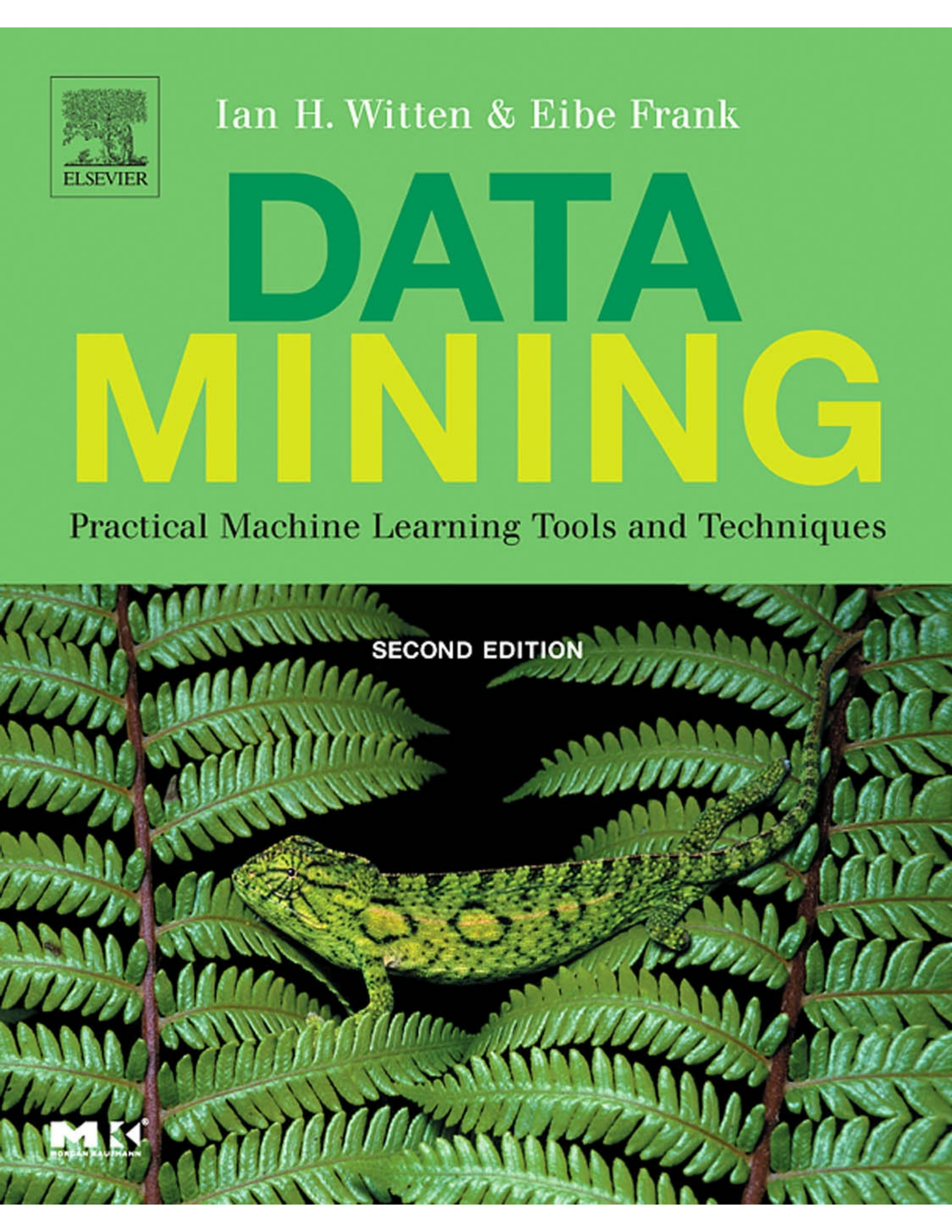

ELSEVIER
Ian H. Witten & Eibe Frank
DATA
MINING
Practical Machine Learning Tools and Techniques
SECOND EDITION
MK
MORGAN KAUFMANN

Data Mining

Practical Machine Learning Tools and Techniques

The Morgan Kaufmann Series in Data Management Systems

Series Editor: Jim Gray, Microsoft Research

Data Mining: Practical Machine Learning Tools and Techniques, Second Edition
Ian H. Witten and Eibe Frank

Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration
Earl Cox

Data Modeling Essentials, Third Edition
Graeme C. Simsion and Graham C. Witt

Location-Based Services
Jochen Schiller and Agnès Voisard

Database Modeling with Microsoft® Visio for Enterprise Architects
Terry Halpin, Ken Evans, Patrick Hallock, and Bill Maclean

Designing Data-Intensive Web Applications
Stefano Ceri, Piero Fraternali, Aldo Bongio, Marco Brambilla, Sara Comai, and Maristella Matera

Mining the Web: Discovering Knowledge from Hypertext Data
Soumen Chakrabarti

Advanced SQL: 1999—Understanding Object-Relational and Other Advanced Features
Jim Melton

Database Tuning: Principles, Experiments, and Troubleshooting Techniques
Dennis Shasha and Philippe Bonnet

SQL: 1999—Understanding Relational Language Components
Jim Melton and Alan R. Simon

Information Visualization in Data Mining and Knowledge Discovery
Edited by Usama Fayyad, Georges G. Grinstein, and Andreas Wierse

Transactional Information Systems: Theory, Algorithms, and the Practice of Concurrency Control and Recovery
Gerhard Weikum and Gottfried Vossen

Spatial Databases: With Application to GIS
Philippe Rigaux, Michel Scholl, and Agnès Voisard

Information Modeling and Relational Databases: From Conceptual Analysis to Logical Design
Terry Halpin

Component Database Systems
Edited by Klaus R. Dittrich and Andreas Geppert

Managing Reference Data in Enterprise Databases: Binding Corporate Data to the Wider World
Malcolm Chisholm

Data Mining: Concepts and Techniques
Jiawei Han and Micheline Kamber

Understanding SQL and Java Together: A Guide to SQLJ, JDBC, and Related Technologies
Jim Melton and Andrew Eisenberg

Database: Principles, Programming, and Performance, Second Edition
Patrick O'Neil and Elizabeth O'Neil

The Object Data Standard: ODMG 3.0
Edited by R. G. G. Cattell, Douglas K. Barry, Mark Berler, Jeff Eastman, David Jordan, Craig Russell, Olaf Schadow, Torsten Stanienda, and Fernando Velez

Data on the Web: From Relations to Semistructured Data and XML
Serge Abiteboul, Peter Buneman, and Dan Suciu

Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations
Ian H. Witten and Eibe Frank

Joe Celko's SQL for Smarties: Advanced SQL Programming, Second Edition
Joe Celko

Joe Celko's Data and Databases: Concepts in Practice
Joe Celko

Developing Time-Oriented Database Applications in SQL
Richard T. Snodgrass

Web Farming for the Data Warehouse
Richard D. Hackathorn

Database Modeling & Design, Third Edition
Toby J. Teorey

Management of Heterogeneous and Autonomous Database Systems
Edited by Ahmed Elmagarmid, Marek Rusinkiewicz, and Amit Sheth

Object-Relational DBMSs: Tracking the Next Great Wave, Second Edition
Michael Stonebraker and Paul Brown, with Dorothy Moore

A Complete Guide to DB2 Universal Database
Don Chamberlin

Universal Database Management: A Guide to Object/Relational Technology
Cynthia Maro Saracco

Readings in Database Systems, Third Edition
Edited by Michael Stonebraker and Joseph M. Hellerstein

Understanding SQL's Stored Procedures: A Complete Guide to SQL/PSM
Jim Melton

Principles of Multimedia Database Systems
V. S. Subrahmanian

Principles of Database Query Processing for Advanced Applications
Clement T. Yu and Weiyi Meng

Advanced Database Systems
Carlo Zaniolo, Stefano Ceri, Christos Faloutsos, Richard T. Snodgrass, V. S. Subrahmanian, and Roberto Zicari

Principles of Transaction Processing for the Systems Professional
Philip A. Bernstein and Eric Newcomer

Using the New DB2: IBM's Object-Relational Database System
Don Chamberlin

Distributed Algorithms
Nancy A. Lynch

Active Database Systems: Triggers and Rules For Advanced Database Processing
Edited by Jennifer Widom and Stefano Ceri

Migrating Legacy Systems: Gateways, Interfaces & the Incremental Approach
Michael L. Brodie and Michael Stonebraker

Atomic Transactions
Nancy Lynch, Michael Merritt, William Weihl, and Alan Fekete

Query Processing For Advanced Database Systems
Edited by Johann Christoph Freytag, David Maier, and Gottfried Vossen

Transaction Processing: Concepts and Techniques
Jim Gray and Andreas Reuter

Building an Object-Oriented Database System: The Story of O_2
Edited by François Bancilhon, Claude Delobel, and Paris Kanellakis

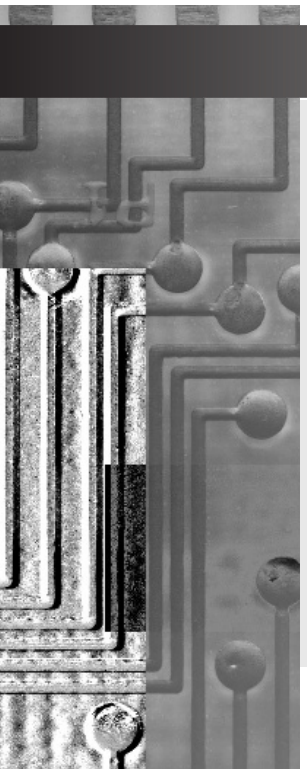
Database Transaction Models For Advanced Applications
Edited by Ahmed K. Elmagarmid

A Guide to Developing Client/Server SQL Applications
Setrag Khoshafian, Arvola Chan, Anna Wong, and Harry K. T. Wong

The Benchmark Handbook For Database and Transaction Processing Systems, Second Edition
Edited by Jim Gray

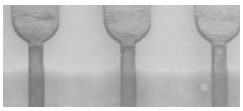
Camelot and Avalon: A Distributed Transaction Facility
Edited by Jeffrey L. Eppinger, Lily B. Mummert, and Alfred Z. Spector

Readings in Object-Oriented Database Systems
Edited by Stanley B. Zdonik and David Maier



Data Mining

Practical Machine Learning Tools and Techniques, Second Edition



Ian H. Witten

Department of Computer Science
University of Waikato

Eibe Frank

Department of Computer Science
University of Waikato

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

MORGAN KAUFMANN PUBLISHERS IS AN IMPRINT OF ELSEVIER

MORGAN KAUFMANN PUBLISHERS

Publisher:	Diane Cerra
Publishing Services Manager:	Simon Crump
Project Manager:	Brandy Lilly
Editorial Assistant:	Asma Stephan
Cover Design:	Yvo Riezebos Design
Cover Image:	Getty Images
Composition:	SNP Best-set Typesetter Ltd., Hong Kong
Technical Illustration:	Dartmouth Publishing, Inc.
Copyeditor:	Graphic World Inc.
Proofreader:	Graphic World Inc.
Indexer:	Graphic World Inc.
Interior printer:	The Maple-Vail Book Manufacturing Group
Cover printer:	Phoenix Color Corp

Morgan Kaufmann Publishers is an imprint of Elsevier.
500 Sansome Street, Suite 400, San Francisco, CA 94111

This book is printed on acid-free paper.

Library of Congress Cataloging-in-Publication Data

Witten, I. H. (Ian H.)
Data mining : practical machine learning tools and techniques / Ian H. Witten, Eibe Frank. – 2nd ed.
p. cm. – (Morgan Kaufmann series in data management systems)
Includes bibliographical references and index.
ISBN: 0-12-088407-0
1. Data mining. I. Frank, Eibe. II. Title. III. Series.

QA76.9.D343W58 2005
006.3–dc22 2005043385

For information on all Morgan Kaufmann publications,
visit our Web site at *www.mkp.com* or *www.books.elsevier.com*

Printed in the United States of America
05 06 07 08 09 5 4 3 2 1

Foreword

Jim Gray, Series Editor
Microsoft Research

Technology now allows us to capture and store vast quantities of data. Finding patterns, trends, and anomalies in these datasets, and summarizing them with simple quantitative models, is one of the grand challenges of the information age—turning data into information and turning information into knowledge.

There has been stunning progress in data mining and machine learning. The synthesis of statistics, machine learning, information theory, and computing has created a solid science, with a firm mathematical base, and with very powerful tools. Witten and Frank present much of this progress in this book and in the companion implementation of the key algorithms. As such, this is a milestone in the synthesis of data mining, data analysis, information theory, and machine learning. If you have not been following this field for the last decade, this is a great way to catch up on this exciting progress. If you have, then Witten and Frank's presentation and the companion open-source workbench, called Weka, will be a useful addition to your toolkit.

They present the basic theory of automatically extracting models from data, and then validating those models. The book does an excellent job of explaining the various models (decision trees, association rules, linear models, clustering, Bayes nets, neural nets) and how to apply them in practice. With this basis, they then walk through the steps and pitfalls of various approaches. They describe how to safely scrub datasets, how to build models, and how to evaluate a model's predictive quality. Most of the book is tutorial, but Part II broadly describes how commercial systems work and gives a tour of the publicly available data mining workbench that the authors provide through a website. This Weka workbench has a graphical user interface that leads you through data mining tasks and has excellent data visualization tools that help understand the models. It is a great companion to the text and a useful and popular tool in its own right.

This book presents this new discipline in a very accessible form: as a text both to train the next generation of practitioners and researchers and to inform lifelong learners like myself. Witten and Frank have a passion for simple and elegant solutions. They approach each topic with this mindset, grounding all concepts in concrete examples, and urging the reader to consider the simple techniques first, and then progress to the more sophisticated ones if the simple ones prove inadequate.

If you are interested in databases, and have not been following the machine learning field, this book is a great way to catch up on this exciting progress. If you have data that you want to analyze and understand, this book and the associated Weka toolkit are an excellent way to start.

Contents

List of Figures

List of Tables

Preface

The convergence of computing and communication has produced a society that feeds on information. Yet most of the information is in its raw form: data. If *data* is characterized as recorded facts, then *information* is the set of patterns, or expectations, that underlie the data. There is a huge amount of information locked up in databases—information that is potentially important but has not yet been discovered or articulated. Our mission is to bring it forth.

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns. Strong patterns, if found, will likely generalize to make accurate predictions on future data. Of course, there will be problems. Many patterns will be banal and uninteresting. Others will be spurious, contingent on accidental coincidences in the particular dataset used. In addition real data is imperfect: Some parts will be garbled, and some will be missing. Anything discovered will be inexact: There will be exceptions to every rule and cases not covered by any rule. Algorithms need to be robust enough to cope with imperfect data and to extract regularities that are inexact but useful.

Machine learning provides the technical basis of data mining. It is used to extract information from the raw data in databases—information that is expressed in a comprehensible form and can be used for a variety of purposes. The process is one of abstraction: taking the data, warts and all, and inferring whatever structure underlies it. This book is about the tools and techniques of machine learning used in practical data mining for finding, and describing, structural patterns in data.

As with any burgeoning new technology that enjoys intense commercial attention, the use of data mining is surrounded by a great deal of hype in the technical—and sometimes the popular—press. Exaggerated reports appear of the secrets that can be uncovered by setting learning algorithms loose on oceans of data. But there is no magic in machine learning, no hidden power, no

alchemy. Instead, there is an identifiable body of simple and practical techniques that can often extract useful information from raw data. This book describes these techniques and shows how they work.

We interpret machine learning as the acquisition of structural descriptions from examples. The kind of descriptions found can be used for prediction, explanation, and understanding. Some data mining applications focus on prediction: forecasting what will happen in new situations from data that describe what happened in the past, often by guessing the classification of new examples. But we are equally—perhaps more—interested in applications in which the result of "learning" is an actual description of a structure that can be used to classify examples. This structural description supports explanation, understanding, and prediction. In our experience, insights gained by the applications' users are of most interest in the majority of practical data mining applications; indeed, this is one of machine learning's major advantages over classical statistical modeling.

The book explains a variety of machine learning methods. Some are pedagogically motivated: simple schemes designed to explain clearly how the basic ideas work. Others are practical: real systems used in applications today. Many are contemporary and have been developed only in the last few years.

A comprehensive software resource, written in the Java language, has been created to illustrate the ideas in the book. Called the Waikato Environment for Knowledge Analysis, or Weka[1] for short, it is available as source code on the World Wide Web at *http://www.cs.waikato.ac.nz/ml/weka.* It is a full, industrial-strength implementation of essentially all the techniques covered in this book. It includes illustrative code and working implementations of machine learning methods. It offers clean, spare implementations of the simplest techniques, designed to aid understanding of the mechanisms involved. It also provides a workbench that includes full, working, state-of-the-art implementations of many popular learning schemes that can be used for practical data mining or for research. Finally, it contains a framework, in the form of a Java class library, that supports applications that use embedded machine learning and even the implementation of new learning schemes.

The objective of this book is to introduce the tools and techniques for machine learning that are used in data mining. After reading it, you will understand what these techniques are and appreciate their strengths and applicability. If you wish to experiment with your own data, you will be able to do this easily with the Weka software.

---

[1] Found only on the islands of New Zealand, the *weka* (pronounced to rhyme with *Mecca*) is a flightless bird with an inquisitive nature.

The book spans the gulf between the intensely practical approach taken by trade books that provide case studies on data mining and the more theoretical, principle-driven exposition found in current textbooks on machine learning. (A brief description of these books appears in the *Further reading* section at the end of Chapter 1.) This gulf is rather wide. To apply machine learning techniques productively, you need to understand something about how they work; this is not a technology that you can apply blindly and expect to get good results. Different problems yield to different techniques, but it is rarely obvious which techniques are suitable for a given situation: you need to know something about the range of possible solutions. We cover an extremely wide range of techniques. We can do this because, unlike many trade books, this volume does not promote any particular commercial software or approach. We include a large number of examples, but they use illustrative datasets that are small enough to allow you to follow what is going on. Real datasets are far too large to show this (and in any case are usually company confidential). Our datasets are chosen not to illustrate actual large-scale practical problems but to help you understand what the different techniques do, how they work, and what their range of application is.

The book is aimed at the technically aware general reader interested in the principles and ideas underlying the current practice of data mining. It will also be of interest to information professionals who need to become acquainted with this new technology and to all those who wish to gain a detailed technical understanding of what machine learning involves. It is written for an eclectic audience of information systems practitioners, programmers, consultants, developers, information technology managers, specification writers, patent examiners, and curious laypeople—as well as students and professors—who need an easy-to-read book with lots of illustrations that describes what the major machine learning techniques are, what they do, how they are used, and how they work. It is practically oriented, with a strong "how to" flavor, and includes algorithms, code, and implementations. All those involved in practical data mining will benefit directly from the techniques described. The book is aimed at people who want to cut through to the reality that underlies the hype about machine learning and who seek a practical, nonacademic, unpretentious approach. We have avoided requiring any specific theoretical or mathematical knowledge except in some sections marked by a light gray bar in the margin. These contain optional material, often for the more technical or theoretically inclined reader, and may be skipped without loss of continuity.

The book is organized in layers that make the ideas accessible to readers who are interested in grasping the basics and to those who would like more depth of treatment, along with full details on the techniques covered. We believe that consumers of machine learning need to have some idea of how the algorithms they use work. It is often observed that data models are only as good as the person

who interprets them, and that person needs to know something about how the models are produced to appreciate the strengths, and limitations, of the technology. However, it is not necessary for all data model users to have a deep understanding of the finer details of the algorithms.

We address this situation by describing machine learning methods at successive levels of detail. You will learn the basic ideas, the topmost level, by reading the first three chapters. Chapter 1 describes, through examples, what machine learning is and where it can be used; it also provides actual practical applications. Chapters 2 and 3 cover the kinds of input and output—or *knowledge representation*—involved. Different kinds of output dictate different styles of algorithm, and at the next level Chapter 4 describes the basic methods of machine learning, simplified to make them easy to comprehend. Here the principles involved are conveyed in a variety of algorithms without getting into intricate details or tricky implementation issues. To make progress in the application of machine learning techniques to particular data mining problems, it is essential to be able to measure how well you are doing. Chapter 5, which can be read out of sequence, equips you to evaluate the results obtained from machine learning, addressing the sometimes complex issues involved in performance evaluation.

At the lowest and most detailed level, Chapter 6 exposes in naked detail the nitty-gritty issues of implementing a spectrum of machine learning algorithms, including the complexities necessary for them to work well in practice. Although many readers may want to ignore this detailed information, it is at this level that the full, working, tested implementations of machine learning schemes in Weka are written. Chapter 7 describes practical topics involved with engineering the input to machine learning—for example, selecting and discretizing attributes—and covers several more advanced techniques for refining and combining the output from different learning techniques. The final chapter of Part I looks to the future.

The book describes most methods used in practical machine learning. However, it does not cover reinforcement learning, because it is rarely applied in practical data mining; genetic algorithm approaches, because these are just an optimization technique; or relational learning and inductive logic programming, because they are rarely used in mainstream data mining applications.

The data mining system that illustrates the ideas in the book is described in Part II to clearly separate conceptual material from the practical aspects of how to use it. You can skip to Part II directly from Chapter 4 if you are in a hurry to analyze your data and don't want to be bothered with the technical details.

Java has been chosen for the implementations of machine learning techniques that accompany this book because, as an object-oriented programming language, it allows a uniform interface to learning schemes and methods for pre- and postprocessing. We have chosen Java instead of C++, Smalltalk, or other

object-oriented languages because programs written in Java can be run on almost any computer without having to be recompiled, having to undergo complicated installation procedures, or—worst of all—having to change the code. A Java program is compiled into byte-code that can be executed on any computer equipped with an appropriate interpreter. This interpreter is called the *Java virtual machine.* Java virtual machines—and, for that matter, Java compilers—are freely available for all important platforms.

Like all widely used programming languages, Java has received its share of criticism. Although this is not the place to elaborate on such issues, in several cases the critics are clearly right. However, of all currently available programming languages that are widely supported, standardized, and extensively documented, Java seems to be the best choice for the purpose of this book. Its main disadvantage is speed of execution—or lack of it. Executing a Java program is several times slower than running a corresponding program written in C language because the virtual machine has to translate the byte-code into machine code before it can be executed. In our experience the difference is a factor of three to five if the virtual machine uses a just-in-time compiler. Instead of translating each byte-code individually, a *just-in-time compiler* translates whole chunks of byte-code into machine code, thereby achieving significant speedup. However, if this is still to slow for your application, there are compilers that translate Java programs directly into machine code, bypassing the byte-code step. This code cannot be executed on other platforms, thereby sacrificing one of Java's most important advantages.

Updated and revised content

We finished writing the first edition of this book in 1999 and now, in April 2005, are just polishing this second edition. The areas of data mining and machine learning have matured in the intervening years. Although the core of material in this edition remains the same, we have made the most of our opportunity to update it to reflect the changes that have taken place over 5 years. There have been errors to fix, errors that we had accumulated in our publicly available errata file. Surprisingly few were found, and we hope there are even fewer in this second edition. (The errata for the second edition may be found through the book's home page at *http://www.cs.waikato.ac.nz/ml/weka/book.html.*) We have thoroughly edited the material and brought it up to date, and we practically doubled the number of references. The most enjoyable part has been adding new material. Here are the highlights.

Bowing to popular demand, we have added comprehensive information on neural networks: the perceptron and closely related Winnow algorithm in Section 4.6 and the multilayer perceptron and backpropagation algorithm

in Section 6.3. We have included more recent material on implementing nonlinear decision boundaries using both the kernel perceptron and radial basis function networks. There is a new section on Bayesian networks, again in response to readers' requests, with a description of how to learn classifiers based on these networks and how to implement them efficiently using all-dimensions trees.

The Weka machine learning workbench that accompanies the book, a widely used and popular feature of the first edition, has acquired a radical new look in the form of an interactive interface—or rather, three separate interactive interfaces—that make it far easier to use. The primary one is the Explorer, which gives access to all of Weka's facilities using menu selection and form filling. The others are the Knowledge Flow interface, which allows you to design configurations for streamed data processing, and the Experimenter, with which you set up automated experiments that run selected machine learning algorithms with different parameter settings on a corpus of datasets, collect performance statistics, and perform significance tests on the results. These interfaces lower the bar for becoming a practicing data miner, and we include a full description of how to use them. However, the book continues to stand alone, independent of Weka, and to underline this we have moved all material on the workbench into a separate Part II at the end of the book.

In addition to becoming far easier to use, Weka has grown over the last 5 years and matured enormously in its data mining capabilities. It now includes an unparalleled range of machine learning algorithms and related techniques. The growth has been partly stimulated by recent developments in the field and partly led by Weka users and driven by demand. This puts us in a position in which we know a great deal about what actual users of data mining want, and we have capitalized on this experience when deciding what to include in this new edition.

The earlier chapters, containing more general and foundational material, have suffered relatively little change. We have added more examples of fielded applications to Chapter 1, a new subsection on sparse data and a little on string attributes and date attributes to Chapter 2, and a description of interactive decision tree construction, a useful and revealing technique to help you grapple with your data using manually built decision trees, to Chapter 3.

In addition to introducing linear decision boundaries for classification, the infrastructure for neural networks, Chapter 4 includes new material on multinomial Bayes models for document classification and on logistic regression. The last 5 years have seen great interest in data mining for text, and this is reflected in our introduction to string attributes in Chapter 2, multinomial Bayes for document classification in Chapter 4, and text transformations in Chapter 7. Chapter 4 includes a great deal of new material on efficient data structures for searching the instance space: *k*D-trees and the recently invented ball trees. These

are used to find nearest neighbors efficiently and to accelerate distance-based clustering.

Chapter 5 describes the principles of statistical evaluation of machine learning, which have not changed. The main addition, apart from a note on the Kappa statistic for measuring the success of a predictor, is a more detailed treatment of cost-sensitive learning. We describe how to use a classifier, built without taking costs into consideration, to make predictions that are sensitive to cost; alternatively, we explain how to take costs into account during the training process to build a cost-sensitive model. We also cover the popular new technique of cost curves.

There are several additions to Chapter 6, apart from the previously mentioned material on neural networks and Bayesian network classifiers. More details—gory details—are given of the heuristics used in the successful RIPPER rule learner. We describe how to use model trees to generate rules for numeric prediction. We show how to apply locally weighted regression to classification problems. Finally, we describe the *X*-means clustering algorithm, which is a big improvement on traditional *k*-means.

Chapter 7 on engineering the input and output has changed most, because this is where recent developments in practical machine learning have been concentrated. We describe new attribute selection schemes such as race search and the use of support vector machines and new methods for combining models such as additive regression, additive logistic regression, logistic model trees, and option trees. We give a full account of LogitBoost (which was mentioned in the first edition but not described). There is a new section on useful transformations, including principal components analysis and transformations for text mining and time series. We also cover recent developments in using unlabeled data to improve classification, including the co-training and co-EM methods.

The final chapter of Part I on new directions and different perspectives has been reworked to keep up with the times and now includes contemporary challenges such as adversarial learning and ubiquitous data mining.

Acknowledgments

Writing the acknowledgments is always the nicest part! A lot of people have helped us, and we relish this opportunity to thank them. This book has arisen out of the machine learning research project in the Computer Science Department at the University of Waikato, New Zealand. We have received generous encouragement and assistance from the academic staff members on that project: John Cleary, Sally Jo Cunningham, Matt Humphrey, Lyn Hunt, Bob McQueen, Lloyd Smith, and Tony Smith. Special thanks go to Mark Hall, Bernhard Pfahringer, and above all Geoff Holmes, the project leader and source of inspi-

ration. All who have worked on the machine learning project here have contributed to our thinking: we would particularly like to mention Steve Garner, Stuart Inglis, and Craig Nevill-Manning for helping us to get the project off the ground in the beginning when success was less certain and things were more difficult.

The Weka system that illustrates the ideas in this book forms a crucial component of it. It was conceived by the authors and designed and implemented by Eibe Frank, along with Len Trigg and Mark Hall. Many people in the machine learning laboratory at Waikato made significant contributions. Since the first edition of the book the Weka team has expanded considerably: so many people have contributed that it is impossible to acknowledge everyone properly. We are grateful to Remco Bouckaert for his implementation of Bayesian networks, Dale Fletcher for many database-related aspects, Ashraf Kibriya and Richard Kirkby for contributions far too numerous to list, Niels Landwehr for logistic model trees, Abdelaziz Mahoui for the implementation of K*, Stefan Mutter for association rule mining, Gabi Schmidberger and Malcolm Ware for numerous miscellaneous contributions, Tony Voyle for least-median-of-squares regression, Yong Wang for Pace regression and the implementation of M5′, and Xin Xu for *JRip*, logistic regression, and many other contributions. Our sincere thanks go to all these people for their dedicated work and to the many contributors to Weka from outside our group at Waikato.

Tucked away as we are in a remote (but very pretty) corner of the Southern Hemisphere, we greatly appreciate the visitors to our department who play a crucial role in acting as sounding boards and helping us to develop our thinking. We would like to mention in particular Rob Holte, Carl Gutwin, and Russell Beale, each of whom visited us for several months; David Aha, who although he only came for a few days did so at an early and fragile stage of the project and performed a great service by his enthusiasm and encouragement; and Kai Ming Ting, who worked with us for 2 years on many of the topics described in Chapter 7 and helped to bring us into the mainstream of machine learning.

Students at Waikato have played a significant role in the development of the project. Jamie Littin worked on ripple-down rules and relational learning. Brent Martin explored instance-based learning and nested instance-based representations. Murray Fife slaved over relational learning, and Nadeeka Madapathage investigated the use of functional languages for expressing machine learning algorithms. Other graduate students have influenced us in numerous ways, particularly Gordon Paynter, YingYing Wen, and Zane Bray, who have worked with us on text mining. Colleagues Steve Jones and Malika Mahoui have also made far-reaching contributions to these and other machine learning projects. More recently we have learned much from our many visiting students from Freiburg, including Peter Reutemann and Nils Weidmann.

Ian Witten would like to acknowledge the formative role of his former students at Calgary, particularly Brent Krawchuk, Dave Maulsby, Thong Phan, and Tanja Mitrovic, all of whom helped him develop his early ideas in machine learning, as did faculty members Bruce MacDonald, Brian Gaines, and David Hill at Calgary and John Andreae at the University of Canterbury.

Eibe Frank is indebted to his former supervisor at the University of Karlsruhe, Klaus-Peter Huber (now with SAS Institute), who infected him with the fascination of machines that learn. On his travels Eibe has benefited from interactions with Peter Turney, Joel Martin, and Berry de Bruijn in Canada and with Luc de Raedt, Christoph Helma, Kristian Kersting, Stefan Kramer, Ulrich Rückert, and Ashwin Srinivasan in Germany.

Diane Cerra and Asma Stephan of Morgan Kaufmann have worked hard to shape this book, and Lisa Royse, our production editor, has made the process go smoothly. Bronwyn Webster has provided excellent support at the Waikato end.

We gratefully acknowledge the unsung efforts of the anonymous reviewers, one of whom in particular made a great number of pertinent and constructive comments that helped us to improve this book significantly. In addition, we would like to thank the librarians of the Repository of Machine Learning Databases at the University of California, Irvine, whose carefully collected datasets have been invaluable in our research.

Our research has been funded by the New Zealand Foundation for Research, Science and Technology and the Royal Society of New Zealand Marsden Fund. The Department of Computer Science at the University of Waikato has generously supported us in all sorts of ways, and we owe a particular debt of gratitude to Mark Apperley for his enlightened leadership and warm encouragement. Part of the first edition was written while both authors were visiting the University of Calgary, Canada, and the support of the Computer Science department there is gratefully acknowledged—as well as the positive and helpful attitude of the long-suffering students in the machine learning course on whom we experimented.

In producing the second edition Ian was generously supported by Canada's Informatics Circle of Research Excellence and by the University of Lethbridge in southern Alberta, which gave him what all authors yearn for—a quiet space in pleasant and convivial surroundings in which to work.

Last, and most of all, we are grateful to our families and partners. Pam, Anna, and Nikki were all too well aware of the implications of having an author in the house ("not again!") but let Ian go ahead and write the book anyway. Julie was always supportive, even when Eibe had to burn the midnight oil in the machine learning lab, and Immo and Ollig provided exciting diversions. Between us we hail from Canada, England, Germany, Ireland, and Samoa: New Zealand has brought us together and provided an ideal, even idyllic, place to do this work.