important topics of customer relationship management. The concept of lift is described. The development of market segmentation by Fingerhut Inc. is reviewed. Chapter 11 describes market-basket analysis, a more qualitative data mining technique. This methodology is described through an example reported in the practitioner literature, and the fundamental data mining concepts of actionability, affinity positioning, and cross-selling are described.

## Part IV: Developing Issues

Chapter 12 presents text and web mining. Chapter 13 discusses ethical issues related to data mining.

*David L. Olson, University of Nebraska–Lincoln*

*Yong Shi, University of Nebraska–Omaha*

# Brief Table

# Contents

PART ONE

# Introduction

CHAPTER ONE

# Initial Description of Data Mining in Business

This chapter:

- Introduces data mining concepts
- Presents typical business data mining applications
- Explains the meaning of key concepts
- Gives a brief overview of data mining tools
- Outlines the remaining chapters of the book

Our culture has developed the ability to generate masses of data. Computer systems expand much faster than the human ability to absorb. Furthermore, Internet connections make it possible to share data in real time on a global basis.

Recent political events emphasize the existence of data to predict. Some blame the system when terrorist strikes are not prevented, because if you dig deep enough, you can always find some data or a memo that pointed to the coming occurrence of these events. However, you would also find a great deal more data predicting things that didn't happen. Obviously, there's a clear need for many organizations to be able to process data faster and more reliably. Data mining involves the use of analysis to detect patterns and allow predictions. Though it's not a perfect science, the intent of data mining is to gain small advantages, because perfect predictions are impossible. These small advantages can be extremely profitable to business. For instance, retail sales organizations have developed sophisticated customer segmentation models to save them from sending sales materials to consumers who likely won't purchase their products, focusing instead on those segments with a higher probability of sales. Banks and other organizations have developed sophisticated customer relationship management programs (supported by data mining) that can predict the value of specific types of customers to that organization, and predict repayment of loans as well. Insurance companies have long applied statistical analysis, which has been extended by data mining tools to aid in the prediction of fraudulent claims. These are only three of many important data mining applications to business.

This book seeks to describe some business applications of data mining. It also will describe the general process of data mining, those database tools needed to support data mining, and the techniques available for data mining.

## Introduction

Data mining refers to the analysis of the large quantities of data that are stored in computers. For example, grocery stores have large amounts of data generated by our purchases. Bar coding has made checking out very convenient for us, and provides retail establishments with masses of data. Grocery stores and other retail stores are able to quickly process our purchases, and use computers to accurately determine product prices. These same computers can help the stores with their inventory management, by instantaneously determining the quantity of items of each product on hand. They are also able to apply computer technology to contact their vendors so they don't run out of the things that we want to purchase. Computers allow the store's accounting system to more accurately measure costs, and determine the profit that store stockholders are concerned about. All of this information is available based upon the bar code information attached to each product. Along with many other sources of information, data gathered through bar coding can be used for data mining analysis.

Data mining is not limited to business. Both major parties in the 2004 U.S. election utilized data mining of potential voters.[1] Data mining has been heavily used in the medical field, to include patient diagnosis records to help identify best practices.[2] The Mayo Clinic worked with IBM to develop an online computer system to identify how that last 100 Mayo patients with the same gender, age, and medical history responded to particular treatments.[3]

Business use of data mining is also impressive. Toyota used the data mining of its **data warehouse** to determine more efficient transportation routes, reducing the time to deliver cars to customers by an average of 19 days. Data warehouses (to be discussed in Chapter 3) are enormous database systems capable of systematically storing all transactional data generated by a business organization, such as WalMart. Toyota also was able to identify sales trends faster, and identify the best locations for new dealerships. Benefits were estimated to be $30 million per year in North America.[4]

Data mining is widely used by banking firms in soliciting credit card customers,[5] by insurance and telecommunication companies in detecting fraud,[6] by manufacturing firms in quality control,[7] and many other applications. Data mining is being applied to improve food product safety,[8] criminal detection,[9] and tourism.[10] Fingerhut has become very successful in **micromarketing,** targeting small groups of highly responsive customers. Media companies such as R. R. Donnelly & Sons provide consumer and life-style data, as well as customized individual publications to firms that use data mining for catalog marketing.

Data mining involves statistical and/or artificial intelligence analysis, usually applied to large-scale data sets. Traditional statistical analysis involves an approach that is usually directed, in that a specific set of expected outcomes exists. This approach is referred to as supervised. However, there is more to data mining than the technical tools used. Data mining involves a spirit of **knowledge discovery** (learning new and useful things), which is referred to as unsupervised. Much of this can be accomplished through automatic means, as we will see in decision tree analysis, for example. But data mining is not limited to automated analysis. Knowledge discovery by humans can be enhanced by graphical tools and the identification of unexpected patterns through a combination of human and computer interaction.

Data mining can be used by businesses in many ways. Three examples are

- **Customer profiling** Identifying those subsets of customers that are most profitable to the business
- **Targeting** Determining the characteristics of profitable customers who have been captured by competitors
- **Market-basket analysis** Determining product purchases by consumers, which can be used for product positioning and for cross-selling.

These are not the only applications of data mining, but they are three of the most important to businesses. Customer profiling is a key part of customer relationship management (CRM), which will be elaborated upon in Chapter 10. Targeting is a key concept in managing **churn,** or customer turnover, also discussed in Chapter 10. Market-basket analysis is an interesting use of data mining that we discuss in Chapter 11.

## What Is Needed to Do Data Mining?

Data mining requires the identification of a problem, along with collection of data that can lead to a better understanding of the market, and computer models to provide statistical or other means of analysis. There are two general types of data mining studies. **Hypothesis testing** involves expressing a theory about the relationship between actions and outcomes. In a simple form, it can be hypothesized that advertising will yield greater profit. This relationship has long been studied by retailing firms in the context of their specific operations. Data mining is applied to identifying relationships based on large quantities of data, which could include testing the response rates to various types of advertising on the sales and profitability of specific product lines. The second form of data mining study is knowledge discovery. In this form of analysis, a preconceived notion may not be present, but rather relationships can be identified by looking at the data. This may be supported by visualization tools that display data, or through fundamental statistical analysis, such as correlation analysis.

A variety of analytic computer models have been used in data mining. Chapters 5 through 9 of this book will discuss various types of these models. Also required is access to data. Quite often, systems including data warehouses and data marts are used to manage large quantities of data (see Chapter 3). Other data mining analyses are done with smaller sets of data, such as can be organized in online analytic processing systems.

## Data Mining

Data mining has been called exploratory data analysis, among other things. Masses of data generated from cash registers, from scanning, and from topic-specific databases throughout the company are explored, analyzed, reduced, and reused. Searches are performed across different models proposed for predicting sales, marketing response, and profit. Classical statistical approaches are fundamental to data mining. Automated AI methods are also used. However, systematic exploration through classical statistical methods is still the basis of data mining. Some of the tools developed by the field of statistical analysis are harnessed through automatic control (with some key human guidance) in dealing with data.