

HANDBOOK OF STATISTICAL  
ANALYSIS AND DATA MINING  
APPLICATIONS

---

*“Great introduction to the real-world process of data mining. The overviews, practical advice, tutorials, and extra DVD material make this book an invaluable resource for both new and experienced data miners.”*

Karl Rexer, Ph.D.  
(President and Founder of Rexer Analytics, Boston, Massachusetts,  
[www.RexerAnalytics.com](http://www.RexerAnalytics.com))

*“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”*

H. G. Wells (1866 – 1946)

*“Today we aren’t quite to the place that H. G. Wells predicted years ago, but society is getting closer out of necessity. Global businesses and organizations are being forced to use statistical analysis and data mining applications in a format that combines **art** and **science–intuition** and **expertise** in collecting and understanding data in order to make **accurate models** that realistically **predict the future** that lead to informed strategic **decisions** thus allowing correct **actions** ensuring **success**, before it is too late . . . today, numeracy is as essential as literacy. As John Elder likes to say: ‘Go data mining!’ It really does save enormous time and money. For those with the patience and faith to get through the early stages of business understanding and data transformation, the cascade of results can be extremely rewarding.”*

Gary Miner, March, 2009

# HANDBOOK OF STATISTICAL ANALYSIS AND DATA MINING APPLICATIONS

---

ROBERT NISBET

*Pacific Capital Bankcorp N.A.  
Santa Barbara, CA*

JOHN ELDER

*Elder Research, Inc., Charlottesville, VA*

GARY MINER

*StatSoft, Inc., Tulsa, Oklahoma*



ELSEVIER

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

Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier  
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA  
525 B Street, Suite 1900, San Diego, California 92101-4495, USA  
84 Theobald's Road, London WC1X 8RR, UK

Copyright © 2009, Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333, E-mail: [permissions@elsevier.com](mailto:permissions@elsevier.com). You may also complete your request online via the Elsevier homepage (<http://elsevier.com>), by selecting "Support & Contact" then "Copyright and Permission" and then "Obtaining Permissions."

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

Nisbet, Robert, 1942-

Handbook of statistical analysis and data mining applications / Robert Nisbet, John Elder, Gary Miner.

p. cm.

Includes index.

ISBN 978-0-12-374765-5 (hardcover : alk. pager) 1. Data mining--Statistical methods. I. Elder, John F. (John Fletcher) II. Miner, Gary. III. Title.

QA76.9.D343N57 2009

006.3'12--dc22

2009008997

### **British Library Cataloguing-in-Publication Data**

A catalogue record for this book is available from the British Library.

ISBN: 978-0-12-374765-5

For information on all Academic Press publications  
visit our Web site at [www.elsevierdirect.com](http://www.elsevierdirect.com)

Printed in Canada

09 10 9 8 7 6 5 4 3 2 1

Working together to grow  
libraries in developing countries

[www.elsevier.com](http://www.elsevier.com) | [www.bookaid.org](http://www.bookaid.org) | [www.sabre.org](http://www.sabre.org)

ELSEVIER

BOOK AID  
International

Sabre Foundation

# Table of Contents

---

Foreword 1	xv
Foreword 2	xvii
Preface	xix
Introduction	xxiii
List of Tutorials by Guest Authors	xxix

## I

---

### HISTORY OF PHASES OF DATA ANALYSIS, BASIC THEORY, AND THE DATA MINING PROCESS

#### 1. The Background for Data Mining Practice

Preamble	3
A Short History of Statistics and Data Mining	4
Modern Statistics: A Duality?	5
Assumptions of the Parametric Model	6
Two Views of Reality	8
Aristotle	8
Plato	9
The Rise of Modern Statistical Analysis: The Second Generation	10
Data, Data Everywhere ...	11
Machine Learning Methods: The Third Generation	11
Statistical Learning Theory: The Fourth Generation	12
Postscript	13

#### 2. Theoretical Considerations for Data Mining

Preamble	15
The Scientific Method	16
What Is Data Mining?	17

A Theoretical Framework for the Data Mining Process	18
Microeconomic Approach	19
Inductive Database Approach	19
Strengths of the Data Mining Process	19
Customer-Centric Versus Account-Centric: A New Way to Look at Your Data	20
The Physical Data Mart	20
The Virtual Data Mart	21
Household Databases	21
The Data Paradigm Shift	22
Creation of the Car	22
Major Activities of Data Mining	23
Major Challenges of Data Mining	25
Examples of Data Mining Applications	26
Major Issues in Data Mining	26
General Requirements for Success in a Data Mining Project	28
Example of a Data Mining Project: Classify a Bat's Species by Its Sound	28
The Importance of Domain Knowledge	30
Postscript	30
Why Did Data Mining Arise?	30
Some Caveats with Data Mining Solutions	31

#### 3. The Data Mining Process

Preamble	33
The Science of Data Mining	33
The Approach to Understanding and Problem Solving	34
CRISP-DM	35
Business Understanding (Mostly Art)	36
Define the Business Objectives of the Data Mining Model	36
Assess the Business Environment for Data Mining	37
Formulate the Data Mining Goals and Objectives	37

## Data Understanding (Mostly Science) 39

Data Acquisition 39

Data Integration 39

Data Description 40

Data Quality Assessment 40

## Data Preparation (A Mixture of Art and Science) 40

## Modeling (A Mixture of Art and Science) 41

Steps in the Modeling Phase of CRISP-DM 41

## Deployment (Mostly Art) 45

## Closing the Information Loop (Art) 46

## The Art of Data Mining 46

Artistic Steps in Data Mining 47

## Postscript 47

## 4. Data Understanding and Preparation

## Preamble 49

## Activities of Data Understanding and Preparation 50

Definitions 50

## Issues That Should be Resolved 51

Basic Issues That Must Be Resolved in Data Understanding 51

Basic Issues That Must Be Resolved in Data Preparation 51

## Data Understanding 51

Data Acquisition 51

Data Extraction 53

Data Description 54

Data Assessment 56

Data Profiling 56

Data Cleansing 56

Data Transformation 57

Data Imputation 59

Data Weighting and Balancing 62

Data Filtering and Smoothing 64

Data Abstraction 66

Data Reduction 69

Data Sampling 69

Data Discretization 73

Data Derivation 73

## Postscript 75

## 5. Feature Selection

Preamble 77

Variables as Features 78

Types of Feature Selections 78

Feature Ranking Methods 78

Gini Index 78

Bi-variate Methods 80

Multivariate Methods 80

Complex Methods 82

Subset Selection Methods 82

The Other Two Ways of Using Feature

Selection in STATISTICA: Interactive  
Workspace 93

STATISTICA DMRecipe Method 93

Postscript 96

6. Accessory Tools for Doing  
Data Mining

Preamble 99

Data Access Tools 100

Structured Query Language (SQL) Tools 100

Extract, Transform, and Load (ETL)

Capabilities 100

Data Exploration Tools 101

Basic Descriptive Statistics 101

Combining Groups (Classes) for Predictive Data  
Mining 105Slicing/Dicing and Drilling Down into Data Sets/  
Results Spreadsheets 106

Modeling Management Tools 107

Data Miner Workspace Templates 107

Modeling Analysis Tools 107

Feature Selection 107

Importance Plots of Variables 108

In-Place Data Processing (IDP) 113

Example: The IDP Facility of STATISTICA Data  
Miner 114

How to Use the SQL 114

Rapid Deployment of Predictive Models 114

Model Monitors 116

Postscript 117

---

## II

---

### THE ALGORITHMS IN DATA MINING AND TEXT MINING, THE ORGANIZATION OF THE THREE MOST COMMON DATA MINING TOOLS, AND SELECTED SPECIALIZED AREAS USING DATA MINING

7. Basic Algorithms for Data Mining: A Brief Overview	8. Advanced Algorithms for Data Mining
Preamble 121	Preamble 151
STATISTICA Data Miner Recipe (DMRecipe) 123	Advanced Data Mining Algorithms 154
KXEN 124	Interactive Trees 154
Basic Data Mining Algorithms 126	Multivariate Adaptive Regression Splines (MARSplines) 158
Association Rules 126	Statistical Learning Theory: Support Vector Machines 162
Neural Networks 128	Sequence, Association, and Link Analyses 164
Radial Basis Function (RBF) Networks 136	Independent Components Analysis (ICA) 168
Automated Neural Nets 138	Kohonen Networks 169
Generalized Additive Models (GAMs) 138	Characteristics of a Kohonen Network 169
Outputs of GAMs 139	Quality Control Data Mining and Root Cause Analysis 169
Interpreting Results of GAMs 139	Image and Object Data Mining: Visualization and 3D-Medical and Other Scanning Imaging 170
Classification and Regression Trees (CART) 139	Postscript 171
Recursive Partitioning 144	
Pruning Trees 144	9. Text Mining and Natural Language Processing
General Comments about CART for Statisticians 144	Preamble 173
Advantages of CART over Other Decision Trees 145	The Development of Text Mining 174
Uses of CART 146	A Practical Example: NTSB 175
General CHAID Models 146	Goals of Text Mining of NTSB Accident Reports 184
Advantages of CHAID 147	Drilling into Words of Interest 188
Disadvantages of CHAID 147	Means with Error Plots 189
Generalized EM and $k$ -Means Cluster Analysis—An Overview 147	Feature Selection Tool 190
$k$ -Means Clustering 147	A Conclusion: Losing Control of the Aircraft in Bad Weather Is Often Fatal 191
EM Cluster Analysis 148	Summary 194
Processing Steps of the EM Algorithm 149	Text Mining Concepts Used in Conducting Text Mining Studies 194
V-fold Cross-Validation as Applied to Clustering 149	Postscript 194
Postscript 150	
	10. The Three Most Common Data Mining Software Tools
	Preamble 197
	SPSS Clementine Overview 197
	Overall Organization of Clementine Components 198
	Organization of the Clementine Interface 199
	Clementine Interface Overview 199
	Setting the Default Directory 201
	SuperNodes 201

Execution of Streams	202
SAS-Enterprise Miner (SAS-EM) Overview	203
Overall Organization of SAS-EM Version 5.3 Components	203
Layout of the SAS-Enterprise Miner Window	204
Various SAS-EM Menus, Dialogs, and Windows Useful During the Data Mining Process	205
Software Requirements to Run SAS-EM 5.3 Software	206
STATISTICA Data Miner, QC-Miner, and Text Miner Overview	214
Overall Organization and Use of STATISTICA Data Miner	214
Three Formats for Doing Data Mining in STATISTICA	230
Postscript	234
<b>11. Classification</b>	
Preamble	235
What Is Classification?	235
Initial Operations in Classification	236
Major Issues with Classification	236
What Is the Nature of Data Set to Be Classified?	236
How Accurate Does the Classification Have to Be?	236
How Understandable Do the Classes Have to Be?	237
Assumptions of Classification Procedures	237
Numerical Variables Operate Best	237
No Missing Values	237
Variables Are Linear and Independent in Their Effects on the Target Variable	237
Methods for Classification	238
Nearest-Neighbor Classifiers	239
Analyzing Imbalanced Data Sets with Machine Learning Programs	240
CHAID	246
Random Forests and Boosted Trees	248
Logistic Regression	250
Neural Networks	251
Naïve Bayesian Classifiers	253
What Is the Best Algorithm for Classification?	256
Postscript	257
<b>12. Numerical Prediction</b>	
Preamble	259
Linear Response Analysis and the Assumptions of the Parametric Model	260
Parametric Statistical Analysis	261
Assumptions of the Parametric Model	262
The Assumption of Independency	262
The Assumption of Normality	262
Normality and the Central Limit Theorem	263
The Assumption of Linearity	264
Linear Regression	264
Methods for Handling Variable Interactions in Linear Regression	265
Collinearity among Variables in a Linear Regression	265
The Concept of the Response Surface	266
Generalized Linear Models (GLMs)	270
Methods for Analyzing Nonlinear Relationships	271
Nonlinear Regression and Estimation	271
Logit and Probit Regression	272
Poisson Regression	272
Exponential Distributions	272
Piecewise Linear Regression	273
Data Mining and Machine Learning Algorithms Used in Numerical Prediction	274
Numerical Prediction with C&RT	274
Model Results Available in C&RT	276
Advantages of Classification and Regression Trees (C&RT) Methods	277
General Issues Related to C&RT	279
Application to Mixed Models	280
Neural Nets for Prediction	280
Manual or Automated Operation?	280
Structuring the Network for Manual Operation	280
Modern Neural Nets Are “Gray Boxes”	281
Example of Automated Neural Net Results	281
Support Vector Machines (SVMs) and Other Kernel Learning Algorithms	282
Postscript	284
<b>13. Model Evaluation and Enhancement</b>	
Preamble	285
Introduction	286
Model Evaluation	286
Splitting Data	287



Avoiding Overfit Through Complexity	
Regularization	288
Error Metric: Estimation	291
Error Metric: Classification	291
Error Metric: Ranking	293
Cross-Validation to Estimate Error Rate and Its Confidence	295
Bootstrap	296
Target Shuffling to Estimate Baseline Performance	297
Re-Cap of the Most Popular Algorithms	300
Linear Methods (Consensus Method, Stepwise Is Variable-Selecting)	300
Decision Trees (Consensus Method, Variable-Selecting)	300
Neural Networks (Consensus Method)	301
Nearest Neighbors (Contributory Method)	301
Clustering (Consensus or Contributory Method)	302
Enhancement Action Checklist	302
Ensembles of Models: The Single Greatest Enhancement Technique	304
Bagging	305
Boosting	305
Ensembles in General	306
How to Thrive as a Data Miner	307
Big Picture of the Project	307
Project Methodology and Deliverables	308
Professional Development	309
Three Goals	310
Postscript	311
14. Medical Informatics	
Preamble	313
What Is Medical Informatics?	313
How Data Mining and Text Mining Relate to Medical Informatics	314
XplorMed	316
ABView: HivResist	317
3D Medical Informatics	317
What Is 3D Informatics?	317
Future and Challenges of 3D Medical Informatics	318
Journals and Associations in the Field of Medical Informatics	318
Postscript	318

## 15. Bioinformatics

Preamble	321
What Is Bioinformatics?	323
Data Analysis Methods in Bioinformatics	326
ClustalW2: Sequence Alignment	326
Searching Databases for RNA Molecules	327
Web Services in Bioinformatics	327
How Do We Apply Data Mining Methods to Bioinformatics?	329
Postscript	332
Tutorial Associated with This Chapter on Bioinformatics	332
Books, Associations, and Journals on Bioinformatics, and Other Resources, Including Online	332
16. Customer Response Modeling	
Preamble	335
Early CRM Issues in Business	336
Knowing How Customers Behaved Before They Acted	336
Transforming Corporations into Business Ecosystems: The Path to Customer Fulfillment	337
CRM in Business Ecosystems	338
Differences Between Static Measures and Evolutionary Measures	338
How Can Human Nature as Viewed Through Plato Help Us in Modeling Customer Response?	339
How Can We Reorganize Our Data to Reflect Motives and Attitudes?	339
What Is a Temporal Abstraction?	340
Conclusions	344
Postscript	345

## 17. Fraud Detection

Preamble	347
Issues with Fraud Detection	348
Fraud Is Rare	348
Fraud Is Evolving!	348
Large Data Sets Are Needed	348
The Fact of Fraud Is Not Always Known during Modeling	348
When the Fraud Happened Is Very Important to Its Detection	349

Fraud Is Very Complex	349
Fraud Detection May Require the Formulation of Rules Based on General Principles, “Red Flags,” Alerts, and Profiles	349
Fraud Detection Requires Both Internal and External Business Data	349
Very Few Data Sets and Modeling Details Are Available	350
How Do You Detect Fraud?	350
Supervised Classification of Fraud	351
How Do You Model Fraud?	352
How Are Fraud Detection Systems Built?	353
Intrusion Detection Modeling	355
Comparison of Models with and without Time-Based Features	355
Building Profiles	360
Deployment of Fraud Profiles	360
Postscript and Prolegomenon	361

### III

## TUTORIALS—STEP-BY-STEP CASE STUDIES AS A STARTING POINT TO LEARN HOW TO DO DATA MINING ANALYSES

### Guest Authors of the Tutorials

<b>A. How to Use Data Miner Recipe</b>	
What is STATISTICA Data Miner Recipe (DMR)?	373
Core Analytic Ingredients	373
<b>B. Data Mining for Aviation Safety</b>	
Airline Safety	378
SDR Database	379
Preparing the Data for Our Tutorial	382
Data Mining Approach	383
Data Mining Algorithm Error Rate	386
Conclusion	387

<b>C. Predicting Movie Box-Office Receipts</b>	
Introduction	391
Data and Variable Definitions	392
Getting to Know the Workspace of the Clementine Data Mining Toolkit	393
Results	396
Publishing and Reuse of Models and Other Outputs	404

### D. Detecting Unsatisfied Customers: A Case Study

Introduction	418
The Data	418
The Objectives of the Study	418
SAS-EM 5.3 Interface	419
A Primer of SAS-EM Predictive Modeling	420
Homework 1	430
Discussions	431
Homework 2	431
Homework 3	431
Scoring Process and the Total Profit	432
Homework 4	438
Discussions	439
Oversampling and Rare Event Detection	439
Discussion	446
Decision Matrix and the Profit Charts	446
Discussions	453
Micro-Target the Profitable Customers	453
Appendix	455

### E. Credit Scoring

Introduction: What Is Credit Scoring?	459
Credit Scoring: Business Objectives	460
Case Study: Consumer Credit Scoring	461
Description	461
Data Preparation	462
Feature Selection	462
STATISTICA Data Miner: “Workhorses” or Predictive Modeling	463
Overview: STATISTICA Data Miner Workspace	464
Analysis and Results	465
Decision Tree: CHAID	465
Classification Matrix: CHAID Model	467

- Comparative Assessment of the Models (Evaluation) 467  
 Classification Matrix: Boosting Trees with Deployment Model (Best Model) 469  
 Deploying the Model for Prediction 469  
 Conclusion 470
- ### F. Churn Analysis
- Objectives 471  
 Steps 472
- ### G. Text Mining: Automobile Brand Review
- Introduction 481  
 Text Mining 482  
 Input Documents 482  
 Selecting Input Documents 482  
 Stop Lists, Synonyms, and Phrases 482  
 Stemming and Support for Different Languages 483  
 Indexing of Input Documents: Scalability of *STATISTICA* Text Mining and Document Retrieval 483  
 Results, Summaries, and Transformations 483  
 Car Review Example 484  
 Saving Results into Input Spreadsheet 498  
 Interactive Trees (C&RT, CHAID) 503  
 Other Applications of Text Mining 512  
 Conclusion 512
- ### H. Predictive Process Control: QC-Data Mining
- Predictive Process Control Using *STATISTICA* and *STATISTICA* Qc-miner 513  
 Case Study: Predictive Process Control 514  
 Understanding Manufacturing Processes 514  
 Data File: ProcessControl.sta 515  
 Variable Information 515  
 Problem Definition 515  
 Design Approaches 515  
 Data Analyses with *STATISTICA* 517  
 Split Input Data into the Training and Testing Sample 517  
 Stratified Random Sampling 517  
 Feature Selection and Root Cause Analyses 517
- Different Models Used for Prediction 518  
 Compute Overlaid Lift Charts from All Models: Static Analyses 520  
 Classification Trees: CHAID 521  
 Compute Overlaid Lift/Gain Charts from All Models: Dynamic Analyses 523  
 Cross-Tabulation Matrix 524  
 Comparative Evaluation of Models: Dynamic Analyses 526  
 Gains Analyses by Deciles: Dynamic Analyses 526  
 Transformation of Change 527  
 Feature Selection and Root Cause Analyses 528  
 Interactive Trees: C&RT 528  
 Conclusion 529
- ### I. Business Administration in a Medical Industry
- ### J. Clinical Psychology: Making Decisions about Best Therapy for a Client
- ### K. Education–Leadership Training for Business and Education
- ### L. Dentistry: Facial Pain Study
- ### M. Profit Analysis of the German Credit Data
- Introduction 651  
 Modeling Strategy 653  
 SAS-EM 5.3 Interface 654  
 A Primer of SAS-EM Predictive Modeling 654  
 Advanced Techniques of Predictive Modeling 669  
 Micro-Target the Profitable Customers 676  
 Appendix 678
- ### N. Predicting Self-Reported Health Status Using Artificial Neural Networks
- Background 681  
 Data 682  
 Preprocessing and Filtering 683

- Part 1: Using a Wrapper Approach in Weka to  
Determine the Most Appropriate Variables for  
Your Neural Network Model 684
- Part 2: Taking the Results from the Wrapper  
Approach in Weka into *STATISTICA* Data  
Miner to Do Neural Network Analyses 691

---

## IV

---

### MEASURING TRUE COMPLEXITY, THE “RIGHT MODEL FOR THE RIGHT USE,” TOP MISTAKES, AND THE FUTURE OF ANALYTICS

#### 18. Model Complexity (and How Ensembles Help)

- Preamble 707
- Model Ensembles 708
- Complexity 710
- Generalized Degrees of Freedom 713
- Examples: Decision Tree Surface with Noise 714
- Summary and Discussion 719
- Postscript 720

#### 19. The Right Model for the Right Purpose: When Less Is Good Enough

- Preamble 723
- More Is not Necessarily Better: Lessons from Nature  
and Engineering 724
- Embrace Change Rather Than Flee from It 725
- Decision Making Breeds True in the Business  
Organism 725
- Muscles in the Business Organism 726
- What Is a Complex System? 726
- The 80:20 Rule in Action 728
- Agile Modeling: An Example of How to Craft  
Sufficient Solutions 728
- Postscript 730

#### 20. Top 10 Data Mining Mistakes

- Preamble 733
- Introduction 734
0. Lack Data 734
1. Focus on Training 735
2. Rely on One Technique 736
3. Ask the Wrong Question 738
4. Listen (Only) to the Data 739
5. Accept Leaks from the Future 742
6. Discount Pesky Cases 743
7. Extrapolate 744
8. Answer Every Inquiry 747
9. Sample Casually 750
10. Believe the Best Model 752
- How Shall We Then Succeed? 753
- Postscript 753

#### 21. Prospects for the Future of Data Mining and Text Mining as Part of Our Everyday Lives

- Preamble 755
- RFID 756
- Social Networking and Data Mining 757
- Example 1 758
- Example 2 759
- Example 3 760
- Example 4 761
- Image and Object Data Mining 761
- Visual Data Preparation for Data Mining: Taking  
Photos, Moving Pictures, and Objects into  
Spreadsheets Representing the Photos, Moving  
Pictures, and Objects 765
- Cloud Computing 769
- What Can Science Learn from Google?* 772
- The Next Generation of Data Mining 772
- From the Desktop to the Clouds ... 778
- Postscript 778

#### 22. Summary: Our Design

- Preamble 781
- Beware of Overtrained Models 782
- A Diversity of Models and Techniques Is Best 783
- The Process Is More Important Than the Tool 783

Text Mining of Unstructured Data Is Becoming Very Important 784	“Smart” Systems Are the Direction in Which Data Mining Technology Is Going 787
Practice Thinking About Your Organization as <i>Organism</i> Rather Than as <i>Machine</i> 784	Postscript 787
Good Solutions Evolve Rather Than Just Appear After Initial Efforts 785	<b>Glossary 789</b>
What You <i>Don’t</i> Do Is Just as Important as What You <i>Do</i> 785	<b>Index 801</b>
Very Intuitive Graphical Interfaces Are Replacing Procedural Programming 786	<b>DVD Install Instructions 823</b>
Data Mining Is No Longer a Boutique Operation; It Is Firmly Established in the Mainstream of Our Society 786	



# Foreword 1

---

This book will help the novice user become familiar with data mining. Basically, data mining is doing data analysis (or statistics) on data sets (often large) that have been obtained from potentially many sources. As such, the miner may not have control of the input data, but must rely on sources that have gathered the data. As such, there are problems that every data miner must be aware of as he or she begins (or completes) a mining operation. I strongly resonated to the material on “The Top 10 Data Mining Mistakes,” which give a worthwhile checklist:

- Ensure you have a response variable and predictor variables—and that they are correctly measured.
- Beware of overfitting. With scads of variables, it is easy with most statistical programs to fit incredibly complex models, but they cannot be reproduced. It is good to save part of the sample to use to test the model. Various methods are offered in this book.
- Don’t use only one method. Using only linear regression can be a problem. Try dichotomizing the response or categorizing it to remove nonlinearities in the response variable. Often, there are clusters of values at zero, which messes up any normality assumption. This, of course, loses information, so you may want to categorize a continuous response variable and use an alternative to regression. Similarly, predictor variables may need to be treated as factors rather than linear predictors. A classic example is using marital status or race as a linear predictor when there is no order.
- Asking the wrong question—when looking for a rare phenomenon, it may be helpful to identify the most common pattern. These may lead to complex analyses, as in item 3, but they may also be conceptually simple. Again, you may need to take care that you don’t overfit the data.
- Don’t become enamored with the data. There may be a substantial history from earlier data or from domain experts that can help with the modeling.
- Be wary of using an outcome variable (or one highly correlated with the outcome variable) and becoming excited about the result. The predictors should be “proper” predictors in the sense that (a) they are measured prior to the outcome and (b) are not a function of the outcome.
- Do not discard outliers without solid justification. Just because an observation is out of line with others is insufficient reason to ignore it. You must check the circumstances that led to the value. In any event, it is useful to conduct the analysis with the observation(s) included and excluded to determine the sensitivity of the results to the outlier.

- Extrapolating is a fine way to go broke—the best example is the stock market. Stick within your data, and if you must go outside, put plenty of caveats. Better still, restrain the impulse to extrapolate. Beware that pictures are often far too simple and we can be misled. Political campaigns oversimplify complex problems (“My opponent wants to raise taxes”; “My opponent will take us to war”) when the realities may imply we have some infrastructure needs that can be handled only with new funding, or we have been attacked by some bad guys.

Be wary of your data sources. If you are combining several sets of data, they need to meet a few standards:

- The definitions of variables that are being merged should be identical. Often they are close but not exact (especially in meta-analysis where clinical studies may have somewhat different definitions due to different medical institutions or laboratories).
- Be careful about missing values. Often when multiple data sets are merged, missing values can be induced: one variable isn’t present in another data set, what you thought was a unique variable name was slightly different in the two sets, so you end up with two variables that both have a lot of missing values.
- How you handle missing values can be crucial. In one example, I used complete cases and lost half of my sample—all variables had at least 85% completeness, but when put together the sample lost half of the data. The residual sum of squares from a stepwise regression was about 8. When I included more variables using mean replacement, almost the same set of predictor variables surfaced, but the residual sum of squares was 20. I then used multiple imputation and found approximately the same set of predictors but had a residual sum of squares (median of 20 imputations) of 25. I find that mean replacement is rather optimistic but surely better than relying on only complete cases. If using stepwise regression, I find it useful to replicate it with a bootstrap or with multiple imputation. However, with large data sets, this approach may be expensive computationally.

To conclude, there is a wealth of material in this handbook that will repay study.

*Peter A. Lachenbruch, Ph.D.,  
Oregon State University  
Past President, 2008, American Statistical Society  
Professor, Oregon State University  
Formerly: FDA and professor at Johns Hopkins University;  
UCLA, and University of Iowa, and  
University of North Carolina Chapel Hill*



# Foreword 2

---

A November 2008 search on Amazon.com for “data mining” books yielded over 15,000 hits—including 72 to be published in 2009. Most of these books either describe data mining in very technical and mathematical terms, beyond the reach of most individuals, or approach data mining at an introductory level without sufficient detail to be useful to the practitioner. The *Handbook of Statistical Analysis and Data Mining Applications* is the book that strikes the right balance between these two treatments of data mining.

This volume is not a theoretical treatment of the subject—the authors themselves recommend other books for this—but rather contains a description of data mining principles and techniques in a series of “knowledge-transfer” sessions, where examples from real data mining projects illustrate the main ideas. This aspect of the book makes it most valuable for practitioners, whether novice or more experienced.

While it would be easier for everyone if data mining were merely a matter of finding and applying the correct mathematical equation or approach for any given problem, the reality is that both “art” and “science” are necessary. The “art” in data mining requires experience: when one has seen and overcome the difficulties in finding solutions from among the many possible approaches, one can apply newfound wisdom to the next project. However, this process takes considerable time and, particularly for data mining novices, the iterative process inevitable in data mining can lead to discouragement when a “textbook” approach doesn’t yield a good solution.

This book is different; it is organized with the practitioner in mind. The volume is divided into four parts. Part I provides an overview of analytics from a historical perspective and frameworks from which to approach data mining, including CRISP-DM and SEMMA. These chapters will provide a novice analyst an excellent overview by defining terms and methods to use, and will provide program managers a framework from which to approach a wide variety of data mining problems. Part II describes algorithms, though without extensive mathematics. These will appeal to practitioners who are or will be involved with day-to-day analytics and need to understand the qualitative aspects of the algorithms. The inclusion of a chapter on text mining is particularly timely, as text mining has shown tremendous growth in recent years.

Part III provides a series of tutorials that are both domain-specific and software-specific. Any instructor knows that examples make the abstract concept more concrete, and these tutorials accomplish exactly that. In addition, each tutorial shows how the solutions were developed using popular data mining software tools, such as Clementine, Enterprise Miner, Weka, and *STATISTICA*. The step-by-step specifics will assist practitioners in learning not only how to approach a wide variety of problems, but also how to use these software

products effectively. Part IV presents a look at the future of data mining, including a treatment of model ensembles and “The Top 10 Data Mining Mistakes,” from the popular presentation by Dr. Elder.

However, the book is best read a few chapters at a time while actively doing the data mining rather than read cover-to-cover (a daunting task for a book this size). Practitioners will appreciate tutorials that match their business objectives and choose to ignore other tutorials. They may choose to read sections on a particular algorithm to increase insight into that algorithm and then decide to add a second algorithm after the first is mastered. For those new to a particular software tool highlighted in the tutorials section, the step-by-step approach will operate much like a user’s manual. Many chapters stand well on their own, such as the excellent “History of Statistics and Data Mining” and “The Top 10 Data Mining Mistakes” chapters. These are broadly applicable and should be read by even the most experienced data miners.

The *Handbook of Statistical Analysis and Data Mining Applications* is an exceptional book that should be on every data miner’s bookshelf or, better yet, found lying open next to their computer.

*Dean Abbott  
President  
Abbott Analytics  
San Diego, California*

# Preface

---

Data mining scientists in research and academia may look askance at this book because it does not present algorithm theory in the commonly accepted mathematical form. Most articles and books on data mining and knowledge discovery are packed with equations and mathematical symbols that only experts can follow. Granted, there is a good reason for insistence on this formalism. The underlying complexity of nature and human response requires teachers and researchers to be extremely clear and unambiguous in their terminology and definitions. Otherwise, ambiguities will be communicated to students and readers, and their understanding will not penetrate to the essential elements of any topic. Academic areas of study are not called *disciplines* without reason.

This rigorous approach to data mining and knowledge discovery builds a fine foundation for academic studies and research by experts. Excellent examples of such books are

- *The Handbook of Data Mining*, 2003, by Nong Ye (Ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2<sup>nd</sup> edition, 2009, by T. Hastie, R. Tibshirani, & J. Friedman. New York: Springer-Verlag.

Books like these were especially necessary in the early days of data mining, when analytical tools were relatively crude and required much manual configuration to make them work right. Early users had to understand the tools in depth to be able to use them productively. These books are still necessary for the college classroom and research centers. Students must understand the theory behind these tools in the same way that the developers understood it so that they will be able to build new and improved versions.

Modern data mining tools, like the ones featured in this book, permit ordinary business analysts to follow a path through the data mining process to create models that are “good enough.” These less-than-optimal models are far better in their ability to leverage faint patterns in databases to solve problems than the ways it used to be done. These tools provide default configurations and automatic operations, which shield the user from the technical complexity underneath. They provide one part in the crude analogy to the automobile interface. You don’t have to be a chemical engineer or physicist who understands moments of force to be able to operate a car. All you have to do is learn to turn the key in the ignition, step on the gas and the brake at the right times, turn the wheel to change direction in a safe manner, and *voila*, you are an expert user of the very complex technology

under the hood. The *other* half of the story is the instruction manual and the driver’s education course that help you to learn how to drive.

This book provides that instruction manual and a series of tutorials to train you how to do data mining in many subject areas. We provide both the right tools and the right intuitive explanations (rather than formal mathematical definitions) of the data mining process and algorithms, which will enable even beginner data miners to understand the basic concepts necessary to understand *what* they are doing. In addition, we provide many tutorials in many different industries and businesses (using many of the most common data mining tools) to show *how* to do it.

---

## OVERALL ORGANIZATION OF THIS BOOK

---

We have divided the chapters in this book into three parts for the same general reason that the ancient Romans split Gaul into three pieces—for the ease of management. The fourth part is a group of tutorials, which serve in principle as Rome served—as the central governing influence. The central theme of this book is the education and training of beginning data mining practitioners, not the rigorous academic preparation of algorithm scientists. Hence, we located the tutorials in the middle of the book in Part III, flanked by topical chapters in Parts I, II, and IV.

This approach is “a mile wide and an inch deep” by design, but there is a lot packed into that inch. There is enough here to stimulate you to take deeper dives into theory, and there is enough here to permit you to construct “smart enough” business operations with a relatively small amount of the right information. James Taylor developed this concept for automating operational decision making in the area of Enterprise Decision Management (Taylor, 2007). Taylor recognized that companies need decision-making systems that are automated enough to keep up with the volume and time-critical nature of modern business operations. These decisions should be deliberate, precise, consistent across the enterprise, smart enough to serve immediate needs appropriately, and agile enough to adapt to new opportunities and challenges in the company. The same concept can be applied to nonoperational systems for Customer Relationship Management (CRM) and marketing support. Even though a CRM model for cross-sell may not be optimal, it may enable several times the response rate in product sales following a marketing campaign. Models like this are “smart enough” to drive companies to the next level of sales. When models like this are proliferated throughout the enterprise to lift all sales to the next level, more refined models can be developed to do even better. This enterprise-wide “lift” in intelligent operations can drive a company through evolutionary rather than revolutionary changes to reach long-term goals.

When one of the primary authors of this book was fighting fires for the U.S. Forest Service, he was struck by the long-term efficiency of Native American contract fire fighters on his crew in Northern California. They worked more slowly than their young “whipper-snapper” counterparts, but they didn’t stop for breaks; they kept up the same pace throughout the day. By the end of the day, they completed far more fire line than the other members of the team. They leveraged their “good enough” work at the moment to accomplish optimal success overall.

Companies can leverage “smart enough” decision systems to do likewise in their pursuit of optimal profitability in their business.

Clearly, use of this book and these tools will not make you experts in data mining. Nor will the explanations in the book permit you to understand the complexity of the theory behind the algorithms and methodologies so necessary for the academic student. But we will conduct you through a relatively thin slice across the wide practice of data mining in many industries and disciplines. We can show you how to create powerful predictive models in your own organization in a relatively short period of time. In addition, this book can function as a springboard to launch you into higher-level studies of the theory behind the practice of data mining. If we can accomplish those goals, we will have succeeded in taking a significant step in bringing the practice of data mining into the mainstream of business analysis.

The three coauthors could not have done this book completely by themselves, and we wish to thank the following individuals, with the disclaimer that we apologize if, by our neglect, we have left out of this “thank you list” anyone who contributed.

Foremost, we would like to thank Acquisitions Editor Lauren Schultz of Elsevier’s Boston office; Lauren was the first to catch the vision and see the need for this book and has worked tirelessly to see it happen. Also, Leah Ackerson, Marketing Manager for Elsevier, and Tom Singer, then Elsevier’s Math and Statistics Acquisitions Editor, who were the first to get us started down this road. Yet, along with Elsevier’s enthusiasm came their desire to have it completed within two months of their making a final decision. . . . So that really pushed us. But Lauren and Leah continually encouraged us during this period by, for instance, flying into the 2008 Knowledge Discovery and Data Mining conference to work out many near-final details.

Bob Nisbet would like to honor and thank his wife, Jean Nisbet, Ph.D., who blasted him off in his technical career by retyping his dissertation five times (before word processing), and assumed much of the family’s burdens during the writing of this book. Bob also thanks Dr. Daniel B. Botkin, the famous global ecologist, for introducing him to the world of modeling and exposing him to the distinction between viewing the world as *machine* and viewing it as *organism*. And, thanks are due to Ken Reed, Ph.D., for inducting Bob into the practice of data mining. Finally, he would like to thank Mike Laracy, a member of his data mining team at NCR Corporation, who showed him how to create powerful customer response models using temporal abstractions.

John Elder would like to thank his wife, Elizabeth Hinson Elder, for her support—keeping five great kids happy and healthy while Dad was stuck on a keyboard—and for her inspiration to excellence. John would also like to thank his colleagues at Elder Research, Inc.—who pour their talents, hard work, and character into using data mining for the good of our clients and community—for their help with research contributions throughout the book. You all make it a joy to come to work. Dustin Hux synthesized a host of material to illustrate the interlocking disciplines making up data mining; Antonia de Medinaceli contributed valuable and thorough edits; Stein Kretsinger made useful suggestions; and Daniel Lautzenheiser created the figure showing a non-intuitive type of outlier.

Co-author Gary Miner wishes to thank his wife, Linda A. Winters-Miner, Ph.D., who has been working with Gary on similar books over the past 10 years and wrote several

of the tutorials included in this book, using real-world data. Gary also wishes to thank the following people from his office who helped in various ways, from keeping Gary's computers running properly to taking over some of his job responsibilities when he took days off to write this book, including Angela Waner, Jon Hillis, Greg Sergeant, Jen Beck, Win Noren, and Dr. Thomas Hill, who gave permission to use and also edited a group of the tutorials that had been written over the years by some of the people listed as guest authors in this book.

Without all the help of the people mentioned here, and maybe many others we failed to specifically mention, this book would never have been completed. Thanks to you all!

*Bob Nisbet* (bob2@rnisbet.com)  
*John Elder* (elder@datamininglab.com)  
*Gary Miner* (miner.gary@gmail.com)  
October 31, 2008

General inquiries can be addressed to: [handbook@datamininglab.com](mailto:handbook@datamininglab.com)

## References

Taylor, J. 2007. *Smart (Enough) Systems*. Upper Saddle River, NJ: Prentice-Hall.

## SAS

To gain experience using SAS® Enterprise Miner™ for the Desktop using tutorials that take you through all the steps of a data mining project, visit HYPERLINK "<http://www.support.sas.com/statandDMapps>" [www.support.sas.com/statandDMapps](http://www.support.sas.com/statandDMapps).

The tutorials include problem definition and data selection, and continue through data exploration, data transformation, sampling, data partitioning, modeling, and model comparison. The tutorials are suitable for data analysts, qualitative experts, and others who want an introduction to using SAS Enterprise Miner for the Desktop using a free 90-day evaluation.

## STATSOFT

To gain experience using *STATISTICA* Data Miner + QC-Miner + Text Miner for the Desktop using tutorials that take you through all the steps of a data mining project, please install the free 90-day *STATISTICA* that is on the DVD bound with this book. Also, please see the "DVD Install Instructions" at the end of the book for details on installing the software and locating the additional tutorials that are only on the DVD.

## SPSS

Call 1.800.543.2185 and mention offer code US09DM0430C to get a free 30-day trial of SPSS Data Mining software (PASW Modeler) for use with the HANDBOOK.

# Introduction

---

Often, data miners are asked, “What are statistical analysis and data mining?” In this book, we will define what data mining is from a procedural standpoint. But most people have a hard time relating what we tell them to the things they know and understand. Before moving on into the book, we would like to provide a little background for data mining that everyone can relate to.

Statistical analysis and data mining are two methods for simulating the unconscious operations that occur in the human brain to provide a rationale for decision making and actions. Statistical analysis is a very directed rationale that is based on norms. We all think and decide on the basis of norms. For example, we consider (unconsciously) what the norm is for dress in a certain situation. Also, we consider the acceptable range of variation in dress styles in our culture. Based on these two concepts, the norm and the variation around that norm, we render judgments like, “That man is inappropriately dressed.” Using similar concepts of mean and standard deviation, statistical analysis proceeds in a very logical way to make very similar judgments (in principle). On the other hand, data mining learns case by case and does not use means or standard deviations. Data mining algorithms build patterns, clarifying the pattern as each case is submitted for processing. These are two very different ways of arriving at the same conclusion: a decision. We will introduce some basic analytical history and theory in Chapters 1 and 2.

The basic process of analytical modeling is presented in Chapter 3. But it may be difficult for you to relate what is happening in the process without some sort of tie to the real world that you know and enjoy. In many ways, the decisions served by analytical modeling are similar to those we make every day. These decisions are based partly on patterns of action formed by experience and partly by intuition.

## PATTERNS OF ACTION

---

A pattern of action can be viewed in terms of the activities of a hurdler on a race track. The runner must start successfully and run to the first hurdle. He must decide very quickly how high to jump to clear the hurdle. He must decide when and in what sequence to move his legs to clear the hurdle with minimum effort and without knocking it down. Then he must run a specified distance to the next hurdle, and do it all over again several times, until he crosses the finish line. Analytical modeling is a lot like that.

The training of the hurdler's "model" of action to run the race happens in a series of operations:

- Run slow at first.
- Practice takeoff from different positions to clear the hurdle.
- Practice different ways to move the legs.
- Determine the best ways to do each activity.
- Practice the best ways for each activity over and over again.

This practice trains the sensory and motor neurons to function together most efficiently.

Individual neurons in the brain are "trained" in practice by adjusting signal strengths and firing thresholds of the motor nerve cells. The performance of a successful hurdler follows the "model" of these activities and the process of coordinating them to run the race. Creation of an analytical "model" of a business process to predict a desired outcome follows a very similar path to the training regimen of a hurdler. We will explore this subject further in Chapter 3 and apply it to develop a data mining process that expresses the basic activities and tasks performed in creating an analytical model.

---

## HUMAN INTUITION

---

In humans, the right side of the brain is the center for visual and aesthetic sensibilities. The left side of the brain is the center for quantitative and time-regulated sensibilities. Human intuition is a blend of both sensibilities. This blend is facilitated by the neural connections between the right side of the brain and the left side. In women, the number of neural connections between right and left sides of the brain is 20% greater (on average) than in men. This higher connectivity of women's brains enables them to exercise intuitive thinking to a greater extent than men. Intuition "builds" a model of reality from both quantitative building blocks and visual sensibilities (and memories).

---

## PUTTING IT ALL TOGETHER

---

Biological taxonomy students claim (in jest) that there are two kinds of people in taxonomy—those who divide things up into two classes (for dichotomous keys) and those who don't. Along with this joke is a similar recognition that taxonomists are divided into the "lumpers" (who combine several species into one) and the "splitters" (who divide one species into many). These distinctions point to a larger dichotomy in the way people think.

In ecology, there used to be two schools of thought: autoecologists (chemistry, physics, and mathematics explain all) and the synecologists (organism relationships in their environment explain all). It wasn't until the 1970s that these two schools of thought learned that both perspectives were needed to understand ecosystems (but more about that later). In business, there are the "big picture" people versus "detail" people. Some people learn by



following an intuitive pathway from general to specific (inductive). Often, we call them “big picture” people. Other people learn by following an intuitive pathway from specific to general (deductive). Often, we call them “detail” people.

This distinction is reflected in many aspects of our society. In Chapter 1, we will explore this distinction to a greater depth in regard to the development of statistical and data mining theory through time.

Many of our human activities involve finding patterns in the data input to our sensory systems. An example is the mental pattern that we develop by sitting in a chair in the middle of a shopping mall and making some judgment about patterns among its clientele. In one mall, people of many ages and races may intermingle. You might conclude from this pattern that this mall is located in an ethnically diverse area. In another mall, you might see a very different pattern. In one mall in Toronto, a great many of the stores had Chinese titles and script on the windows. One observer noticed that he was the only non-Asian seen for a half-hour. This led to the conclusion that the mall catered to the Chinese community and was owned (probably) by a Chinese company or person.

Statistical methods employed in testing this “hypothesis” would include

- Performing a survey of customers to gain empirical data on race, age, length of time in the United States, etc.;
- Calculating means (averages) and standard deviations (an expression of the average variability of all the customers around the mean);
- Using the mean and standard deviation for all observations to calculate a metric (e.g., student’s  $t$ -value) to compare to standard tables;

If the metric exceeds the standard table value, this attribute (e.g., race) is present in the data at a higher rate than expected at random.

More advanced statistical techniques can accept data from multiple attributes and process them in combination to produce a metric (e.g., average squared error), which reflects how well a subset of attributes (selected by the processing method) predicts desired outcome. This process “builds” an analytical equation, using standard statistical methods. This analytical “model” is based on averages across the range of variation of the input attribute data. This approach to finding the pattern in the data is basically a deductive, top-down process (general to specific). The general part is the statistical model employed for the analysis (i.e., normal parametric model). This approach to model building is very “Aristotelian.” In Chapter 1, we will explore the distinctions between Aristotelian and Platonic approaches for understanding truth in the world around us.

Both statistical analysis and data mining algorithms operate on patterns: statistical analysis uses a predefined pattern (i.e., the Parametric Model) and compares some measure of the observations to standard metrics of the model. We will discuss this approach in more detail in Chapter 1. Data mining doesn’t start with a model; it *builds* a model with the data. Thus, statistical analysis uses a model to characterize a pattern in the data; data mining uses the pattern in the data to build a model. This approach uses *deductive reasoning*, following an Aristotelian approach to truth. From the “model” accepted in the beginning (based on the mathematical distributions assumed), outcomes are deduced. On the other hand, data

mining methods discover patterns in data *inductively*, rather than *deductively*, following a more Platonic approach to truth. We will unpack this distinction to a much greater extent in Chapter 1.

Which is the best way to do it? The answer is . . . it depends. It depends on the data. Some data sets can be analyzed better with statistical analysis techniques, and other data sets can be analyzed better with data mining techniques. How do you know which approach to use for a given data set? Much ink has been devoted to paper to try to answer that question. We will not add to that effort. Rather, we will provide a guide to general analytical theory (Chapter 2) and broad analytical procedures (Chapter 3) that can be used with techniques for either approach. For the sake of simplicity, we will refer to the joint body of techniques as *analytics*.

In Chapters 4 and 5, we introduce basic process and preparation procedures for analytics.

Chapters 6–9 introduce accessory tools and some basic and advanced analytic algorithms used commonly for various kinds of analytics projects, followed by the use of specialized algorithms for the analysis of textual data.

Chapters 10–12 provide general introductions to three common analytics tool packages and the two most common application areas for those tools (classification and numerical prediction).

Chapter 13 discusses various methods for evaluating the models you build. We will discuss

- Training and testing activities
- Resampling methods
- Ensemble methods
- Use of graphical plots
- Use of lift charts and ROC curves

Additional details about these powerful techniques can be found in Chapter 5 and in Witten and Frank (2006).

Chapters 14–17 guide you through the application of analytics to four common problem areas: medical informatics, bioinformatics, customer response modeling, and fraud.

One of the guiding principles in the development of this book is the inclusion of many tutorials in the body of the book and on the DVD. There are tutorials for SAS-Enterprise Miner, SPSS Clementine, and *STATISTICA* Data Miner. You can follow through the appropriate tutorials with *STATISTICA* Data Miner. If you download the free trials of the other tools (as described at the end of the Preface), you can follow the tutorials based on them. In any event, the overall principle of this book is to provide enough of an introduction to get you started doing data mining, plus at least one tool for you to use in the beginning of your work.

Chapters 18–20 discuss the issues in analytics regarding model complexity, parsimony, and modeling mistakes.

Chapter 18, on how to measure true complexity, is the most complex and “researchy” chapter of the book, and can be skipped by most readers; but Chapter 20, on classic analytic

mistakes, should be a big help to anyone who needs to implement real models in the real world.

Chapter 21 gives you a glimpse of the future of analytics. Where is data mining going in the future? Much statistical and data mining research during the past 30 years has focused on designing better algorithms for finding faint patterns in “mountains” of data. Current directions in data mining are organized around how to link together many processing instances rather than improving the mathematical algorithms for pattern recognition. We can see these developments taking shape in at least these major areas:

- RAID (Radio Frequency Identification Technologies)
- Social networks
- Visual data mining: object identification, video and audio, and 3D scanning
- Cloud computing

It is likely that even these global processing strategies are not the end of the line in data mining development. Chapter 1 ends with the statement that we will discover increasingly novel and clever ways to mimic the most powerful pattern recognition engine in the universe: the human brain. Chapter 22 wraps up the whole discussion with a summary.

Much concern in the business world now is organized around the need for effective *business intelligence* (BI) processes. Currently, this term refers just to business reporting, and there is not much “intelligence” in it. Data mining can bring another level of “intelligence” to bear on problem solving and pattern recognition. But even the state that data mining may assume in the near future (with cloud computing and social networking) is only the first step in developing truly intelligent decision-making engines.

One step further in the future could be to drive the hardware supporting data mining to the level of nanotechnology. Powerful biological computers the size of pin heads (and smaller) may be the next wave of technological development to drive data mining advances. Rather than the sky, the atom is the limit.

## References

Whitten, I. H. & Frank, E. (2006). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan-Kaufmann.



# List of Tutorials by Guest Authors

---

Tutorials are located in three places:

1. Tutorials A–N are located in Part III of this book.
2. Tutorials O–Z, AA–KK are located on the DVD bound with this book.
3. Additional Tutorials are located on the book’s companion Web site:  
<http://www.elsevierdirect.com/companions/9780123747655>

**Tutorials in Part III of the printed book, with accompanying datasets and results located on the DVD that is bound with this book:**

**Tutorial A** (*Field: General*)

How to Use Data Miner Recipe *STATISTICA* Data Miner Only  
Gary Miner, Ph.D.

**Tutorial B** (*Field: Engineering*)

Data Mining for Aviation Safety Using Data Mining Recipe “Automatized Data Mining”  
from *STATISTICA*  
Alan Stolzer, Ph.D.

**Tutorial C** (*Field: Entertainment Business*)

Predicting Movie Box-Office Receipts Using SPSS Clementine Data Mining Software  
Dursun Delen, Ph.D.

**Tutorial D** (*Field: Financial–Business*)

Detecting Unstatisfied Customers: A Case Study Using SAS Enterprise Miner Version  
5.3 for the Analysis  
Chamont Wang, Ph.D.

**Tutorial E** (*Field: Financial*)

Credit Scoring Using *STATISTICA* Data Miner  
Sachin Lahoti and Kiron Mathew

**Tutorial F** (*Field: Business*)

Churn Analysis using SPSS-Clementine  
Robert Nisbet, Ph.D.

**Tutorial G** (*Field: Customer Satisfaction–Business*)

Text Mining: Automobile Brand Review Using STATISTICA Data Miner and Text Miner

Sachin Lahoti and Kiron Mathew, edited by Gary Miner, Ph.D.

**Tutorial H** (*Field: Industry Quality Control*)

Predictive Process Control: QC-Data Mining Using STATISTICA Data Miner and QC-Miner

Sachin Lahoti and Kiron Mathew, edited by Gary Miner, Ph.D.

**Tutorials I, J, and K****Three Short Tutorials Showing the Use of Data Mining and Particularly C&RT to Predict and Display Possible Structural Relationships among Data**

edited by Linda A. Miner, Ph.D.

**Tutorial I** (*Field: Business Administration*)

Business Administration in a Medical Industry: Determining Possible Predictors for Days with Hospice Service for Patients with Dementia

Linda A. Miner, Ph.D., James Ross, MD, and Karen James, RN, BSN, CHPN

**Tutorial J** (*Field: Clinical Psychology & Patient Care*)

Clinical Psychology: Making Decisions about Best Therapy for a Client: Using Data Mining to Explore the Structure of a Depression Instrument

David P. Armentrout, Ph.D. and Linda A. Miner, Ph.D.

**Tutorial K** (*Field: Leadership Training–Business*)

Education–Leadership Training for Business and Education Using C&RT to Predict and Display Possible Structural Relationships

Greg S. Robinson, Ph.D., Linda A. Miner, Ph.D., and Mary A. Millikin, Ph.D.

**Tutorial L** (*Field: Dentistry*)

Dentistry: Facial Pain Study Based on 84 Predictor Variables (Both Categorical and Continuous)

Charles G. Widmer, DDS, MS.

**Tutorial M** (*Field: Financial–Banking*)

Profit Analysis of the German Credit Data using SAS-EM Version 5.3

Chamont Wang, Ph.D., edited by Gary Miner, Ph.D.

**Tutorial N** (*Field: Medical Informatics*)

Predicting Self-Reported Health Status Using Artificial Neural Networks

Nephi Walton, Stacey Knight, MStat, and Mollie R. Poynton, Ph.D., APRN, BC

---

**Tutorials on the DVD bound with this book including datasets, data mining projects, and results:**

**Tutorial O** (*Field: Demographics*)

Regression Trees Using Boston Housing Data Set  
Kiron Mathew, edited by Gary Miner, Ph.D.

**Tutorial P** (*Field: Medical Informatics & Bioinformatics*)

Cancer Gene  
Kiron Mathew, edited by Gary Miner, Ph.D.

**Tutorial Q** (*Field: CRM – Customer Relationship Management*)

Clustering of Shoppers: Clustering Techniques for Data Mining Modeling  
Kiron Mathew, edited by Gary Miner, Ph.D.

**Tutorial R** (*Field: Financial–Banking*)

Credit Risk using Discriminant Analysis in a Data Mining Model  
Sachin Lahoti and Kiron Mathew, edited by Gary Miner, Ph.D.

**Tutorial S** (*Field: Data Analysis*)

Data Preparation and Management  
Kiron Mathew, edited by Gary Miner, Ph.D.

**Tutorial T** (*Field: Deployment of Predictive Models*)

Deployment of a Data Mining Model  
Kiron Mathew and Sachin Lahoti

**Tutorial U** (*Field: Medical Informatics*)

Stratified Random Sampling for Rare Medical Events: A Data Mining Method to Understand Pattern and Meaning of Infrequent Categories in Data  
David Redfearn, Ph.D., edited by Gary Miner, Ph.D.  
*[This Tutorial is not included on the DVD bound with the book, but instead is on this book's companion Web site.]*

**Tutorial V** (*Field: Medical Informatics–Bioinformatics*)

Heart Disease Utilizing Visual Data Mining Methods  
Kiron Mathew, edited by Gary Miner, Ph.D.

**Tutorial W** (*Field: Medical Informatics–Bioinformatics*)

Type II Diabetes Versus Assessing Hemoglobin A1c and LDL, Age, and Sex: Examination of the Data by Progressively Analyzing from Phase 1 (Traditional Statistics) through Phase 4 (Advanced Bayesian and Statistical Learning Theory) Data Analysis Methods, Including Deployment of Model for Predicting Success in New Patients  
Dalton Ellis, MD and Ashley Estep, DO, edited by Gary Miner, Ph.D.

**Tutorial X** (*Field: Separating Competing Signals*)

Independent Component Analysis

Thomas Hill, Ph.D, edited by Gary Miner, Ph.D.

**Tutorial Y** (*Fields: Engineering–Air Travel–Text Mining*)

NTSB Aircraft Accident Reports

Kiron Mathew, by Thomas Hill, Ph.D., and Gary Miner, Ph.D.

**Tutorial Z** (*Field: Preventive Health Care*)Obesity Control in Children: Medical Tutorial Using *STATISTICA* Data Miner Recipe—  
Childhood Obesity Intervention Attempt

Linda A. Miner, Ph.D., Walter L. Larimore, MD, Cheryl Flynt, RN, and Stephanie Rick

**Tutorial AA** (*Field: Statistics–Data Mining*)

Random Forests Classification

Thomas Hill, Ph.D. and Kiron Mathew, edited by Gary Miner, Ph.D.

**Tutorial BB** (*Field: Data Mining–Response Optimization*)

Response Optimization for Data Mining Models

Kiron Mathew and Thomas Hill, Ph.D., edited by Gary Miner, Ph.D.

**Tutorial CC** (*Field: Industry–Quality Control*)

Diagnostic Tooling and Data Mining: Semiconductor Industry

Kiron Mathew and Sachin Lahoti, edited by Gary Miner, Ph.D.

**Tutorial DD** (*Field: Sociology*)

Visual Data Mining: Titanic Survivors

Kiron Mathew and Thomas Hill, Ph.D., edited by Gary Miner, Ph.D.

**Tutorial EE** (*Field: Demography–Census*)

Census Data Analysis: Basic Statistical Data Description

Kiron Mathew, edited by Gary Miner, Ph.D.

**Tutorial FF** (*Field: Environment*)

Linear and Logistic Regression (Ozone Data)

Jessica Sieck, edited by Gary Miner, Ph.D.

**Tutorial GG** (*Field: Survival Analysis–Medical Informatics*)R-Integration into a Data Miner Workspace Node: R-node Competing Hazards Program  
Named `cmprsk` from the R-LibraryIvan Korsakov and Wayne Kendal, MD., Ph.D., FRCSC, FRPCP, edited by Gary Miner,  
Ph.D.



**Tutorial HH** (*Fields: Social Networks–Sociology & Medical Informatics*)

Social Networks among Community Organizations: Tulsa Cornerstone Assistance Network Partners Survey of Satisfaction Derived From this Social Network by Members of Cornerstone Partners: Out of 24 Survey Questions, Which are Important in Predicting Partner Satisfaction?

Enis Sakirgil, MD and Timothy Potter, MD, edited by Gary Miner, Ph.D.

**Tutorial II** (*Field: Social Networks*)

Nairobi, Kenya Baboon Project: Social Networking among Baboon Populations in Kenya on the Laikipia Plateau

Shirley C. Strum, Ph.D., edited by Gary Miner, Ph.D.

**Tutorial JJ** (*Field: Statistics Resampling Methods*)

Jackknife and Bootstrap Data Miner Workspace and MACRO for STATISTICA Data Miner

Gary Miner, Ph.D.

**Tutorial KK** (*Field: Bioinformatics*)

Dahlia Mosaic Virus: A DNA Microarray Analysis on 10 Cultivars From a Single Source: Dahlia Garden in Prague, Czech Republic

Hanu R. Pappu, Ph.D., edited by Gary Miner, Ph.D.

**Tutorials that are on the book's companion Web site:**

<http://www.elsevierdirect.com/companions/9780123747655>

**Tutorial LL** (*Field: Physics–Meteorology*)

Prediction of Hurricanes from Cloud Data

Jose F. Nieves, Ph.D. and Juan S. Lebron

**Tutorial MM** (*Field: Education–Administration*)

Characteristics of Student Populations in Schools and Universities of the United States

Allen Mednick, Ph.D. Candidate

**Tutorial NN** (*Field: Business–Psychology*)

Business Referrals based on Customer Service Records

Ronald Mellado Miller, Ph.D.

**Tutorial OO** (*Field: Medicine–Bioinformatics*)

Autism: Rating From Parents and Teachers Using an Assessment of Autism Measure and Also Linkage with Genotype: Three Different Polymorphisms Affecting Serotonin Functions Versus Behavioral Data

Ira L. Cohen, Ph.D.

**Tutorial PP** (*Field: Ecology*)

Human Ecology: Clustering Fiber Length Histograms by Degree of Damage  
Mourad Krifa, Ph.D.

**Tutorial QQ** (*Field: Finance–Business*)

Wall Street: Stock Market Predictions  
Gary Miner, Ph.D.

**Tutorial RR** (*Field: Scorecards–Business Financial*)

Developing Credit Scorecards Using SAS® Enterprise Miner™  
R. Wayne Thompson, SAS, Carey, North Carolina

**Tutorial SS** (*Field: Astronomy*)

Astro-Statistics: Data Mining of SkyLab Project: Hubble Telescope Star and  
Galaxy Data  
Joseph M. Hilbe, JD, Ph.D., Gary Miner, Ph.D., and Robert Nisbet, Ph.D.

**Tutorial TT** (*Field: Customer Response–Business Commercial*)

Boost Profitability and Reduce Marketing Campaign Costs with PASW Modeler by  
SPSS Inc.  
David Vergara